COPULA-BASED MODELS IN RAILROAD MAINTENANCE AND SAFETY ANALYSIS

by

Emmanuel Nii Martey

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Civil Engineering

Summer 2018

COPULA-BASED MODELS IN RAILROAD MAINTENANCE AND SAFETY ANALYSIS

by

Emmanuel Nii Martey

Approved: _____

Sue McNeil, Ph.D.
Chair of the Department of Civil and Environmental Engineering

Approved: _____

Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Nii Attoh-Okine, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Sue McNeil, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Allan Zarembski, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Steven Chrismer, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

First and foremost, I want to thank the Almighty God for how far He has brought me. I thank Him for His mercies and guidance throughout the entire period of my Ph.D. study.

My sincerest gratitude goes to my graduate advisor, Prof. Nii Attoh-Okine for his patience, support and encouragement throughout my study. I would also like to express my gratitude to the other members of my dissertation committee for their invaluable advice and support. I wish to thank past and present members of our research group including Dr. Offei Amanor Adarkwa, Dr. Silvia Galvan Nunez and Ahmed Lasisi for their constructive criticisms and contributions.

Finally, I wish to thank my family and all my loved ones who have encouraged and supported me throughout my Ph.D. study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The American railroad industry has been a primary stakeholder in the economic development of the nation for close to two centuries. The railroads account for over two-fifths of freight revenue ton-miles and transports about a third of all national exports. To ensure good operable conditions of rail infrastructure particularly the track, the railroads have spent more than 40% of their revenue on capital expenditure and maintenance since industry deregulation. Due to budgetary and high logistical constraints, there has been a gradual shift to predictive maintenance strategies with railroads planning track geometry maintenance activities in advance. To employ such strategies, there is the need to know beforehand the effectiveness of maintenance activities which can be evaluated by the amount of improvement or recovery in track geometry condition.

Well executed maintenance invariably improves operational efficiency and safety which are primary objectives of the railroads. The huge investment in maintenance led to all-time lows in train derailment rate, accident rate, and collision rate recorded in recent years. Despite their relatively low frequency, derailments remain a major concern for the railroads due to their high consequences which include loss of life and property, disruption of services, injury, and destruction to the natural environment. It is therefore important to carefully examine train derailment severity in order to minimize these ramifications.

In many railroad applications of data analysis; non-normality of data occurs in several forms. For example, exploratory data analysis of both derailment data and track geometry data showed that the marginal and joint distributions of the variables were not normal. Conventional correlation analysis is generally not suitable for analyzing the dependencies between variables with non-normality, tail dependence, asymmetric

dependence, skewness and other nonlinearities. Furthermore, conventional correlation analysis also fails to consider the underlying dependence between multiple response variables which may be skewed or discrete in nature. This dissertation focuses on the formulation of copula-based methodologies to analyze railroad maintenance and safety applications considering the underlying dependence between the variables of interest. Copulas allow for the separate modeling of arbitrary marginal distributions and the dependence structure. Copulas are suitable for modeling various forms of dependence and can be employed in the generation of large volumes of data.

Three railroad engineering case studies are undertaken in this dissertation. In the first case study, a bivariate copula-based approach is developed to evaluate the tamping recovery of track geometry parameters such as surface, alignment, cross level, gage, and warp considering the underlying dependence between the variables of interest. In the second case study, a mixed copula-based regression model is developed which simultaneously models the monetary damage and number of derailed cars conditional on a set of covariates that might affect both derailment severity outcomes. Marginal generalized linear regression models are combined with a bivariate copula which characterizes the dependence between the two responses. In the third and final case study, vine copula models, a cascade of bivariate copulas as building blocks, are used to model high-dimensional dependencies within the derailment severity data.

Results from this dissertation provide greater insight and comprehension of the train derailment severity and track geometry recovery phenomena considering various forms of dependence between the variables of interest. These results will aid decision making which would help reduce the consequences of train derailments as well as improve track maintenance strategies.

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Rail transportation has been a primary stakeholder in the economic development of the United States for more than 185 years by connecting various businesses domestically and internationally. The United States has the largest railroad network in the world with Class I railroads (freight railroads with 2016 operating revenues of $447.6 million or more) accounting for about 140,000 miles of track which is about 69% of U.S. freight rail mileage (AAR, 2017). This essential mode of transportation accounted for about 42.7% of freight revenue ton-miles which makes up the biggest proportion of inter-city freight (Peng, 2011; He et al., 2015).

Operation efficiency (travel comfort, reliability) and safety are primary objectives of the railroads. Thus, the railroad infrastructure (particularly the track) is regularly maintained in order to achieve these objectives. Since industry deregulation in 1980, American freight railroads have spent more than $635 billion on capital expenditure and maintenance expenses which amounts to more than 40% of their revenue (AAR, 2017). In 2015 alone, Class I railroads spent $9.7 billion on maintenance and an additional $17.4 billion on expansion and modernization which are 15% and 27% respectively of their annual expenditure respectively (AAR, 2016). As a result of budgetary and high logistical constraints, railroads plan track geometry maintenance activities in advance (Quiroga and Schnieder, 2012; Caetano and Teixeira, 2016). This has resulted in a gradual shift towards prognostic maintenance strategies which require the need to know beforehand the effectiveness of maintenance activities (such as tamping) which can be estimated by the amount of improvement or recovery in the condition of track geometry (Famurewa et al., 2013).

The aforementioned investments by the railroads have translated into great strides being made in area of safety. Train accident rate in 2016 was at an all-time low with a 42% reduction from 2000. In the same time frame, employee injury rate and grade crossing collision rate have decreased by 46% and 38% respectively. Train derailment rate, accident rate and collision rate caused by track defects were also at an all-time low in 2016 making it arguably the safest year in American rail history (AAR, 2017). Despite the relatively low derailment rate, the high ramifications of their occurrence which include disruption of services, injury, loss of life and property and damage to the natural environment remain a primary concern of the railroads. Thus, there is the need to carefully examine train derailment severity in order to minimize the consequences.

1.2 Statement of the Problem

In many infrastructure applications including railroad applications such as track geometry recovery and derailment severity, non-normality of data transpires in various forms. These include non-normality of the marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables (Yan, 2006; Attoh-Okine, 2013). Conventional correlation analysis is generally not suitable for analyzing the dependencies between variables with non-normality, tail dependence, asymmetric dependence, skewness and other nonlinearities. Furthermore, conventional correlation analysis also fails to take into account the underlying dependence between multiple response variables which may be skewed or discrete in nature.

This dissertation seeks to address the limitations of conventional correlation analysis using the copula approach to model the underlying dependences between the variables of interest in railroad engineering applications by taking into consideration various forms of dependence. Copulas are used to describe the dependence between random variables and can be defined as functions that combine arbitrary marginal distributions to form a joint distribution. Copulas can be employed as standalone

models (such as bivariate and vine copula models) or combined with other models (such as copula regression models, copula bayesian networks and copula autoregressive models). Copula modeling is an emerging statistical method which has been widely used in the financial industry and is gaining traction in engineering. However, its application in the railroad industry is very limited.

1.3 Objective of the Study

The main objectives of this research are to apply copula methodology to train derailment severity data and track geometry maintenance data, determine the underlying dependences between the variables of interest and develop probabilistic models as decision tools in railroad maintenance and safety analysis. The main objectives of this research will be achieved through the following sub-objectives:

- To study the derailment severity and track geometry recovery phenomena to identify the factors that influence or affect them.
- To develop copula models for determining the underlying dependence between different variables that contribute to the derailment severity and track geometry recovery phenomena.
- To combine the copula approach with existing models such as generalized linear models.
- To evaluate alternative copula-based models.
- To compare the copula-based models with widely used statistical models such as linear regression and independent multivariate regression models

1.4 Research Approach

The first stage of research involved a comprehensive (state-of-the-art) background review to establish the potential areas in railroad engineering research where copula-based methodologies can be applied as standalone models or in tandem with other alternate models. Data was subsequently obtained for some of the identified areas. Track geometry inspection data was obtained from a Class I U.S. railroad

which contains information on various track geometry parameters such as surface, alignment, cross level, gage and warp. Accident/derailment data was obtained from the Rail Equipment Accident/Incident (REA) database which is maintained by the Federal Railroad Administration (FRA) of U.S. Department of Transportation (U.S. DOT). The database contains detailed track accident information such as accident cause, number of derailed cars, total monetary damage, track type, track class, train length and derailment speed.

Exploratory data analysis was conducted on both datasets that revealed non-normality of the marginal distributions of the variables of interest as well as their joint distributions. In addition, asymmetric and tail dependences between some of these variables were also observed. Copula dependence modeling is suitable for analyzing non-normal variables as well as non-linear, asymmetric and tail dependences.

The implementation of the copula-based methodologies to various railroad safety and track geometry maintenance applications were subsequently conducted. One approach, bivariate copula modeling, was applied to track geometry maintenance data (tamping recovery) whereas two other approaches, copula-based regression modeling and vine copula modeling, were applied to derailment severity data. Bivariate copula modeling was employed to analyze the tamping recovery of track geometry parameters taking into consideration the underlying dependence between the variables of interest. It was subsequently used to generate a set of data points with similar characteristics to the observed data points. Multivariate copula modeling based on vine copulas was used to analyze and model high-dimensional complex dependences within derailment severity data. Results show that some pairwise dependencies were found to show asymmetric and tail dependences violating the multivariate normal assumption. Vine copula modeling was subsequently employed in the generation of multivariate derailment severity data. Other potential applications of copula-based methodologies in the railroad industry were suggested. Figure 1.1 shows the main research approach of the dissertation.

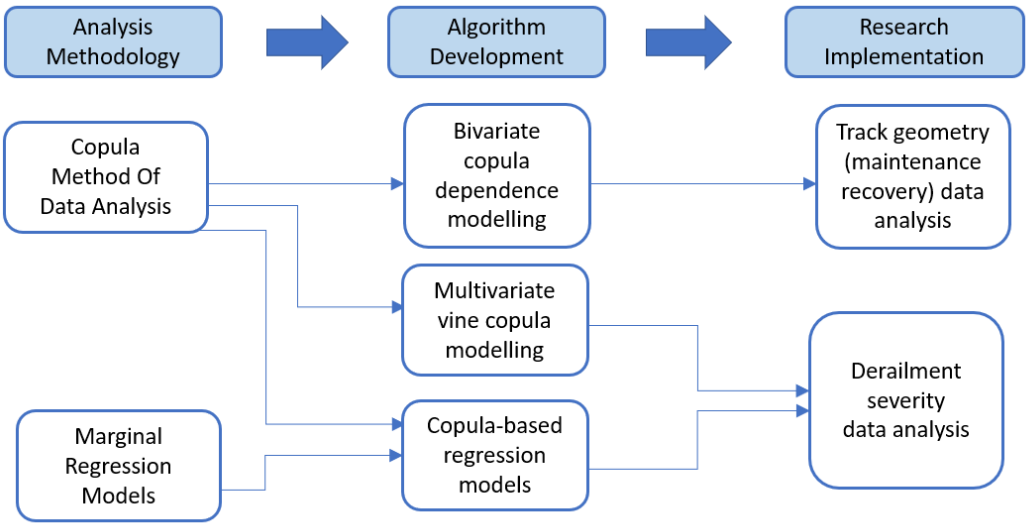


Figure 1.1: Research Approach.

1.5 Dissertation Structure

This dissertation is divided into seven (7) main chapters. The dissertation has been organized in such a way that it first provides knowledge on the various railroad maintenance and safety concepts as well as the data before proceeding to the statistical concepts relating to the various copula-based models. At the end of each chapter related to a specific copula-based model is a railroad engineering case study. Below are the summaries of the various chapters:

1.5.1 Chapter 1: Introduction

This introductory chapter which outlines the need to analyzing railroads concepts such as track geometry maintenance recovery and derailment severity. It also outlines the growing importance of copula-based models in civil infrastructure applications particularly railroad engineering applications. It also presents the motivation of this research, as well as the structure and the contributions of this dissertation.

1.5.2 Chapter 2: Railroad Background

In chapter 2, a literature review on various railroad safety and maintenance concepts is provided. An overview of the various components of a railroad track is presented. The importance of track geometry is outlined and the various track geometry parameters are discussed. The various types of track geometry maintenance activities are subsequently reviewed. The importance of analyzing derailment severity and tamping recovery of track geometry is discussed. Finally, a literature review of existing tamping recovery and derailment severity models is provided. The gaps in the literature are also discussed in this chapter.

1.5.3 Chapter 3: Exploratory Data Analysis

In chapter 3, the various track geometry and derailment severity data sets utilized in the dissertation are described. This chapter also discusses the findings of the exploratory data analysis of the data sets.

1.5.4 Chapter 4: Copula Models

In chapter 4, a detailed overview of copula models is provided. The basic concepts of copula function theory are introduced. The various classes of copulas, dependence concepts and measures, statistical inference (parameter estimation) of copulas and copula selection techniques are subsequently discussed. Finally, case study is presented in which the tamping recovery of various track geometry parameters are modeled using a copula-based approach.

1.5.5 Chapter 5: Copula-based Regression Models

This chapter provides an overview of copula-based regression models which combine several marginal regression models with a bivariate parametric copula which characterizes the underlying dependence between the response variables. Various types of marginal regression models are reviewed. Past applications of copula-based regression models in several transportation fields including modeling automobile crash severity

are discussed. The model formulation of the mixed-copula based regression model including estimation and inference are also provided. Finally, the chapter concludes with a case study on the application of copula-based regression models in predicting bivariate train derailment severity outcomes namely the number of derailed cars and total monetary damage incurred.

1.5.6 Chapter 6: Vine Copula Models

In chapter 6, a detailed overview of vine copula models is provided. The various types of multivariate dependence modeling based on copulas are reviewed. The theory of pair-copula construction upon which vine copulas are developed is explained and the graphical representation of vine copulas known as regular vines is also discussed. The various vine structure selection methods, parameter estimation techniques and pair-copula families selection procedures of vine copulas are also reviewed. This chapter concludes with a case study in which high-dimension dependence of derailment severity data is modeled using vine copulas.

1.5.7 Chapter 7: Concluding Remarks

This is the concluding chapter of the dissertation. In this chapter, results of the research conducted are summarized. Recommendations and future work are also discussed in this chapter.

1.6 Contributions of the Dissertation

Contributions of this dissertation can be found in the following journal publications, conference proceedings and conference presentations.

1.6.1 Journal Publications

1. E. N. Martey and N. O. Attoh-Okine, "Modeling Tamping Recovery of Track Geometry using the copula-based Approach," Journal of Rail and Rapid Transit 0(0), 2018. DOI: 10.1177/0954409718757556.

2. E. N. Martey and N. O. Attoh-Okine, "Bivariate Severity Analysis of Train Derailments using Copula-based Regression Models," ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng, 2018 (Accepted).
3. E. N. Martey and N. O. Attoh-Okine, "Modelling Dependence of Train Derailment Severity using Vine Copula Models," Submitted to Transportation Research Part C: Emerging Technologies, 2018.

The railroad background and literature review of all three papers can be found in chapter 2. The methodology and case study of the first paper can be found in chapter 4. The methodology and case study of the second paper can be found in chapter 5. The methodology and case study of the third paper can be found in chapter 6. The derailment dataset was used in both the second and third case studies (papers) leading to the description of the dataset appearing more than once in the dissertation.

1.6.2 Conference Papers

1. E. N. Martey, A.O. Lasisi and N. O. Attoh-Okine, "Track Geometry Big Data Analysis: A Machine Learning Approach" 2017 IEEE International Conference on Big Data, Boston, MA, USA, Pages: 3800-3809 DOI: 10.1109/BigData.2017.8258381.

The railroad background and literature review of this paper can be found in chapter 2.

1.6.3 Conference Presentations

1. E. N. Martey and N. O. Attoh-Okine, "Bivariate Severity Analysis of Train Derailments using Copula-based Regression Models," 14th Annual Inter-University Symposium on Infrastructure Management, Newark, DE, June 16, 2018.
2. E. N. Martey and N. O. Attoh-Okine, "Bivariate Severity Analysis of Train Derailments using Copula-based Regression Models," 8th Annual Graduate Research Forum, University of Delaware, April 20, 2018.
3. E. N. Martey and N. O. Attoh-Okine, "Modeling Tamping Recovery using Copula-based approach,," 8th Annual Graduate Research Forum, University of Delaware, April 20, 2018.

4. E. N. Martey, A.O. Lasisi and N. O. Attoh-Okine, "Track Geometry Big Data Analysis: A Machine Learning Approach" 2017 IEEE International Conference on Big Data, Boston, MA, USA, December 11, 2017.
5. E. N. Martey and N. O. Attoh-Okine, "Modeling Tamping Recovery using Copula-based approach," 13th Annual Inter-University Symposium on Infrastructure Management, West Lafayette, IN, June 23, 2017.
6. E. N. Martey and N. O. Attoh-Okine, "Severity Analysis of Train Derailments Using Vine Copula Models," 7th Annual Graduate Research Forum, University of Delaware, April 13, 2017.
7. E. N. Martey and N. O. Attoh-Okine,"Severity Analysis of Train Derailments Using Copula Models," Informs 2016 Annual Conference, Nashville, TN, November 13, 2016.
8. E. N. Martey and N. O. Attoh-Okine, "Severity Analysis of Train Derailments Using Vine Copula Models," 12th Annual Inter-University Symposium on Infrastructure Management, Stillwater, OK, June 11, 2016.

REFERENCES

AAR, . Total Annual Spending 2015 Data, 2016. URL https://www.aar.org/FactSheets/Safety/AARAnnualSpending{_}2016Update{_}7.15.16.pdf.

AAR, . An Overview of America's Freight Railroads, 2017. URL https://www.aar.org/BackgroundPapers/OverviewofAmerica'sFreightRRs.pdf.

Attoh-Okine, Nii O. Pair-copulas in infrastructure multivariate dependence modeling. *Construction and Building Materials*, 49:903–911, 2013. ISSN 09500618. doi: 10.1016/j.conbuildmat.2013.06.055.

Caetano, Luis Filipe and Teixeira, Paulo Fonseca. Predictive Maintenance Model for Ballast Tamping. *Journal of Transportation Engineering*, 142(4):4016006, 2016. ISSN 0733-947X. doi: 10.1061/(ASCE)TE.1943-5436.0000825.

Famurewa, S. M.; Xin, T.; Rantatalo, M., and Kumar, U. Optimisation of maintenance track possession time: A tamping case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(1):12–22, 2013. ISSN 0954-4097. doi: 10.1177/0954409713495667.

He, Qing; Li, Hongfei; Bhattacharjya, Debarun; Parikh, Dhaivat P, and Hampapur, Arun. Track geometry defect rectification based on track deterioration modelling and derailment risk assessment. *Journal of the Operational Research Society*, 66(3):392–404, 2015. ISSN 0160-5682. doi: 10.1057/jors.2014.7.

Peng, Fan. *Scheduling of track inspection and maintenance activities in railroad networks*. PhD thesis, University of Illinois at Urbana-Champaign, 2011.

Quiroga, L. M. and Schnieder, E. Monte Carlo simulation of railway track geometry deterioration and restoration. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(3):274–282, 2012. ISSN 1748-006X. doi: 10.1177/1748006X11418422.

Yan, Jun. Multivariate Modeling with Copulas and Engineering Applications. In *Springer Handbook of Engineering Statistics*, pages 973–990. Springer London, London, 2006. doi: 10.1007/978-1-84628-288-1_51.

Chapter 2

RAILROAD BACKGROUND

2.1 Introduction

Railroad transportation allows for the movement of passengers and freight from one location to another on wheeled vehicles on rails, also referred to as tracks. This chapter provides background on the various railroad engineering concepts discussed in the dissertation. An overview of the various track superstructure and substructure components is presented. The importance of track geometry and the various track geometry maintenance activities undertaken to preserve and enhance track geometry quality are discussed. In addition, the importance of analyzing track geometry maintenance recovery and derailment severity are also highlighted. A review of the current state-of-the-art of both tamping recovery models and derailment severity models is presented and the gaps in the literature are subsequently highlighted and addressed.

2.2 Track Characterization

Track is the most fundamental element of the railroad infrastructure. Track provides support to rolling stock through the distribution of wheel loads from the track superstructure to the track substructure. There has been a great evolution of the railroad track structure since its creation more than a century and a half ago resulting in a far stronger and durable track structure (Li et al., 2015). However, there has not been considerable change regarding the principle of the (ballasted) track structure. The enhancements after the Second World include the development of concrete ties, continuous welded rail, advanced measuring equipment, maintenance management systems, mechanized maintenance, heavier rail-profiles and innovative elastic fastenings. Ballasted track also known as "classical or conventional track" comprises of a

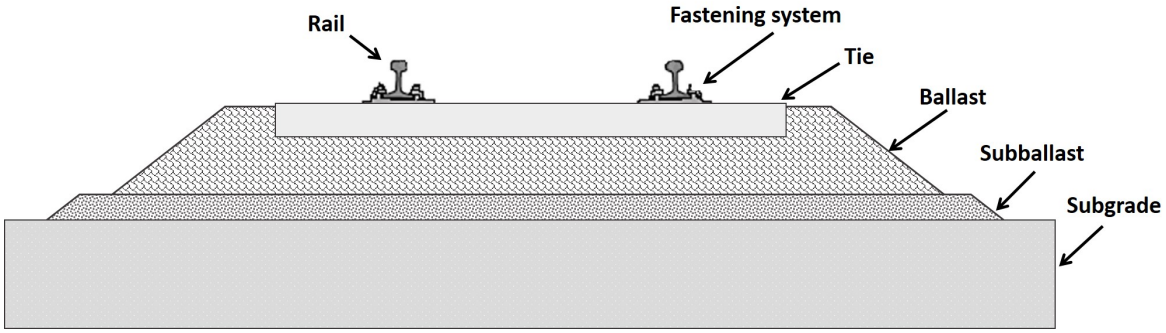


Figure 2.1: Ballasted track structure (Attoh-Okine, 2017)

flat framework consisting of rails and ties which is supported on ballast (Esveld, 2001). The various components of the ballasted track are illustrated in figure 2.1.

2.2.1 Track Superstructure

The track superstructure comprises of the primary load-supporting components of the track that react and distribution train loads to the track substructure. It consists of the rail, fastening system and ties which function in tandem with support rolling stock through the reduction of high stresses at the wheel-level interface to bearable magnitudes for the track substructure layers (Li et al., 2015). The superstructure is separated from the substructure by the tie-ballast interface (Selig and Waters, 1994).

2.2.1.1 Rail

Rail is the most important track structure element (Esveld, 2001). Rails are the longitudinal steel members which offer uniform and constant guidance to the wheels of rolling stock. The rails must be stiff enough to act as beam for the transfer of concentrated wheel loads to the spaced tie supports without huge deflection between the supports (Selig and Waters, 1994). The functions of the rail are as follows (Esveld, 2001):

- Provides support for wheel loads and transfer these loads to the ties

- Offers a smooth-running surface and disseminates acceleration and braking forces via adhesion
- Provides lateral guidance to the wheels
- Serves as an electrical conductor (in the case of electrified lines)
- Serves as a conductor of signal currents

2.2.1.2 Ties (Cross-ties, Sleepers)

Ties or sleepers transfer the wheel loads from the rails to the ballast and transversely secure the rail to maintain the right width of the gage (gauge) (Parvez and Foster, 2017). The ties along with the rail comprise the built-up section of the superstructure. In ballasted track, the rail rests on the ties (Esveld, 2001). Wooden and concrete ties are the most popular with composite/plastic and steel ties less widely used (Li et al., 2015).

Wood ties are the most popular in the States not only because of its lower cost but also for its high resiliency making it better suited for dynamic track interaction. However, they are more sensitive to drainage issues due to their vulnerability to rot and decay (Li et al., 2015). Concrete ties have a much more secure fastening system than wood ties. Concrete ties are heavier and more durable than wooden ties. However, concrete ties are tougher to handle and require pads to provide sufficient resiliency (Selig and Waters, 1994). Concrete ties are less susceptible to climatic conditions but more prone to impact loads. To guarantee stability, it is advisable to support the tie only in areas underneath the rails. The functions of the ties are as follows (Esveld, 2001):

- Accommodate loads from the rail and transfer them over the underlying ballast at tolerable pressure levels for the ballast.
- Offer support and fixing possibilities for the rail foot and fastening.
- Restrain the relative movement of the rail vertically, horizontally and laterally by anchoring the substructure in the ballast.
- Preserve proper rail inclinations and track gage

- Offer sufficient electrical insulation between the rails.

2.2.1.3 Fastenings

Fastening systems (fastenings, fasteners) restrain the relative movement of the rail and tie vertically, horizontally and laterally thereby maintaining the position of the rail (Selig and Waters, 1994; Li et al., 2015). Fastenings include all the components which together form the structural connection between rail and tie. The selection of fasteners is highly dependent on the structure and characteristic of the tie (Esveld, 2001).

The plate and cut spike fasteners are still the most popular fastening system for wood ties with elastic fasteners increasingly being used. On the other hand, elastic fasteners are the only fastenings employed with concrete ties. Plate and cute spike fasteners must be employed with suitable rail anchorage in order to restrain the longitudinal movement of the rail.

Elastic fasteners on the other hand offers resilient restraint of the rail not only longitudinally but also laterally and vertically. This resilience develops with deformation similar to spring stiffness, although not all fasteners provide a linear stiffness variation with deflection (Li et al., 2015).

The functions of the fastening system is as follows (Esveld, 2001):

- Absorb the rail forces elastically and distribute them to the ties
- Offer electrical insulation between the rails and ties, particularly in the case of concrete and steel ties.
- Retain the track gage and rail inclination within certain limits
- Dampen vibrations and impacts due to traffic as much as possible

2.2.1.4 Special Track Work

Special trackwork is also a component of the superstructure which comprises of locations of unique track constructions which demand additional care and attention since they usually experience faster deterioration. It may be described as “all rail,

track structures and fittings, apart from plain unguarded track, that is not curved or fabricated prior to laying". This includes turnouts (switches), crossing diamonds, insulated joints and along with grade-crossings. These components can produce high dynamic loads that are transmitted to the substructure. Thus, support of special trackwork is essential in track design since dynamic loads from rail discontinuities can be more than twice as great as that of open track (Li et al., 2015).

2.2.2 Track Substructure

The track substructure has the greatest effect on track performance. The substructure comprises of ballast, subballast, subgrade, and drainage arrangements. Other terms used to describe the track substructure include trackbed, roadbed, track foundation, and formation. Track substructure comprises of the foundation layers that offer support to the track superstructure and the drainage structures (arrangements). The foundation layer comprises of the subgrade, subballast and subgrade in the case of ballast track and subbase and subgrade in the case of slab track (Li et al., 2015).

2.2.2.1 Ballast

Ballast is the top layer of the substructure which is made up of loose, large, angular, coarse-grained and uniformly graded crushed rock aggregate (such as granite and basalt) and has direct contact with the ties (Silvast et al., 2010; Li et al., 2015).

The ballast bed can withstand high compressive stresses due to the internal friction between the grain. However, it cannot withstand high tensile stresses. The ballast has bearing strength which is substantial in the vertical direction but much lower in the lateral direction (Esveld, 2001).

The ballast has the following functions (Silvast et al., 2010; Li et al., 2015):

- Transfers wheel/rail forces from tie to levels tolerable for lower structural layers.
- Facilitate surfacing and lining activities.
- Offers efficient drainage of precipitation from the track.

- Support the rail-fastener-tie track panel by offering sufficient vertical, lateral and longitudinal resistance in order to maintain vertical and horizontal geometry of the track
- Provides appropriate resiliency together with other track components as well as damping of dynamic wheel/rail forces.

The ballast can be classified into four areas namely (Selig and Waters, 1994):

1. Crib - the granular material between the ties
2. Shoulder - material beyond the end of the tie down to the bottom of the ballast
3. Top ballast - the upper section of the supporting ballast bed which is disturbed by tamping
4. Bottom ballast - the lower section of the supporting ballast bed which is disturbed by tamping

2.2.2.2 Subballast (Blanket)

Subballast is the granular intermediate layer placed between the ballast and subgrade in order to promote good filtering action. The subballast separates the coarse-grained ballast from the fine-grained subgrade (Esveld, 2001). This intermediate layer complements the ballast by improving load distribution thereby decreasing the applied stresses on the subgrade. The subballast also offers protection against frost action (Esveld, 2001; Li et al., 2015). The subballast has the following functions not fulfilled by ballast (Li et al., 2015):

- Prevents penetration and mixture of subgrade and ballast through separation
- Prevents subgrade attrition by ballast, when the upper subgrade is made of clay stone or shale that may be abraded by large ballast particles.

The subballast usually comprises of broadly-graded naturally occurring or processed sand-gravel mixtures or broadly graded crushed natural aggregates or slags. The subballast particles must be durable and fulfil the filter/separation requirements for ballast and subgrade (Selig and Waters, 1994).

2.2.2.3 Subgrade

The subgrade is the platform/foundation upon which the track structure (from the subballast upwards) is constructed and comprises of either soil or rock (Selig and Waters, 1994; Li et al., 2015). The subgrade can be either part of an embankment that is constructed with fill materials or can be natural ground in cut sections of track where subgrade may comprise of the natural soil or placed soil layer(s). The main function of the subgrade is to act as the track foundation by offering uniform and adequate support for the track structure. The subgrade must offer a suitable working base for the construction of the overlying substructure layers and accommodate wheel loads without failure or excessive deformation (Li et al., 2015). The subgrade must also provide sufficient bearing strength and stability, show reasonable settlement behavior, and provide good drainage of precipitation (rain and melted snow) from the ballast (Esveld, 2001).

2.2.2.4 Geosynthetics

Geosynthetics are a family of products manufactured from synthetic polymers employed in a vast array of civil engineering applications (including railroad track applications). Geosynthetics include geotextiles, geocells, geomembranes, geogrids, geosynthetic liners and geo-composites. Track geometry maintenance activities such as tamping and stoneblowing are ineffective if they do not tackle the main cause of track deformation such as subgrade failure. Deformed soft subgrades (as a result of overstressing from imposed loading) can be stabilized by stiffening of the subballast layer through the installation of geocell or geogrids.

Geocell and geogrids are geosynthetics which offer soil stabilization and strength through structural reinforcement. The tensile strength of geogrid aids in the reduction of stresses transferred to the subgrade. On the other hand, the composite action of the subballast being restricted by the geocell offers greater stiffness of the layer in comparison to the subballast only. The stiffening of the overlying layer decreases the vertical stresses acting on the surface of the subgrade.

Geotextiles are permeable (woven or non-woven) geosynthetic which are usually employed to offer filtration and separation between different graded substructure layers. They permit the departure of water from the fine-grained layer without permitting small soil particles to pass through into the voids of the coarse-grained layer. Geotextiles offer separation of the ballast from lower substructure layers containing fine material that may contaminate the ballast. However, this may not be effective due to the abrasion and puncturing of the geotextile fabric by the ballast thereby losing protection against separation and confinement. Additionally, geotextiles typically do not prevent upward migration of fine grained soil subgrade (such as silt and clay particles) into ballast. The openings in the fabric of these geotextiles are too big thus defeating the purpose of preventing infiltration. Furthermore, problems can occur with geotextile becoming caked with fine particles impeding drainage. Lastly, geotextile removal during maintenance or rehabilitation can be complicated since it binds up on the undercutter chain, teeth, and sprockets (Li et al., 2015). To prevent or minimize damage, it is essential to place a fine-grain protection layer beneath and above the geotextile fabric (Esveld, 2001).

Geomembranes or geosynthethic liners are employed if the aim is to establish an impermeable layer. Geomembranes are impermeable, flexible, geosynthetic sheets typically made from synthetic polymers such as neoprene, polyvinyl chloride (PVC), chlorinated polyethylene, or bitumen. Geomembranes are applied in track substructure to prevent the passage of water from one side of it to another thus forming an impermeable layer (Li et al., 2015).

2.3 Track Geometry

2.3.1 General

Track Geometry may be defined as the three-dimensional geometry of track layouts and related measurements used in design, construction and maintenance of

railroad tracks. To identify defects prior to their development beyond acceptable standards, track geometry condition are regularly evaluated during track inspection (Caetano and Teixeira, 2015, 2016). Track Geometry is influenced by climatic conditions, traffic conditions such as loads and speed, construction materials and techniques as well as maintenance history (Audley and Andrews, 2013). Track geometry analysis and maintenance are imperative from cost reduction and track availability enhancement perspectives (Famurewa et al., 2016).

2.3.2 Track Geometry Quality

Track geometry quality can be defined as the "assessment of deviations (excursions) from the mean or designed geometrical characteristics of specified parameters in the vertical and lateral planes which give rise to safety concerns or have a correlation with ride quality". Track geometry condition can be assessed by indicators such as the standard deviation (SD) over a specified length, mean value of the section or extreme (peak) values of isolated defects of the track geometry parameters. The main geometric parameters used to evaluate the quality and irregularity of track geometry include surface (longitudinal level, profile or vertical alignment), alignment (horizontal alignment), gage (gauge), cross level (cant) and warp (twist) (Vale et al., 2012; Famurewa et al., 2013; Khouy, 2013). The primary track geometry components are illustrated in figure 2.2. Surface, cross level and warp are vertical geometric parameters whereas alignment and gage are horizontal geometric parameters (Soleimanmeigouni et al., 2016b).

Surface and alignment can be defined as the track geometry of railroad track center-line projected onto longitudinal vertical and horizontal planes respectively. Surface (also known as longitudinal level) can be termed as the elevation along the longitudinal axis of the rail. Gage is the distance between two rail heads at right angles to the rails in a plane 5/8" below the top of the rail head. Gage variation along with alignment have been found to play important roles in the operational quality of the

railroad track substructure. Cross level on the other hand is the difference in elevation between the adjacent running rails computed from the angle between the running surface and a horizontal reference plane (Khouy, 2013). Warp (twist) is a measure of the crosslevel variation (Audley and Andrews, 2013). Warp can also be defined as the algebraic difference between two cross levels (in inches) taken at any two points within a specified chord length.

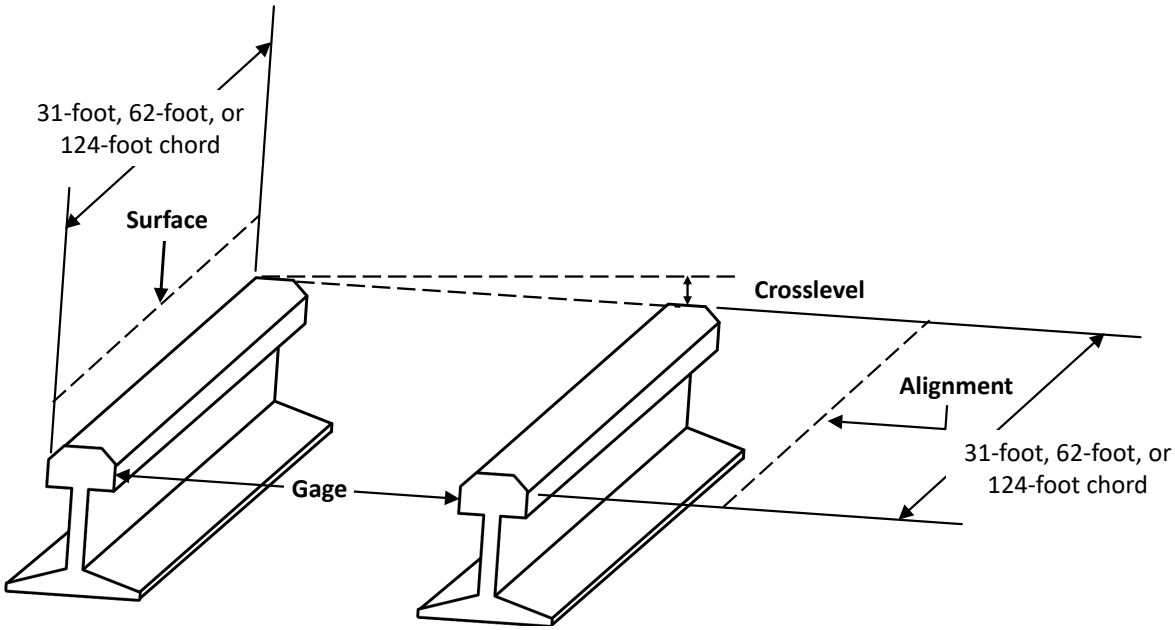


Figure 2.2: Track Geometry Components (Galvan-Nunez, 2017)

Track geometry deterioration is often evaluated by the irregularities or defects of these parameters: surface defects, horizontal alignment defects, cross level defects, gage deviations and warp (track twist) deviations. Infrastructure managers often combine these parameters (defects) into an artificial track quality index (TQI) as a representative measure of the different track geometry parameters and is employed as a decisive metric for maintenance planning. TQI can be quantified as a function of the standard deviations (SD) of each irregularity and allowable train speed. TQI can either be a track geometry index (TGI) or track structure index (TSI) (Andrade and Teixeira,

2012; Caetano and Teixeira, 2015; Soleimanmeigouni et al., 2016b). However, the standard deviation of short wavelength of the surface parameter (defect) is still regarded as the most decisive criterion for maintenance decisions (Andrade and Teixeira, 2012; Caetano and Teixeira, 2015).

According to the International Union of Railways (UIC) and previous research, the standard deviation of the short wavelength surface and alignment are the key parameters used to trigger preventive maintenance procedures. They have been found to be good predictors of the maintenance needs for the rest of the track geometry parameters (Andrade and Teixeira, 2013; Caetano and Teixeira, 2015, 2016). Track geometry irregularities can be categorized into short wavelength irregularities and long wavelength irregularities. Long wavelength track irregularities have an adverse influence on ride comfort. However, short wavelength irregularities generate more vibration on axles and wheels (Soleimanmeigouni et al., 2016b). Thus, short wavelengths have a much greater effect on ride quality (Audley and Andrews, 2013).

The surface parameter is considered to be the most representative of the track quality (Audley and Andrews, 2013). It is the main factor for determining track maintenance expenses and often triggers the need for maintenance intervention (Khouy, 2013). The large proportion of research in the fields of track geometry deterioration and maintenance modeling employ the short wavelength surface parameter as the decisive factor. Reasons for the utilization of the surface parameter include the fact that vertical defects develop more quickly than horizontal defects as well as the automatic recuperation of the horizontal and cross-level defects during track maintenance (Soleimanmeigouni et al., 2016b).

Surface defects can be defined as the vertical geometric deviation measured in inches from the rail top on the running surface to the ideal mean line of the longitudinal profile. Shortwave surface defects have been found to recover very well during tamping. Experimental studies have verified a linear dependence between standard deviation of surface irregularities and accumulated tonnage. Despite surface being the most prominent parameter, disregarding the other parameters during the evaluation

of track geometry condition may result in erroneous assessment leading to ineffective maintenance planning. For instance, warp is a crucial factor that is considered during derailment risk assessment and thus must not be ignored during track geometry evaluation (Andrade and Teixeira, 2012; Caetano and Teixeira, 2016; Soleimanmeigouni et al., 2016b).

A track section is said to have a track geometry defect when the amplitude of track geometry parameter exceeds a given safety threshold. Track geometry defects are therefore severe ill-conditioned geometry parameters. Geometry cars usually classify defects into two severity levels namely "red tags" and "yellow tags". Red tags are defects whose amplitudes violate Federal Railroad Administration (FRA) track safety standards and need to be rectified as soon as possible to avoid fines. Yellow tags on the other hand, are defects that are below FRA thresholds and may or may not exceed the railroad's own safety limits for remediation. Railroads rectify red tags within a due date upon detection however decisions are made on remedying yellow tags are based on field experience considering factors such as track geometry condition, defect history, rail tonnage, track curvature and consequential derailment cost (He et al., 2015).

There is the need for regular inspection or monitoring of track geometry condition or quality using track geometry inspection cars. Track geometry inspection cars assess track irregularities using both an inertia measurement system and an optical system. The vertical and lateral deviation of the track is computed for consecutive 1-foot measurements by means of recorded vehicle accelerations measured by an accelerometer. Limits for track quality are defined based on travelling comfort and safety criteria. The measurement and enhancement of track quality are essential in the establishment of both the restoration period and maintenance cost (Vale and Ribeiro, 2014; Khouy et al., 2012; Vale et al., 2012).

2.4 Track Geometry Maintenance

2.4.1 General

Track geometry deteriorates under traffic loading and undergoes condition-based maintenance (Caetano and Teixeira, 2016). Assessing and maintaining track geometry within acceptable limits are key components of railroad infrastructure maintenance operations (Quiroga and Schnieder, 2012). Maintenance is pivotal in guaranteeing safety, punctuality and effective utilization of capacity. Appropriate maintenance planning is necessary to keep acceptable conditions of infrastructure that economic and social activities largely depend on. However, such an exercise is intricate and challenging to undertake due to various factors such as terrestrial factors, topographical factors, track alignment, atmospheric conditions, rolling stocks, monetary or budget constraints and track availability (Wen et al., 2016).

Maintenance may be defined as a group of activities focused on the enhancement of the overall reliability and availability of a system that are usually classified into preventive and corrective maintenance activities. Preventive maintenance actions comprise of scheduled maintenance actions conducted to guarantee safety and avoid abrupt system failures. Corrective maintenance actions are however conducted following system failure or breakdown to return the system to operable conditions (Gustavsson, 2015). Aside comfort and safety, economic or budget constraints is another reason for preventive maintenance since track maintenance makes up a large proportion of railroad management expenditure (Vale et al., 2012).

Track Geometry rectification is one of the leading track maintenance costs for track maintenance planning (by infrastructure managers) and is thus considered as a cost-driving factor of overall maintenance cost for most passenger operations (Gustavsson, 2015). Track maintenance costs make up about 55% of overall maintenance costs of high-speed lines (Andrade and Teixeira, 2012). Track maintenance encompasses all procedures aimed at the preservation and restoration of the nominal state (Vale et al., 2012). Intervention measures can be classified into track maintenance and track renewal activities (Famurewa et al., 2013).

Intervention thresholds have been established as adjustable parameters for maintenance strategies calibration. Dynamic Intervention thresholds related to the track geometry deviation have been found to be more cost-effective than constant thresholds (Quiroga et al., 2012). Dynamic thresholds are dependent on parameters such as the age of the track and the number of interventions are ideal from a lifecycle viewpoint (Famurewa et al., 2013). However, such calibration requires the availability of models which characterize the actual track geometry degradation and restoration process (Quiroga et al., 2012).

2.4.2 Track Maintenance Activities

Common track maintenance activities include (Vale et al., 2012):

- Tamping which corrects the surface profile, cross level and alignment of the track.
- Ballast injection or stoneblowing to restore the surface profile.
- Rail grinding to rectify rail corrugations, fatigue and restore the profile of the rail.
- Rail replacement
- Track stabilization to return the lateral resistance to the initial level through track vibration.

Tamping can be classified into conventional tamping and design over-lift tamping. Design over-lift tamping, improved tamper control systems and stoneblowing can all be considered as enhanced track geometry maintenance methods in comparison to conventional tamping. Other track maintenance activities include ballast shoulder cleaning, ditching, ballast undercutting/cleaning and track vacuum (Li et al., 2015).

2.4.2.1 Conventional Tamping

Tamping is a maintenance activity employed to rectify track geometry deviations such as incorrect surface profile (vertical deviation) and incorrect alignment (lateral deviation) by rearranging and compacting the ballast (Khouy et al., 2012; Audley and Andrews, 2013). Tamping is the main maintenance activity employed to restore

track geometry condition and is one of the most essential yet costly track maintenance activities (Caetano and Teixeira, 2016; Wen et al., 2016). Tamping is the most common procedure employed to rectify the surface profile which is the geometric parameter which significantly affects rolling stock and the track dynamics in the vertical direction (Vale et al., 2012). Tamping results in a significant decrease in the track geometry irregularity measurements and alters the track deterioration (Soleimanmeigouni et al., 2016a).

Tamping has a significant influence on the effective capacity of a railway network as a result of its distinct needs such as track possession duration, track quality demand, scheduling constraints and heavy equipment utilization. Thus, it is important for optimize the scheduling of this maintenance task (Famurewa et al., 2013; Gustavsson, 2015). However, the execution of tamping more often is not optimally planned. Tamping are at times performed at very low (standard deviation) levels and thus are not influenced by travel comfort (Khouy et al., 2012). Hasty tamping may result in shorter life cycle and track design capacity may not be attained given ineffective tamping procedures (Quiroga et al., 2012; Famurewa et al., 2013).

Tamping maintenance involves heavy machinery and substantial labor resources (Caetano and Teixeira, 2016). Tamping can be executed either mechanically or manually (Audley and Andrews, 2013). Tamping operations can be performed as either preventive or corrective maintenance (Khouy et al., 2012). Corrective tamping is performed to rectify isolated defects whereas preventive tamping can be performed at stations, turnouts (switches) and crossings, and open lines. These two kinds of tamping procedures are planned in different ways (Wen et al., 2016).

Tamping can also be classified into complete and partial tamping procedures. Complete tamping intervention is executed on the entire length of track section whereas partial tamping is carried out on a fraction of the segment. Complete and partial tamping have different effects on the track geometry condition. Thus, separate analysis of these kinds of interventions can result in a drastic decrease in the variation of recovery values of track quality after tamping (Soleimanmeigouni et al., 2016a,b). Conventional

tamping is conducted as follows (Selig and Waters, 1994):

A) The tamper is positioned over the tie to be tamped

B) The lifting rollers of the tamper lifts the tie to the desired level thereby creating a void between the tie base and ballast.

C) The tamping arms (tines) are subsequently inserted into the ballast on either side of a tie.

D) The tines squeeze the ballast together moving the ballast from the crib region to the void underneath the tie thereby filling the void and maintaining the elevated position of the tie

E) The tines are removed from the ballast, the track is lowered and the tamper proceeds to the subsequent tie.

The tamping process is shown on figure 2.3.

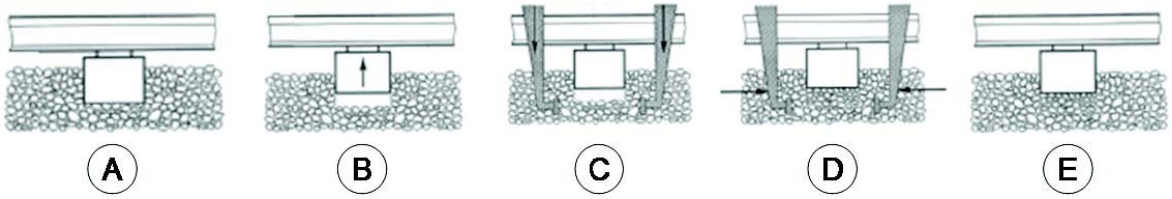


Figure 2.3: Conventional tamping process (Selig and Waters, 1994)

Tamping does not offer durable geometry rectification if it does not tackle the main cause of track deformation such as subgrade failure. However, tamping provides a lasting remedy if the rate of subgrade deformation is relatively low. Conventional tamping has limitations such as its tendency to smoothen the geometry error rather than restore track to its original track geometry condition (designed shape of profile and alignment) (Li et al., 2015). Additionally, tamping has a damaging effect on ballast with the tamping arms crushing the ballast particles (Soleimanmeigouni et al., 2016b; Audley and Andrews, 2013) over the insertion depth thereby decreasing its bulk density (Li et al., 2015).

The resumption of traffic allows the ballast to recover to its original bulk density prior to maintenance and is responsible for the quick initial settlement. This phenomenon is known as "ballast memory" since the track is usually returned to its profile prior to tamping (Esveld, 2001; Khouy, 2013; Li et al., 2015). Improved alternative methods to conventional tamping offer enhanced quality and more durable geometry rectifications. These methods include design over-lift tamping and stone blowing (or injection). Other track maintenance activities include ballast cleaning (or undercutting) and track vacuum (Li et al., 2015).

2.4.2.2 Design Over-lift Tamping

Tampers rectify the vertical track geometry such that an enhanced geometry is achieved (such as a straight line) without taking into consideration subsequent degradation. However, each passing rolling stock attempts to return the track to its original position (Esveld, 2001). Tamping causes a disturbance to the micromechanical structure of the ballast causing significant settlement (Le Pen et al., 2014) and reverts to its prior rough shape upon resumption of traffic (reloading of the track) (Li et al., 2015). The ballast is said to have a "memory" of the shape to which it had degraded prior to each tamping procedure which is a common disadvantage of conventional tamping. This process can never tend towards the ideal straight line unless it is based on an overlift which caters for the expected deformations (Esveld, 2001). Thus, it is best tamping practice to employ design overlift tamping (Le Pen et al., 2014).

Design over-lift tamping offers a track lift greater than that required to return a dipped track to a level position. The amount of over-lift is intended to cater for the ensuing quick ballast settlement with resumption of traffic in order to allow settlement into a design smooth profile. The amount of over-lift applicable to each tie is dependent on the amount of dip at each location. Design over-lift tamping is better at eliminating ballast memory resulting in a more durable geometry rectification. Geometry rectification durability measurements show that design over-lift tamping often lasts three times longer than conventional tamping. These results are similar irrespective of level

of traffic loading (Li et al., 2015). One way of applying overlift is by the addition of a certain portion of the length between the existing and ideal geometry to the lift height. This can only be conducted if the real track geometry is known (Esveld, 2001).

2.4.2.3 Stone Blowing/Injection

Stone blowing is the injection of small stones into the gap between the tie base (lifted to a target level) and the ballast surface (Sol-Sánchez et al., 2017). This is performed without the disturbance of the ballast maintaining its compaction and stability (Esveld, 2001). Stoneblowing is derived from the old practice of shovel packing where voids were created by jacks used to raise the rail and ties and shovels were used to manually place measured small grade ballast between the raised tie base and the ballast. Stone blowing employs a similar concept to remedy track profile error, however instead of a shovel, stone is blown using compressed air from pneumatic injector tubes (Li et al., 2015).

The stone-blowing process is as follows (Selig and Waters, 1994):

A) The tie rests in the ballast prior to adjustment.

B) The tie is initially raised to create a void underneath it,

C) The tubes of the pneumatic ballast injector are inserted into the ballast along the side of the lifted tie to a depth that offers the stones with a flow path and access underneath the tie.

D) A measured amount of stone is injected by compressed air into the void beneath the tie

E) The tubes are removed from the ballast.

F) The tie is lowered onto the added stone where it is subsequently compacted by traffic

The stone-blowing process is illustrated on figure 2.4.

The amount of stone blown is dependent on the target track elevation. The stone must be small enough to disperse well when injected and fill the void without any obstructions (blockages) within the tube, but must be big enough for interlocking,

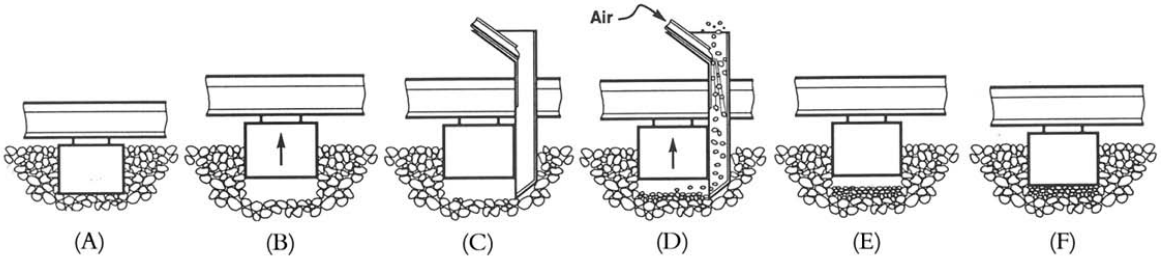


Figure 2.4: The stone-blowing process (Selig and Waters, 1994)

provide support to the tie and not drop into the voids within the existing ballast (Li et al., 2015). The advantages of stoneblowing include decrease in ballast deterioration and maintenance frequency due to the avoidance of the need for decompaction resulting from ballast disturbance (Sol-Sánchez et al., 2017). Settlement after stone blowing is limited due to the minimal disturbance of the tubes inserted into the ballast. This is because the stone blown is placed on top of compacted ballast and the stability of the stone which is usually of high quality in order to withstand the high traffic load stresses (Li et al., 2015).

The pneumatic ballast injector can be applied to sections of track that are inaccessible for machine maintenance. Unlike a tamper, the stoneblower operates in a design mode rather than a smoothing mode (Esveld, 2001). Stone-blowing is more appropriate in areas which demand higher frequency of tamping since it is less destructive to ballast (Khouy, 2013). Thus, stone blowing is suitable at recurring dips at track transitions such as bridge approaches (Li et al., 2015). Stone blowing is appropriate for relatively low lift used in remedying dips less than an inch associated with short wavelength geometry defects (faults) whereas tamping is suitable for high lift used in rectifying dips greater than an inch which are normally related to large wavelength geometry faults (Khouy, 2013; Li et al., 2015).

Despite its advantages, there are concerns related with stoneblowing such as the stiffening of the granular layer and its limited capacity to damp loads. One way of

improving stoneblowing effectiveness is the combination of stoneblowing with elastic elements such as Under Sleeper Pads (USPs) which enhances the durability of the ballast layer and also offers reduction in the maintenance frequency. However, this method introduces increased cost due to the fixing of pads to the tie bottom. Another novel alternative solution known as stone-rubber blowing involves the partial replacement of small stones with rubber particles obtained from waste tires which serve as flexible aggregates which provide enhanced capacity to dampen loads (Sol-Sánchez et al., 2017).

2.4.2.4 Ballast Shoulder Cleaning

Shoulder ballast is the region between the end of the ties and the bottom of the ballast layer. Fouling (of the shoulder ballast) impedes drainage of the ballast in the crib region. Mechanized shoulder cleaning involves the excavation of the shoulder ballast and subsequent replacement with new, clean ballast. Ballast shoulder cleaning offers enhanced drainage from the crib region, given that the crib ballast is moderately (not highly) fouled [fouling index is less than 30 percent]. In this case, shoulder cleaning offers an escape route for any water being retained in the crib region permitting the washing away of fouling material. However, shoulder cleaning is ineffective given a high degree of fouling due to the very low permeability of the crib ballast. In such a situation, a total removal of the highly fouled ballast across the full breadth of the track is needed (Li et al., 2015).

2.4.2.5 Ditching

Whereas shoulder cleaning improves internal draining, the ditching enhances external drainage by offering outlets for quick drainage away from the track. This involves the lateral departure of water from the track and subsequent longitudinal removal from the railroad right of way to low-lying regions. Ditches must have sufficient lateral and longitudinal gradient and an invert elevation of adequate depth beneath the subgrade under track to prevent the return of water to the track. Ditching machines are

used to offer ditches the adequate slopes and contours. However, it is hard to establish and maintain ditches due to the presence of utilities, structures, or other features along the track. In spite of these challenges, it necessary to have them addressed in order to prevent long retention of water in the track (Li et al., 2015).

2.4.2.6 Ballast Undercutting/Cleaning

Ballast, similar to all track components, has a finite life and must be eventually replaced. The end of ballast life usually coincides with the voids being filled with fouling material, which reduces the permeability to the point that the drainage function is lost. Ballast undercutting or cleaning involves the excavation of fouled ballast and its subsequent separation into large ballast aggregate and fine material by means of shaking and sieving (Li et al., 2015). Maintenance activities such as tamping and stone blowing are ineffective when the ballast is heavily fouled. Ballast cleaning or ballast renewal is needed in such a case. However, these activities are expensive and time consuming leading to disruptions and thus are not frequently conducted. The decision as to which maintenance activity is suitable should be based on the site condition and an in-situ investigation of the track layers, including the sub-surface profile (Tennakoon, 2012). Ballast undercutting (renewal) attempts to manage the long-term development of track roughness which is due to the progressive deterioration of ballast (Scanlan et al., 2017).

The effect of ballast undercutting is significant due to the replacement of degraded or fouled ballast with clean ballast resulting in a return of the ballast's damping properties to its ideal state. This leads to a decrease in dynamic loading effects of moving trains on bridge structures (Mohammadzadeh et al., 2017). The ballast undercutter (cleaner) excavates the ballast to a minimum depth below the ties using chain with "excavating teeth" attached which conveys it upwards to a system of vibrating sieves where fine material is wasted. The clean coarse material is reclaimed and returned to the track. Instead of the traditional method of dumping wasted ballast to the side of the track, modern techniques employ waste loaders which run concurrently with ballast

undercutter (Esveld, 2001). The amount of potentially recoverable ballast to reclaim or waste should take into consideration the financial implications of either action (Li et al., 2015). Heavy ballast fouling may require complete replacement with fresh ballast instead of ballast cleaning (Tennakoon, 2012).

Correction of track geometrical misalignment is also conducted during ballast cleaning. This results in better track quality and less impact forces between the track and the wheel (Mohammadzadeh et al., 2017). Ballast undercutting has also been found to considerably decrease track roughness over mineral subgrades such as sand, clay, till and silt but has been shown to be ineffective when applied over soft organic subgrades (Scanlan et al., 2017).

2.4.2.7 Track Vacuum

Fouled ballast around the ties can be eliminated by employing strong track vacuum machines. Track vacuum machines excavate fouled ballast from the track via their large hoses into a holding tank with vacuum offered by large motors. Some vacuums can crush the fouled dense ballast into smaller particles via a rotating bit at the end of the hose. This enables easier excavation from beneath the tie and further below into the ballast. However, the rate of vacuum excavation is low, thus it is mostly employed along short sections of track such as creating a passage for the undercutter chain or excavation around a rail joint. Track vacuuming is suitable in third rail electrified track territory where ballast undercutting procedure is complicated by the power rail and appurtenances. The switch undercutter is more appropriate for excavation of fouled ballast along longer sections of track (Li et al., 2015).

2.5 Tamping Recovery

2.5.1 General

The lifespan of track structure and track quality at any given period, can be characterized in terms of deterioration and recovery events (Famurewa et al., 2013).

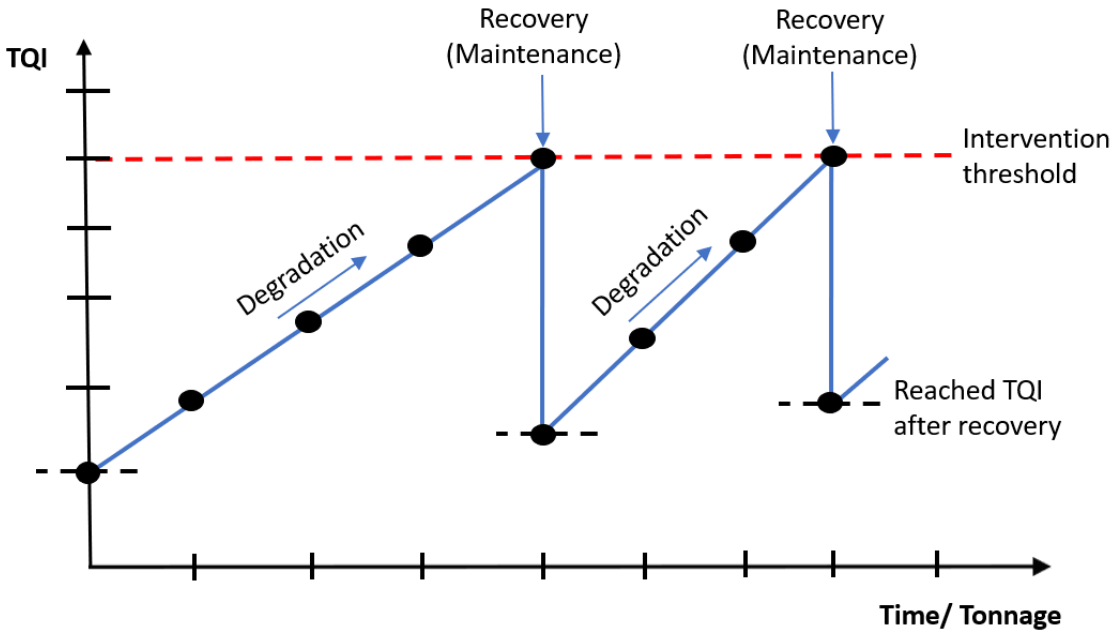


Figure 2.5: Illustration of track geometry degradation and recovery

Thus, track geometry degradation and recovery are key components of any track geometry maintenance model and are crucial for long-term forecast of track geometry performance (Soleimanmeigouni et al., 2016a,b). The illustration of track geometry degradation and recovery is shown in figure 2.5. Much focus has been made on the analysis of track geometry degradation with various deterministic and stochastic techniques employed. These techniques include linear and exponential regression models, polynomial models, multi-stage linear models, neural networks, grey models, path analysis, data mining, models with random coefficient, Markov models, time series models and stochastic processes. For extensive literature on track deterioration models, refer to Soleimanmeigouni et al. (2016b). On the other hand, in comparison relatively little research has been conducted that covers the restoration of track quality due to tamping (Veit, 2007; Lichtberger, 2005; Andrade and Teixeira, 2012, 2013).

The recovery in track geometry condition may be dependent on several factors including track quality prior to tamping, frequency of previous tamping operations

(maintenance history), subsurface (ballast) conditions, tamping procedure, age of track components, operational speeds and human factors (Famurewa et al., 2013; Audley and Andrews, 2013). The dominant factor that influences tamping efficiency or tamping recovery of track geometry is the track condition just before tamping. The recovery of the standard deviation of the surface profile depends on the track geometric quality just prior to maintenance according to Office for Research and Experiments of the International Union of Railways (UIC). The higher the standard deviation of the surface profile, the higher the variability of the track recovery (Vale et al., 2012).

Tamping recovery is dependent on previous tamping procedures since tamping has a damaging effect on the ballast (the tamping machine arms crush the ballast particles) which is the major factor of track stability. This leads to the resultant quality in the current tamping being lower than the resultant quality of preceding tamping (Wen et al., 2016). Tamping recovery also reduces with increasing number of accumulated tamping interventions due to the ballast deterioration with traffic loads as well as the ballast damage due to successive tamping procedures (Caetano and Teixeira, 2016). Tamping efficiency decreases with increase in ballast service life leading to a reduction in the durability of track quality and increased frequency of tamping to maintain track condition at acceptable standards (Zhao et al., 2006).

2.5.2 Tamping Recovery Models

There are two main of modeling restoration (or recovery) after tamping namely deterministic or probabilistic (stochastic) approaches. The choice of methodology to employ should be chosen based on the degree of uncertainty in the recovery values after tamping.

2.5.2.1 Deterministic Models

In deterministic techniques, tamping recovery is directly evaluated in relation to influencing factors such as track quality prior to tamping, the operational speeds and maintenance history. The model parameters are treated as unknown constants with

uncertainty incorporated using confidence intervals. Majority of studies have evaluated tamping recovery using deterministic techniques such as linear regression models and have assumed that tamping effectiveness is mainly dependent on the track geometry quality prior to tamping. Linear regression models are highly popular due to their simplicity and have been employed in the development of track geometry maintenance models and optimization scheduling models (Soleimanmeigouni et al., 2016a,b).

Miwa (2002) and Oyama and Miwa (2006) both applied linear regression restoration models to predict the maintenance effectiveness of tamping with the amount of recovery dependent on the track condition prior to tamping. Their restoration models were combined with an exponential smoothing degradation model which were subsequently used in developing an optimization track maintenance scheduling model. Andrade and Teixeira (2012) employed linear tamping restoration models as well as a linear track deterioration model which were subsequently used in the development of a biobjective model to optimize planned maintenance and renewal activities related to track geometry. Vale et al. (2012) employed linear tamping restoration model as well as a linear track deterioration model which were subsequently used in the development of a mathematical maintenance model (formulated as integer (mixed 0-1 linear) programming) which optimizes tamping operations in ballasted track as preventive maintenance.

Meier-Hirmer et al. (2009) developed a maintenance strategy model comprising of three sub-models namely an intervention efficiency model, a gamma process track deterioration model and a maintenance cost model. This model was used to establish the long-term costs of various maintenance strategies and optimize these costs based on various parameters such as intervention threshold or inspection interval. The authors observed that the maintenance efficiency or recovery appeared to be normally distributed and employed linear regression to characterize the intervention benefit which was assumed to be dependent on the deterioration prior to intervention. Famurewa et al. (2013) developed an empirical regression model for recovery after tamping intervention based on previous (longitudinal level) data on examined routes. The empirical

recovery model was combined with an exponential track degradation model to optimize the tamping intervention schedule through the minimization of the total intervention cost particularly the track possession cost.

Wen et al. (2016) evaluated the tamping recovery based on both the track condition before tamping and the frequency/number of previously performed tamping procedures. This restoration model was subsequently employed in a Mixed Integer Linear Programming (MILP) model formulated for the scheduling optimization of preventive condition-based tamping through the minimization of net present costs considering several factors. Caetano and Teixeira (2016) evaluated the effect of the age of track sections (segments) operations on tamping recovery by comparing renewed sections (ages of approximately 10 years) and nonrenewed sections (approximately 20 years). Despite the variation in track geometry deterioration rates due to loss of tamping effectiveness, the average number of maintenance tamping procedures were found to be greater in older track sections. This is similar to findings by Audley and Andrews (2013).

Khouy et al. (2012) evaluated the effectiveness of tamping by examining the track condition (longitudinal level) before and after tamping which was subsequently categorized using a tamping intervention graph into bad, good or excellent in relation to the level of improvement in track condition after maintenance. A large proportion of the sections were found to be either in the good or bad category. Due to the high variation in recovery observed, factors such as the effect of ballast age on tamping efficiency were evaluated. However, no clear effect of ballast age was noted contrary to findings by Caetano and Teixeira (2015) and Audley and Andrews (2013). Soleimanmeigouni et al. (2017) proposed two-level piecewise linear model to characterize the track geometry recovery and deterioration with possible spatial dependencies within deterioration parameters captured using Autoregressive Moving Average models. Multivariate linear regression was employed to tie various explanatory variables with response variables such as recovery values and changes in deterioration rates after tamping. Tamping recovery was dependent on both track condition before tamping

and tamping type (partial or complete) with the interaction effect between the two covariates also considered.

2.5.2.2 Probabilistic Models

Linear regression models are highly popular due to their simplicity. However, they assume linear dependency and assume normality of the random variables and joint distribution. Non-normality transpires in various forms: non-normality of marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables (Yan, 2006; Attoh-Okine, 2013). Furthermore, in most cases there exists a high degree of uncertainty in recovery values even in instances where track quality is identical prior to tamping which cannot be accounted for using deterministic techniques. This variation is even higher at the end of the life-cycle than at the beginning. For this reason, probabilistic techniques have increasingly been employed to cater for this variation by assuming the recovery after tamping is a random variable with a given probability distribution. A unique distribution for the recovery values after tamping is selected given a group of influencing variables with the parameters (or measures) of the distribution assumed to be a function of the inputs (Soleimanmeigouni et al., 2016b).

Quiroga and Schnieder (2012) developed a simulation approach for modelling the recovery and degradation of track geometry. The stochastic model statistically characterizes the phenomena given historical data and employs Monte Carlo method to attain simulated process realizations. The tamping recovery was assumed to be dependent on the number of accumulated tamping interventions. The track quality (longitudinal mean deviation) after tamping was assumed to be lognormally distributed stochastic variable dependent on the number of accumulated tamping interventions. It was observed that the variance of the track quality (longitudinal mean deviation) after tamping increased with greater number of accumulated tamping interventions. It was also observed that the deterioration rate (quality loss rate) increased considerably after

each tamping intervention. Quiroga et al. (2012) combined the Monte Carlo Simulation approach developed by Quiroga and Schnieder (2012) and a heuristic algorithm for maintenance intervention planning to evaluate the optimization of two maintenance strategies namely adaptive (dynamic) and constant intervention thresholds.

Audley and Andrews (2013) evaluated the effect of tamping on track geometry condition degradation taking into consideration two probability distributions which characterize the track quality for periods between tamping. Firstly, the authors analyzed the distributions of times for the track geometry to degrade to specified states or levels of performance following tamping given the line speed and the maintenance history. The two-parameter Weibull distribution was found to best model the times to degradation despite the better fit of its three-parameter counterpart since the extra parameter (location parameter or failure-free parameter) provided a better fit but no physical reason to justify a non-zero location parameter. Results of the analysis corroborated the theory that tamping damages the ballast and results in faster deterioration of the track geometry which was evident by the reduction of the characteristic life parameter with the frequency of tamping interventions. Additionally, it was observed that the more the track geometry degrades, the greater the rate of degradation which was evident by the increase in the shape parameter with track quality measurement (standard deviation of the vertical alignment). Secondly, the authors analyzed the track geometry quality after intervention. Despite the three-parameter lognormal distribution having the best fit, two-parameter lognormal distribution with a slightly lower fit was selected due to its ease of use to model the recovery values after tamping (probability of achieving the track quality condition after tamping) given operational speeds and maintenance history. Tamping efficiency was found to decrease with increasing number of accumulated tamping interventions which provides further proof that tamping damages ballast. Tamping efficiency was also found to reduce with increase in operational speed.

Soleimanmeigouni et al. (2016a) evaluated the effect of tamping on several (different) track geometry parameters such as surface (longitudinal) level, alignment and

crosslevel (cant) analyzing both the tamping recovery as well as the change in degradation rate after tamping. A probabilistic model was used to model tamping recovery of the geometry parameters which was assumed to be dependent on the track geometry condition prior to tamping. The track geometry deterioration was modelled using linear regression and Wiener process. The recovery values of the crosslevel (cant) and alignment were assumed to follow a three-parameter lognormal distribution with the recovery values of the surface profile was assumed to follow a three-parameter Weibull distribution. Tamping was found to have a negative effect (impact) on the deterioration rate with the increase in the degradation rate evident by the observed increase in the regression slope and drift coefficient of the Wiener process. Complete and partial tamping interventions were also clustered and examined separately since they have different effects on track geometry condition. Complete tamping interventions were found to have a considerably greater effect on track geometry condition compared to partial tamping. Additionally, a linear correlation analysis conducted showed a moderate dependence between the recovery of surface (longitudinal) level and that of the crosslevel (cant) and a weak dependence between the surface (longitudinal) level and that of the alignment. However, Pearson's correlation coefficient assumes linear dependence between the random variables and assumes normality of these random variables and their joint distribution. Thus, it will be more appropriate to employ concordance measures which are suitable for measuring both linear and non-linear dependence. These measures are scale-invariant and measure dependence irrespective of assumed distributions.

In summary, the vast majority of tamping recovery models do not take into consideration the underlying dependence between the variables of interest which may exhibit tail dependency, asymmetric dependence and other non-linear dependencies. However, copula-based approaches take into account these nonlinearities by allowing for the separate modeling of the arbitrary univariate marginal distributions and the dependence structure which are subsequently combined to form a joint distribution with the underlying dependence.

2.6 Derailment Severity

2.6.1 General

Despite the relatively low frequency of train derailments, they have been a major concern due to their high consequence justifying the need to critically examine the severity of train derailments in order to minimize and mitigate the resulting damage (Jeong et al., 2007; Liu et al., 2013). Derailments may result in loss of life and property, interruption of services and destruction of the environment (Liu et al., 2013), and are the most frequent kind of Federal Railroad Administration (FRA)-reportable mainline train accident in the United States (Barkan et al., 2003; Liu et al., 2012; Liu, 2015). Derailments made up about three-quarters of freight-train accidents in the United States from 2001 to 2010. Therefore, analyzing the magnitude and variability of derailment severity is as important as estimating the likelihood of derailment (Liu et al., 2013).

Derailment severity may be influenced by factors like car mass, derailment speed, residual train length (number of cars after the point of derailment), derailment cause, ground friction, rail friction, derailment cause, proportion of loaded railcars in the train (loading factor) and train power distribution. Estimation of these variables is often established through exact estimation or the determination of statistical distributions and time history of the examined factors (Mohammadzadeh and Ghahremani, 2010).

Metrics that may be used to assess the severity of train derailments include the number of derailed cars (Nayak et al., 1983; Saccomanno and El-Hage, 1989, 1991; Toma, 1998; Barkan et al., 2003; Anderson, 2005; Liu et al., 2011, 2012, 2013), monetary damage (Barkan et al., 2003; Liu, 2015) or casualties (Liu, 2015). The number of derailed cars is the most suitable and most popular metric for evaluating severity. The term "cars" is generically used and refers to all vehicles including railcars and locomotives (unless categorically stated otherwise) (Liu et al., 2013). Monetary damage is prone to considerable variations due to factors such as cost difference between locomotives and railcars and differences in repair cost between regular track and special track such as turnouts and crossings (Barkan et al., 2003; Liu et al., 2013).

Casualties on the other hand are more appropriate when dealing with solely passenger train derailments. Other derailment outcome measures studied include duration of the event, peak collision forces, cars involved in peak collision, accident scene dimensions, maximum closing velocities (i.e., relative velocities between impacting cars), and peak coupler forces (Yang et al., 1972; Toma, 1998; Jeong et al., 2007).

The severity of a train derailment is influenced by factors such as derailment speed, derailment cause, residual train length, derailment cause, ground friction, rail friction, car mass, proportion of loaded railcars in the train (loading factor) and train power distribution. FRA track classes specify certain characteristics associated with track quality. Thus, track class is representative of track quality and minimum standards are specified by regulation for each class with higher classes having more stringent requirements (Anderson and Barkan, 2004, 2005). Due to the lack of a more appropriate batch of causal parameters for track quality as well as its ubiquitous usage in the American railroad industry, track class has been employed as proxy variable for statistical estimation of derailment probability (Nayak et al., 1983; Dennis, 2002; Anderson and Barkan, 2004, 2005) as well as derailment severity (Anderson and Barkan, 2005). Despite the apparent positive strong correlation between speed and track class, Anderson and Barkan (2005) found no clear relationship on average between higher track class and greater number of derailing cars. The authors highlighted the majority of derailments on higher track class being initiated at less than normal operational speeds or variations in derailment severity for different accident causes, which are likely correlated with track quality as possible reasons for this. Simulation models and statistical analysis are the two main methods of modeling train derailment severity.

2.6.2 Derailment Severity Models

2.6.2.1 Simulation Models

Simulation models are commonly built on comprehensive non-linear wheel-rail interaction models. These mechanistic models visualize the reaction of railroad vehicles to certain operational and environmental conditions (Liu et al., 2013). Yang et al.

(1972) developed an analytical simulation model with the point of derailment as the only initial assumption. This model was used to investigate the influence of various factors such as train length, derailment speed, ground friction, coupler moment, braking, car length and car weight on the behavior and severity of train derailments. All the aforementioned factors were found to influence the number of derailed cars with the exception of coupler moment characteristics which were found to have negligible effect.

An improved simulation model published by Anderson (1994) called DERAIL was originally part of an overall derailment disaster model (Coppens et al., 1988; Birk et al., 1990a,b) called Derailment Accident Simulation (DERACS) comprising of several sub-models which simulate the consequences of train derailments. Unlike the previous model developed by Yang et al. (1972), the train derailment simulation software package allowed for coupler failure, vehicle uncoupling; vehicle roll, collision of cars, independent car motion and modeling of curved track. However, this improved model was found to suffer frequent numerical instabilities. Simulation models presented by Johnson (1991), Guran et al. (1992), Gracie (1991) and Roorda and Gracie (1992) examined the fundamental kinematics of the derailment process but did not enhance existing state-of-the-art models (Toma, 1998). However, Roorda et al. (1993) related the number of derailed cars to the number of loaded cars in the train, the latter being representative of both overall train length and weight.

Toma (1998) developed a comprehensive planar model which included a detailed rail car and coupler model, ground reaction force model, collision model, and allowed uncoupling and derailment of cars, which were all improvements upon previous models. The model was based on coupled sets of 5 degrees of freedom sub-system models for each rail car. The model investigated the effect of train speed, car mass, number of cars, braking force, ground reaction force, and derailment quotient on outcome measures such as the number of derailed cars, duration of the event, peak collision force, cars involved in peak collision, and accident scene dimensions. A composite measure, called accident severity was formulated based on the number of derailed cars,

the peak collision force, and the accident scene dimensions. Train speed, car mass and train length were found to have significant effect on the number of derailed cars and the peak collision force with braking force and ground reaction force having relatively little effect and derailment quotient having negligible effect.

Commercially available simulation models such as Dynamic Analysis Design Simulation (DADS) (Han and Koo, 2003) and Automatic Dynamic Analysis of Mechanical Systems (ADAMS) (Paetsch et al., 2006) have also been used to analyze the severity of train derailments. Jeong et al. (2007) developed a purpose-built simulation model to investigate the influence of various variables such as train length, car mass, initial translational and rotational velocities, and coefficients of friction on the derailment outcomes. Outcomes considered include the number of derailed cars, maximum closing velocities (i.e., relative velocities between impacting cars), and peak coupler forces. The computational times to run this model was found to be significantly less than commercial models such as ADAMS model (minutes versus hours).

2.6.2.2 Statistical Analytical Models

Statistical analysis of train derailment severity are conducted using historical derailment data. Estimation of these variables is often established through exact estimation or the determination of statistical distributions and time history of the examined factors.

Nayak et al. (1983) proposed a positive non-linear relationship between derailment severity (in terms of average number of derailed cars) and derailment speed (expressed in mph). They expressed the mean number of derailed cars as a function of the square root of the derailment speed.

$$M_{nd} = 1.7\sqrt{Speed} \tag{2.1}$$

where M_{nd} is the mean number of derailed cars. The derailment speed expresses the volume of kinetic energy produced during the derailment that has to be dispersed prior

to the re-establishment of the car-track stability (Bagheri et al., 2011). All other things being equal, the greater the derailment speed the higher the number of cars derailing.

Saccomanno and El-Hage (1989, 1991) proposed an equation for estimating the mean number of derailed cars based on a truncated geometric distribution which takes into account the joint effects of accident cause, derailment speed and residual train length. Subsequent work by Anderson and Barkan (2005) revealed inaccuracies in the proposed equation for which modifications were made (equation 2) to ensure that the number of derailed cars lied within the range of one and the residual train length. Residual train length can be defined as the number of cars after the point of derailment (POD).

For the same reason, Bagheri (2009) and Bagheri et al. (2011) also made modifications to the model proposed by Saccomanno and El-Hage (1989, 1991) such that for any train length L and position j, the truncated geometric distribution for the probability of k cars derailing is given by

$$P_r(x\ cars\ derailing \mid POD\ at\ position\ j) = \begin{cases} \frac{p(1-p)^{x-1}}{1-(1-p)^{L_r}} & if x = 1, ..., L_r \\ 0, & otherwise \end{cases} \tag{2.2}$$

Where $L_r = L - j + 1$ is the residual length (i.e. the number of cars after POD), and $1 - p$ is the probability of derailment given a position after POD.

Both modifications resulted in similar models for the mean number of cars derailing (M_{nd}) expressed as

$$D = \frac{1}{p} - \frac{L_r(1-p)^{L_r}}{1-(1-p)^{L_r}} \tag{2.3}$$

Where D - the mean of the truncated geometric distribution (i.e. the estimated number of cars derailed), p = logistic function of continuous "success" probability.

The probability, p, is assumed to be related to the factors/ covariates through

the logit link function

$$\frac{1}{1+e^{-z}} \tag{2.4}$$

where z is a linear function of speed, residual length and derailment causes.

Liu et al. (2013) investigated the effects of train power distribution and proportion of loaded railcars in the train (or loading factor) on derailment severity. The study revealed that a higher loading factor corresponds to greater kinetic energy during a derailment resulting in a greater derailment severity, ceteris paribus. The study also revealed that a derailed train with a higher loading factor is more likely to have distributed train power. The authors proposed quantile regression analysis of derailment severity in which they investigated other distributional statistics such as conditional quantiles in order to provide further comprehension of the derailment severity distribution. Prior to this, all previous models had focused on mean derailment severity analysis. The authors also proposed a zero-truncated negative binomial (ZTNB) model which is expressed as follows:

$$Z = exp\left[1.38 - 0.03R - 0.26S - 0.57L - 0.06\left(R^2\right) + 0.24R \times S + 0.21 \times L\right] \tag{2.5}$$

where Z is the estimated number of derailed cars, R is the logarithmic residual train length, S is the logarithmic derailment speed and L is the loading factor. The ZTNB model was found to result in a greater likelihood and mean value of the response variable (number of cars derailing) in comparison to traditional count data models such as Poisson and negative binomial models.

Majority of the existing literature have failed to consider the multivariate nature of derailment severity and have instead focused primarily on only one severity outcome namely the number of derailed cars. However, it is also important to concurrently analyze the monetary damage incurred by railroads during derailments. A multivariate

derailment severity model can be developed to jointly model multiple severity outcomes given a set of covariates taking into consideration the underlying dependence between the responses. To achieve this, a copula-based regression model of number of derailed cars and monetary damage is proposed for their joint analysis with a set of covariates that might influence both responses. Copulas used in a multivariate regression framework have been found to addresses endogeneity due to similar unobserved or omitted variables that may affect both outcomes. Furthermore, majority of the statistical analytical models do not consider the underlying dependence between the various variables of interest which may exhibit nonlinear dependence, tail dependence or asymmetric dependence. These models are also not flexible in evaluating high dimensional dependence structures. To address these limitations, a vine-copula based approach is proposed to model the high-dimensional dependence between the derailment severity variables.

2.7 Key Observations from the Literature

Based on the literature review, the following conclusions can be made:

- The main track geometry parameters used to evaluate track geometry quality include surface (longitudinal level, profile or vertical alignment), alignment (horizontal alignment), gage (gauge), cross level (cant) and warp (twist).
- Track geometry condition can be assessed by indicators such as the standard deviation (SD) over a specified length, mean value of the section or extreme (peak) values of isolated defects of the track geometry parameters.
- Infrastructure managers often combine these parameters (defects) into an artificial track quality index (TQI) as a representative measure of the different track geometry parameters and is employed as a decisive metric for maintenance planning. However, standard deviation of the short wavelength variation of the surface parameter has been regarded as the most decisive criterion for maintenance decisions. Disregarding the other parameters during the evaluation of

track geometry quality may lead to erroneous assessment resulting in ineffective maintenance planning.

- Conventional tamping is the most common track geometry maintenance activity. However, improved track geometry maintenance activities such as design over-lift tamping, enhanced tamper control systems and stone blowing are recommended. Design over-lift tamping is considered as best tamping practice.
- Tamping recovery models can be classified into two main categories namely deterministic models and probabilistic models. The choice of methodology is based on the degree of degree of uncertainty in the recovery values after tamping. Deterministic models are employed given low uncertainty whereas probabilistic models are utilized given high uncertainty.
- Deterministic models such as linear regression models are widely used because of their simplicity. However, linear regression models assume linear dependence and assume normality of the random variables and their joint distribution.
- Furthermore, in most cases there exists a high level of uncertainty in recovery values even in instances where track condition is identical before tamping. This variation is even higher at the end of the life-cycle than at the beginning. For this reason, probabilistic models are increasingly being utilized to cater for this variation by assuming the recovery after tamping is a random variable with a given probability distribution.
- The vast majority of tamping recovery models do not take into consideration the underlying dependence between the variables of interest which may exhibit tail dependency, asymmetric dependence and other non-linear dependencies.
- Derailment severity models can be classified into two main groups namely simulation (mechanistic) models and statistical analysis models.
- Most statistical models do consider the multivariate nature of derailment severity and have instead focused mainly on only one severity outcome namely the number of derailed cars. However, it is also important to concurrently analyze the monetary damage incurred by railroads during derailments. To simultaneously

model multiple severity outcomes, a copula-based regression model is proposed for their joint analysis given a set of covariates taking into account the underlying dependence between the outcomes.

- Furthermore, most of these statistical models do not take into consideration the underlying dependence between the variables which may exhibit nonlinear dependence, tail dependence or asymmetric dependence. These models are also not flexible in evaluating high dimensional dependence structures. To address these limitations, a vine-copula model is proposed to evaluate the high-dimensional dependence between the variables of interest.

REFERENCES

Anderson, R. J. DERAIL. A Train Derailment Simulation Software Package. Dynamics Laboratory Report No. DL/94/RJA/4. Technical report, Queen's University, Kingston, Canada, 1994.

Anderson, Robert. T. and Barkan, Christopher P. L. Railroad Accident Rates for Use in Transportation Risk Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 1863(-1):88–98, 2004. ISSN 0361-1981. doi: 10.3141/1863-12.

Anderson, Robert Thomas. *Quantitative analysis of factors affecting railroad accident probability and severity*. PhD thesis, University of Illinois at Urbana-Champaign,, 2005.

Anderson, RT and Barkan, CPL. Derailment probability analysis and modeling of mainline freight trains. In *Proceedings of the 8th International Heavy Haul Railway Conference*, 2005.

Andrade, A. R. and Teixeira, P. F. Biobjective Optimization Model for Maintenance and Renewal Decisions Related to Rail Track Geometry. *Transportation Research Record: Journal of the Transportation Research Board*, 2261(-1):163–170, 2012. ISSN 0361-1981. doi: 10.3141/2261-19.

Andrade, A. R. and Teixeira, P. F. Hierarchical Bayesian modelling of rail track geometry degradation. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(4):364–375, 2013. ISSN 0954-4097. doi: 10.1177/0954409713486619.

Attoh-Okine, Nii O. Pair-copulas in infrastructure multivariate dependence modeling. *Construction and Building Materials*, 49:903–911, 2013. ISSN 09500618. doi: 10.1016/j.conbuildmat.2013.06.055.

Attoh-Okine, Nii O. *Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering*. John Wiley & Sons, Inc., 2017.

Audley, M. and Andrews, J. The effects of tamping on railway track geometry degradation. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(4):376–391, 2013. ISSN 0954-4097. doi: 10.1177/0954409713480439.

Bagheri, Morteza. *Risk-Based Model for Effective Marshalling of Dangerous Goods Railway Cars*. PhD thesis, University of Waterloo, Ontario, 2009.

Bagheri, Morteza; Saccomanno, Frank; Chenouri, Shojaeddin, and Fu, Liping. Reducing the threat of in-transit derailments involving dangerous goods through effective placement along the train consist. *Accident Analysis and Prevention*, 43(3):613–620, 2011. ISSN 00014575. doi: 10.1016/j.aap.2010.09.008.

Barkan, Christopher P. L.; Dick, C. Tyler, and Anderson, Robert. T. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 1825(9):64–74, 2003. ISSN 03611981.

Birk, A. M.; Anderson, R. J., and Coppens, A. J. A computer simulation of a derailment accident Part I - Model Basis. *Journal of Hazardous Materials*, 25(1-2):121–147, 1990a. ISSN 03043894. doi: 10.1016/0304-3894(90)85075-E.

Birk, A. M.; Anderson, R. J., and Coppens, A. J. A computer simulation of a derailment accident Part II - sample simulation. *Journal of Hazardous Materials*, 25(1-2):149–165, 1990b.

Caetano, Luis Filipe and Teixeira, Paulo Fonseca. Optimisation model to schedule railway track renewal operations: a life-cycle cost approach. *Structure and Infrastructure Engineering*, 11(11):1524–1536, 2015. ISSN 1573-2479. doi: 10.1080/15732479.2014.982133.

Caetano, Luis Filipe and Teixeira, Paulo Fonseca. Predictive Maintenance Model for Ballast Tamping. *Journal of Transportation Engineering*, 142(4):4016006, 2016. ISSN 0733-947X. doi: 10.1061/(ASCE)TE.1943-5436.0000825.

Coppens, A.J.; Wong, J.D.E; Bibby, A.; Birk A.M., , and Anderson R.J., . Development of a Derailment Accident Computer Simulation Model. Technical report, Transport Canada Report No. TP 9254E, 1988.

Dennis, Scott M. Changes in railroad track accident rates. *Journal of the Transportation Research Forum*, 56(4), 2002.

Esveld, Coenraad. *Modern Railway Track, 2nd Edition*. 2001. ISBN 9080032433.

Famurewa, S. M.; Xin, T.; Rantatalo, M., and Kumar, U. Optimisation of maintenance track possession time: A tamping case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(1):12–22, 2013. ISSN 0954-4097. doi: 10.1177/0954409713495667.

Famurewa, S. M.; Juntti, U.; Nissen, A., and Kumar, U. Augmented utilisation of possession time: Analysis for track geometry maintenance. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 230(4):1118–1130, 2016. ISSN 0954-4097. doi: 10.1177/0954409715583890.

Galvan-Nunez, Silvia. *Hybrid Bayesian-Wiener Process in Track Geometry Degradation Analysis*. PhD thesis, University of Delaware, 2017.

Gracie, B.J. *Train Derailment Mechanics*. PhD thesis, University of Waterloo, Canada, 1991.

Guran, A.; Vakakis, A., and Ossia, K. Effect of mass distribution on jack-knifing of a train of cars. In *Proceedings of CSME Forum. "Transport 1992+"*, volume 3, pages 710–713, 1992.

Gustavsson, Emil. Scheduling tamping operations on railway tracks using mixed integer linear programming. *EURO Journal on Transportation and Logistics*, 4(1):97–112, 2015. ISSN 2192-4376. doi: 10.1007/s13676-014-0067-z.

Han, Hyung-Suk and Koo, Jeong-Seo. Simulation of Train Crashes in Three Dimensions. *Vehicle System Dynamics*, 40(6):435–450, 2003. ISSN 0042-3114. doi: 10.1076/vesd.40.6.435.17906.

He, Qing; Li, Hongfei; Bhattacharjya, Debarun; Parikh, Dhaivat P, and Hampapur, Arun. Track geometry defect rectification based on track deterioration modelling and derailment risk assessment. *Journal of the Operational Research Society*, 66(3): 392–404, 2015. ISSN 0160-5682. doi: 10.1057/jors.2014.7.

Jeong, D.Y.; Lyons, M.L.; Orringer, O, and Perlman, A.B. Equations of motion for train derailment dynamics. *Proceedings of the 2007 ASME Rail Transportation Division Fall Technical Conference, September 11-12, 2007 Chicago, IL*, RTDF2007-4: 1–7, 2007. ISSN 10788883. doi: 10.1115/RTDF2007-46009.

Johnson, W.A. Simple model for the jack-knifing of a train of coaches and Samuel Vince (1749-1821). *International Journal of Mechanical Engineering Education*, 19 (3):159–169, 1991.

Khouy, Iman Arasteh. *Cost-Effective Maintenance of Railway Track Geometry*. PhD thesis, Lulea University of Technology, 2013.

Khouy, Iman Arasteh K; Schunnesson, Hakån; Nissen, Arne, and Juntti, Ulla J. Evaluation of track geometry degradation in swedish heavy haul railroad - A case study. *International Journal of COMADEM*, 15(2):11–16, 2012. ISSN 13637681. doi: 10.1177/0954409713482239.

Le Pen, Louis; Watson, Geoff; Powrie, William; Yeo, Graeme; Weston, Paul, and Roberts, Clive. The behaviour of railway level crossings: Insights through field monitoring. *Transportation Geotechnics*, 1(4):201–213, 2014. ISSN 22143912. doi: 10.1016/j.trgeo.2014.05.002.

Li, Dingqing; Hyslip, James P.; Sussmann, Theodore R., and Chrismer, S. M. *Railway geotechnics*. CRC Press, New York, 2015. ISBN 9780415695015.

Lichtberger, Bernhard. *Track compendium : formation, permanent way, maintenance, economics*. Eurailpress, 2005. ISBN 3777103209.

Liu, Xiang. Statistical Temporal Analysis of Freight-Train Derailment Rates in the United States : 2000 to 2012. 2476(1):119–125, 2015.

Liu, Xiang; Barkan, Christopher, and Saat, M. Analysis of Derailments by Accident Cause. *Transportation Research Record: Journal of the Transportation Research Board*, 2261:178–185, 2011. doi: 10.3141/2261-21.

Liu, Xiang; Saat, M., and Barkan, Christopher. Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*, 2289(2289):154–163, 2012. ISSN 0361-1981. doi: 10.3141/2289-20.

Liu, Xiang; Saat, M. Rapik; Qin, Xiao, and Barkan, Christopher P L. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. *Accident Analysis and Prevention*, 59:87–93, 2013. ISSN 00014575. doi: 10.1016/j.aap.2013.04.039.

Meier-Hirmer, C; Riboulet, G; Sourget, F, and Roussignol, M. Maintenance optimization for a system with a gamma deterioration process and intervention delay: application to track maintenance. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 223(3):189–198, 2009. ISSN 1748-006X. doi: 10.1243/1748006XJRR234.

Miwa, Masashi. Mathematical Programming Model Analysis for the Optimal T rack Track Maintenance Schedule. *Quart Rep RTRI*, 43(3), 2002.

Mohammadzadeh, Saeed and Ghahremani, Soodabeh. Estimation of train derailment probability using rail profile alterations. *Structure and Infrastructure Engineering*, 2479(August 2013):1–20, 2010. ISSN 1573-2479. doi: 10.1080/15732479.2010.500670.

Mohammadzadeh, Saeed; Miri, Amin, and Nouri, Mehrdad. Assessing ballast cleaning as a rehabilitation method for railway masonry arch bridges by dynamic load tests. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 0(0):095440971771004, 2017. ISSN 0954-4097. doi: 10.1177/0954409717710047.

Nayak, P Ranganath; Rosenfield, Donald B; Hagopian, John H; Lillie, Arthur D, and Park, Acorn. Event Probabilities and Impact Zones for Hazardous Materials Accidents on Railroads. Technical report, Report DOT/FRA/ORD- 83/20. FRA, U.S. Department of Transportation, 1983.

Oyama, Tatsuo and Miwa, Masashi. Mathematical modeling analyses for obtaining an optimal railway track maintenance schedule. *Japan Journal of Industrial and Applied Mathematics*, 23(2):207–224, 2006. ISSN 0916-7005. doi: 10.1007/BF03167551.

Paetsch, C. R.; Perlman, A. B., and Jeong, D. Y. Dynamic Simulation of Train Derailments. In *Rail Transportation*, volume 2006, pages 105–114. ASME, 2006. ISBN 0-7918-4778-0. doi: 10.1115/IMECE2006-14607.

Parvez, Ahsan and Foster, Stephen James. Fatigue of steel-fibre-reinforced concrete prestressed railway sleepers. *Engineering Structures*, 141:241–250, 2017. ISSN 18737323. doi: 10.1016/j.engstruct.2017.03.025.

Quiroga, L. M. and Schnieder, E. Monte Carlo simulation of railway track geometry deterioration and restoration. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(3):274–282, 2012. ISSN 1748-006X. doi: 10.1177/1748006X11418422.

Quiroga, L. M.; Schnieder, E., and Antoni, M. Holistic long term optimization of maintenance strategies on ballasted railway track. In *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012*, 2012.

Roorda, J. and Gracie, B.J. Train Derailment Mechanics A Simple Model. *Proceedings of Canadian Society for Mechanical Engineering CSME Forum "Transport 1992+*, 3:714–719, 1992.

Roorda, J; Gracie, B; Energy, Atomic; Limited, Canada, and River, Chalk. Derailment of trains. *International Journal of Mechanical Engineering Education*, 22(3):165–176, 1993.

Saccomanno, F. F. and El-Hage, S. M. Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. *Transportation Research Record*, (1245):34–51, 1989.

Saccomanno, F. F. and El-Hage, S. M. Establishing derailment profiles by position for corridor shipments of dangerous goods. *Canadian Journal of Civil Engineering*, 18 (1):67–75, 1991. ISSN 0315-1468. doi: 10.1139/l91-009.

Scanlan, Kirk M; Hendry, Michael T, and Martin, C Derek. Evaluating the impact of ballast undercutting on the roughness of track geometry over different subgrade conditions. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 0(0):095440971772034, 2017. ISSN 0954-4097. doi: 10.1177/0954409717720347.

Selig, E. T. (Ernest Theodore) and Waters, John M. *Track geotechnology and substructure management*. T. Telford, 1994. ISBN 0727720139.

Silvast, M; Nurmikolu, A; Wiljanen, B, and Levomaki, M. An inspection of railway ballast quality using ground penetrating radar in Finland. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 224(5): 345–351, 2010. ISSN 0954-4097. doi: 10.1243/09544097JRRT367.

Sol-Sánchez, Miguel; Moreno-Navarro, Fernando; Martínez-Montes, German, and Rubio-Gámez, Mª Carmen. An alternative sustainable railway maintenance technique based on the use of rubber particles. *Journal of Cleaner Production*, 142: 3850–3858, 2017. ISSN 09596526. doi: 10.1016/j.jclepro.2016.10.077.

Soleimanmeigouni, I.; Ahmadi, A.; Arasteh Khouy, I., and Letot, C. Evaluation of the effect of tamping on the track geometry condition: A case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:408–420, 2016a. ISSN 0954-4097. doi: 10.1177/0954409716671548.

Soleimanmeigouni, I.; Ahmadi, A., and Kumar, U. Track geometry degradation and maintenance modelling: A review. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:73–102, 2016b. ISSN 0954-4097. doi: 10.1177/0954409716657849.

Soleimanmeigouni, Iman; Xiao, Xun; Ahmadi, Alireza; Xie, Min; Nissen, Arne, and Kumar, Uday. Modelling the evolution of ballasted railway track geometry by a two-level piecewise model. *Structure and Infrastructure Engineering*, 2479(June):1–13, 2017. ISSN 1573-2479. doi: 10.1080/15732479.2017.1326946.

Tennakoon, Nayoma Chulani. *Geotechnical study of engineering behaviour of fouled ballast*. PhD thesis, University of Wollongong, Australia, 2012.

Toma, Elton Edward. *A Computer Model of a Train Derailment*. PhD thesis, Queen's University, Kingston, Ontario, Canda, 1998.

Vale, Cecília and Ribeiro, Isabel M. Railway condition-based maintenance model with stochastic deterioration. *Journal of Civil Engineering and Management*, 20(5):686–692, 2014. ISSN 1392-3730. doi: 10.3846/13923730.2013.802711.

Vale, Cecília; Ribeiro, Isabel M, and Calçada, Rui. Integer Programming to Optimize Tamping in Railway Tracks as Preventive Maintenance. *Journal of Transportation Engineering*, 138(January):123–132, 2012. ISSN 0733947X. doi: 10.1061/(ASCE)TE.1943-5436.0000296.

Veit, Peter. Track quality: luxury or necessity? Technical Report July, 2007.

Wen, M.; Li, R., and Salling, K. B. Optimization of preventive condition-based tamping for railway tracks. *European Journal of Operational Research*, 252(2):455–465, 2016. ISSN 03772217. doi: 10.1016/j.ejor.2016.01.024.

Yan, Jun. Multivariate Modeling with Copulas and Engineering Applications. In *Springer Handbook of Engineering Statistics*, pages 973–990. Springer London, London, 2006. doi: 10.1007/978-1-84628-288-1_51.

Yang, T H; MANOS, W P, and Johnstone, B. Dynamic analysis of train derailments. In *1972 Winter Annual Meeting of ASME*, 1972.

Zhao, Jianmin; Chan, Andrew H C, and Stirling, Alan B. Risk analysis of derailment induced by rail breaks-a probabilistic approach. *Annual Reliability and Maintainability Symposium*, 00(C):486–491, 2006.

Chapter 3

DATA SOURCES AND EXPLORATORY DATA ANALYSIS

3.1 Introduction

Prior to the implementation of the copula-based methodologies, exploratory data analysis was conducted on the datasets. Exploratory Data Analysis (EDA) initially championed by Tukey (1977) offers conceptual and computational instruments for uncovering patterns to support hypothesis development and refinement and supplements confirmatory data analysis (CDA) which employs significance and hypothesis testing (Behrens, 1997).

EDA provides further insight into the datasets uncovering patterns, data characteristics, relationships and underlying structure in the data through visualization without making any initial assumptions. Thus, EDA helps corroborate any assumptions that are made in the formulation of the problem or that are needed when implementing certain methodologies. In this case, EDA allows one to ascertain the non-normality of marginal and joint distributions of the various variables in the data as well as the underlying dependences between the variables making copulas a suitable methodology. Furthermore, EDA also allows for the identification of essential variables, missing data, outliers and anomalies.

3.2 Data Set Description

3.2.1 Track Geometry Data Set

Track geometry inspection data was obtained from a Class I U.S. railroad which contains information on various track geometry parameters. One mile of track was used for the analysis. Data was measured and collected for every 1 foot of track using a track geometry car. The track geometry car records several geometry parameters and

non-geometric attributes. However, the surface, alignment, cross level, gage and warp (including their wavelength variations) were used for the exploratory data analysis. 62-foot and 124-foot wavelength variations of the surface and alignment parameters of the left and right rails were considered as well as the 62-foot variation of the warp parameter. Thus, a total of 11 track geometry variations were initially investigated during EDA of the raw data. The inspection data used in this case study were from 28 inspection dates spanning the years 2013 to 2016. Figure 3.1 shows the spatial variation (foot-by-foot measurements) of the track geometry at a given inspection date along the track for the surface right (62-ft), alignment right (62-ft), cross level and warp (62-ft) parameters.

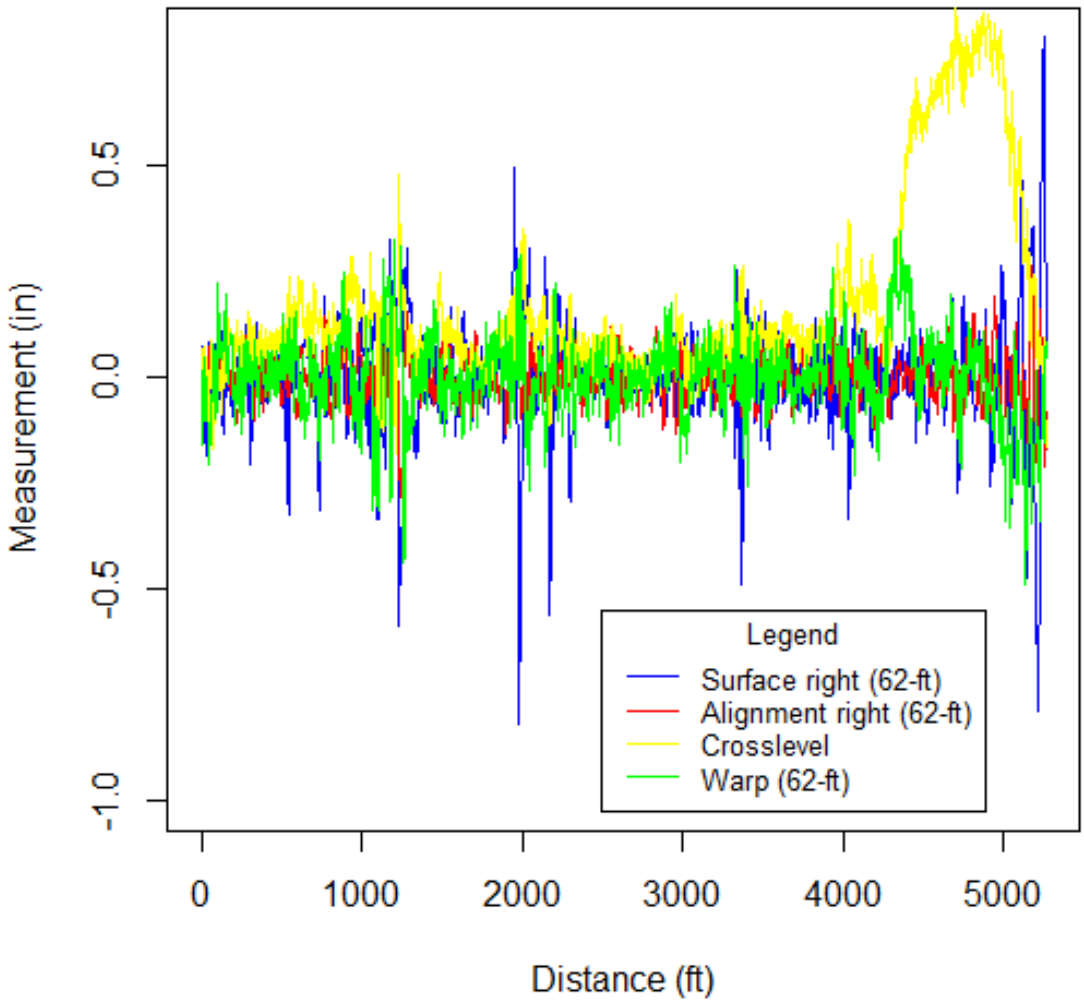


Figure 3.1: Illustration of spatial variation of some track geometry parameters at a given inspection date

Track geometry maintenance activities such as tamping were conducted to rectify track geometry deviations such as incorrect surface level (vertical deviation) and incorrect alignment (lateral deviation) by rearranging and compacting the ballast (Khouy, 2013; Audley and Andrews, 2013). Tamping results in a jump reduction

in the track geometry irregularity measurements and alters the track deterioration (Soleimanmeigouni et al., 2016a). However, these activities do not offer durable geometry rectification if the underlying cause of track deformation such as subgrade failure is not addressed. The subballast layer was strengthened through the placement of a geocell along 800 feet section of the track during track reconstruction. Track renewal activities are considered as intervention measures but are not categorized under track maintenance activities. Figure 3.2 shows an illustration of the surface right (62-ft) track geometry parameter at multiple inspection dates.

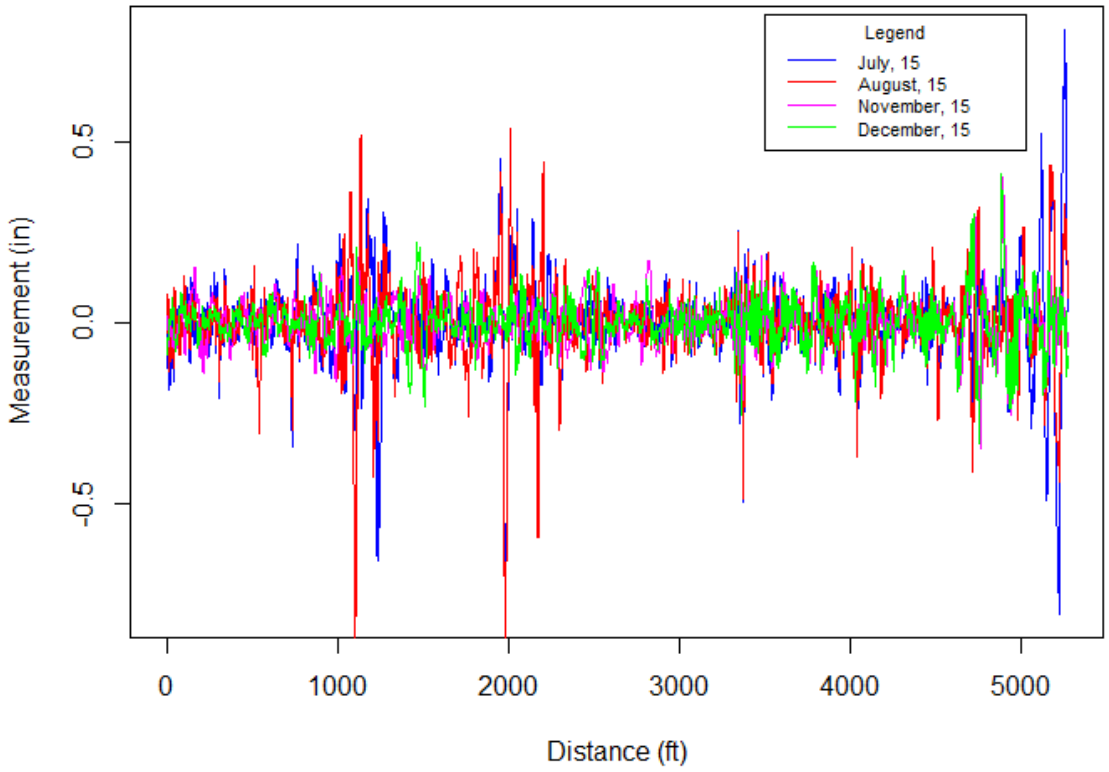


Figure 3.2: Illustration of surface right (62-ft) track geometry parameter at multiple inspecton dates

The inspection data characterized in the form of signals was initially cleaned and preprocessed. No missing data was identified in the dataset. Long wavelength track irregularities have an adverse influence on ride comfort. However, short wavelength irregularities generate more vibration on axles and wheels (Soleimanmeigouni et al., 2016b). Thus, short wavelengths have a much greater effect on ride quality (Audley and Andrews, 2013). Furthermore, the standard deviation of the short wavelength surface and alignment are the key parameters used to trigger preventive maintenance procedures (Caetano and Teixeira, 2015) and have been found to be good predictors of the maintenance needs for the rest of the track geometry parameters (Andrade and Teixeira, 2013). Thus, the 62-foot variations of the surface and alignment parameters were used in favor of their 124-foot counterparts. Subsequently, the surface and alignment parameters of the left and right rails were averaged as a representative of the whole track (Audley and Andrews, 2013) and were termed as "surface" and "alignment" respectively to avoid long wordy descriptions. The track quality index (TQI) represented by the standard deviation (SD) of each of the track geometry parameters namely surface, alignment, crosslevel, warp and gage was subsequently computed for track segments with 100 feet of length. The degradation and recovery plot for the various track geometry parameter for a given track segment is shown on figure 3.3. The tamping recovery values for each parameter were obtained by computing the difference between the standard deviation (SD) of the track geometry parameters before tamping and the corresponding standard deviation after tamping. Recovery values after tamping cannot be depicted appropriately by measurement data obtained with a long inspection interval and thus should not be considered (Soleimanmeigouni et al., 2016a).

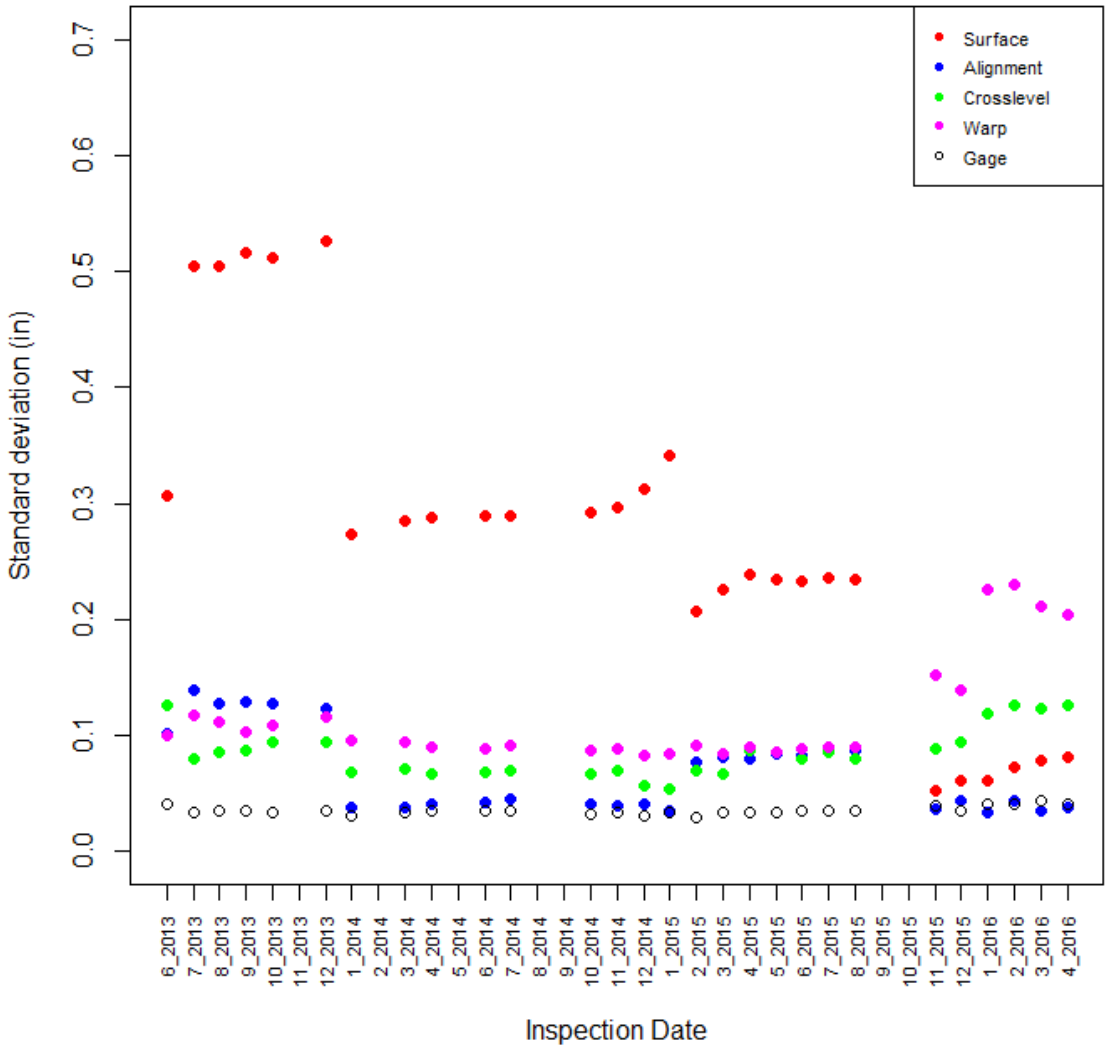


Figure 3.3: Degradation and recovery plot for various track geometry parameters at a given 100-foot track segment

3.2.2 Derailment Data Set

Data was obtained from the Rail Equipment Accident/Incident (REA) database managed by the Federal Railroad Administration (FRA) of U.S. Department of Transportation (U.S. DOT). A "rail equipment accident/incident" is a collision, derailment,

fire, explosion, act of God, or other event involving the operation of railroad on-track equipment (standing or moving). U.S. railroads are mandated to present detailed reports (Form 6180.54) to the FRA on all accidents or incidents whose damage costs exceed a specified monetary value. The damage incurred includes damage caused to the railroad track, signals, on-track equipment, track structures and roadbed as well as labor costs and the costs for acquiring new equipment and material. The reporting threshold is periodically altered to account for inflation and other adjustments and has increased from \$5700 in 1990 to \$10,700 in 2017 (FRA, 2016). The relatively low threshold results in most accidents being reported to the FRA (Barkan et al., 2003).

The database contains detailed track accident information such as accident cause, number of derailed cars, total monetary damage, track type, track class, train length and derailment speed. The database contains 4990 accidents and incidents for the year 2005. The breakdown of these accidents and incidents is shown in figure 3.4. Derailments made up the largest proportion of REAs (2614, about 52.4%) followed by "other impacts" (726, 14.5%), side collisions (347, 6.9%) and highway-rail collisions (295, 5.9%).

The types of train consists involved in these accidents include freight trains, passenger trains, commuter trains, works trains, yard/switching and maintenance/inspections cars. On the other hand, the types of tracks involved in these accidents include mainline track, yard, siding and industry. The breakdown of train consist types can be found in figure 3.5 whereas the breakdown of track types is presented in figure 3.6. Freight trains was the popular consist type (about 34.3%). On the other hand, about 49% of accidents/incidents occurred in the yard whereas about 30% occurred on mainline track.

Similar to previous derailment severity research, 690 freight-train derailments occurring on Class I mainline track in the year 2005 were initially considered for exploratory data analysis after cleaning and preprocessing of the data. The variables considered include monetary damage, the number of derailed cars, derailment speed, residual train length and proportion of loaded railcars in the train (loading factor).

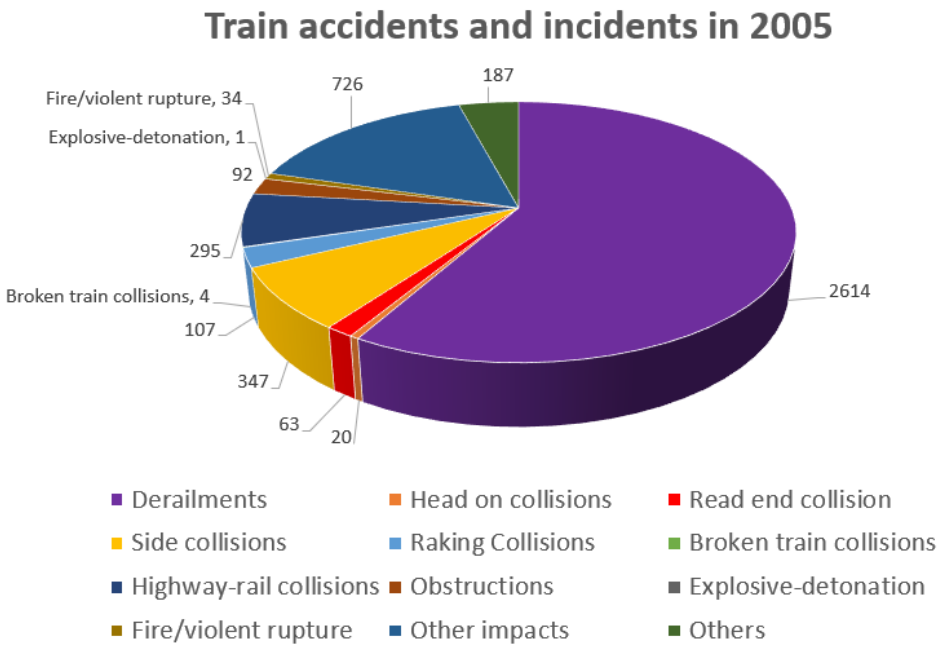


Figure 3.4: Breakdown of train accidents and incidents in 2005

The Federal Railroad Administration (FRA) classifies accidents/derailments into five major accident cause categories namely:

- Track, Roadbed and Structures (T)
- Signal and Communication (S)
- Mechanical and Electrical Failures (E)
- Train Operations - Human Factors (H)
- Miscellaneous causes not otherwise listed (M)

The breakdown of these freight-train derailments based on major accident cause category is given in figure 3.7. Track, Roadbed and Structures (T) was the most popular category (about 47.5%) followed by Mechanical and Electrical Failures (about 25.1%). The subcategory breakdown of the various major cause categories is presented in appendix B.

Accident cause has been found to influence the severity of train derailments. To

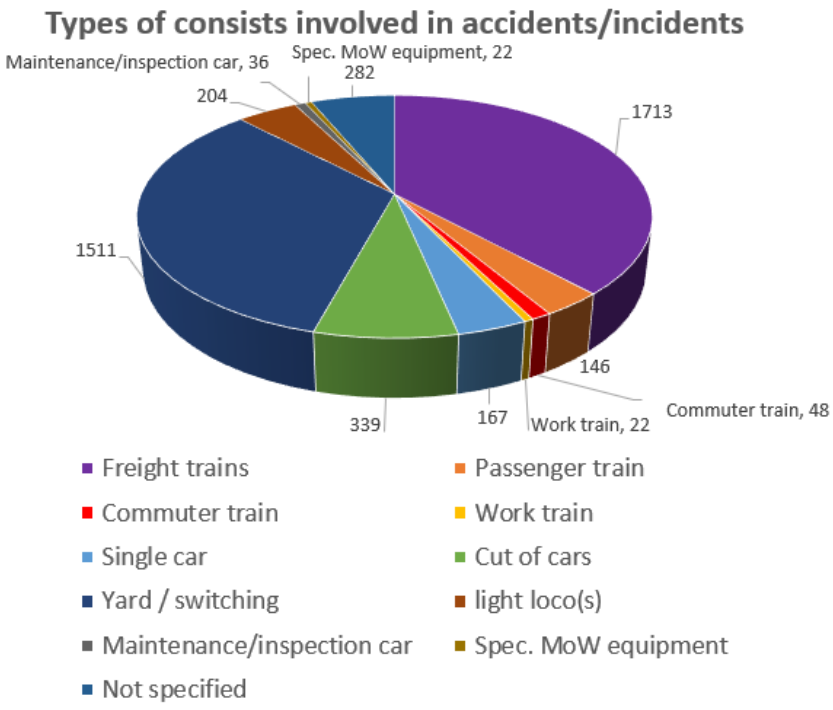


Figure 3.5: Types of train consists involved in accidents/incidents

cater for the effect (and variations) due to derailment cause, 124 derailments caused by broken rail were subsequently considered. Broken rails are the most frequent cause of freight-train derailment on Class I mainlines in the United States (Barkan et al., 2003; Liu et al., 2013). Broken rail falls under the "Rail, Joint Bar and Rail Anchoring" subcategory of the Track, Roadbed and Structures major cause category. Broken rails have been found to result in a higher derailment severity in comparison with other causes such as bearing failure with the former causing twice as many derailed cars on average as that of the latter (Barkan et al., 2003). Due to their high frequency and severity, broken rails are more likely to present higher risk than other causes.

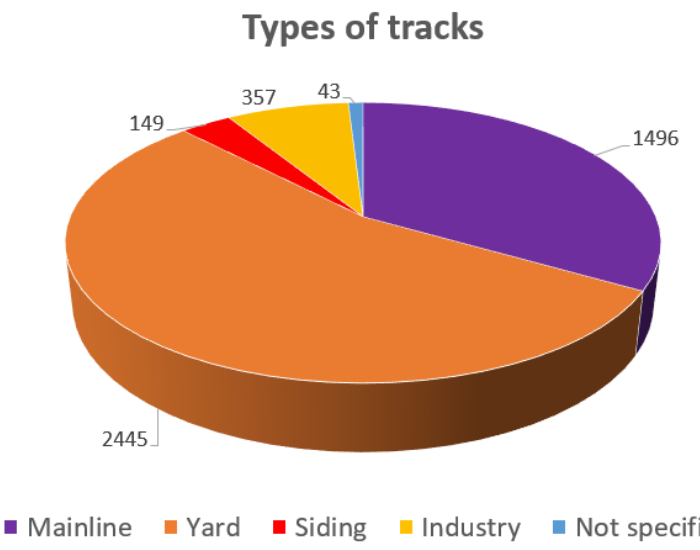


Figure 3.6: Types of tracks involved in accidents/incidents

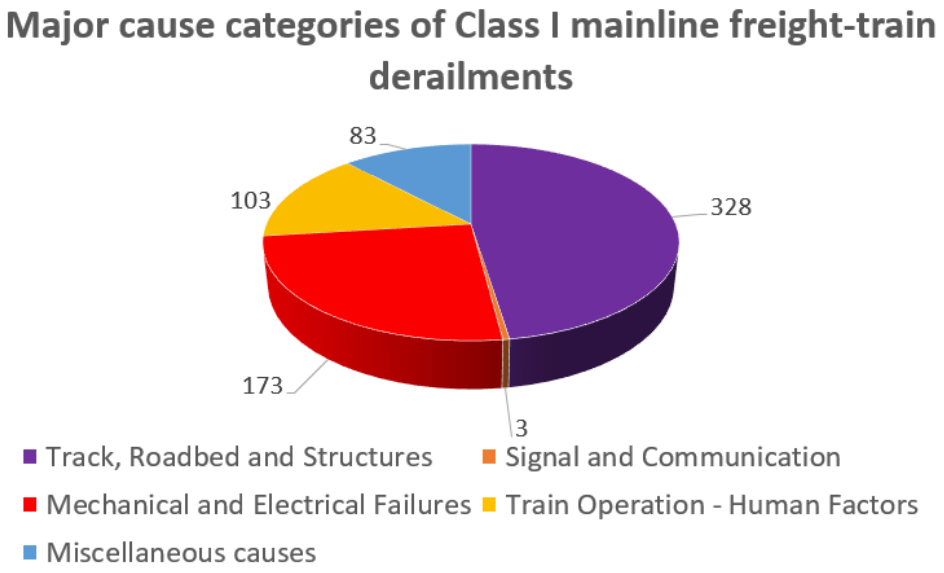


Figure 3.7: Major accident cause category breakdown of Class I mainline freight train derailments

3.3 Histogram and Quantile-Quantile Plot

A histogram provides an illustrative representation of the frequency distribution of a continuous univariate numerical dataset over a continuous interval or given time frame. Histograms provide an estimate of the underlying distribution while offering information on the center, spread, skewness, shape and possible outliers in the data. Quantile-quantile plot (also known as Q-Q plot) is a probability scatterplot used to compare two probability distributions by plotting their quantiles relative to each other. Thus, Q-Q plots can be used to assess the similarity of the empirical distribution of a sample with common theoretical distributions such as normal and exponential distributions. Normal Q-Q plots can be used to test the assumption that a random variable is normally distributed. If the empirical distribution of the examined sample comes from the normal distribution the points roughly form a straight line along a 45-degree reference line. However, the nonlinearity of the points or increase in deviation from this line indicates non-normality.

3.3.1 Track Geometry Set

Histograms and Q-Q plots illustrating the data points of several track geometry parameters across all the inspection dates as well as a single given inspection date were examined. Figure 3.8 shows the histograms and Q-Q plots for all the data points for surface right (62-ft), alignment right (62-ft) and cross level across all inspection dates from 2013 to 2016 whereas figure 3.9 shows that of a given inspection date.

The histograms in these figures generally exhibit atypical Gaussian shapes and the Q-Q plots showed that points deviated from the reference line. For the histograms of figures 3.8 and 3.9, the surface (62-ft) and alignment (62-ft) parameters appeared to be fairly symmetric however they tend to deviate from the Gaussian shape in the tail regions. They have heavy tails compared to the normal distributions which serves as a violation of normality. This is confirmed by their respective Q-Q plots with deviations from the reference line in both tails leading to the formation of an "inverted-S" shape with the initial sample quantiles being much lower than the initial theoretical quantiles

with their last quantiles being much higher than the last theoretical quantiles. The crosslevel parameter however can be said to have a bimodal type distribution with two distinct peaks with majority of the points located on the left half with a mode of about 0 inches and minority on the right half with a mode of about 0.7 inches. This is as a result of the shallow curvature observed at some sections of the track.

Subsequently histograms illustrating the TQI (standard deviation) of the various track geometry parameters as well as the recovery values, TQI before and after tamping for the surface parameter were analyzed. Figure 3.10 presents the histograms and Q-Q plots for all the data points for standard deviation (SD) surface, SD alignment and SD cross level across all inspection dates from 2013 to 2016. On the other hand, figure 3.11 shows the histograms and Q-Q plots for SD surface recovery values, SD surface before tamping and SD surface after tamping. From figures 3.10 and 3.11, the histograms of these parameters were found to be right skewed (a violation of normality) with with majority of the points located on the left half with a heavy right tail of data. This is confirmed by their respective Q-Q plots with a concave plot created with the largest values larger than would be expected under normality. This indicates higher concentration of data beyond the right-hand side of a normal distribution. Figures related to the histograms of other track geometry parameters are shown in Appendix A.

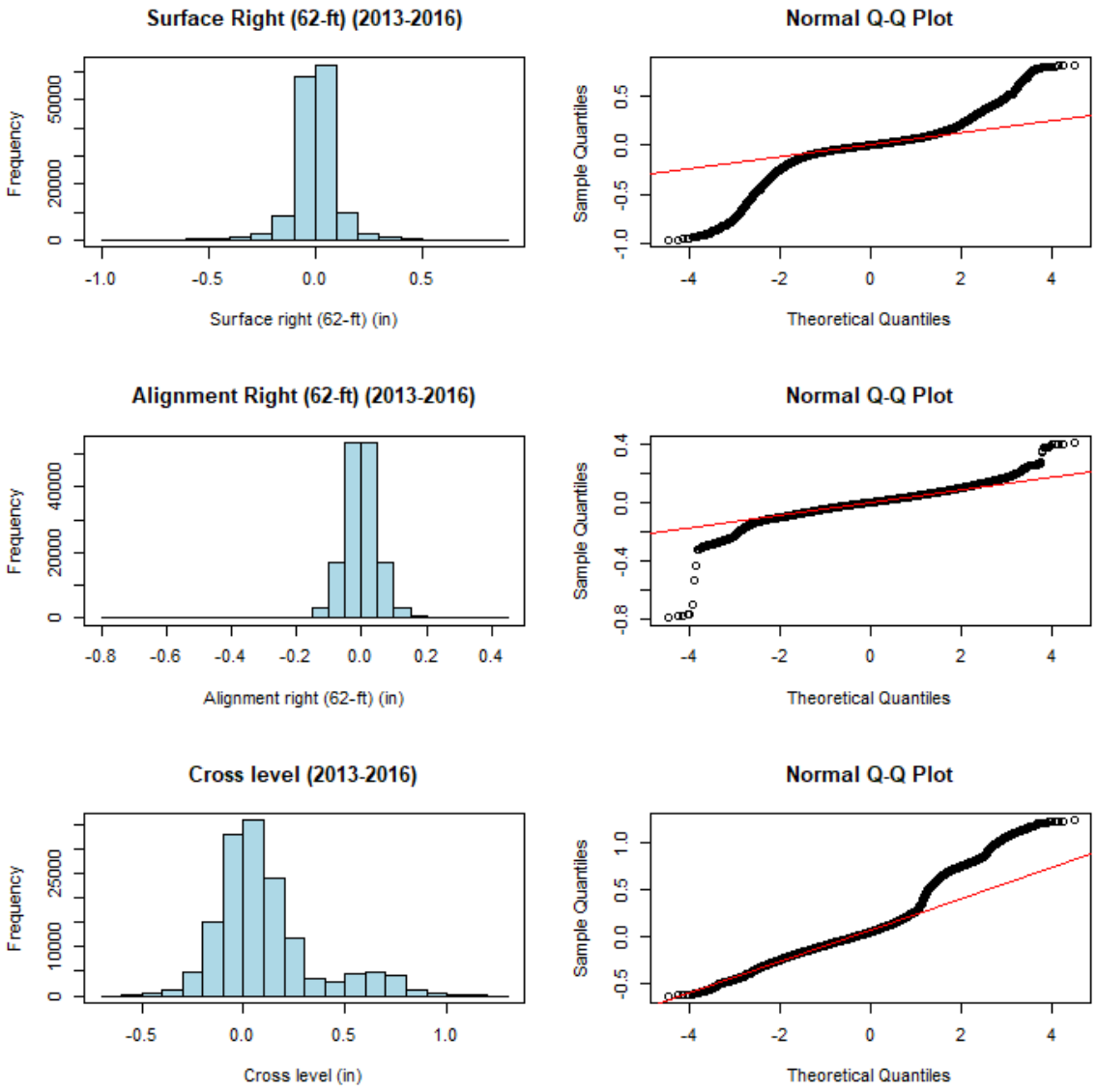


Figure 3.8: Histograms and Q-Q plots for surface right, alignment right and crosslevel data points from 2013 to 2016

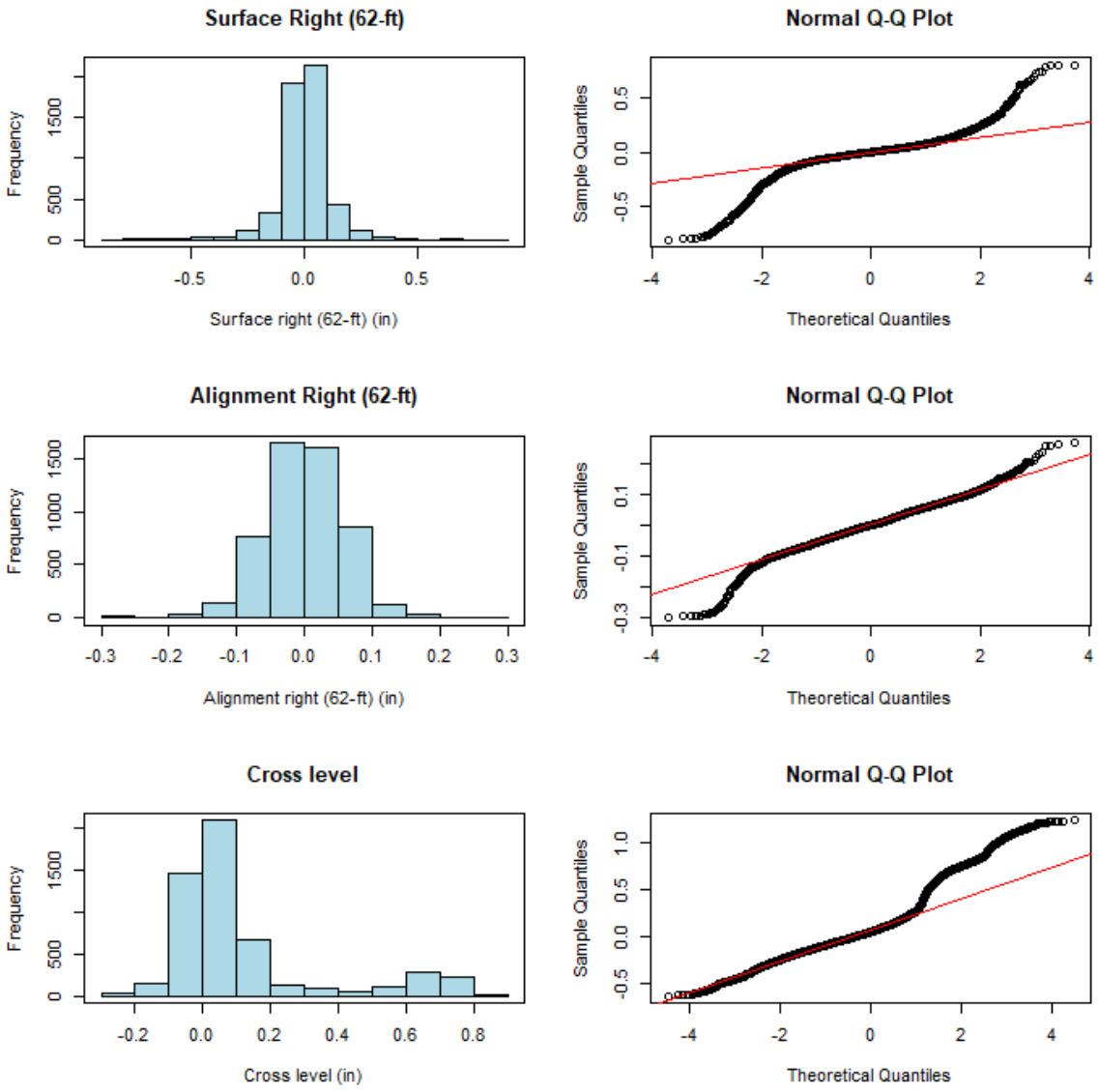


Figure 3.9: Histograms and Q-Q plots for surface right, alignment right and crosslevel data points for a given inspection date

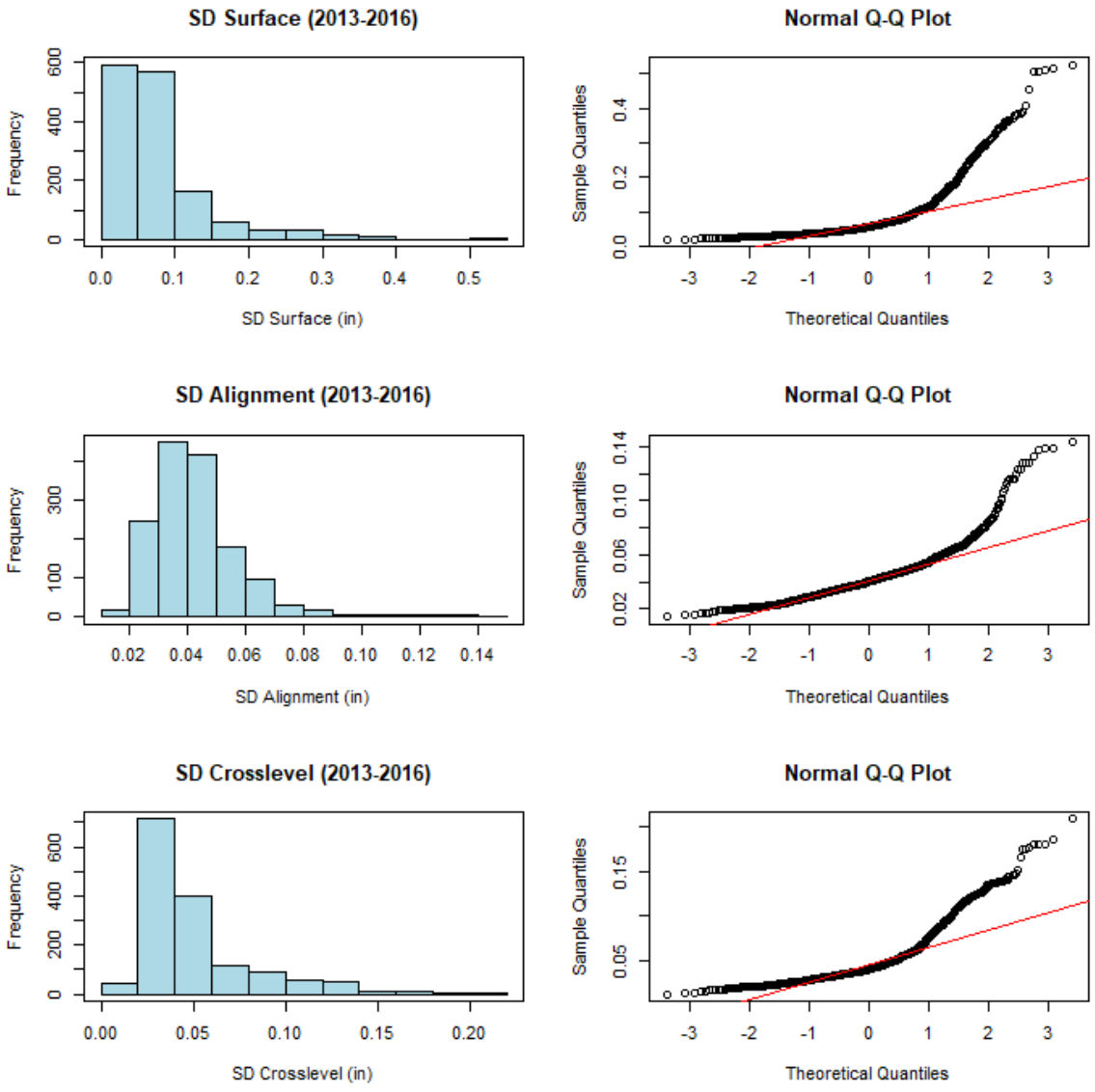


Figure 3.10: Histograms and Q-Q plots for SD surface, SD alignment and SD crosslevel data points from 2013 to 2016

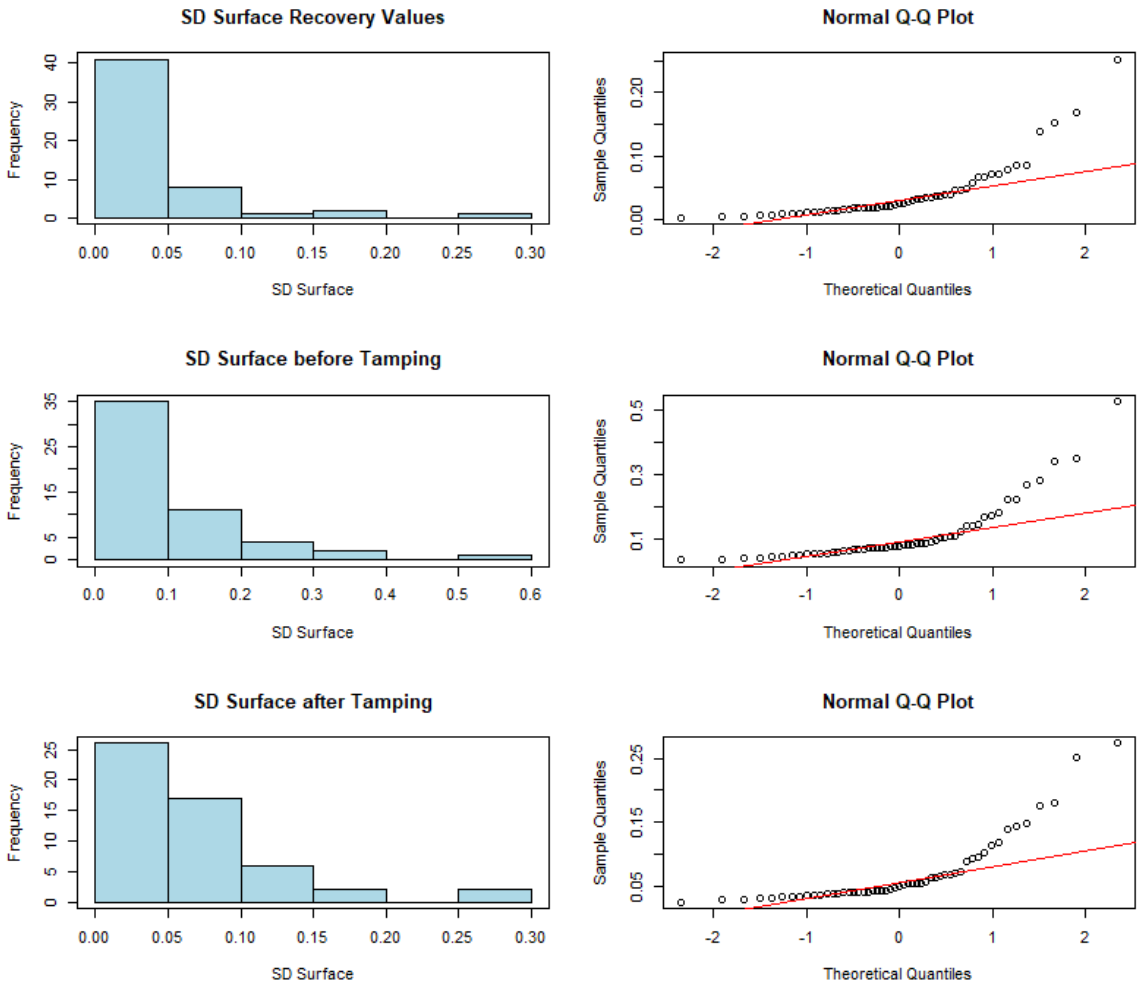


Figure 3.11: Histograms and Q-Q plots for SD surface recovery values, SD surface before tamping and SD surface after tamping

3.3.2 Derailment Data Set

Histograms and Q-Q plots illustrating the derailment severity outcomes (such as monetary damage and number of derailed cars) and covariates (such as derailment speed, residual train length, loading factor) were examined. Figure 3.12 shows the histograms and Q-Q plots of the monetary damage, number of derailed cars and derailment speeds for all freight-train derailments occurring on Class I mainline track in the year 2005 whereas figure 3.13 shows that of broken-rail caused freight-train derailments. The histograms in these figures generally exhibit non-Gaussian shapes. As shown in figures 3.12 and 3.13, the histograms of the monetary damage, number of derailed cars and derailment speed were all found to be right skewed with most data points located on the left half with a long right tail of data. This is corroborated by the concave plots shown in their respective Q-Q plots. Figures related to the histograms of other derailment severity covariates such as residual train length and loading factor are shown in Appendix B.

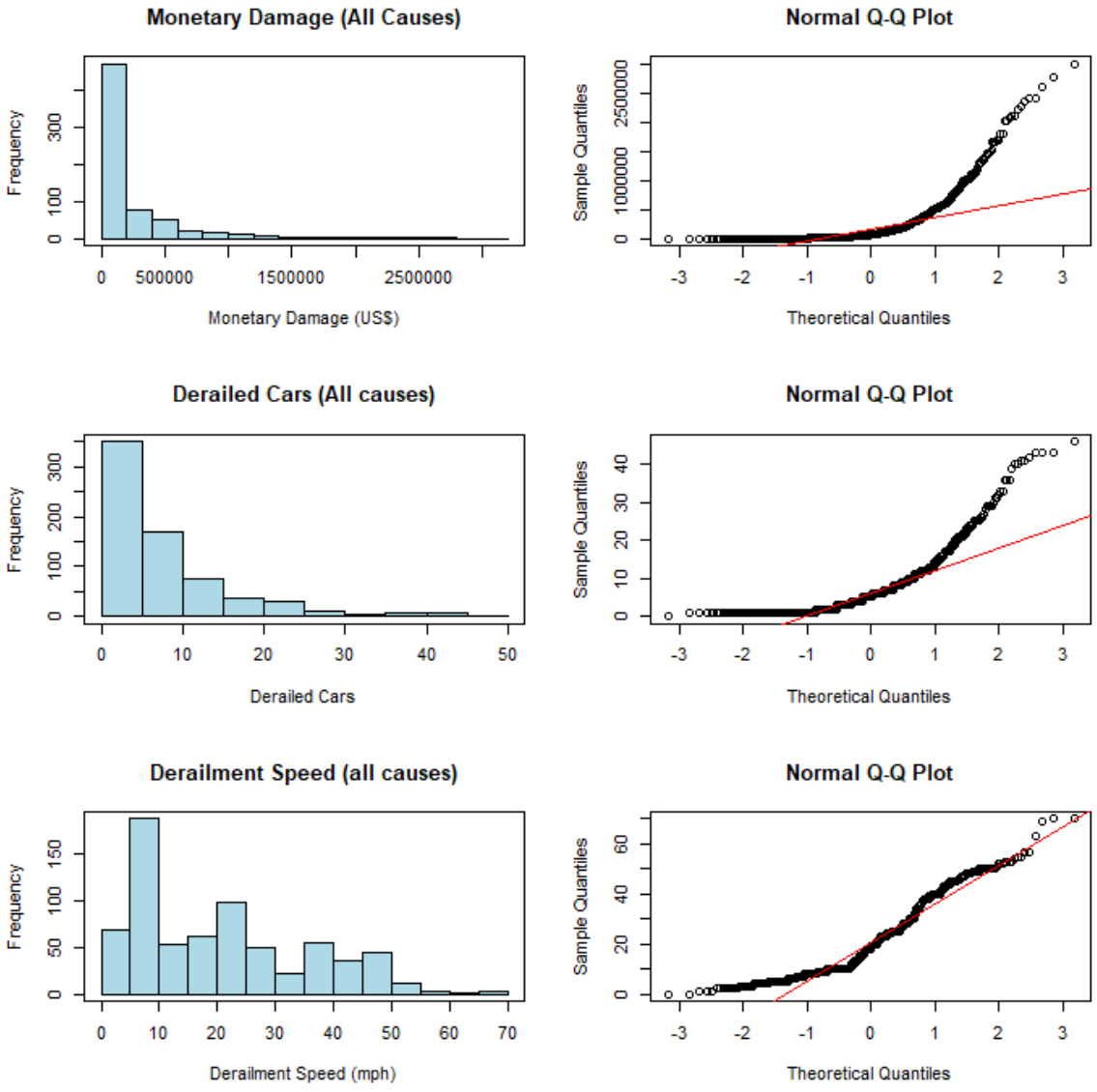


Figure 3.12: Histograms and Q-Q plots for monetary damage, derailed cars and derailment speed for all freight-train derailments occurring on Class I mainline track

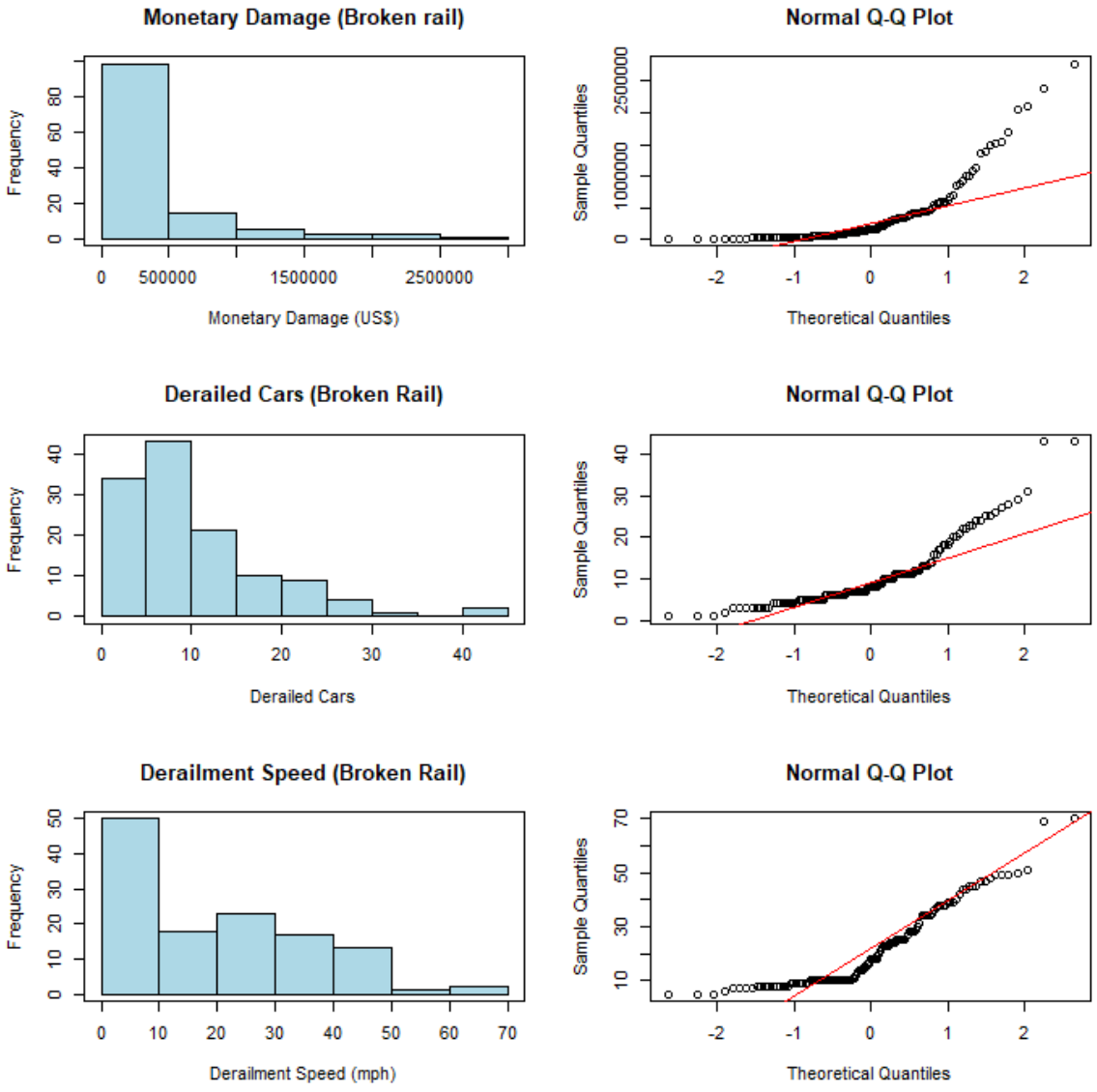


Figure 3.13: Histograms and Q-Q plots for monetary damage, derailed cars and derailment speed for broken-rail caused freight-train derailments occurring on Class I mainline track

3.4 Box and Whisker Plot

The box and whisker diagram (popularly known as box plot) is a standardized graphical representation of the distribution of numerical data through their quartiles or a five-number summary. The five-number summary consists of the minimum, first lower quartile (Q1), median, upper quartile (Q3) and maximum. Some box plots include an extra character to denote the mean of the variable. It is called a box and whisker plot since a box is drawn from the lower (first) quartile to the upper quartile (known as the interquartile range (IQR)) whereas whiskers are drawn from the minimum to the lower quartile and the maximum to the upper quartile. The whiskers are indicative of the variability beyond these quartiles. Box plots aid in the comparison of distributions and provide information on the skewness and spread of the data. Box plots also aid in the identification of outliers. Data points greater than $Q3 + (1.5 \times IQR)$ and less than $Q1 - (1.5 \times IQR)$ are considered as outliers.

3.4.1 Track Geometry Set

The box and whisker diagrams of all the observed data points of the various parameters during the study time frame were initially considered. Figures 3.14 and 3.15 illustrate the box plot for surface right (62-ft) and crosslevel across all the inspection dates. The median values of the surface right (62-ft) were found to be relatively constant throughout the whole duration compared to the crosslevel. High variability of the surface right (62-ft) was observed for most inspection dates with several potential outliers. The variability was found to reduce drastically after August, 2015 when track reconstruction and geocell placement along a 800-foot section were undertaken.

The median of the crosslevel level generally changed from one inspection to another but was more stable after August, 2015 when track reconstruction and geocell replacement were undertaken. High variability was observed for most inspection dates. The variability was found to reduce drastically after August, 2015 when track reconstruction and geocell replacement were undertaken.

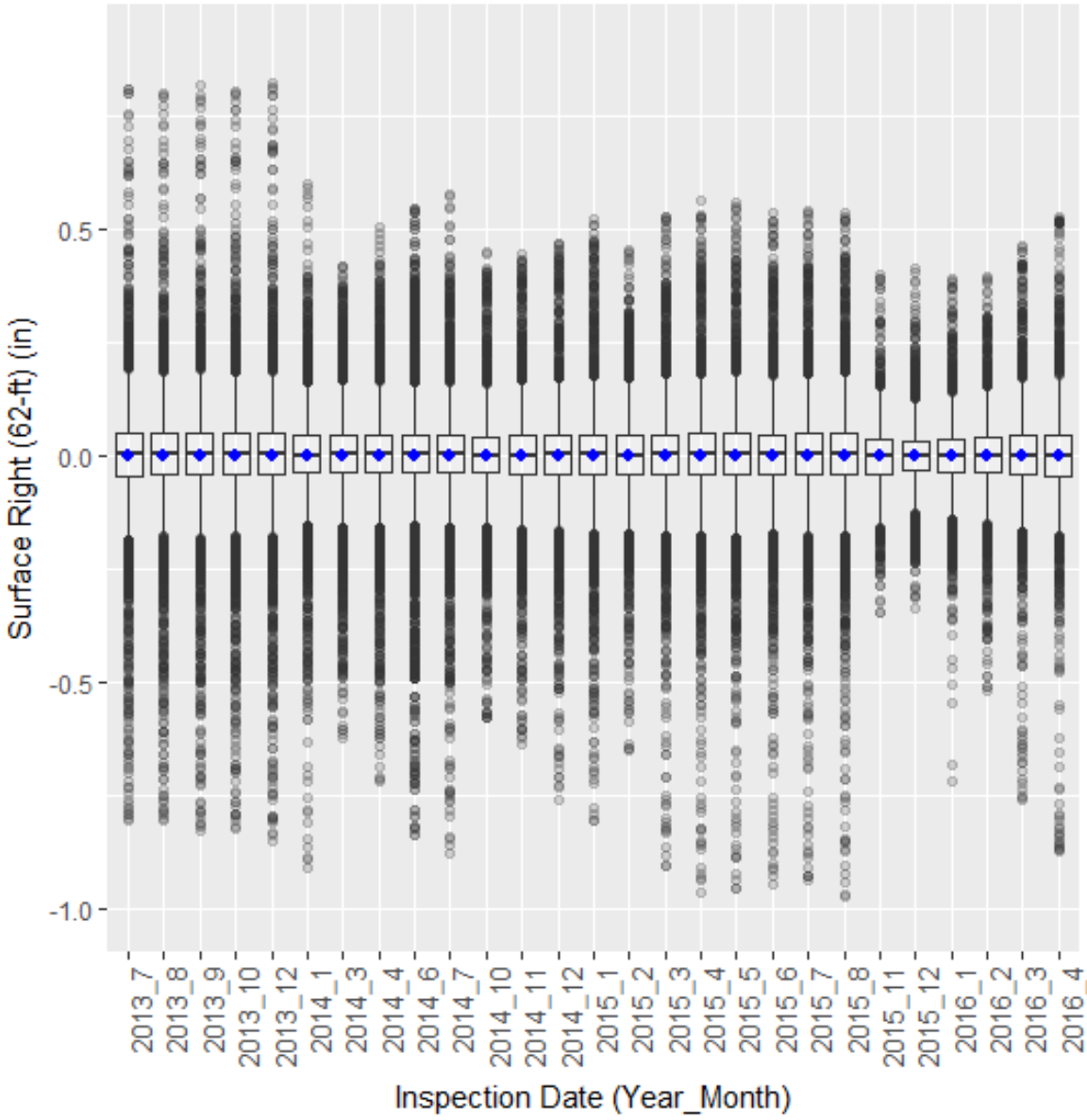


Figure 3.14: Box plot of surface right (62-ft) data points across all the inspection dates

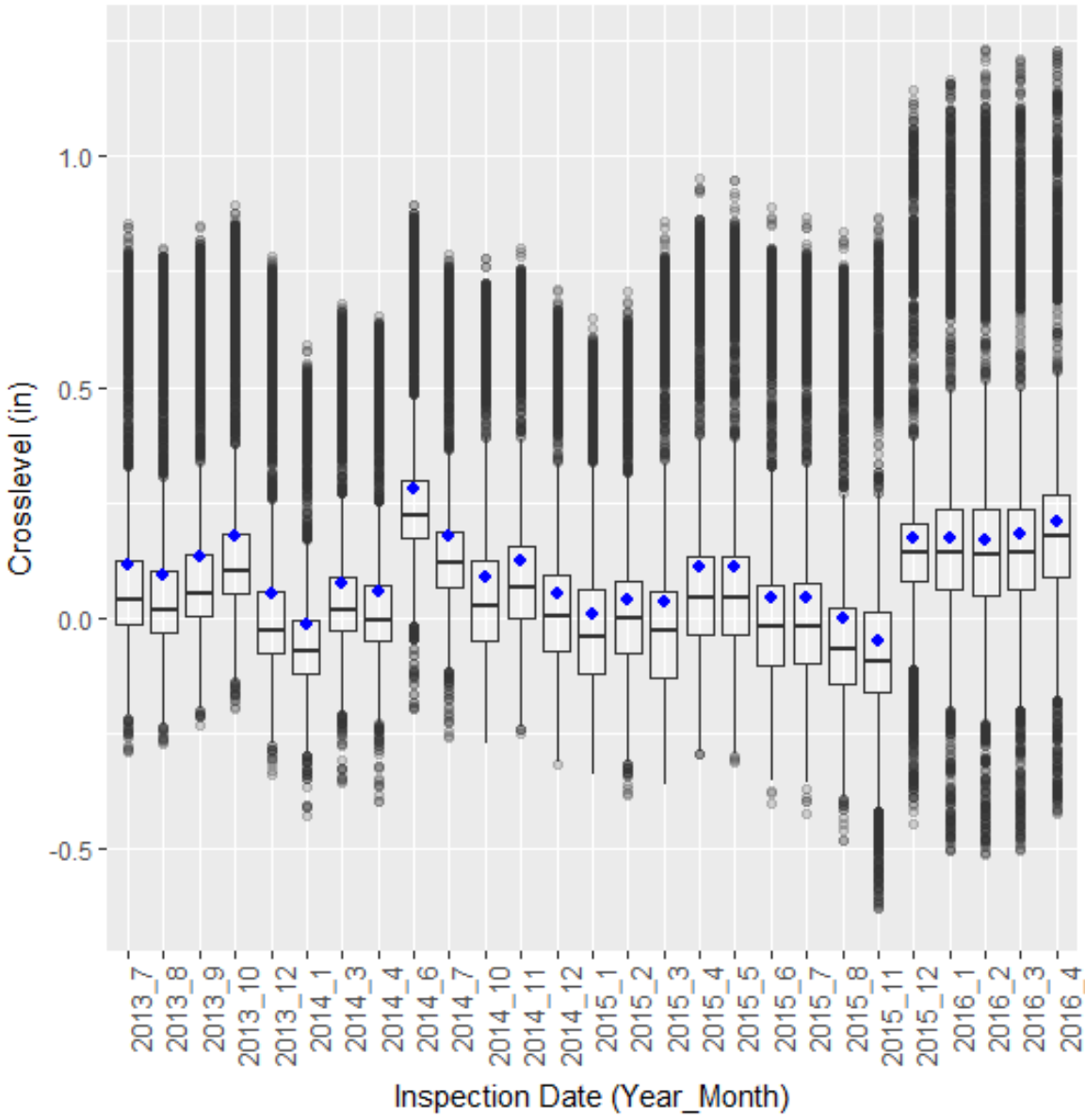


Figure 3.15: Box plot of crosslevel data points across all the inspection dates

The box plots of the track quality index (standard deviation) of the various parameters were subsequently considered. Figures 3.16 illustrates the box plot of the TQI before tamping, TQI after tamping and recovery values for standard deviation (SD) surface with figures 3.17 and 3.18 illustrating that of SD alignment and SD crosslevel respectively. As shown in all three plots, the TQI (SD of the parameter) generally reduced after tamping maintenance with an observed reduction in median, upper and lower quartiles, maximum and minimum. The reduction was found to be greatest in the surface parameter. Additionally, the variability in the TQI was also found to considerably reduce in the surface parameter but similar reductions were not observed in the alignment and crosslevel. These observations seem to corroborate the fact that the shortwave surface parameter tends to recover relatively better during tamping than other parameters (Lichtberger, 2005; Soleimanmeigouni et al., 2016b). Several potential outliers were also identified in the box plots and are represented by circles. The box plots for SD gage and SD warp can be found in Appendix A.

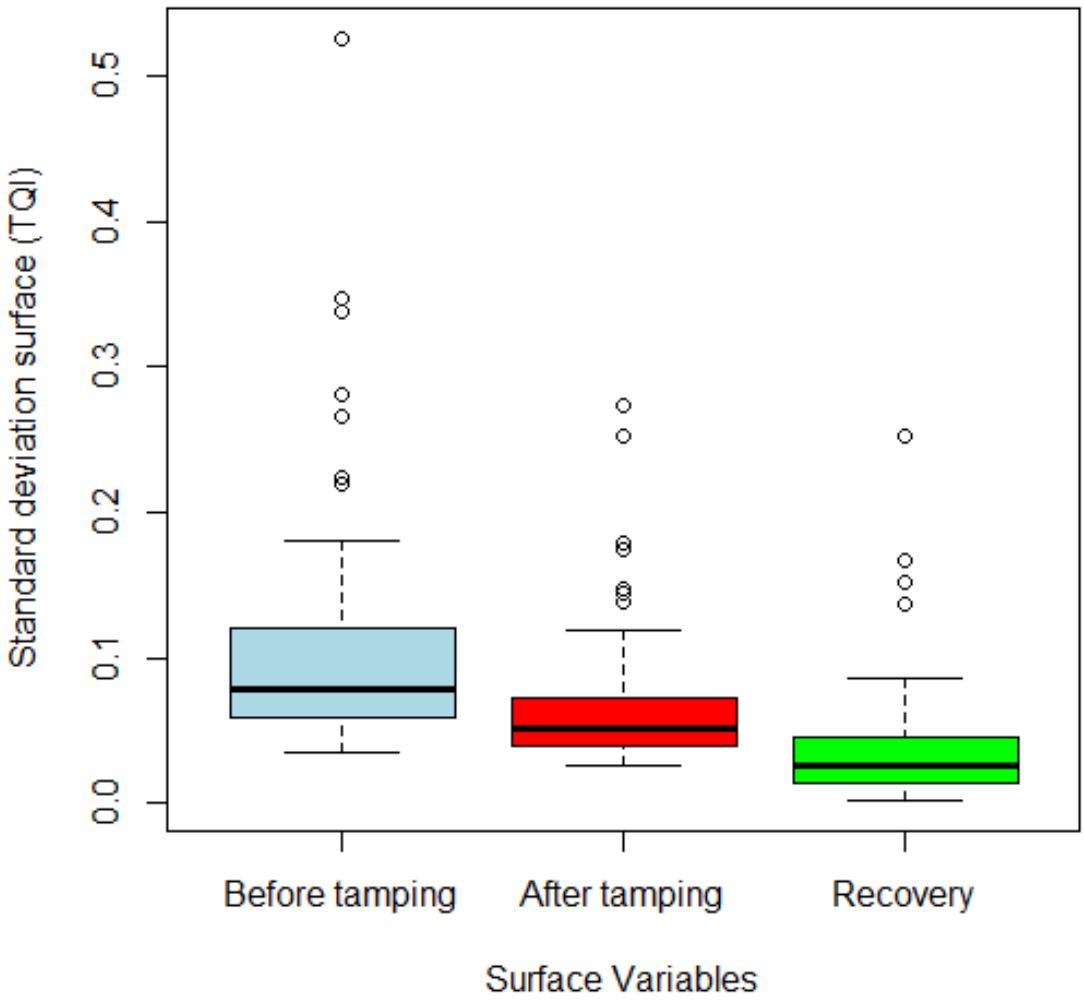


Figure 3.16: Box plot of TQI before tamping, TQI after tamping and recovery values for SD surface

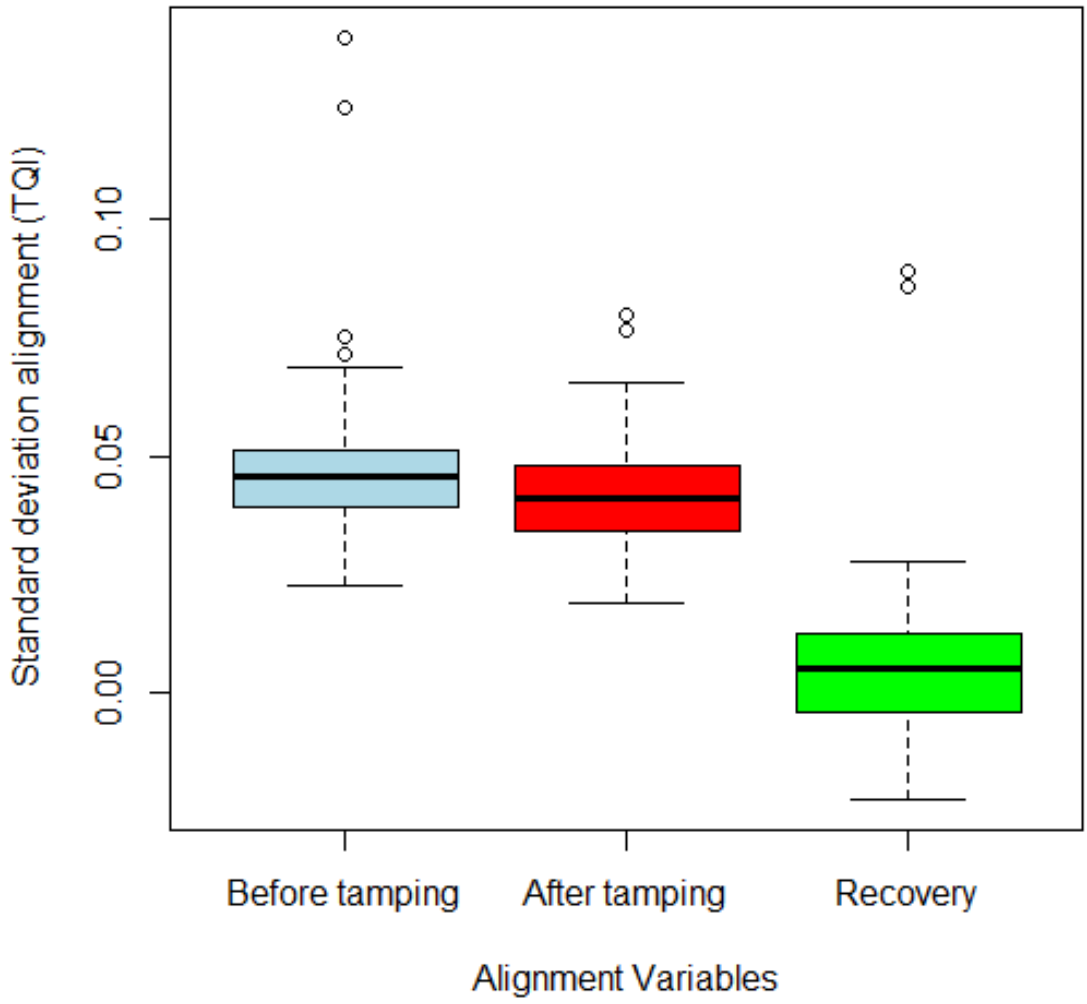


Figure 3.17: Box plot of TQI before tamping, TQI after tamping and recovery values for SD alignment

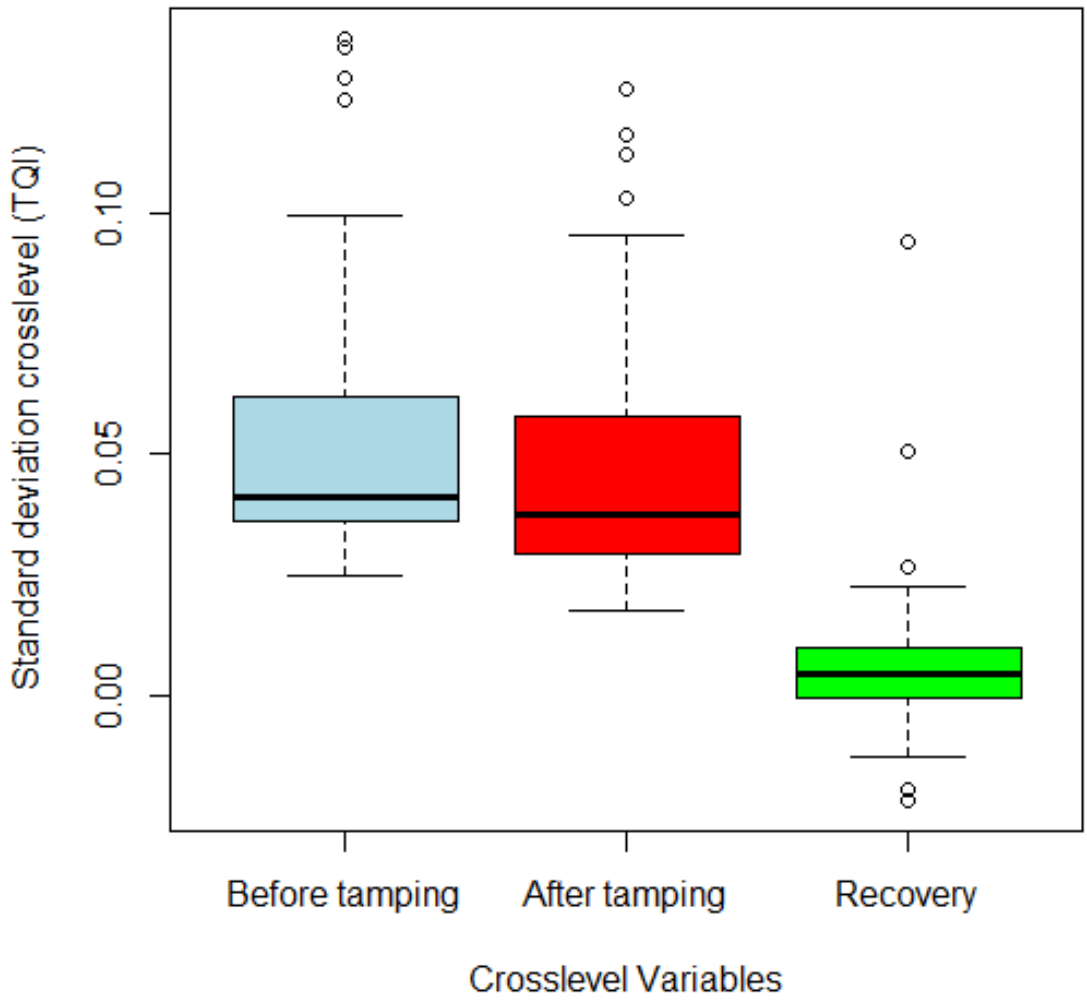


Figure 3.18: Box plot of TQI before tamping, TQI after tamping and recovery values for SD crosslevel

3.4.2 Derailment Data Set

The box and whisker diagrams of the variables of interest for all freight-train derailments on Class I mainline track were initially examined based on their major accident cause category type. Figure 3.19 presents the distribution of monetary damage incurred across the various major accident cause categories. All categories show asymmetry with the median splitting the boxes into unequal halves, unequal length of both whiskers as well as unequal number of outliers on both sides of the whiskers. They are all said to be right skewed because the mean is greater than the median and the median is closer to the lower quartile than the upper quartile making the right hand side of the box greater than the left hand side. On average, the signal and communication cause category was found to have the highest monetary damage with human factor have the least monetary damage (mean is denoted by the blue diamond shape). Additionally, the signal and communication cause was found to have the lowest median and the only category without any potential outliers. Mechanical and electrical failures was found to have the highest variability (determined by the entire range of data points) among the category types.

Figure 3.20 illustrates the distribution of derailed cars across the various major accident cause categories. Similar to monetary damage plot, all categories were found to be right skewed with mean values greater than median values. Additionally, signal and communication category was found to have the highest number of derailed cars on average with mechanical and electrical failures having the least. The box plot of the derailment severity covariates across the various categories can be found in Appendix B. Subsequently, the distribution of the derailment severity outcomes were examined based on major accident cause sub-category type. Figures 3.22 and 3.21 shows the distribution of monetary damage and derailed cars across Track, Roadbed and Structures causes sub-category respectively. For both outcomes, "Rail, Joint Bar and Rail Anchoring" sub-category was found to have the highest mean and median values (even higher than that of the Signal and communication category) as well as the highest variability. The variation of derailment severity outcomes (monetary damage

and number of derailed) across accident cause categories and sub-categories emphasizes the need to analyze and take into account the actual derailment cause. The distribution of severity outcomes across other sub-categories can also be found in Appendix B.

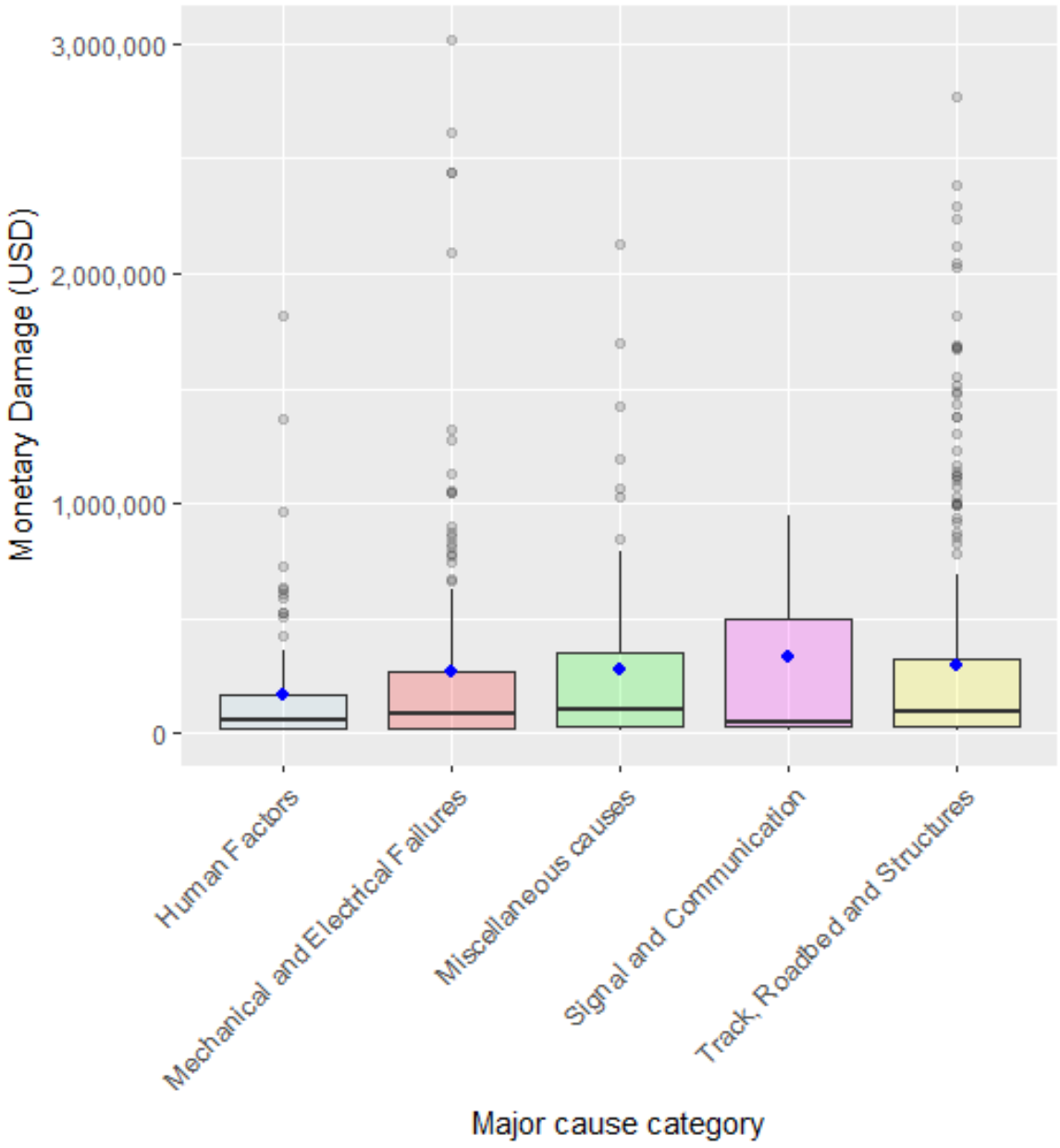


Figure 3.19: Box plot illustrating distribution of monetary damage across all major accident cause categories

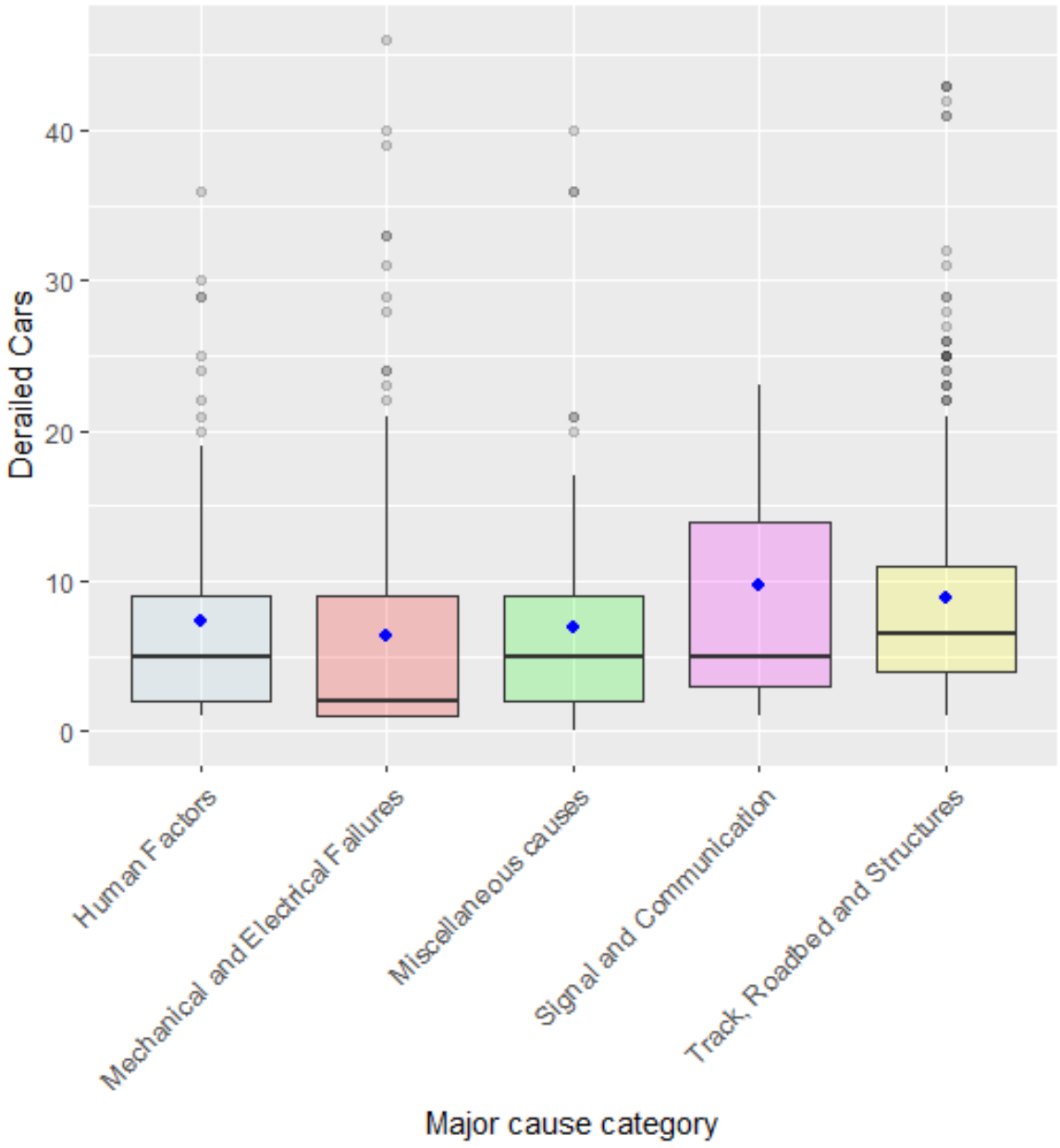


Figure 3.20: Box plot illustrating distribution of derailed cars across all major accident cause categories

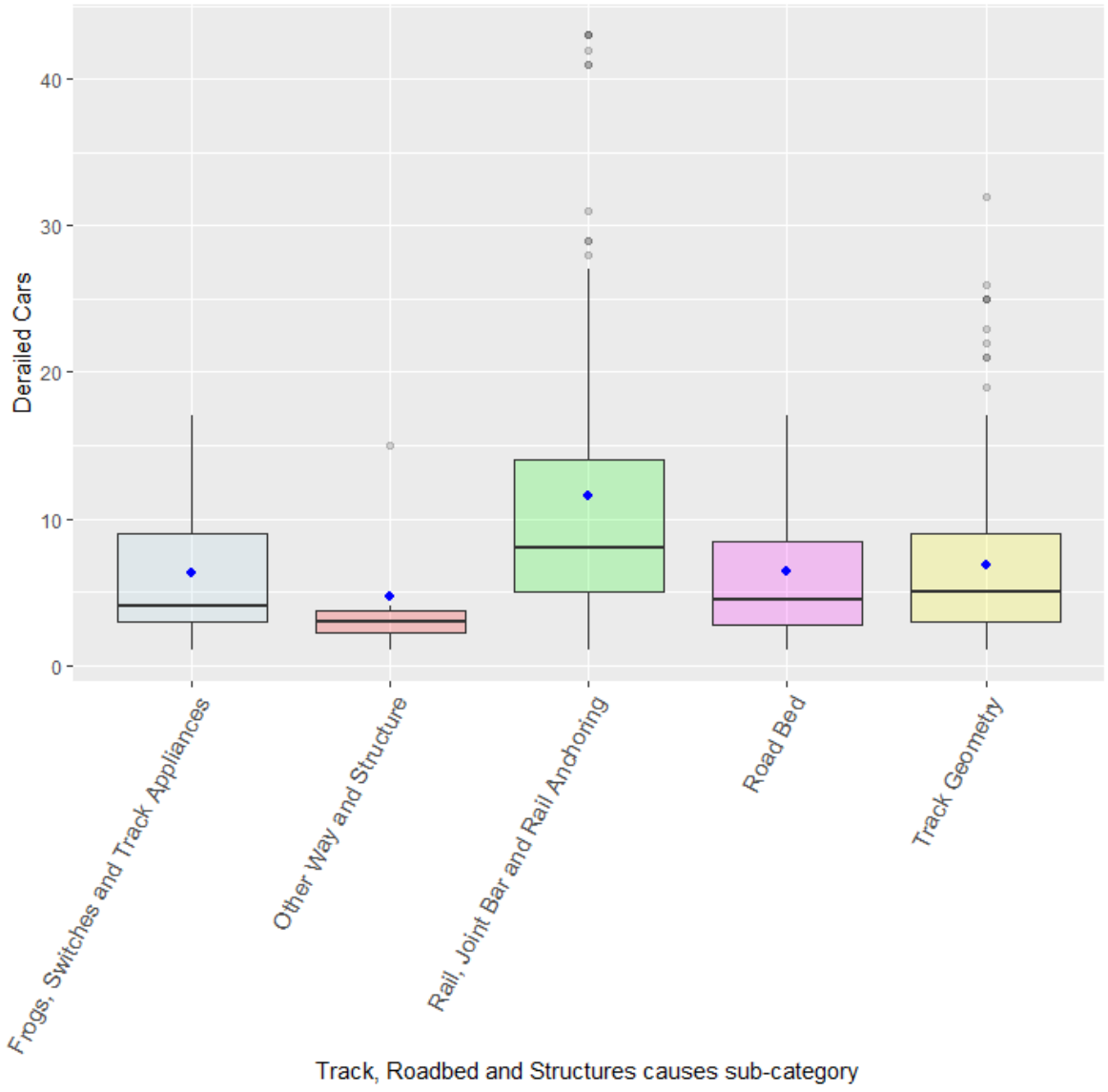


Figure 3.21: Box plot illustrating distribution of derailed cars across Track, Roadbed and Structures causes sub-category

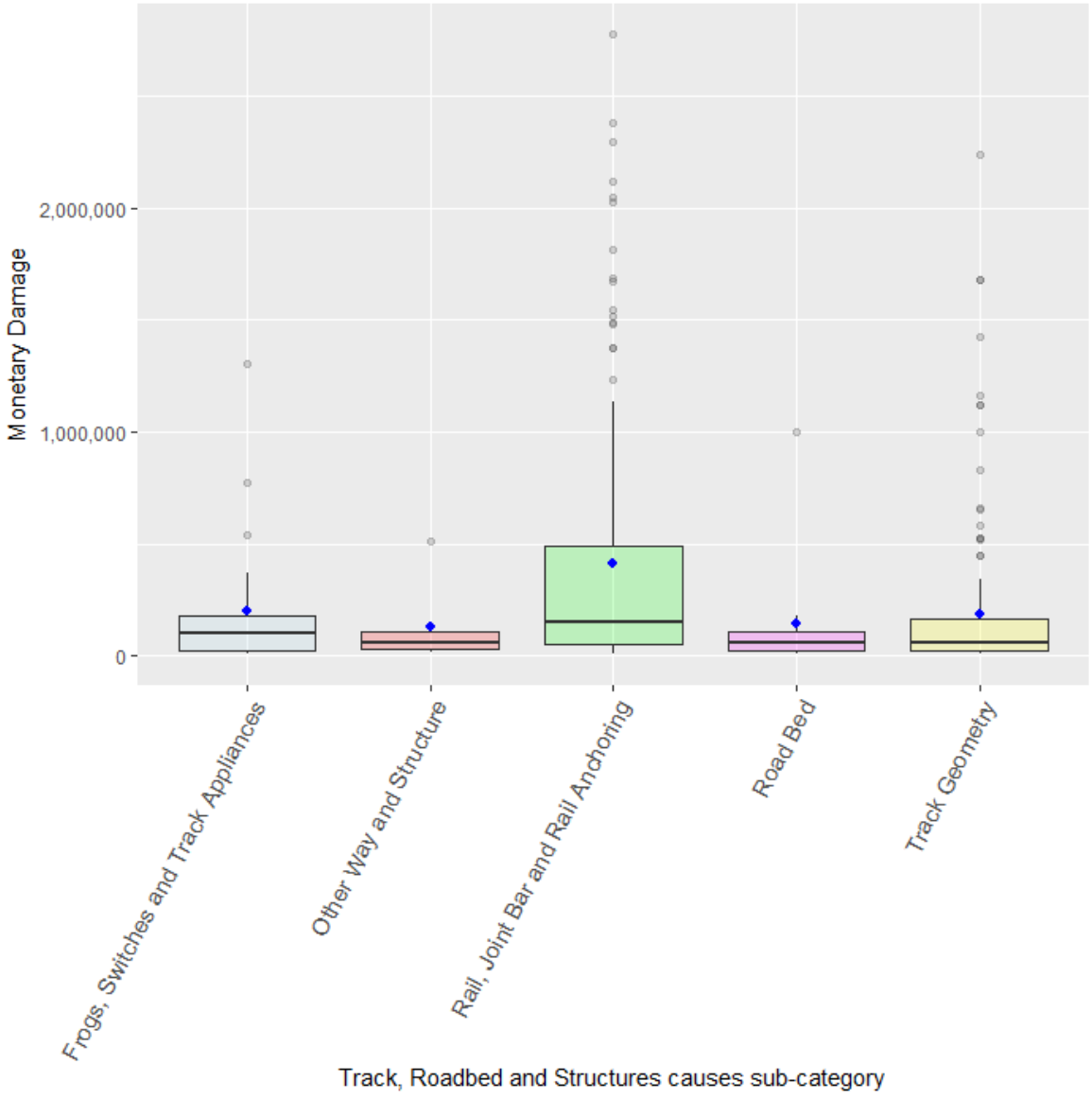


Figure 3.22: Box plot illustrating distribution of monetary across Track, Roadbed and Structures causes sub-category

3.5 Scatter Plots

Scatter plots are plots of the data points of a pair of variables. Scatter plots offer an illustrative depiction of the correlation or relationship between the pair of variables.

3.5.1 Track Geometry Data Set

Figure 3.23 presents a correlation plot matrix of the pairs of several geometry parameters for a given inspection date. On the other hand, figure 3.24 shows a correlation plot matrix of the TQI of the parameters at a given inspection date whereas figure 3.25 presents a correlation plot matrix of the recovery values of the parameters. Bivariate scatter plots are shown below the diagonal, histograms of the individual parameters on the diagonal and Kendall's correlation coefficient of each pair above the diagonal. Kendall's Tau was preferred over Pearson's correlation coefficient since it measures dependence independent of the assumed distribution and dependence whereas the latter assumes normality and linear dependence. Generally, very weak correlations were found between the pair of variables of interest when examining the "raw" data as shown in figure 3.23. However, an examination of the TQI (standard deviation) of the same track geometry parameters reveals some dependence between the variables due to the aggregated nature of the TQI. As shown in figure 3.24, SD crosslevel and SD warp were found to have the highest correlation between the various pairs of variables. This may be attributed to the fact that warp is a measure of the crosslevel variation. Furthermore, SD surface appears to be more correlated with SD crosslevel and SD warp (vertical parameters) than SD alignment and SD gage (horizontal parameters). SD alignment (a longitudinal horizontal parameter) was found to be most correlated with SD gage (a transverse horizontal parameter) followed by SD surface (a longitudinal vertical parameter). Similar to the TQI, the crosslevel and warp recovery values was found to have the highest correlation as shown in figure 3.25. Detailed explanation of the correlation analysis between the recovery values of the various track geometry parameter using several dependence measures can be found in section 4.6.9.

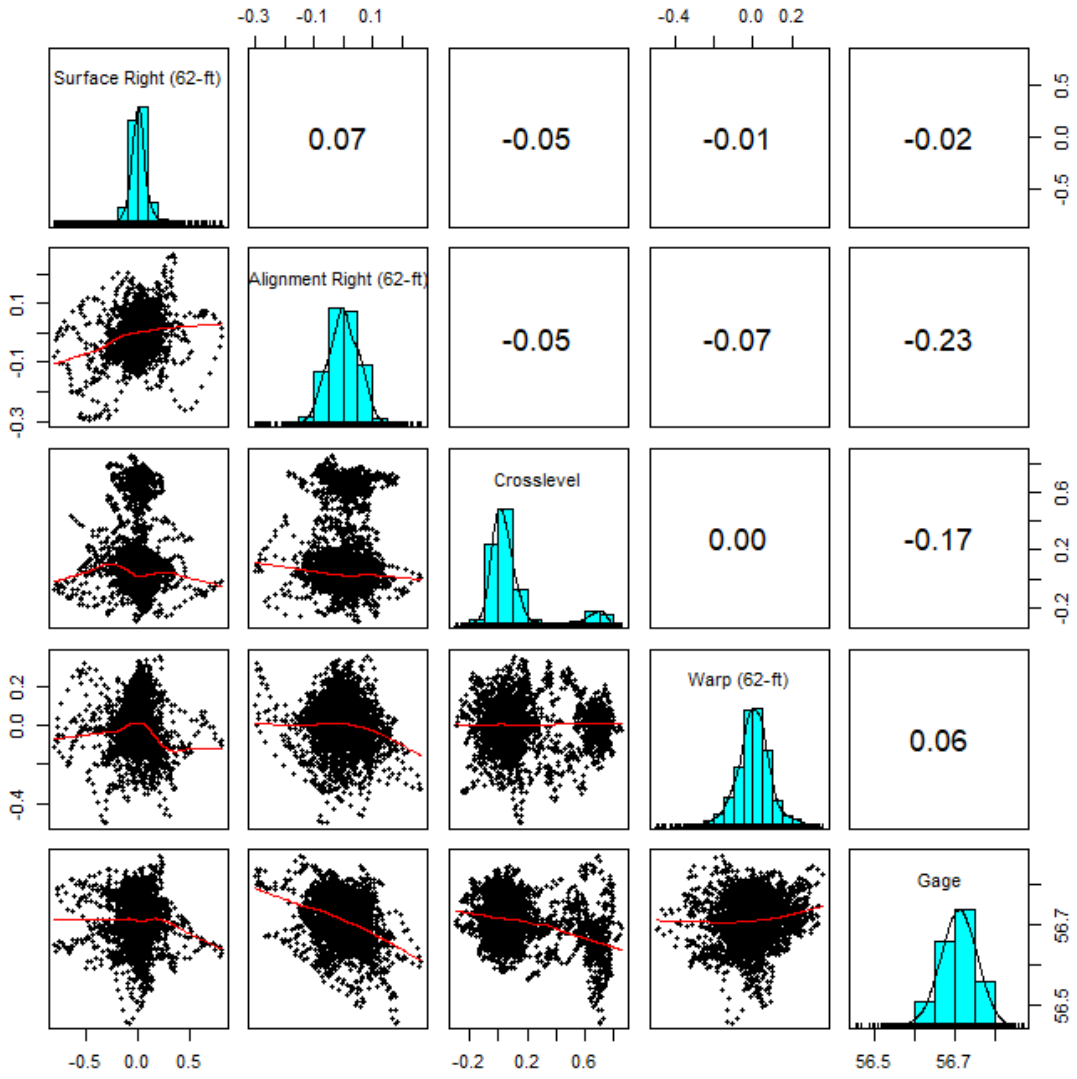


Figure 3.23: Correlation plot matrix of selected track geometry parameters at a given inspection date

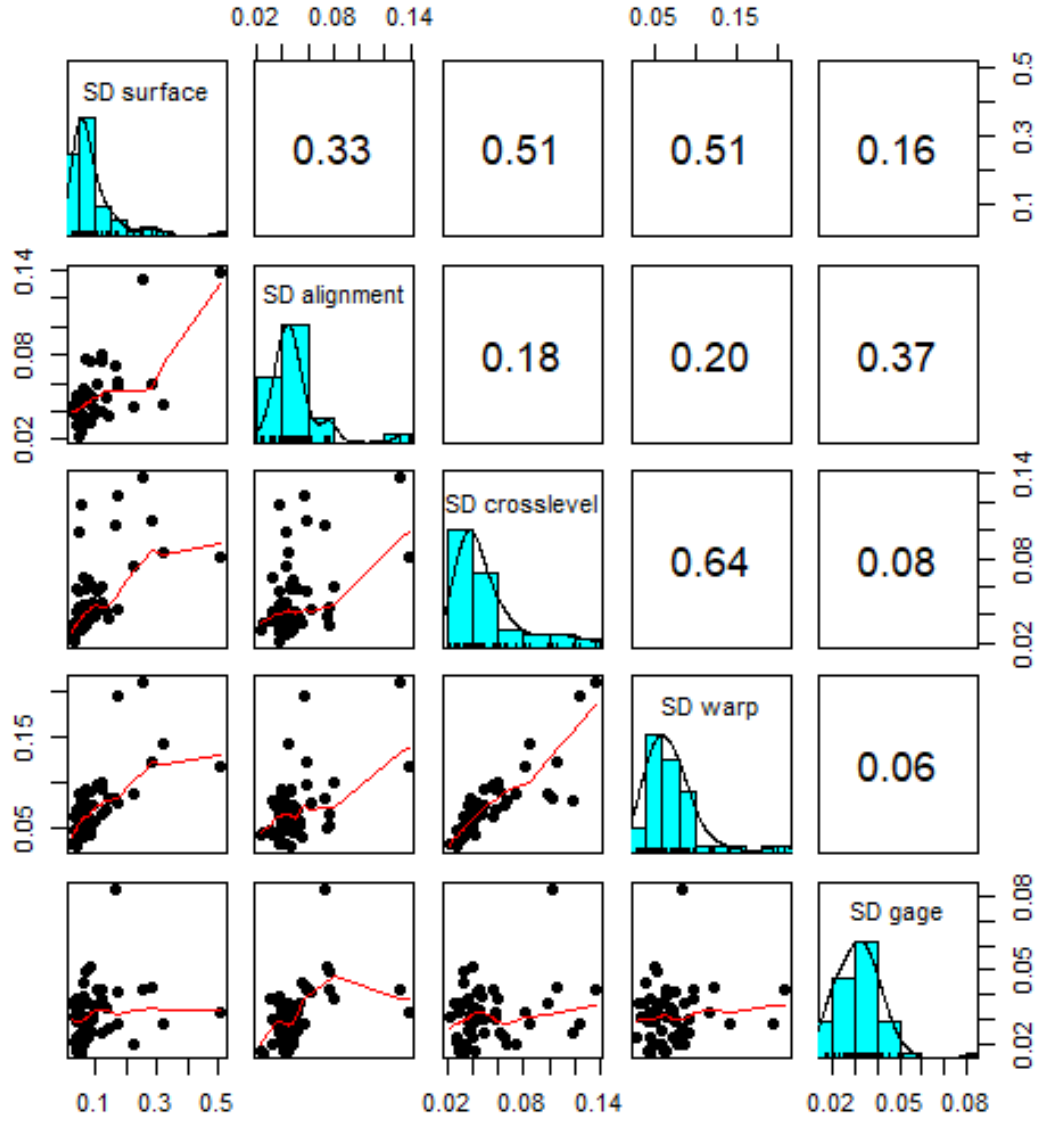


Figure 3.24: Correlation plot matrix of TQI of selected track geometry parameters at a given inspection date

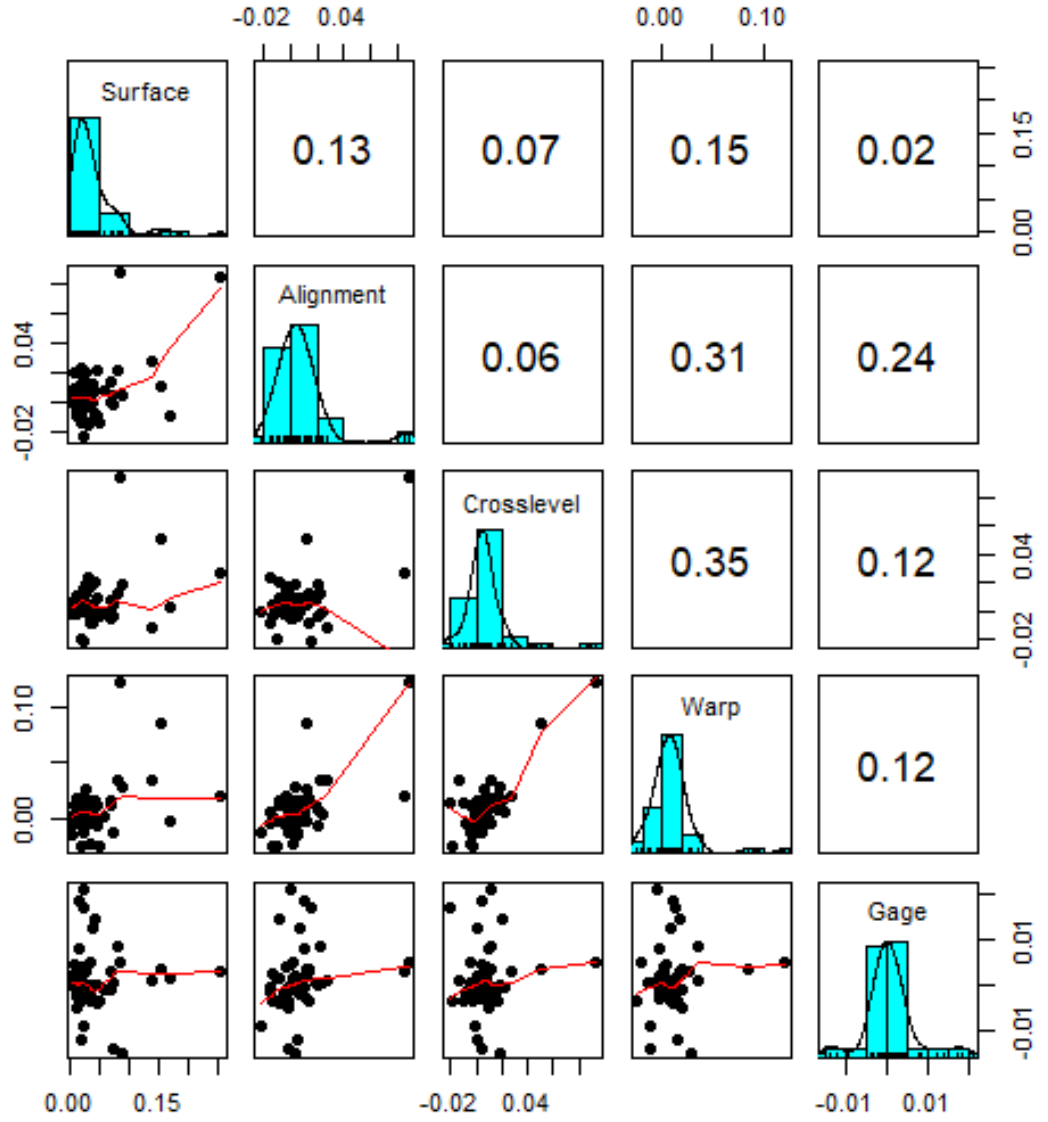


Figure 3.25: Correlation scatter plot matrix of recovery values of selected track geometry parameters at a given inspection date

3.5.2 Derailment Data Set

Figure 3.26 presents a correlation plot matrix of the pairs of variables for all freight train derailments whereas figure 3.27 presents that of only broken-rail caused derailments. Bivariate scatter plots are presented beneath the diagonal, histograms of the individual variables on the diagonal and Kendall's Tau correlation coefficient above the diagonal. Kendall's Tau was preferred over Pearson's correlation coefficient since it is scale-invariant measure of dependence which remains unaltered under monotonically increasing transformations. Furthermore, it measures dependence independent of the assumed distribution and dependence whereas the latter assumes normality and linear dependence. For both datasets, Monetary damage and number of derailed cars was found to have the highest correlation with a moderate dependence (0.48) observed for all freight-train derailments and a relatively strong dependence (0.59) when examining only broken-rail caused derailments. On the other hand, Residual Train Length and Loading Factor was found to have the least correlation with no dependence observed between the pair for both datasets. In general, weak or no correlations were observed between the pairs of covariates.

Generally, the outcomes (Monetary Damage, Number of Derailed Cars) were found to be positively correlated with the covariates (Derailment Speed, Residual Train Length, Loading Factor). For all freight train derailments dataset, both outcomes exhibit relatively low correlations with the covariates. However, for broken-rail caused derailments, the outcomes were found to have moderate correlations with both derailment speed and residual train length but low correlations with loading factor.

Derailment severity has been found to increase exponentially with derailment speed and residual train length. Thus, logarithm transformation of these variables have been found to offer a better fit (Saccomanno and El-Hage, 1989, 1991; Liu et al., 2013). This was confirmed and adopted during the confirmatory (copula) data analysis. However, unlike Pearson's coefficient, Kendall's Tau is a scale-invariant dependence measure thus the logarithmic transformations of these covariates do not affect the measure of dependence observed.

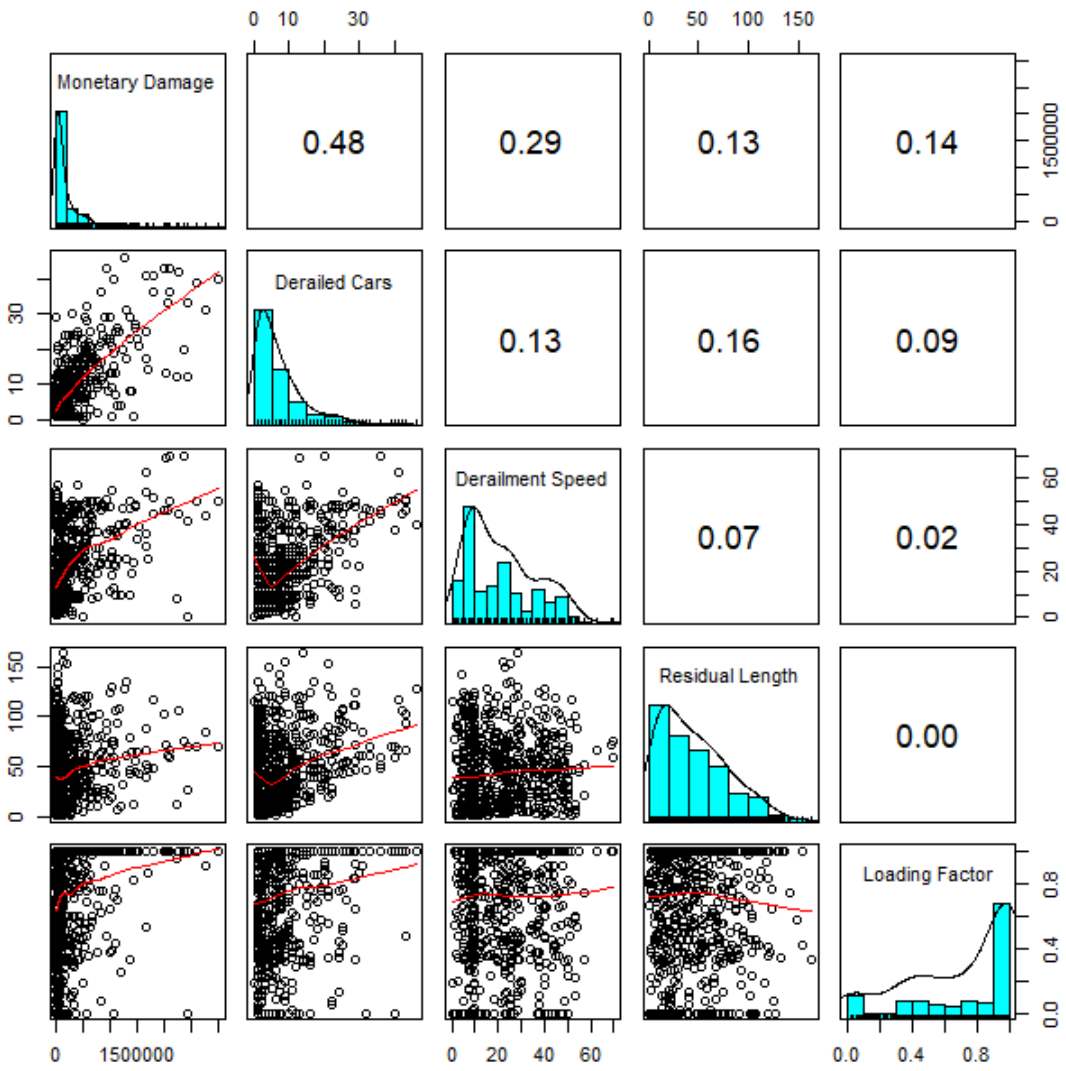


Figure 3.26: Correlation plot matrix of variables for all freight-train derailments

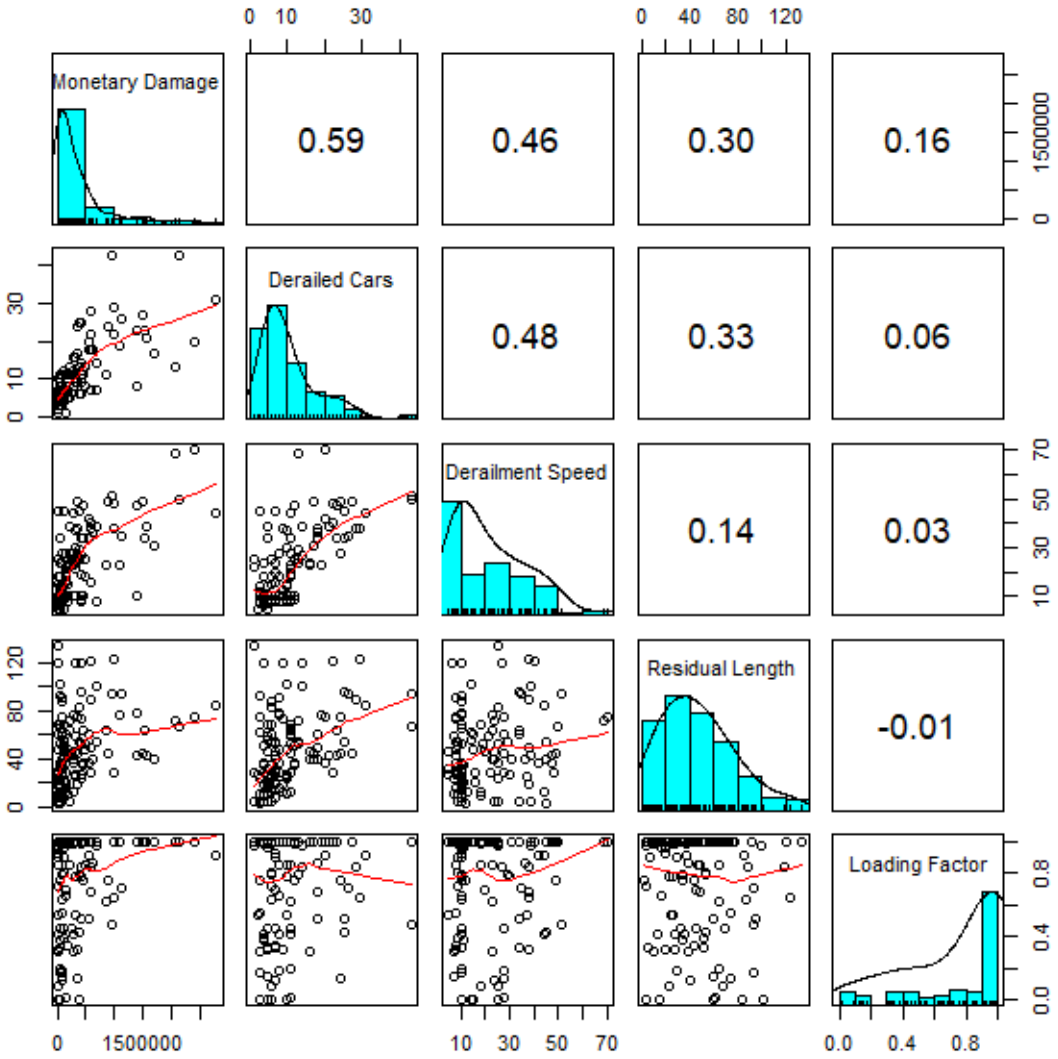


Figure 3.27: Correlation plot matrix of variables for broken-rail caused freight-train derailments

3.6 Concluding Remarks

Based on the exploratory data analysis conducted in this chapter, the following conclusions can be made:

1. High degree of uncertainty due to high variation in track geometry condition including recovery values of the parameters was observed as shown by their respective box plots. Due to this variation, a probabilistic (stochastic) approach to modeling tamping recovery of track geometry appears to be suitable.
2. The variables of interests in both track geometry and derailment severity data sets were found to be non-normally distributed. Various violations of normality were exhibited including skewness and fat tails (as shown in their respective histograms, Q-Q plots and box plots). Due to the non-normality of the marginal and joint distribution of the variables of interest, a copula-based approach to modeling tamping recovery of track geometry recovery and train derailment severity appears to be suitable.
3. The derailment severity outcomes namely monetary damage and number of derailed cars appears to be (highly) correlated (as shown in the scatter/correlation plot matrix). Therefore, it does not appear appropriate to dismiss the underlying dependence between the severity outcomes in a multivariate regression framework. Thus, it appears appropriate to jointly analyze their relationship with a set of covariates that might affect both outcomes using a copula-based regression model in order to enhance severity prediction.

REFERENCES

Andrade, A. R. and Teixeira, P. F. Hierarchical Bayesian modelling of rail track geometry degradation. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(4):364–375, 2013. ISSN 0954-4097. doi: 10.1177/0954409713486619.

Audley, M. and Andrews, J. The effects of tamping on railway track geometry degradation. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(4):376–391, 2013. ISSN 0954-4097. doi: 10.1177/0954409713480439.

Barkan, Christopher P. L.; Dick, C. Tyler, and Anderson, Robert. T. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 1825(9):64–74, 2003. ISSN 03611981.

Behrens, John T. Principles and procedures of exploratory data analysis. *Psychological Methods*, 2(2):131–160, 1997. ISSN 1082-989X. doi: 10.1037/1082-989X.2.2.131.

Caetano, Luis Filipe and Teixeira, Paulo Fonseca. Optimisation model to schedule railway track renewal operations: a life-cycle cost approach. *Structure and Infrastructure Engineering*, 11(11):1524–1536, 2015. ISSN 1573-2479. doi: 10.1080/15732479.2014.982133.

FRA, . Monetary Threshold for Reporting Rail Equipment Accidents/Incidents for Calendar Year 2017. *Federal Registry*, 81(247):57–60, 2016.

Khouy, Iman Arasteh. *Cost-Effective Maintenance of Railway Track Geometry*. PhD thesis, Lulea University of Technology, 2013.

Lichtberger, Bernhard. *Track compendium : formation, permanent way, maintenance, economics*. Eurailpress, 2005. ISBN 3777103209.

Liu, Xiang; Saat, M. Rapik; Qin, Xiao, and Barkan, Christopher P L. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. *Accident Analysis and Prevention*, 59:87–93, 2013. ISSN 00014575. doi: 10.1016/j.aap.2013.04.039.

Saccomanno, F. F. and El-Hage, S. M. Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. *Transportation Research Record*, (1245):34–51, 1989.

Saccomanno, F. F. and El-Hage, S. M. Establishing derailment profiles by position for corridor shipments of dangerous goods. *Canadian Journal of Civil Engineering*, 18 (1):67–75, 1991. ISSN 0315-1468. doi: 10.1139/l91-009.

Soleimanmeigouni, I.; Ahmadi, A.; Arasteh Khouy, I., and Letot, C. Evaluation of the effect of tamping on the track geometry condition: A case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:408–420, 2016a. ISSN 0954-4097. doi: 10.1177/0954409716671548.

Soleimanmeigouni, I.; Ahmadi, A., and Kumar, U. Track geometry degradation and maintenance modelling: A review. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:73–102, 2016b. ISSN 0954-4097. doi: 10.1177/0954409716657849.

Tukey, John Wilder. *Exploratory data analysis*. Addison-Wesley Pub. Co, 1977. ISBN 0201076160.

Chapter 4

COPULA MODELS

This chapter provides a detailed overview on copulas. The basic concepts of copula function theory are introduced. The various classes of copulas, dependence concepts and measures, statistical inference (parameter estimation) of copulas and copula selection techniques are subsequently discussed. Finally, a case study is presented in which the tamping recovery of various track geometry parameters are modeled using a copula-based approach.

4.1 General

Copulas are functions that combine or link multivariate distribution functions to their univariate marginal distribution functions. An n-dimensional copula is a multivariate distribution function $C(u_1, ..., u_n)$ defined on the unit hypercube $[0, 1]^n$, with n-random variables as uniformly distributed marginals (Nelsen, 2006; Czado et al., 2012; Zilko et al., 2016). C is a bivariate copula if $C : [0, 1]^2 \rightarrow [0, 1]$ and meets the following conditions:

1. $C(u, 0) = C(0, v) = 0$ for any $u, v \in [0, 1]$
2. $C(u, 1) = u$ and $C(1, v) = v$ for any $u, v \in [0, 1]$
3. $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0 \quad \forall \quad 0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$

The copula approach via Sklar's theorem (Sklar, 1959) permits the separation of the multivariate distribution into univariate margins, and the dependence structure which is modelled via the copula function without loss of information (Dalla Valle et al., 2016). Sklar's theorem (Sklar, 1959) offers the link between univariate marginals and copula to the multivariate joint distribution. Sklar's Theorem states that for any

n-dimensional distribution function with given marginals $F_1, ..., F_n$, there exists an n-dimensional copula $C : [0,1]^n \to [0,1]$ such that for all $(x_1, ..., x_n) \in \mathbb{R}^n$

$$F(x_1, ..., x_n) = C\{F_1(x_1), ..., F_n(x_n)\} \tag{4.1}$$

holds. C is unique if each $F_i(x)$ is continuous; otherwise it is uniquely determined by the product of their ranges (Range of $F_1 \times ... \times$ Range of F_n).

Sklar's theorem offers a useful means of constructing copulas given the marginals $F_1, ..., F_n$ such that

$$C(x_1, ..., x_n) = F\left(F_1^{-1}(x_1), ..., F_n^{-1}(x_n)\right) \tag{4.2}$$

If F is absolutely continuous, then the copula density c is well defined and can be written as

$$c(u_1, ..., u_n) = \frac{\partial^n C(u_1, ..., u_n)}{\partial u_1, ..., \partial u_n} \tag{4.3}$$

The density f of the multivariate distribution F given the copula density c can be expressed as

$$f(x_1, ..., x_n) = c\{F_1(x_1), ..., F_n(x_n)\} \prod_{i=1}^{n} f_i(x_i) \tag{4.4}$$

Via recursive conditioning, a four-dimensional density function can be expressed/ decomposed as follows:

$$
\begin{aligned}
f(x_1, x_2, x_3, x_4) &= f_1(x_1) \cdot f(x_2|x_1) \cdot f(x_3|x_1, x_2) \cdot f(x_4|x_1, x_2, x_3) && (4.5)\\
f(x_2|x_1) &= \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) f_2(x_2)}{f_1(x_1)} && (4.6)\\
f(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2)) \cdot f_2(x_2) && (4.7)\\
f(x_3|x_1, x_2) &= \frac{f(x_2, x_3|x_1)}{f(x_2|x_1)} = \frac{c_{23|1}(F(x_2|x_1), F(x_3|x_1)) f(x_2|x_1) f(x_3|x_1)}{f(x_2|x_1)} && (4.8)\\
f(x_3|x_1, x_2) &= c_{23|1}(F(x_2|x_1), F(x_3|x_1)) f(x_2|x_1) f(x_3|x_1) && (4.9)
\end{aligned}
$$

But

$$
\begin{aligned}
f(x_3|x_1) &= \frac{f(x_1, x_3)}{f_1(x_1)} = \frac{c_{13}(F_1(x_1), F_3(x_3)) \cdot f_1(x_1) f_3(x_3)}{f_1(x_1)} && (4.10)\\
f(x_3|x_1) &= = c_{13}(F_1(x_1), F_3(x_3) \cdot f_3(x_3)) && (4.11)
\end{aligned}
$$

Thus

$$
\begin{aligned}
f(x_4|x_1, x_2, x_3) &= \frac{f(x_3, x_4|x_1, x_2)}{f(x_3|x_1, x_2)} && (4.12)\\
f(x_4|x_1, x_2, x_3) &= = \frac{c_{34|12} F(x_3|x_1, x_2) F(x_4|x_1, x_2) f(x_3|x_1, x_2) f(x_4|x_1, x_2)}{f(x_3|x_1, x_2)} && (4.13)\\
f(x_4|x_1, x_2, x_3) &= c_{34|12}(F(x_3|x_1, x_2) F(x_4|x_1, x_2)) f(x_4|x_1, x_2) && (4.14)
\end{aligned}
$$

But

$$
f(x_4|x_1, x_2) = c_{24|1}(F(x_2|x_1), F(x_4|x_1)) f(x_2|x_1) f(x_4|x_1) \quad (4.15)
$$

Thus finally

$$f(x_4|x_1,x_2,x_3) = \\ c_{34|12}(F(x_3|x_1,x_2)F(x_4|x_1,x_2))c_{24|1}(F(x_2|x_1),F(x_4|x_1))c_{14}(F_1(x_1),F_4(x_4))f_4(x_4) \tag{4.16}$$

Copula-based modeling is an emerging statistical method which has been widely used in the financial industry and is gaining traction in civil engineering. Copula-based methodologies have been applied in hydrology or water resources (Salvadori and De Michele, 1992; Grimaldi and Serinaldi, 2006; Genest and Favre, 2007), travel behavior modeling (Bhat and Eluru, 2009) and vehicle axle weight modeling (Srinivas et al., 2006). Other areas include infrastructure (pavement) dependence modeling (Attoh-Okine, 2013), pipeline data analysis (Atique and Attoh-Okine, 2016), and automobile injury severity studies (Eluru et al., 2010; Nashad et al., 2016). However, its application in the railroad industry is very limited with Zilko et al. (2016) modeling railroad disruption lengths using Copula Bayesian Networks.

4.2 Classes of Copulas

There are two popular classes of Copulas namely elliptical copulas and Archimedean copulas with 3rd less common class called Extreme-value copulas (Yan, 2006).

4.2.1 Elliptical Copulas

Elliptical (or meta-elliptical) copulas are copulas of elliptical distributions. Elliptical contoured distributions are radially symmetric. Members of this family include bivariate normal, bivariate Pearson type II and type VII distributions (the latter including bivariate t and Cauchy distributions as special cases) (Nelsen, 2006). A multivariate elliptical distribution of random vector $(X_1, ..., X_n)$ centered at zero has density of the form $\phi(w) = \psi(w^\top \Sigma w)$, where $w \in \mathbb{R}^n$ and Σ is a $n \times n$ dispersion matrix, which can be parameterized such that $\Sigma_{ij} = Cov(X_i, X_j)$ (Yan, 2006). The margins of elliptical distributions are all of the same type (Embrechts et al., 2003).

Elliptical copulas are directly obtained by the inversion of Sklar's Theorem and thus can be expressed in the form

$$C(u_1, ..., u_n) = F\left(F_1^{-1}(u_1), ..., F_n^{-1}(u_n)\right) \tag{4.17}$$

The two most common elliptical copulas are the Normal or Gaussian copula and the student t-copula which are related to the multivariate normal and multivariate student-t distributions respectively. Both of these copulas are radially symmetric and tail-symmetric. However, student-t copula has tail dependence whereas Gaussian copula has no tail dependence. The bivariate Gaussian copula can be expressed as

$$C(u_1, u_2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \tag{4.18}$$

where Φ_ρ and Φ are the bivariate and univariate standard normal distribution functions respectively and $\rho \in (-1, 1)$ is the dependence parameter. The bivariate Student-t copula can be expressed as

$$C(u_1, u_2) = t_{\rho,\nu}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)) \tag{4.19}$$

where $t_{\rho,\nu}$ and t_ν are the bivariate and univariate Student-t distribution functions respectively and the degrees of freedom parameter $\nu > 2$.

Similar to elliptical distributions, simulation from elliptical copulas are easy. However, elliptical copulas do not have closed form expressions and are only radially symmetrical (Embrechts et al., 2003). Properties of bivariate elliptical copula families including parameter range, Kendall's tau and tail dependence is given in Table 4.1.

4.2.2 Archimedean Copulas

Archimedean Copulas are constructed by means of a complete monotonic function without the need for distribution functions or random variables (Yan, 2006). Unlike elliptical copulas, Archimedean copulas have closed form expressions and are not

Table 4.1: Properties of bivariate elliptical copula families (Brechmann and Schepsmeier, 2013)

Elliptical copula	Parameter Range	Kendall's Tau	Tail Dependence
Gaussian/ Normal	$\rho \in (-1, 1)$	$\frac{2}{\pi} \arcsin(\rho)$	0
Student-t	$\rho \in (-1, 1), \nu > 2$	$\frac{2}{\pi} \arcsin(\rho)$	$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$

restricted to only radial symmetry. Archimedean copulas can be expressed as

$$C(u_1, ..., u_n) = \varphi^{-1}\left(\varphi(u_1), ..., \varphi(u_n)\right) \tag{4.20}$$

where the generator of the copula, $\varphi : [0, 1] \to [0, \infty]$ is a continuous strictly decreasing convex function such that $\varphi(0) = \infty$ and $\varphi(1) = 0$ and φ^{-1} is its pseudo-inverse which is given as

$$\varphi^{-1}(\nu) = \begin{cases} \varphi^{-1}(\nu) & 0 \leq \nu \leq \varphi(0) \\ 0 & \varphi(0) \leq \nu \leq \infty \end{cases} \tag{4.21}$$

Common one-parameter Archimedean copulas include Gumbel, Clayton, Frank and Joe Copulas. Gumbel and Joe copulas are suitable for modeling upper tail dependence whereas Clayton performs well with lower tail dependence. The Joe Copula has an even stronger positive upper tail dependence in comparison to the Gumbel copula and can be observed by tighter clustering of observations in the upper tail. Frank copula is suitable for radially symmetric dependence with very weak tail dependencies (even weaker than the Gaussian copula) (Bhat and Eluru, 2009). Common two-parameter Archimedean copula families include Clayton-Gumbel (BB1), Joe-Gumbel (BB6), Joe-Clayton (BB7) and Joe-Frank (BB8) which are more flexible. The more flexible structures of Clayton-Gumbel (BB1) and Joe-Clayton (BB7) permit different non-zero lower and upper tail dependence coefficients. The properties of various one-parametric and two-parametric bivariate Archimedean copulas are given in Table 4.2.

Table 4.2: Properties of Archimedean bivariate copula families (Brechmann and Schepsmeier, 2013)

Name	Generator function	Parameter range	Kendall's τ	Tail dependence (lower,upper)
Clayton	$\frac{1}{\theta}(t^{-\theta}-1)$	$\theta > 0$	$\frac{\theta}{\theta+2}$	$(2^{-\frac{1}{\theta}}, 0)$
Gumbel	$(-\log t)^{\theta}$	$\theta \geq 1$	$1-\frac{1}{\theta}$	$(0, 2-2^{\frac{1}{\theta}})$
Frank	$-\log\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right]$	$\theta \in \mathbb{R} \setminus \{0\}$	$1-\frac{4}{\theta}+4\frac{D_1(\theta)}{\theta}$	$(0,0)$
Joe	$-\log[1-(1-t)^{\theta}]$	$\theta > 1$	$1+\frac{4}{\theta^2}\int_0^1 t\log(t)(1-t)^{\frac{2(1-\theta)}{\theta}}dt$	$(0, 2-2^{\frac{1}{\theta}})$
Clayton-Gumbel (BB1)	$(t^{-\theta}-1)^{\delta}$	$\theta > 0$, $\delta \geq 1$	$1-\frac{2}{\delta(\theta+2)}$	$(2^{-\frac{1}{\theta\delta}}, 2-2^{\frac{1}{\delta}})$
Joe-Gumbel (BB6)	$(-log[1-(1-t)^{\theta})^{-\delta}$	$\theta \geq 1$, $\delta \geq 1$	$1+\frac{4}{\delta\theta}\int_0^1(-\log(1-(1-t)^{\theta}) \times(1-t)(1-(1-t)^{-\theta}))dt$	$(0, 2-2^{\frac{1}{\delta\theta}})$
Joe-Clayton (BB7)	$(1-(1-t)^{\theta})^{-\delta}-1$	$\theta \geq 1$, $\delta > 0$	$1+\frac{4}{\theta\delta}\int_0^1(-(1-(1-t)^{\theta})^{\delta+1} \times\frac{(1-(1-t)^{\theta})^{-\delta})-1}{(1-t)^{\theta-1}}\Big)dt$	$(2^{-\frac{1}{\delta}}, 2-2^{\frac{1}{\theta}})$
Joe-Frank (BB8)	$-log\left[\frac{1-(1-\delta t)^{\theta}}{1-(1-\delta)^{\theta}}\right]$	$\theta \geq 1$, $\delta \in (0,1]$	$1+\frac{4}{\theta\delta}\int_0^1\left(-\log(\frac{(1-t\delta)^{\theta}-1}{(1-\delta)^{\theta}-1}\right) \times(1-t\delta)(1-(1-t\delta)^{-\theta}))\,dt$	(0,0)

4.3 Dependence Measures

4.3.1 Dependence Concepts

Two random variables (X, Y) are said to be dependent if they do not satisfy the condition of probabilistic independence i.e. $F(x, y) \neq F_X(x)F_Y(y)$ (Liu, 2011). Global correlation coefficients measure the average dependence over the domain of the variables of interest (Lewandowski, 2008). A global dependence measure summarizes the dependence structure of two random variables in a single number. Examples include Pearson's correlation coefficient, Kendall's tau and Spearman's rho. Given that $\delta(\cdot, \cdot)$ is a scalar measure of dependence, the following are desirable characteristics of a dependence measure (Embrechts et al., 2002):

1. $\delta(X, Y) = \delta(Y, X)$ (symmetry)
2. $-1 \leq \delta(X, Y) \leq +1$ (normalization)
3. $\delta(X, Y) = 1 \iff (X, Y)$ (comonotonic)
   $\delta(X, Y) = -1 \iff (X, Y)$ (countermonotonic)
4. For $T : \mathbb{R} \to \mathbb{R}$ strictly monotonic on the range of X:
   $$\delta(T(X), Y) = \begin{cases} \delta(X, Y) & T \textit{ increasing}, \\ -\delta(X, Y) & T \textit{ decreasing}. \end{cases}$$

Dependence can be evaluated using several concepts such as linear correlation, concordance (or rank correlation) and tail dependence (Liu, 2011). Linear correlation fulfills only properties 1 and 2. Rank correlation fulfills not only properties 1 and 2 but also properties 3 and 4 if the variables are continuous.

4.3.2 Linear Correlation

The most popular dependence between two random variables (X, Y) is the Pearson product-moment correlation coefficient (also known as Pearson's coefficient or linear correlation coefficient) which is associated with linear dependence and multivariate normal distribution. Linear correlation is a natural dependence measure for multivariate normality and, more generally, elliptically distributed distributions (Embrechts

et al., 2002). Pearson's coefficient can be mathematically expressed as:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.22}$$

where $cov(X,Y)$ is the covariance between X and Y, $cov(X,Y) = E(X,Y) - E(X)E(Y)$ and σ_X and σ_Y are the standard deviations of X and Y respectively. If two variables are independent $\rho(X,Y) = 0$ since $cov(X,Y) = 0$. Given perfect linear dependence of the variables i.e. $Y = aX + b$, $\rho(X,Y) \pm 1$ whereas $-1 < \rho(X,Y) < +1$ given imperfect linear dependence (Liu, 2011).

Pearson's coefficient is highly popular for the following reasons (Embrechts et al., 2002)

1. It is easy to compute since it is easy to find the second moments (variances and covariances) for most bivariate distributions.
2. It is invariant given linear transformations of the variables.
3. It is a natural dependence measure for multivariate spherical and elliptical distributions.

Assumption of multivariate normality allows for the use of multivariate methods such as principal component analysis and factor analysis however this does not apply to all cases with non-normality occurring in many applications making the aforementioned techniques unsuitable. Non-normality transpires in various forms: non-normality of marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables (Yan, 2006). Linear correlation has the following limitations (Embrechts et al., 2002; Liu, 2011):

1. Pearson's coefficient is undefined if the second moments of the variables do not exist. Thus, they do not always exist and this is common in heavy-tailed distributions.
2. Pearson's coefficient is invariant given strictly increasing nonlinear transformations of the variables.

3. Independence between two random variables indicates zero linear correlation (uncorrelation). However, uncorrelation does not generally imply independence and only applies to multivariate normality.
4. Pearson's coefficient is not a robust measure since one observation can have an arbitrary high influence on the linear correlation.

The shortcomings of linear correlation have led to the consideration of other dependence concepts such as concordance.

4.3.3 Rank Correlation/Concordance Measure

Rank correlation coefficients measure the correspondence between two rankings and assesses its significance (Liu, 2011). This form of dependence that evaluates the consistency or agreement of two or more sets of rankings is also known as concordance. There are several advantages of using rank correlations over ordinary product moment correlations such as Pearson's correlation coefficient. These advantages include(Bedford and Cooke, 2001):

1. they always exist.
2. they are independent of marginal distributions meaning they can take any value in the [-1, 1] interval
3. they are invariant under monotonic increasing transformations of the marginals .

Since copulas are also invariant under monotone transformations, scale-invariant measures of dependence such as Kendall's Tau and Spearman's Rho are more suitable for evaluating the degree of dependence. They are both rank correlations and remain unaltered under strictly increasing transformations (Nelsen, 2006; Yan, 2006; Ayuso et al., 2016). Both can be defined via a concordance function, Q which is the difference between concordance and discordance probabilities of two continuous vectors (X_a, Y_a) and (X_b, Y_b) with likely differing joint distributions A and B but similar margins (F_a) and (F_b). The concordance function Q can be expressed as:

$$Q = P_C - P_D = Pr((X_a - X_b)(Y_a - Y_b) > 0) - Pr((X_a - X_b)(Y_a - Y_b) < 0) \quad (4.23)$$

where P_C is the concordance probability and P_D is the discordance probability. Given the copulas of the joint distributions A and B, the concordance function Q can be also expressed as:

$$Q = Q(C_A, C_B) = 4 \int_0^1 \int_0^1 C_A(u, v) dC_B(u, v) - 1 \tag{4.24}$$

Other concordance measures include Gini's measure of association, Blomqvist's measure of association (or medial correlation coefficient) and Moran's coefficient (see (Nelsen, 2006; Dorey and Joubert, 2005).

4.3.3.1 Kendall's Tau

Kendall's tau measure [τ or $Q(C, C)$] can be defined as the difference between the concordance and discordance probabilities of two independent and identically distributed pairs of observations. Thus given a bivariate random vector (X_a, Y_a) with copula C, it can be expressed as (Yan, 2006):

$$\tau = Q(C, C) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \tag{4.25}$$

The sample version of Kendall's tau can be defined as the difference between the probabilities of concordance and discordance for a pair of observations that is chosen randomly from the sample. This can be expressed as (Nelsen, 2006):

$$\tau = \frac{N_C - N_D}{N} \tag{4.26}$$

Where N_C is the number of concordant pairs, N_D is the number of discordant pairs and N is the number of distinct pairs of observations from a random sample which is equal to $\binom{n}{2}$ where n is the number of observation from a vector of continuous random variables. The population version of Kendall's tau can be defined as the difference

between the concordance and discordance probabilities and can be expressed as:

$$Q = P_C - P_D = Pr((X_a - X_b)(Y_a - Y_b) > 0) - Pr((X_a - X_b)(Y_a - Y_b) < 0) \quad (4.27)$$

Non-linear dependence is usually evaluated using Kendall's tau (Czado et al., 2012). Kendall's tau measures dependence independent of the assumed distribution and thus is suitable when linking various (non-Gaussian) copula families (Dissmann et al., 2013). For Archimedean copulas, the closed form expression of Kendall's tau is based on the copula-specific generator function whereas their computation for Elliptical copulas are more complicated (Schepsmeier and Czado, 2016). The Kendall's tau for various bivariate elliptical copulas and bivariate Archimedean copulas are shown in Tables 4.1 and 4.2 respectively.

4.3.3.2 Spearman's Rho

Spearman's Rho [ρ or $3Q(C, \Pi)$] is often referred to as the 'grade correlation coefficient' where grades are population analogs of ranks (Nelsen, 2006). The population version of Spearman's Rho is proportional to the difference between the concordance and discordance probabilities for the vectors (X_a, Y_a) and (X_b, Y_c) with similar margins however one vector has a distribution function A, whereas the elements of the other are independent. The population version of Spearman's Rho can be expressed as:

$$\rho_{X,Y} = 3\left(Pr[(X_a - X_b)(Y_a - Y_c) > 0] - Pr[(X_a - X_b)(Y_a - Y_c) < 0]\right) \quad (4.28)$$

where 3 is a normalization factor that scales ρ into the range of $[-1, 1]$ and (X_c, Y_b) can equally be used in place of (X_b, Y_c). In terms of copula, Spearman's rho is proportional to the difference between concordance and discordance probabilities of two vectors: both with the same margins however one has copula C and the other product copula Π obtained under independence. This can be expressed as:

$$\rho = 3Q(C, \Pi) = 12 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 3 \quad (4.29)$$

Spearman's Rho is equivalent to Pearson's product-moment correlation coefficient for the probability-integral-transformed variables or the grades of a pair of continuous random variables.

4.3.4 Tail Dependence

All the aforementioned correlation coefficients measure the average dependence over the domain of the variables of interest. Tail dependence is the measure that tries to capture the dependence more locally rather than globally, in the tails (lower and/or upper) of distribution (Lewandowski, 2008). It can be defined as the measure of the co-movements in the tails of the distributions of random variables. Tail dependence also relates to the conditional probability that one variable exceeds some value given that another exceeds some value. For continuous marginal distributions, tail dependence is a copula property; hence it is invariant under monotonic transformations (Liu, 2011). Let $Y = (Y_1, Y_2)$ be a pair of random variables. The pair is said to be upper tail dependent if

$$\lambda_U = \lim_{\upsilon \to 1} P\{Y_1 > F_1^{-1}(\upsilon) | Y_2 > F_2^{-1}(\upsilon)\} > 0 \tag{4.30}$$

if the limit λ_U exists. This is the probability that Y_1 reaches extremely large values, given that Y_2 attains extremely large values. Similarly, the pair is said to be lower tail dependent if

$$\lambda_L = \lim_{\upsilon \to 0} P\{Y_1 \leq F_1^{-1}(\upsilon) | Y_2 \leq F_2^{-1}(\upsilon)\} > 0 \tag{4.31}$$

if the limit λ_L exists. The lower and upper tail dependence coefficients of various elliptical and Archimedean copula families can be found in tables 4.1 and 4.2 respectively. If the lower and upper tail coefficients differ, the dependence can be said to be asymmetric. Asymmetric dependence is dependence that is not identical on both sides of a central line or line of symmetry over the domain of the variables of interest. On the other hand, symmetric dependence is dependence that is identical on both sides of a central line or line of symmetry.

4.4 Statistical Inference of Copulas

Estimation approaches of copulas can be categorized into four groups namely (Patton, 2012):

- Parametric estimation methods
- Semi-parametric estimation methods
- Non-parametric estimation methods
- Other estimation methods

Estimation of copulas has essentially been developed in the context of independent and identically distributed random variables (i.i.d.) samples (Fermanian and Scaillet, 2003). Since most copula estimation research refers to independent samples of a vector of random variables, care is needed when applying these techniques to other data such as time series data (Morettin et al., 2011). Copula estimation methods have also been modified or developed for time series data (Patton, 2012).

4.4.1 Parametric Estimation Methods

Parametric estimation methods assume both parametric marginal distributions and parametric copulas. Parametric estimation methods include method of moments estimation, maximum likelihood techniques such as full (exact) maximum likelihood estimation and inference of function for margins.

4.4.1.1 Full Maximum Likelihood Estimation

Full maximum likelihood estimation techniques involve the simultaneous maximization of joint distribution model parameters (Nicoloutsopoulos, 2005). Full maximum likelihood estimation is usually preferred for estimation or inference due to its widely known optimality characteristics (Kim et al., 2007). In this method, there is the need to assume parametric marginal distributions. Given the right specification of marginals, the estimator has the usual optimality characteristics of the maximum likelihood estimator (Weiß, 2011). Thus, estimation of copulas are performed fully parametrically by assuming parametric models for both the marginals distributions

and copula and then conducting maximum likelihood estimation (Chen and Huang, 2007).

Given the parameter of interest is $\theta = (\beta^T, \alpha^T)^T$ where β is the marginals parameter vector and α is the association or dependence parameter vector for the copula, the exact log-likelihood of the parameter vector θ expressed from equation 4.4 is given as (Yan, 2006):

$$l(\theta) = \sum_{i=1}^{n} \log c[F_1(X_{i1};\beta), ..., F_p(X_{ip};\beta);\alpha] + \sum_{i=1}^{n}\sum_{j=1}^{p} \log f_i(X_{ij};\beta) \tag{4.32}$$

The full-approach maximum likelihood estimator of θ is given by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \tag{4.33}$$

where Θ is the parameter space. Under regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is consistent and asymptotically efficient with limiting distribution

$$\sqrt{N}(\hat{\theta} - \theta_0) \to N[0, I^{-1}(\theta_0)] \tag{4.34}$$

where θ_0 is the true parameter value and I is the Fisher information matrix. When all components of the multivariate model are parametric, maximum likelihood is the most efficient estimation method (Patton, 2012). However, the full or exact approach of maximum likelihood (ML) estimation can create a computation burden even for relatively simple bivariate models with large number of parameters to be estimated. This burden becomes much larger in higher dimensions (Patton, 2012).

4.4.1.2 Inference of Functions for Margins

Separation of margins and copula suggests that one may estimate the marginal parameters and association parameters in two steps, leading to the possible use of a stepwise or multistage approach. For this reason as well as the computation burden of exact maximum likelihood, the Inference of Functions for Margins (IFM) proposed by

Joe and Xu (1996) is suitable. IFM estimation also known as stepwise or multistage maximum likelihood estimation is one of the most widely used estimation methods (Oh and Patton, 2013). The IFM estimation method estimates the marginal parameters β initially in a first step by (Yan, 2006):

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{p} \log f_i(X_{ij}; \beta) \tag{4.35}$$

and subsequently estimates the association (copula) parameters given the marginal parameters by

$$\hat{\alpha} = \arg\max_{\theta} \sum_{i=1}^{n} \log c \times [F_1(X_{i1}; \hat{\beta}), ..., F_p(X_{ip}; \hat{\beta}); \alpha] \tag{4.36}$$

In the case where each margin has its own parameters β_i resulting in $\beta = (\beta_1^T, ..., \beta_p^T)^T$, the initial step comprises of maximum likelihood estimation of each marginal distribution $j = 1, ..., p$ which is given by

$$\hat{\beta} = \arg\max_{\beta_j} \sum_{i=1}^{n} \log f_i(X_{ij}; \beta_j) \tag{4.37}$$

In this instance, each maximization task has a few number of parameters, thereby drastically decreasing the computational burden. In comparison with the full ML estimator, the IFM estimator has advantages in numerical computations and is asymptotically efficient (Yan, 2006). The two estimation methods have been found to be almost equally efficient in many cases (Kim et al., 2007). IFM is asymptotically less efficient than full MLE (except in the special case where the variables are independent) (Patton, 2012). In finite samples, the IFM estimator has been found to be highly efficient relative to the full ML estimator. The IFM estimate can be employed as an initial value in a full ML estimation (Yan, 2006). Thus, inference function for margins (IFM) method has become the preferred fully parametric method since it is close to maximum likelihood

(ML) in approach and is easier to implement. Since IFM is a fully parametric, misspecification of the marginal distributions may have an impact on the performance of the estimator (Kim et al., 2007).

4.4.2 Semi-parametric Estimation Methods

Semiparametric copula-based models employ a nonparametric model for estimation the marginal distributions parameters and a parametric model for the copula parameters (Patton, 2012). In pseudo-maximum likelihood estimation (PML), parametric marginal distributions are replaced by empirical cumulative distribution functions (CDF) and the copula parameters are subsequently estimated using maximum likelihood estimation (Weiß, 2011). Pseduo-maximum likelihood (also known as canonical maximum likelihood) employs the empirical CDF of each margin to transform the observations $(X_{i1}, ..., X_{ip})^T$ into pseudo-observations with uniform margins $(U_{i1}, ..., U_{ip})^T$ (Yan, 2006). The estimator $\hat{\theta}$ is given by

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log c(U_{i1}, ..., U_{ip}; \theta) \tag{4.38}$$

The PML estimator is consistent, asymptotically normal, and fully efficient at independence (Yan, 2006). Pseudo-maximum likelihood estimation (PMLE) has been found to be not as efficient as the MLE in general and PMLE has been found to be asymptotically efficient under certain conditions (Kim et al., 2007). Kim et al. (2007) found that PML estimation was much better suited for parameter estimation than various parametric estimation techniques.

The PMLE estimation proposed by Genest et al. (1995) can be considered as the semiparametric equivalent of IFM estimation. Empirical margins are employed which are scaled using an asymptotically negligible factor $\frac{n}{n+1}$ to avoid "difficulties arising from the potential unboundedness of the $\log(c(x, y|\theta))$ as some of the x,y tend to one" (Nicoloutsopoulos, 2005). This general approach by Genest et al. (1995) has been examined in different contexts with various modifications (Kim et al., 2007). An

example is the "rank approximate Z-estimator" (RAZ) proposed by Tsukahara (2005) which considers the pseudo-maximum likelihood estimator by Genest et al. (1995) as a special case of RAZ-estimator.

4.4.3 Non-parametric Estimation Methods

Non-parametric estimation treats both the copula and the marginals parameter-free and thus provides the greatest generality. Non-parametric estimation can offer initial information required for revealing and subsequent formulation of an underlying parametric copula model (Chen and Huang, 2007). Copula can be estimated using non-parametric estimation methods based on empirical distributions or copulas proposed by Deheuvels (1979). These empirical copulas resemble usual multivariate empirical cumulative distribution functions and are highly discontinuous (constant on some data-dependent pavements). Thus, they cannot be employed as a graphical tool (Fermanian and Scaillet, 2003).

Fermanian and Scaillet (2003) proposed a nonparametric estimation for copulas for time series employing a kernel based approach. This methodology has the advantage of offering smooth (differentiable) reconstitution of the copula function without assuming a parametric dependence structure and without losing the usual parametric rate of convergence. Morettin et al. (2011) proposed a non-parametric approach for estimating copulas for time series using wavelets following a similar approach by Fermanian and Scaillet (2003). Densities are initially estimated, followed by distribution functions and quantiles and finally the copula is estimated.

4.4.4 Other Estimation Methods

Other estimation methods include method of moments estimation, minimum distance estimation, simulation techniques and Bayesian estimation.

4.4.4.1 Method of moments estimation

Despite pseudo-maximum likelihood estimation being the recognized standard for rank-based estimation of copula parameters, the method can be be computationally

intensive and its application is restricted to instances where the copula has a density with respect to Lebesgue measure. Thus, nonparametric analogues of the method of moments are usually employed (Genest et al., 2013). Method of moments estimation takes advantage of known invertible one-to-one mapping between the parameter(s) of certain copulas and certain measures of dependence such as concordance measures (Patton, 2012; Oh and Patton, 2013). Method of moments estimation is suitable in cases where mapping is known in closed form (Oh and Patton, 2013). Also, moment-based methods are suitable in cases where the copula does not have a density or where the maximization of the log pseudo-likelihood function is computationally intensive. Thus, providing a quick estimation of the copula parameter or at least an initial value for pseudo maximum-likelihood estimation (Genest et al., 2013). However, method of moments estimation does not seem to perform well when the copula parameter is a vector with efficiency issues relating to its extension to multiparameter families such as log-copulas (Nicoloutsopoulos, 2005).

Common methods of moments estimators include inversion of Kendall's tau (estimate based on Kendall's Tau) and inversion of Spearman's rho (estimate based on Spearman's rho). Since they are ranked-based, they can be referred to as a non-parametric adaptation of the celebrated method of moments (Genest and Favre, 2007). They are also known to be consistent and asymptotically normal under weak regularity conditions (Genest et al., 2013).

The most popular method of moment technique is the inversion of Kendall's Tau (estimate based on Kendall's tau) primarily due to its form being often explicit. The inversion of Kendall's Tau's consists of solving the equation $\tau(C_\theta) = \tau_n$ for θ (Genest et al., 2013). Given that the copula parameter θ is equal to a smooth function g of the population version of Kendall's tau τ then $\tilde{\theta} = g(\tau_n)$ provides the Kendall-based estimate of θ (Genest and Favre, 2007). The asymptotic behavior of the estimate based of Kendall's tau can be derived through the characteristics of the concordance measure using the theory of U-statistics (Genest et al., 2013). An adaptation of Proposition 3.1

of Genest and Rivest (1993) implies that

$$\sqrt{N}\frac{\tau_n - \tau}{4s} \approx \mathcal{N}(0,1) \tag{4.39}$$

where

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(W_i + \tilde{W} - 2\bar{W})^2 \tag{4.40}$$

and

$$\tilde{W}_i = \frac{1}{n}\sum_{j=1}^{n} I_{ji} = \frac{1}{n}\#\{j : X_i \leq X_j, Y_i \leq Y_j\} \tag{4.41}$$

Delta method, an application of Slutsky's theorem implies that as $n \to \infty$

$$\tilde{\theta} \approx \mathcal{N}[\theta, \frac{1}{n}\{4Sg'(\tau_n)\}^2] \tag{4.42}$$

The estimate based on Spearman's rho can be derived in a similar fashion as that of Kendall's tau. Given that the copula parameter θ is equal to a smooth function h of the population version of Spearman's rho ρ then $\breve{\theta} = g(\rho_n)$ provides the Spearman-based estimate of θ. From standard convergence results about empirical processes it can be shown that

$$\rho_n \approx \mathcal{N}(\rho, \frac{\sigma^2}{n}) \tag{4.43}$$

where the asymptotic variance ρ^2 depends on the underlying copula C (Genest and Favre, 2007).

Genest et al. (2013) examined an even simpler rank-based estimator for the dependence parameter based on the inversion of the medial correlation coefficient, also known as Blomqvist's beta β. The estimator is obtained by computing the equation $\beta = \beta_n$ for the copula parameter where β_n is the ranked-based estimate of β derived from a random sample of size n. Despite being less efficient, the computation of β_n involves only $\mathcal{O}(n)$ operations, as opposed to $\mathcal{O}(n^2)$ for the empirical version of Kendall's tau and Spearman's rho. Furthermore, the population version of Blomqvist's beta is available in closed form for many common copula families.

General method of moments estimation is employed for overidentified models where the number of (implied) dependence measures may be larger than the number of unknown parameters (Patton, 2012). Simulated method of moment estimation considers overidentified models as well as dependence measures with unknown closed-form functions of copula parameter by employing simulations in place of mapping (Oh and Patton, 2013).

4.4.4.2 Minimum Distance Estimation

Copula estimation can be conducted through the minimization of distance metrics initially developed for Goodness of Fitness (GoF) testing. Thus, each GoF-test produces a minimum-distance (MD) estimator for the copula parameters. MD estimators are less used in literature (Weiß, 2011). Some MD estimators have been developed based on empirical copula process. Biau and Wegkamp (2005) developed a minimum L_1 distance estimator for parametric copula densities based on empirical copula process. Tsukahara (2005) examined the performance of two minimum distance estimators derived from Cramr-von-Mises (CvM) and Kolmogorov-Smirnov (KS) distances. The empirical asymptotic behavior of these distances between the hypothesized and empirical copula were explored in a simulation study.

In addition to minimum-distance estimators based on the empirical copula process, Weiß (2011) examined MD estimators based on Rosenblatt's transform and Kendall's dependence function. The MD estimators were subsequently compared to maximum likelihood (ML) estimators. It was found that ML estimators produce smaller estimation bias with less computation effort in comparison to MD-estimators.

4.5 Copula Model Selection

Copula model selection can be conducted using the following (Krämer and Schepsmeier, 2011):

- Akaike and Bayesian Information Criteria
- Formal goodness-of-fit tests

- Likelihood ratio tests (Vuong and Clarke tests)
- Graphical diagnostic tools
- Bivariate Asymptotic Independence Test

4.5.1 Akaike and Bayesian Information Criteria

Due to the wide range of copula families available, criteria such as the Akaike information criterion (AIC), Bayesian information criterion (BIC) and root mean square error (RMSE) are usually employed to select appropriate families as well as other multi-dimensional models by estimating their fitting biases (Ma et al., 2013). The Akaike Information Criterion (AIC) proposed by Akaike (1974) is one of the most widely used model selection criteria. AIC is not a hypothesis test but a test between competing models and thus a tool for model selection. AIC corrects the log likelihood of a copula for the number of parameters. It can also be defined as the negative log-likelihood with a number of parameters as a punitive term. Generally, AIC can be expressed as

$$AIC = -2\ln(L) + 2k \tag{4.44}$$

where L is the likelihood and k is the number of parameters.

In terms of bivariate copulas, given observations $u_{i,j}, i = 1, ..., N, j = 1, 2$, the AIC of a bivariate copula family c with parameter(s) θ is defined as

$$AIC = -2\sum_{i=1}^{N} \ln[c(u_{i,1}, u_{i,2}|\theta)] + 2k \tag{4.45}$$

where k=1 for one-parametric bivariate copulas and k=2 for two-parametric bivariate copulas.

Similarly, the Bayesian Information Criterion (BIC) proposed by Schwarz (1978) is given as

$$BIC = -2\sum_{i=1}^{N} \ln[c(u_{i,1}, u_{i,2}|\theta)] + \ln(N)k \tag{4.46}$$

Thus, a greater penalty is applied to two-parametric bivariate copulas when using BIC in comparison to AIC. Given a data set, several competing copula models are ranked according to their AIC or BIC, with the one having the lowest value being selected as the best.

4.5.2 Formal Goodness-of-fit Tests

As previously mentioned, criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are usually employed to select appropriate copula families by estimating their fitting biases. However, relatively small fitting biases do not always provide an acceptable depiction of the observations. Thus, the adequacy or competency of copula or a parametric family of copulas for the description of the dependence structures in the historical data can be examined using specialized goodness-of-fit (GoF) tests for copulas. (Ma et al., 2013).

Copula goodness-of-fit (or specification) tests involve the testing of the null hypothesis ($H_0 : C \in C_0$) that the dependence structure of a multivariate distribution is adequately characterized by a specific parametric family C_0 of copulas. This differs from the estimation of the dependence parameter of a copula ($C_0 : \theta \in \mathcal{O}$ where $\mathcal{O}$ is an open subset of $\mathbb{R}^p$ given integer $p \geq 1$ (Genest et al., 2009). There are no distributional assumptions for the marginals since only the fit of the dependence structure is of interest (Berg, 2009). Thus, the marginals are considered as (infinite-dimensional) nuisance parameters (Genest et al., 2009). Rather the testing is conducted using rank data (Berg, 2009). Since copulas are invariant given monotonic increasing transformations (of its components), testing of the null hypothesis usually having the inference being dependent on the maximally invariant statistics in relation to this group of transformations, i.e., the ranks (Genest et al., 2009).

Based on model assumptions, inference for GoF tests can be classified into two groups, parametric and semi-parametric (Wei, 2014). Copula goodness-of-fit (or specification) tests can be categorized into three types (Genest et al., 2009).

- Tests developed for testing specific dependence structures such as the Gaussian copula and Clayton copula family (also known as gamma frailty model in survival analysis).
- Procedures developed for testing any class of copulas (applicable to all copulas) but whose implementation involves arbitrary categorization of the data or strategic choice of smoothing parameter, weight function, kernel, window or arbitrary parameter
- Blanket tests - tests which are applicable to all copulas and require no ad hoc (arbitrary) categorization or strategic choice for their use.

Only a few tests have the desirable properties of blanket tests (Huang and Prokhorov, 2014). Examples of blanket tests include tests based on empirical copula (such as ranked-based version of Cramervon Mises and KolmogorovSmirnov statistics), tests based on Kendall's and Rosenblatt's probability integral transformation of the data as well as test based on a sample equivalent of Spearman's dependence function (Genest et al., 2009; Huang and Prokhorov, 2014). Another test proposed by Huang and Prokhorov (2014) is based on the information matrix equality which equates the copula Hessian and the outer-product of copula score. This test is considerably less difficult computationally than the aforementioned blanket tests.

However, most blanket tests have been found to have difficulty making a distinction between Gaussian and Student's t copulas, both symmetric copulas with differing tail properties. Others rely on probability integral transformation which may be hard to derive analytically in models such as Student's t copula and vine copulas. To tackle these limitations, Zhang et al. (2016) proposed a specification test for semi-parametric copula models. This proposed test is based on a ratio constructed using "in-sample" and "out-of-sample" pseudo-likelihoods which is computationally simple and numerically stable.

4.5.3 Vuong and Clarke tests

The Vuong and Clarke tests proposed by Vuong (1989) and Clarke (2007) respectively are likelihood-ratio tests used to compare non-nested models. These tests are also based on the Kullback-Leibner information criterion (KLIC). KLIC is the measure of the distance between two statistical models and can be expressed as follows (Schepsmeier, 2010):

$$KLIC := E_0[\log h_0(Y_i|x_i)] - E_0[\log f(Y_i|x_i, \hat{\beta})] \tag{4.47}$$

where $h_0(\cdot|\cdot)$ is the unknown true conditional probability function of Y_i given x_i, E_0 is the expected value under the true model and $\hat{\beta}$ is the parameter estimator of the imperfect model $f(Y_i|x_i, \hat{\beta})$. The model with the smallest KLIC is selected as the best model. Model 1 is better than model 2 if

$$E_0\left[\log \frac{f_1(Y_i|x_i, 1, \hat{\beta}_1)}{f_2(Y_i|x_i, 2, \hat{\beta}_2}\right] > 0 \tag{4.48}$$

where $f_1(Y_i|x_i, 1, \hat{\beta}_1)$ and $f_2(Y_i|x_i, 2, \hat{\beta}_2$ are the probability functions of models 1 and 2 respectively.

Given that $\ell^{(j)}, \ell^{(k)} \in \mathbb{R}^n$ are the vectors of pointwise loglikelihoods of the models with copula family j and k respectively, the differences of the pointwise loglikelihood can be computed as follows:

$$d_i = \ell_i^{(j)} - \ell_i^{(k)}, \quad i = 1, .., n. \tag{4.49}$$

The expected value of the differences $d =: (d_1, ..., d_n)^t$ is given as

$$E_0[d] = \mu_0^d = (\mu_1^d, ..., \mu_n^d)^t$$

The null-hypothesis of the Vuong test is

$$H_0 : \mu_0^d = 0 \text{ versus } H_1 : \mu_0^d \neq 0,$$

where μ_0^d is known. The Vuong test statistic can be mathematically defined as

$$T_V = \frac{\sqrt{n} \cdot \bar{d}}{\sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2}} \tag{4.50}$$

where $\bar{d}$ is the mean of the differences. The test statistic is asymptotically normally distributed with zero mean and unit variance. The null hypothesis is rejected at significance level α if $|T_V| \geq z_{1-\frac{\alpha}{2}}$. Thus, the smallest α for which the null hypothesis can be rejected is equal to $2\Phi(-|T_V|)$ where Φ is the standard normal distribution function. The p-value of the Vuong test is equal to $2\Phi(-|T_V|)$.

One disadvantage of the conventional Vuong test is its failure to account for the number of parameters in models. Thus, the log-likelihood ratio can be corrected with either the correction term of Akaike's information criteria (AIC) or Schwarz's Bayesian information criteria (BIC). The correction term for AIC is $p-q$ whereas the correction term for BIC is $\frac{p}{2n}\log n - \frac{q}{2n}\log n$ where p and q are the number of parameters of models 1 and 2 respectively and n is the number of observations (Schepsmeier, 2010). The log-likelihood ratio with Schwarz correction can be expressed as follows:

$$\log f_1(Y_i, x_{i,1}, \hat{\beta}_1) - \log f_2(Y_2, x_{i,2}, \hat{\beta}_2) - \left(\frac{p}{2n}\log n - \frac{q}{2n}\log n\right) \tag{4.51}$$

The Clarke test proposed by Clarke (2007) is a simple distribution-free test for non-nested model selection. It is similar to the Vuong test but differs in its null hypothesis and has been found to be asymptotically more efficient than the latter when the distribution of individual log-likelihood ratios is highly peaked. The null hypothesis of the Clarke test is given as:

$$H_0 : P\left(\log\left(\frac{f_1(Y_i, x_{i,1}, \hat{\beta}_1)}{f_2(Y_2, x_{i,2}, \hat{\beta}_2)}\right) > 0\right) = p \tag{4.52}$$

If $p = 0.5$ then the models are equivalent. The intuition behind this null hypothesis is that the log-likelihood ratios should be evenly distributed around zero and in

expectation a half of the ratios should be greater than zero and other half less than zero.

The Clarke test statistic is given as

$$B = \sum_{i=1}^{n} I_{0,+\infty}(d_i), \tag{4.53}$$

where $I_{(\cdot)}$ is the indicator function and B is a binomial distributed random variable with parameter n and $p = 0.5$. Thus the two models are equivalent if $B = E(np) = \frac{n}{2}$. Similar to the Vuong test, the Clarke test statistic can be corrected for the number of parameters with the Akaike or Schwarz corrections, which correspond to the penalty terms in the AIC and BIC respectively.

4.5.4 Graphical Diagnostic Tools

Chi-plots (χ-plots) were initially developed by Fisher and Switzer (1985) and expanded upon graphically by Fisher and Switzer (2001). The chi-plot was developed to tackle the issue of detecting patterns or randomness in scatter plots. Their development is influenced by control charts and is based on the chi-square statistic for independence in a two-way table. Chi-plots augment scatterplots of data by offering a graph that has characteristic patterns depending on whether the variables (i) are independent, (ii) have some degree of monotonic relationship (i.e., nonzero grade correlation), or (iii) have more complex dependence structure. Thus, chi-plots can visualize a large range of dependence forms and is reliant on ranked data.

Given observations $u_{ij}, i = 1, ..., N, j = 1, 1$, the chi-plot is a scatter plot of the pairs of two quantities namely chi-statistics

$$\chi_i = \frac{\hat{F}_{1,2}(u_{i,1}, u_{i,2}) - \hat{F}_1(u_{i,1})\hat{F}_2(u_{i,2})}{\sqrt{F_1(u_{i,1})(1 - F_1(u_{i,1}))F_2(u_{i,2})(1 - F_2(u_{i,2}))}}$$

and lambda-statistics.

$$\lambda_i = 4sgn\left(\tilde{F}_1(u_{i,1}), \tilde{F}_2(u_{i,2})\right) \cdot \max\left(\tilde{F}_1(u_{i,1})^2, \tilde{F}_2(u_{i,2})^2\right)$$

where $\hat{F}_1, \hat{F}_2$, and $\hat{F}_{1,2}$, are the empirical distribution functions of the uniform random variables U_1 and U_2 and (U_1, U_2), respectively; $\tilde{F}_1 = \hat{F}_1 - 0.5$ and $\tilde{F}_2 = \hat{F}_2 - 0.5$.

Both quantities are reliant on ranked data and are scaled to the unit interval. χ_i corresponds to a correlation coefficient between dichotomized values of U_1 and U_2 whereas λ_i is a measure of the distance of the data point (x_i, y_i) from the center of the bivariate dataset. The pairs of the two quantities (δ_i, χ_i) are often situated above zero for positively dependent margins and vice versa for negatively dependent margins.

Kendall's plot (K-plot) proposed by Genest and Boies (2003) is another rank-based graphical tool for visualizing dependence. Kendall's plot is the bivariate copula equivalent to quantile-quantile (Q-Q) plots and is based on probability integral transformation. K-plots are generally easier to interpret than chi-plots and maintain the latter's property of invariance with respect to monotone transformations of the marginal distributions.

Two variables U_1 and U_2 are considered independent if the points of a K-plot generally lie on the main diagonal (y = x). Any deviation from the main diagonal is an indicator of dependence with greater deviation indicating greater dependence. Points situated above the diagonal line indicate positive dependence and vice-versa for negative dependence. Illustration of the Kendall's plot is shown on figure 6.10 (left panel).

Similar to chi-plot, Kendall's plot is based on two quantities: H-statistics - the ordered values of the empirical bivariate distribution function $H_i := \hat{F}_{U_1U_2}(u_{i,1}, u_{i,2})$ and W-statistics $W_{i:n}$ - the expected values of the order statistics from a random sample of size N of the random variable $W = C(U_1, U_2)$ under the null hypothesis of independence between U_1 and U_2.

$W_{i:n}$ can be computed as follows:

$$W_{i:n} = \binom{N-1}{i-1} \int_0^1 \omega k_0(\omega)(K_0(\omega))^{i-1}(1-K_0(\omega))^{N-i} d\omega$$

where

$$K_0(\omega) = \omega - \omega \log(\omega)$$

Illustration of the chi plot is shown on figure 6.10 (central panel).

The lambda function (λ-function) plot proposed by Genest and Rivest (1993) is also a graphical tool for visualizing dependence. The lambda function is distinctive for each bivariate copula family and is defined by Kendall's cumulative distribution function K:

$$\lambda(v, \theta) := v - K(v, \theta) \tag{4.54}$$

where $K(v, \theta) := P(C_\theta(U_1, U_2) \leq v), v \in [0, 1]$.

The theoretical λ-function of Archimedean copulas can be expressed in closed form as a function of its generator function.

$$\lambda(v, \theta) = \frac{\varphi(v)}{\varphi'(v)} \tag{4.55}$$

where $\varphi'(v)$ is the derivative of φ. On the other hand, the closed-form expression of elliptical copulas does not exist. Instead, the theoretical λ-function of elliptical copulas is usually simulated based on samples of size 1000. The theoretical λ-function plot shows the limits of the -function corresponding to Kendall's $\tau = 0$ and Kendall's $\tau = 1$ (Schepsmeier, 2010). The theoretical lambda function closed-form expressions of various bivariate Archimedean copula families and their limits can be found in Schepsmeier (2010). For large data sets, the λ-function can be easily estimated empirically by using the empirical copula function. Illustration of the lambda function plot is shown on figure 6.10 (right panel).

4.5.5 Bivariate Asymptotic Independence Test

The Bivariate Asymptotic Independence Test based on Kendall's Tau was proposed by Genest and Favre (2007). Prior to selection of the other bivariate copula, the test can be performed to determine the independence of the pair of variables. The null hypothesis states that the variables are independent and the alternative hypothesis states that the variables are not independent.

Under the null hypothesis, the statistic is close to normal with zero mean and variance $\frac{9N(N-1)}{2(2N+5)}$. Thus, the test exploits the asymptotic normality of the test statistic

$$T = \sqrt{\frac{9N(N-1)}{2(2N+5)}} \times |\hat{T}|, \tag{4.56}$$

where N is the number of observations and $\hat{T}$ is the empirical Kendall's tau of the two variables. The p-value of the null hypothesis of bivariate independence hence is asymptotically

$$p.value = 2 \times (1 - \Phi(T))$$

where Φ is the standard normal distribution function. The independence copula is selected for the pair of variables if the p-value of the test is higher than 5% meaning the null hypothesis is accepted.

4.6 Case Study (Tamping Recovery of Track Geometry)

4.6.1 Introduction

Railroad track deteriorates with age and usage (tonnage) with decreasing performance over time which may eventually lead to failure. Railroad infrastructure components often have a service life of more than 30 years justifying the need for an optimal long-term maintenance strategy. Due to budget restrictions and high logistical cost constraints, railroads plan most track geometry maintenance activities up to a year in advance (Quiroga and Schnieder, 2012; Soleimanmeigouni et al., 2016a; Caetano and Teixeira, 2016).

Track Geometry is a key feature of railroad construction (Esveld, 2001; Khouy, 2013). The condition of track geometry is important for various reasons. Riding comfort and safety (risk of derailment) are dependent on the track geometry condition (Quiroga and Schnieder, 2012). Well-maintained track geometry not only guarantees ride comfort and safety but also increases the life of the track as well as track availability for train operations. Thus, track geometry maintenance is imperative in relation to cost reduction and availability of track (Famurewa et al., 2016). Furthermore, the deterioration of many other track components is closely linked to track geometry condition (Jovanovic, 2004; Khouy, 2013).

Track geometry maintenance activities are regularly conducted in order to maintain track geometry condition to achieve good riding quality and safety (Miwa, 2002). These activities such as tamping, stoneblowing and ballast undercutting are conducted to control track deterioration and recover damaged track sections to operable conditions. They enhance the track geometry quality but fail to return the track geometry to a good-as-new condition (Soleimanmeigouni et al., 2016b). If prognostic (predictive) tamping strategies are to be employed, there is the need to know beforehand the effectiveness of tamping which can be evaluated by the amount of improvement or recovery in track geometry condition (Famurewa et al., 2013).

Majority of studies have evaluated tamping recovery using deterministic techniques such as linear regression models and have assumed that tamping effectiveness is mainly dependent on the track geometry quality prior to tamping. However, in most cases there exists a high degree of uncertainty due to high variation in the restoration values after tamping even for similar track geometry condition. This variation is even higher at the end of the life-cycle than at the beginning. For this reason, probabilistic or stochastic techniques have been employed to cater for this variation by assuming the recovery value after tamping is a random variable with a given probability distribution (Soleimanmeigouni et al., 2016b).

Furthermore, most tamping recovery models do not take into account the underlying dependence between the tamping recovery values and the influencing factors

such as track geometry condition before tamping. In this case study, a copula-based approach is employed which takes into consideration various forms of dependences by allowing for the separate modeling of the arbitrary marginal distributions and the dependence structure which are subsequently combined to form a joint distribution with the underlying dependence.

4.6.2 Track Information and Data Collection

One mile of track of a Class 1 U.S. railroad was used for the analysis. Inspection data was measured and collected for every 1 foot of track using a track geometry car. The track geometry car records several geometry parameters. However, the surface, alignment, cross level, gage and warp parameters were used for this case study. The inspection data used in this case study were from 28 inspection dates spanning the years 2013 to 2016.

The inspection data was initially cleaned and preprocessed. The standard deviation (SD) of each of the track geometry parameters was subsequently computed for track segments with 100 feet of length. The tamping recovery values for each parameter were obtained by calculating the difference between the standard deviation (SD) of the track geometry parameters before tamping and the corresponding standard deviation after tamping.

4.6.3 Analysis

4.6.3.1 Marginal fitting

In order to select the best-fit for the marginal distributions for the recovery values after tamping, track quality before tamping and track quality after tamping; the Kolmogorov-Smirnov, Anderson-Darling and Chi-squared tests were chosen as the goodness-of-fit criteria. These test statistics evaluates how well the data (stochastic variable) follow a specific (an a priori) distribution. The smaller the statistic, the better the distribution fits the given data. In order to select the best distribution, the statistic should be considerably lower than the others, else additional criteria such

as probability plots need to be employed. The null hypothesis states that the data followed a specific distribution with the alternative hypothesis states that the data does not follow a specific distribution. The null hypothesis is rejected if the p-value is lower than a significance level of 5%.

The Kolmogorov-Smirnov test (KS test) is a nonparametric statistical test of the equality of two probability distributions namely the empirical distribution of the data and a reference probability distribution. The Anderson-Darling tests offers more weighting to the tails compared to Kolmogorov-Smirnov test.

4.6.3.2 Copula fitting

The underlying dependence between track quality before tamping and recovery values as well as the dependence of the track quality before tamping and track quality after tamping were characterized using copulas. In order to select the best-fit of bivariate copula that describes the underlying dependence, the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information (BIC) (Schwarz, 1978) were used. AIC corrects the log-likelihood of a copula for the number of parameters. AIC is often favored for bivariate copula selection ahead of other alternative criteria such as Vuong (1989) and Clarke (2007) goodness-of-fit tests and BIC. This is as a result of its high performance in simulation analysis and its greater reliability (Dissmann et al., 2013; Dalla Valle et al., 2016).

Prior to selection of the bivariate copula, the Genest and Favre bivariate asymptotic independence test based on Kendall's Tau is performed to determine the independence of the pair of variables. The null hypothesis states that the variables are independent and the alternative hypothesis states that the variables are not independent. The independence copula is selected for the pair of variables if the p-value of the test is higher than 5% meaning the null hypothesis is accepted.

The pair-copula families considered during the analysis were the *independence copula*, *elliptical bivariate Gaussian (Normal)* and *Student t-copulas* as well as the single parameter Archimedean copulas such as *bivariate Clayton, Gumbel, Frank and*

Joe copulas. Others include the two-parameter Archimedean copulas such as *Clayton-Gumbel (BB1), Joe-Gumbel (BB6), Joe-Clayton (BB7) and Joe-Frank (BB8)* copulas. The Clayton-Gumbel (BB1) and Joe-Clayton (BB7) permit different non-zero lower and upper tail dependence coefficients.

Rotated versions (90^0 and 270^0) of these Archimedean copulas can be used to fit negative dependences (with the exception of Frank copula which has no rotated version). However, no negative dependences were observed during exploratory analysis so these rotations were not considered during further analysis. This catalogue for the implementation of copula family choice address a vast range of dependence behavior. Properties of these copulas are found in Table 4.3.

Table 4.3: Properties of pair-copula families considered

Copula	**Properties**
Normal/ Gaussian (N)	Tail symmetric, no tail dependence
Student t-copula (t)	Tail-symmetric, tail dependence
Clayton (C)	Tail-asymmetric, reflection-asymmetric, suitable for modelling lower tail dependence (no upper tail dependence)
Gumbel (G)	Tail-asymmetric, reflection-asymmetric, suitable for modelling upper tail dependence (no lower tail dependence), suitable for highly correlated variables at high values and less correlated values at low levels
Joe (J)	Tail-asymmetric, suitable for modelling upper tail dependence (no lower tail dependence)
Frank (F)	Tail-symmetric, no tail dependence, tends to work well when tail dependence is very weak.
Clayton-Gumbel (BB1)	Tail-asymmetric, suitable for different non-zero upper and lower tail dependence
Joe- Clayton (BB7).	Tail-asymmetric, suitable for different non-zero upper and lower tail dependence
Rotations of Archimedean copulas	Suitable for modelling various forms of negative dependence

4.6.4 Surface (Longitudinal Level)

4.6.4.1 Marginal fitting

The three-parameter lognormal distribution was found to have the best fit of the recovery values of the standard deviation (SD) surface producing the lowest statistic for all three tests namely the Kolomogorov-Smirnov, Anderson-Darling and Chi-squared tests as shown in Table 4.4. It also had a p-value far greater than 0.05 for Kolmogorov-Smirnov and Chi-squared tests meaning the null hypothesis that it follows the distribution can be accepted. The closed form expression for the p-value of the 3-parameter lognormal distribution however does not exist for the Anderson-Darling test. Audley and Andrews (2013) found the three-parameter lognormal distribution to have the best fit of recovery values of the SD surface profile but employed its two-parameter counterpart due to its ease of use. The two-parameter lognormal distribution has been used to model the recovery values of the surface profile (longitudinal level) by several researchers (Quiroga and Schnieder, 2012; Quiroga et al., 2012; Audley and Andrews, 2013). However, in this case study, the three-parameter lognormal distribution was used to model the recovery value of the surface profile based on the aforementioned results.

Similarly, the three-parameter lognormal distribution was also found to have the best fit for both the standard deviation (SD) surface values before tamping and SD surface values after tamping as shown in Tables 4.5 and 4.6 respectively.

4.6.4.2 Copula fitting

The Gumbel copula was found to provide the best fit of the underlying dependence between the SD Surface values before tamping and the recovery values. The Gumbel copula produced both the lowest AIC and BIC values as shown in Table 4.7.

The selection of the Gumbel copula suggests an asymmetric dependence (specifically an upper tail dependence) between the track quality (standard deviation surface) before tamping and the recovery value. Upper tail dependence means that the pair are

Table 4.4: Results for the fitted distribution to recovery values for SD Surface.

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.054	0.992	0.156	*	2.91	0.709
Lognormal	0.057	0.99555	0.153	0.953	2.94	0.714
Weibull (3P)	0.111	0.491	4.66	0.143	N/A	N/A
Weibull	0.098	0.656	0.962	0.023	2.91	0.573
Gamma (3P)	0.092	0.730	4.39	*	N/A	N/A
Gamma	0.150	0.166	1.66	0.038	9.67	0.085
Normal	0.227	0.007	4.86	0.003	10.1	0.017
Logistic	0.229	0.006	2.91	0.030	8.42	0.038
Exponential	0.096	0.678	0.903	0.151	7.01	0.220
Exponential (2P)	0.097	0.668	0.998	0.231	1.92	0.860

highly correlated at high values (upper tail of the distributions) but poorly correlated at lower values.

Simulated values were generated given the 3-parameter lognormal marginals (for both track quality before tamping and recovery value) and Gumbel copula. An illustrative comparison between the real and simulated values for recovery values against track condition before tamping for SD surface is shown in figure 4.1.

The Joe-Clayton (popularly known as BB7) copula was found to offer the best fit of the underlying dependence between the track quality (standard deviation surface) before tamping and the track quality after tamping. The BB7 copula was found to produce the lowest AIC and BIC values as shown in Table 4.8.

The Joe-Clayton copula consists of the Joe copula and Clayton copula which are suitable for modeling upper tail and lower tail dependence respectively. The selection of the BB7 copula suggests an asymmetric dependence (with different non-zero lower and upper tail dependence coefficients) between the SD surface values before tamping and SD surface values after tamping. Similarly, simulated values were generated given the 3-parameter lognormal marginals (for both track quality before tamping and track quality after tamping) and Joe-Clayton (BB7) copula. An illustrative comparison

Table 4.5: Results for the fitted distribution to values before tamping for SD Surface

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi squared	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.086	0.797	0.345	*	5.15	0.397
Lognormal	0.148	0.175	1.17	0.282	4.69	0.320
Weibull (3P)	0.104	0.576	4.71	0.035	N/A	N/A
Weibull	0.189	0.040	3.30	0.038	10.2	0.037
Gamma (3P)	0.109	0.523	4.71	*	N/A	N/A
Gamma	0.187	0.042	2.84	0.064	15.0	0.005
Normal	0.255	0.002	5.16	0.002	12.3	0.006
Logistic	0.258	0.001	4.90	0.015	11.8	0.019
Exponential	0.264	9.4E-04	4.62	0.004	21.5	2.6E-04
Exponential (2P)	0.138	0.239	1.35	0.100	8.27	0.141

between the real and simulated values for track condition after tamping against track condition before tamping for SD surface is shown in figure 4.2.

4.6.5 Alignment

4.6.5.1 Marginal fitting

Similar to the surface profile results, the 3-parameter lognormal distribution was found to have the best fit for the recovery values of SD alignment after tamping as shown in Table 4.9. The 3-parameter lognormal distribution has previously been used to model the recovery values of SD alignment by Soleimanmeigouni et al. (2016a). The 3-parameter lognormal distribution was also found to have the best fit of track quality (SD alignment) values before tamping and after tamping as shown as in Tables 4.10 and 4.11 respectively.

4.6.5.2 Copula fitting

The Joe copula provided the best fit of the underlying dependence between the SD alignment values before tamping and the tamping recovery values producing both the lowest AIC and BIC values as shown in Table 4.12. The selection of the Joe copula suggests an upper tail dependence between the SD alignment values before tamping

Table 4.6: Results for the fitted distribution to values after tamping for SD Surface

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi squared	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.100	0.628	0.404	*	4.84	0.436
Lognormal	0.154	0.146	1.88	0.109	4.01	0.405
Weibull (3P)	0.114	0.467	4.92	0.007	N/A	N/A
Weibull	0.186	0.044	3.81	0.021	10.2	0.037
Gamma (3P)	0.121	0.392	1.40	*	5.04	0.284
Gamma	0.210	0.016	3.03	0.032	11.3	0.024
Normal	0.244	0.003	5.16	0.002	9.73	0.021
Logistic	0.245	0.003	3.82	0.011	11.5	0.009
Exponential	0.323	2.0E-05	5.70	0.001	26.1	3.0E-05
Exponential (2P)	0.129	0.310	1.60	0.044	12.0	0.018

and tamping recovery values. The Joe Copula has an even stronger positive upper tail dependence in comparison to the Gumbel copula and can be observed by tighter clustering of observations in the upper tail (Bhat and Eluru, 2009).

Simulated values were generated given the 3-parameter lognormal margins (for SD alignment values before tamping and recovery values) and Joe Copula. Figure 4.3 shows the comparison between the observed and simulated values for recovery values against SD alignment before tamping.

The Gaussian (or Normal) copula offered the best fit of the underlying dependence between the SD alignment values before tamping and SD alignment values after tamping as shown in Table 4.4. The selection of the Gaussian copula suggests that the underlying dependence between the pair is radially-symmetric with strong central dependence and very weak tail dependency. Similarly, simulated values were produced given the 3-parameter lognormal marginals (for both SD alignment before tamping and SD alignment after tamping) and Gaussian copula. An illustrative comparison between the real and simulated values for track condition after tamping against track condition before tamping for SD alignment is shown in figure 4.4.

Table 4.7: Results for the fitted bivariate copula between values before tamping and recovery values for SD Surface

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
Gumbel	$\theta = 3.31$	-	-79.95	-77.98	0.7
BB7	$\theta = 4.17$	$\delta = 1.2$	-78.95	-75.01	0.68
BB6	$\theta = 1.59$	$\delta = 2.45$	-78.53	-74.59	0.69
BB1	$\theta = 0$	$\delta = 3.31$	-77.95	-74.01	0.7
Joe	$\theta = 4.61$	-	-77.23	-75.26	0.65
BB8	$\theta = 4.61$	$\delta = 1$	-75.23	-71.29	0.65
Gaussian/Normal	$\rho = 0.89$	-	-74.50	-72.53	0.7
Student-t copula	$\rho = 0.89$	$\nu = 30$	-72.42	-68.48	0.7
Frank	$\theta = 9.98$	-	-65.65	-63.68	0.67
Clayton	$\theta = 2.35$	-	-47.97	-46.00	0.54

4.6.6 Cross level

4.6.6.1 Marginal fitting

The three-parameter log-logistic distribution was found to best fit the recovery values of SD cross level as shown in Table 4.14. The 3-parameter log-logistic distribution has an identical shape to the 3-parameter log-normal distribution (which was found to be the next best distribution) but has heavier tails. The 3-parameter lognormal distribution was also found to have the best fit of track quality (SD cross level) values before tamping and after tamping as shown as in Tables 4.15 and 4.16 respectively.

4.6.6.2 Copula fitting

The bivariate asymptotic independence test performed prior to copula fitting and selection determined that the recovery values of the cross level and the track quality (SD cross level) before tamping were independent. The p-value of 0.43 was found higher than the 0.05 significance level. Thus the null hypothesis that the variables are independent was accepted and the independence copula was selected for the pair of variables. Ignoring the test would have led to the selection of the Joe Copula of parameter value of 1.38 and Kendall's tau of 0.18. Simulated values were generated given

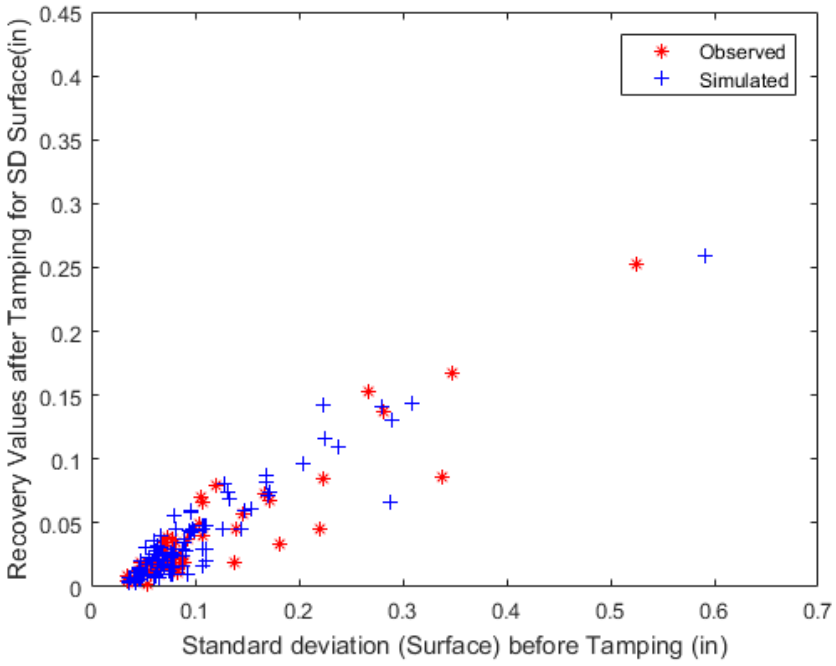


Figure 4.1: Comparison between real and simulated values for SD Surface given 3-parameter Lognormal marginals (Before tamping and Recovery values) and Gumbel copula.

3-parameter lognormal marginal (Before tamping) 3-parameter log-logistic marginal (Recovery values) and Independence copula. The simulated values are illustrated in figure 4.5.

On the other hand, the Joe-Clayton (BB7) copula was found to best fit the underlying dependence between track quality (SD cross level) before tamping and track quality after tamping. The BB7 copula was found to have the lowest AIC and BIC values as shown in Table 4.17. Simulated values of SD Crosslevel given 3-parameter lognormal marginals (Before tamping and after tamping) and Joe-Clayton (BB7) copula are illustrated in figure 4.6.

Table 4.8: Results for the fitted bivariate copula between values before and after tamping for SD Surface

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
BB7	$\theta = 4.36$	$\delta = 2.3$	-89.26	-85.32	0.72
Gumbel	$\theta = 3.56$	-	-85.24	-83.27	0.72
BB1	$\theta = 0.18$	$\delta = 3.3$	-83.54	-79.60	0.72
BB6	$\theta = 1.32$	$\delta = 2.97$	-83.42	-79.48	0.72
Student-t copula	$\rho = 0.89$	$\nu = 2$	-82.60	-78.67	0.69
Joe	$\theta = 4.84$	-	-80.77	78.80	0.67
BB8	$\theta = 4.84$	$\delta = 1$	-78.77	-74.83	0.67
Gaussian/Normal	$\rho = 0.89$	-	-76.72	-74.75	0.7
Frank	$\theta = 10.56$	-	-66.71	-64.74	0.68
Clayton	$\theta = 2.89$	-	-58.85	-56.88	0.59

Table 4.9: Results for the fitted distribution to recovery values for SD Alignment

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi squared	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.100	0.625	0.760	*	4.22	0.518
Weibull (3P)	0.13478	0.266	1.75	<0.005	7.84	0.098
Gamma (3P)	0.113	0.475	1.00	*	5.15	0.397
Normal	0.180	0.056	2.95	0.029	10.4	0.034
Logistic	0.161	0.115	0.794	0.485	7.38	0.117
Exponential (2P)	0.258	0.001	6.26	9.1E-4	23.7	2.9E-05

4.6.7 Warp

4.6.7.1 Marginal distribution fitting

Similar to the cross level results, the 3-parameter log-logistic distribution was found to have the best fit for the tamping recovery values of SD warp as shown in Table 4.18. Furthermore, the 3-parameter lognormal distribution was also found to have the best fit of track quality (SD warp) values before tamping and after tamping as shown in Tables 4.19 and 4.20 respectively.

Table 4.10: Results for the fitted distribution to values before tamping for SD Alignment

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.121	0.388	0.949	*	6.57	0.255
Lognormal	0.143	0.208	1.10	0.294	5.26	0.385
Weibull (3P)	0.146	0.190	1.96	*	13.3	0.010
Weibull	0.168	0.088	3.82	0.011	5.18	0.269
Gamma (3P)	0.140	0.228	1.28	*	6.55	0.256
Gamma	0.165	0.100	1.71	0.133	14.2	0.007
Normal	0.218	0.011	3.69	0.013	23.7	9.2E-05
Logistic	0.215	0.012	1.28	0.238	20.2	4.5E-04
Exponential	0.404	3.2E-08	12.5	1.1E-5	92.8	0
Exponential (2P)	0.263	9.0E-04	5.50	0.002	33.2	1.1E-06

Table 4.11: Results for the fitted distribution to values after tamping for SD Alignment

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.065	0.966	0.250	*	2.42	0.789
Lognormal	0.073	0.924	0.317	0.919	3.69	0.595
Weibull (3P)	0.078	0.876	0.385	0.409	3.68	0.596
Weibull	0.074	0.914	0.654	0.520	1.20	0.945
Gamma (3P)	0.068	0.954	0.268	*	2.39	0.793
Gamma	0.071	0.933	0.277	0.963	2.64	0.756
Normal	0.096	0.676	0.556	0.144	2.01	0.848
Logistic	0.089	0.765	0.293	0.943	3.15	0.677
Exponential	0.404	3.4E-08	12.5	1.1E-5	72.5	1.2E-15
Exponential (2P)	0.251	0.002	5.46	0.002	32.4	1.6E-06

Table 4.12: Results for the fitted bivariate copula between values before tamping and recovery values for SD Alignment

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
Joe	$\theta = 1.93$	-	-18.41	-16.44	0.34
BB8	$\theta = 1.93$	$\delta = 1$	-16.41	-12.47	0.34
BB6	$\theta = 1.93$	$\delta = 1$	-16.41	-12.47	0.34
BB7	$\theta = 1.93$	$\delta = 0$	-16.40	-12.46	0.34
Gumbel	$\theta = 1.52$	-	-14.52	-12.55	0.34
BB1	$\theta = 0$	$\delta = 1.52$	-12.51	-8.57	0.34
Student-t copula	$\rho = 0.45$	$\nu = 2.11$	-9.70	-5.76	0.29
Gaussian/Normal	0.44	-	-7.33	-5.36	0.29
Frank	$\theta = 2.74$	-	-6.81	-4.84	0.28
Clayton	$\theta = 0.38$	-	-0.48	1.49	0.16

Table 4.13: Results for the fitted bivariate copula between values before tamping and after tamping for SD Alignment

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
Gaussian/Normal	$\rho = 0.59$	-	-17.03	-15.06	0.4
Student-t copula	$\rho = 0.58$	$\nu = 30$	-14.68	-10.74	$\rho = 0.39$
Frank	$\theta = 3.71$	-	-14.60	-12.63	0.37
Clayton	$\theta = 0.98$	-	-14.56	-12.59	0.33
BB1	$\theta = 0.56$	$\delta = 1.25$	-13.93	-9.99	0.38
BB7	$\theta = 1.35$	$\delta = 0.8$	-13.63	-9.69	0.37
Gumbel	$\theta = 1.53$	-	-13.20	-11.23	0.35
BB8	$\theta = 6$	$\delta = 0.49$	-12.43	-8.49	0.36
BB6	$\theta = 1$	$\delta = 1.53$	-11.19	-7.25	0.35
Joe	$\theta = 1.68$	-	-9.24	-7.27	0.28

Table 4.14: Results for the fitted distribution to recovery values for SD Cross level

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.094	0.705	0.502	*	3.11	0.683
Lognormal (3P)	0.134	0.276	1.39	*	13.3	0.021
Weibull (3P)	0.164	0.103	2.69	*	20.2	1.5E-4
Gamma (3P)	0.147	0.220	1.63	*	13.3	0.020
Normal	0.206	0.019	3.16	0.023	16.3	0.003
Logistic	0.194	0.031	0.857	0.441	14.4	0.006
Exponential (2P)	0.338	6.4E-6	8.38	1.0E-4	39.4	2.7E-9

Table 4.15: Results for the fitted distribution to values before tamping for SD Cross level

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.125	0.347	0.723	*	4.09	0.536
Lognormal	0.183	0.049	2.06	0.087	17.7	0.003
Weibull (3P)	0.135	0.267	1.27	*	2.59	0.628
Weibull	0.202	0.022	4.00	0.018	7.63	0.106
Gamma (3P)	0.170	0.192	1.22	*	2.08	0.720
Gamma	0.170	0.081	2.60	0.036	17.5	0.002
Normal	0.209	0.017	4.40	0.006	12.2	0.016
Logistic	0.226	0.008	3.39	0.017	10.9	0.028
Exponential	0.389	1.3E-07	8.74	5.8E-5	70.2	2.1E-14
Exponential (2P)	0.158	0.127	1.80	0.171	1.58	0.813

Table 4.16: Results for the fitted distribution to values after tamping for SD Cross level

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Lognormal (3P)	0.085	0.805	0.348	*	2.53	0.772
Lognormal	0.130	0.309	0.947	0.396	6.60	0.159
Weibull (3P)	0.135	0.267	1.27	*	2.59	0.628
Weibull	0.202	0.022	4.00	0.091	7.63	0.106
Gamma (3P)	0.103	0.589	0.600	*	4.17	0.525
Gamma	0.127	0.330	1.40	0.177	13.3	0.010
Normal	0.197	0.028	3.01	0.027	12.3	0.006
Logistic	0.207	0.018	2.96	0.066	10.1	0.018
Exponential	0.346	3.54-06	7.35	2.4E-4	12.1	0.017
Exponential (2P)	0.153	0.152	1.69	0.186	2.93	0.569

Table 4.17: Results for the fitted bivariate copula between values before tamping and after tamping for SD Cross level

Copula	**Parameter 1**	**Parameter 2**	**AIC**	**BIC**	**Kendall's Tau**
BB7	2.86	1.51	-55.37	-51.43	0.62
BB1	0.4	2.23	-54.30	-50.36	0.63
Student-t copula	0.82	2.86	-54.23	-50.287	0.61
Gaussian/Normal	0.82	-	-52.42	-50.45	0.61
BB8	5.28	0.88	-52.03	-48.09	0.61
Frank	8.06	-	-48.73	-46.76	0.6
Gumbel	2.62	-	-54.83	-52.86	0.62
BB6	1	2.62	-52.82	-48.88	0.62
Joe	3.29	-	-48.08	-46.11	0.55
Clayton	2.1	-	-42.58	-40.61	0.51

Table 4.18: Results for the fitted distribution to recovery values for SD Warp

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.093	0.713	0.552	*	2.83	0.727
Lognormal (3P)	0.132	0.285	1.24	*	5.49	0.359
Weibull (3P)	0.166	0.096	2.19	*	9.72	0.045
Gamma (3P)	0.146	0.191	1.45	*	5.50	0.36
Normal	0.207	0.018	3.032	0.027	18.6	9.6E-4
Logistic	0.187	0.043	0.833	0.457	13.9	0.008
Exponential (2P)	0.303	8.3E-5	6.50	9.7E-4	18.5	3.5E-4

Table 4.19: Results for the fitted distribution to values before tamping for SD Warp

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.071	0.937	0.188	*	1.337	0.931
Log-Logistic	0.080	0.863	0.451	0.800	0.186	0.999
Lognormal (3P)	0.0601	0.985	0.200	*	1.05	0.958
Lognormal	0.092	0.722	0.647	0.602	0.191	0.999
Weibull (3P)	0.088	0.772	0.579	*	1.30	0.934
Weibull	0.125	0.349	2.43	0.062	3.70	0.448
Gamma (3P)	0.077	0.885	0.415	*	0.545	0.990
Gamma	0.141	0.221	1.61	0.248	4.33	0.503
Normal	0.184	0.049	3.11	0.024	7.48	0.113
Logistic	0.182	0.053	1.63	0.149	5.63	0.228
Exponential	0.373	4.6E-07	8.92	4.8E-5	37.1	1.70E-07
Exponential (2P)	0.148	0.180	1.80	0.161	5.60	0.231

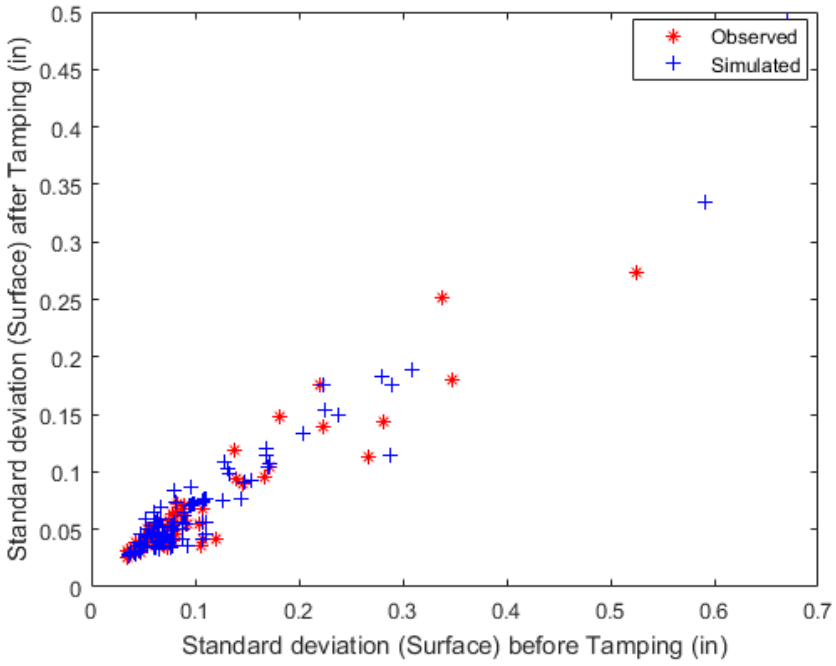


Figure 4.2: Comparison between real and simulated values for SD Surface given 3-parameter lognormal marginals (Before tamping and after tamping) and Joe-Clayton (BB7) copula.

4.6.7.2 Copula fitting

The Joe copula provided the best fit of the underlying dependence between the SD warp values before tamping and the tamping recovery values. The Joe copula produced both the lowest AIC and BIC values as shown in Table 4.21. Simulated values were generated given 3-parameter lognormal marginal (Before tamping) 3-parameter loglogistic marginal (Recovery values) and Joe copula. The simulated values are illustrated in figure 4.7.

The Gumbel copula provided the best fit of the underlying dependence between track quality (SD cross level) before tamping and track quality after tamping. The Gumbel copula was found to have the lowest AIC and BIC values as shown in Table 4.22. Simulated values were generated given 3-parameter lognormal marginals (track condition before and after tamping) and Gumbel copula. The simulated values are

Table 4.20: Results for the fitted distribution to values after tamping for SD Warp

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.084	0.816	0.285	*	5.894	0.317
Log-Logistic	0.085	0.803	0.349	0.934	6.66	0.247
Lognormal (3P)	0.073	0.923	0.263	*	4.22	0.519
Lognormal	0.073	0.918	0.264	0.961	4.00	0.550
Weibull (3P)	0.099	0.641	0.525	*	5.41	0.368
Weibull	0.094	0.705	1.09	0.338	12.0	0.018
Gamma (3P)	0.074	0.912	0.320	*	4.469	0.484
Gamma	0.097	0.663	0.534		4.18	0.523
Normal	0.137	0.248	1.33	0.222	5.218	0.390
Logistic	0.142	0.217	0.871	0.432	2.47	0.781
Exponential	0.351	2.4E-6	8.93	4.8E-5	24.6	1.9E-05
Exponential (2P)	0.259	0.001	4.58	0.006	20.2	4.5E-4

Table 4.21: Results for the fitted bivariate copula between SD values before tamping and SD recovery values after tamping for Warp

Copula	**Parameter 1**	**Parameter 2**	**AIC**	**BIC**	**Kendall's Tau**
Joe	$\theta = 1.67$	-	-11.62	-9.65	0.27
Gumbel	$\theta = 1.42$	-	-10.43	-8.46	0.29
BB7	$\theta = 1.63$	$\delta = 0.15$	-9.96	-6.01	0.3
BB8	$\theta = 1.67$	$\delta = 1$	-9.62	-5.68	0.27
BB6	$\theta = 1.67$	$\delta = 1$	-9.62	-5.68	0.27
BB1	$\theta = 0$	$\delta = 1.42$	-8.43	-4.48	0.29
Student-t copula	$\rho = 0.33$	$\nu = 2.3$	-7.06	-3.12	0.22
Gaussian/Normal	$\rho = 0.43$	-	-6.97	-5.00	0.29
Frank	$\theta = 2.24$	-	-4.08	-2.11	0.24
Clayton	$\theta = 0.47$	-	-2.65	-0.68	0.19

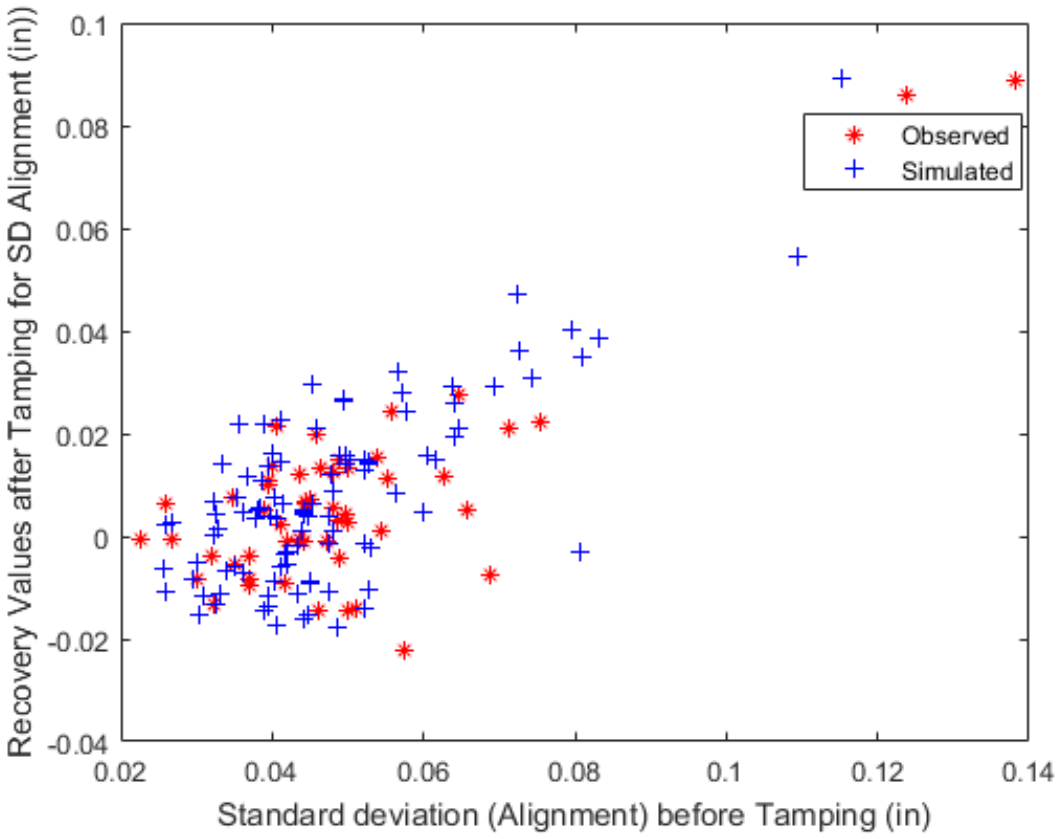


Figure 4.3: Comparison between real and simulated values for SD Alignment given 3-parameter Lognormal marginals (Before tamping and Recovery values) and Joe copula.

illustrated in figure 4.8.

4.6.8 Gage

4.6.8.1 Marginal Distribution fitting

The 3-parameter log-logistic distribution was found to have the best fit for the tamping recovery values of SD gage. The 3-parameter log-logistic distribution offered the lowest statistic for all three tests as shown in Table 4.23. It also had a p-value far greater than 0.05 for Kolmogorov-Smirnov and Chi-squared tests meaning the null hypothesis that it follows the distribution can be accepted. The closed form expression

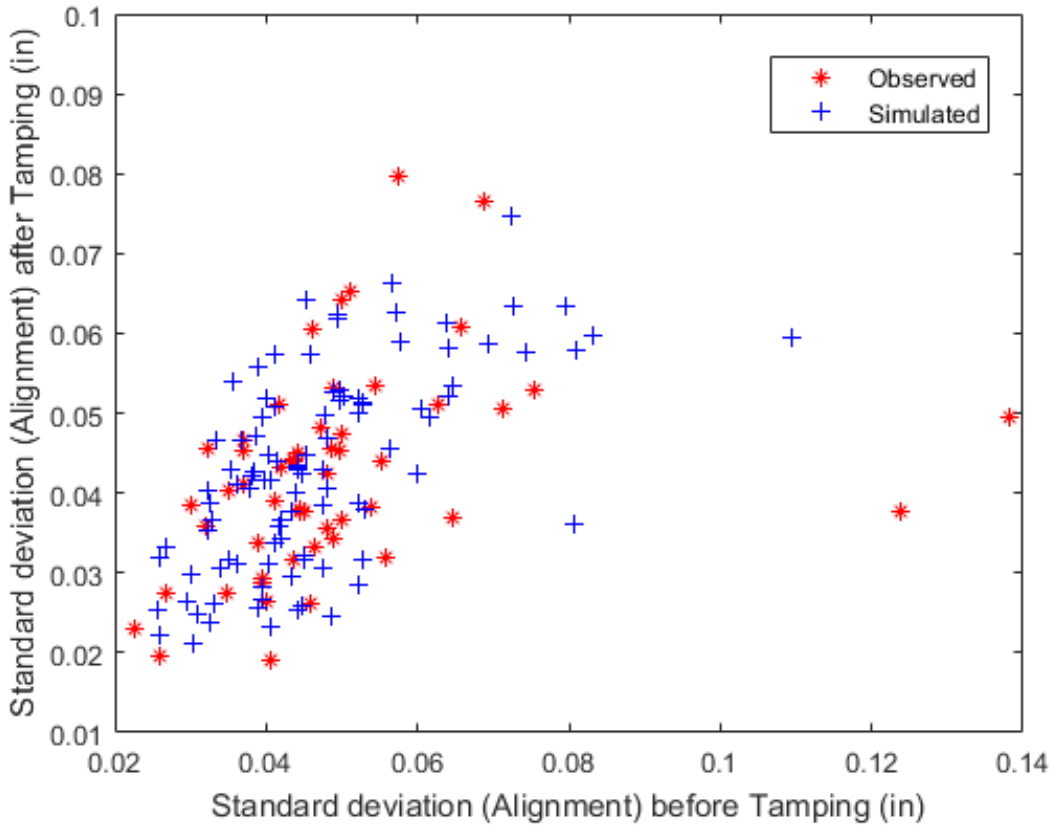


Figure 4.4: Comparison between real and simulated values for SD Alignment given 3-parameter lognormal marginals (Before tamping and after tamping) and Gaussian (Normal) copula.

for the p-value of the 3-parameter log-logistic distribution however does not exist for the Anderson-Darling test. The 2-parameter lognormal distribution and the 3-parameter log-logistic distribution were found to provide the best fit for both SD Gage values before tamping and SD Gage values after tamping as shown in Tables 4.24 and 4.25 respectively.

4.6.8.2 Copula fitting

The Joe-Frank (BB8) Copula was found to offer the best fit of the underlying dependence between the SD gage values before tamping and tamping recovery values

Table 4.22: Results for the fitted bivariate copula between values before tamping and after tamping for SD Warp

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
Gumbel	$\theta = 2.61$	-	-54.78	-52.81	0.62
BB8	$\theta = 5.72$	$\delta = 0.85$	-53.33	-49.39	0.62
BB6	$\theta = 1.08$	$\delta = 2.49$	-52.79	-48.85	0.62
BB1	$\theta = 0$	$\delta = 2.61$	-52.78	-48.84	0.62
BB7	$\theta = 3.18$	$\delta = 0.75$	-52.40	-48.46	0.6
Joe	$\theta = 3.4$	-	-52.08	-50.11	0.56
Student-t copula	$\rho = 0.82$	$\nu = 2.53$	-51.54	-47.60	0.61
Frank	$\theta = 8.37$	-	-49.53	-47.56	0.62
Gaussian/Normal	$\rho = 0.8$	-	-48.01	-46.04	0.59
Clayton	$\theta = 1.78$	-	-32.41	-30.44	0.47

Table 4.23: Results for the fitted distribution to recovery values for SD Gage

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi squared	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.100	0.617	1.01	*	10.4	0.064
Lognormal (3P)	0.141	0.223	1.75	*	15.327	0.002
Weibull (3P)	0.163	0.106	2.22	*	15.3	0.002
Gamma (3P)	0.143	0.210	1.78	*	15.3	0.002
Normal	0.157	0.131	2.00	0.092	15.7	0.001
Logistic	0.137	0.247	1.15	0.285	10.4	0.034
Exponential (2P)	0.401	4.1E-08	9.93	2.3E-5	38.8	3.8E-9

Table 4.24: Results for the fitted distribution to values before tamping for SD Gage

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.065	0.969	0.225	*	1.21	0.944
Log-Logistic	0.076	0.900	0.250	0.992	0.996	0.963
Lognormal (3P)	0.061	0.982	0.194	*	0.937	0.967
Lognormal	0.059	0.987	0.173	0.996	0.625	0.987
Weibull (3P)	0.067	0.956	0.285	*	0.835	0.975
Weibull	0.077	0.884	1.04	0.254	0.978	0.964
Gamma (3P)	0.065	0.969	0.213	*	0.571	0.989
Gamma	0.071	0.935	0.431	0.946	2.15	0.828
Normal	0.091	0.740	1.010	0.351	1.80	0.876
Logistic	0.100	0.975	0.464	0.783	3.06	0.690
Exponential	0.404	3.3E-8	11.7	1.1E-5	79.4	2.2E-16
Exponential (2P)	0.193	0.033	3.31	0.024	14.8	0.005

Table 4.25: Results for the fitted distribution to values before tamping for SD Gage

Distribution	**Kolmogorov Smirnov**		**Anderson Darling**		**Chi squared**	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Log-Logistic (3P)	0.067	0.957	0.273	*	1.48	0.916
Log-Logistic	0.092	0.725	0.410	0.964	1.88	0.865
Lognormal (3P)	0.075	0.909	0.403	*	2.94	0.709
Lognormal	0.06878	0.949	0.337	0.901	3.23	0.665
Weibull (3P)	0.090	0.749	0.707	*	4.10	0.398
Weibull	0.065	0.605	0.895	0.210	2.19	0.822
Gamma (3P)	0.079	0.871	0.503	*	4.72	0.451
Gamma	0.094	0.705	0.335	0.909	5.15	0.399
Normal	0.1289	0.3159	0.9029	0.412	3.629	0.6059
Logistic	0.1119	0.499	0.587	0.654	2.29	0.807
Exponential	0.408	2.3E-8	12.4	1.1E-5	70.6	3.2E-15
Exponential (2P)	0.209	0.017	3.4	0.024	16.5	0.002

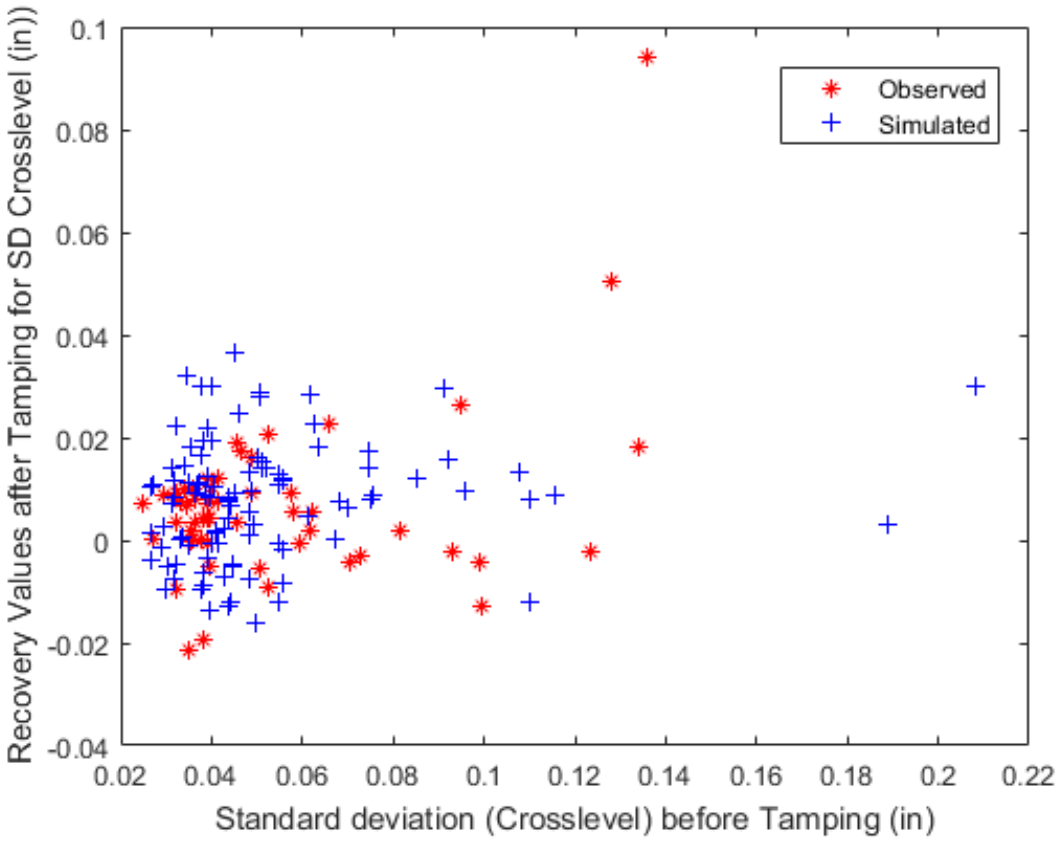


Figure 4.5: Comparison between real and simulated values for SD Crosslevel given 3-parameter Lognormal marginal (Before tamping) 3-parameter loglogistic marginal (Recovery values) and Independent copula.

producing both the lowest AIC and BIC values as shown in Table 4.26. The BB8 copula consists of the Joe Copula and Frank Copula. The Joe copula is suitable for strong upper tail dependence whereas Frank copula is suitable for very strong central dependence with very weak tail dependence. The Frank copula has stronger central dependence than the Gaussian copula (denoted by significant central clustering) and even weaker tail dependence than the Gaussian copula(denoted by fanning out at the tails) (Bhat and Eluru, 2009). Simulated values were generated given 2-parameter log-normal (values before tamping), 3-parameter log-logistic distribution (recovery values) and Joe-Frank (BB8) copula. The comparison of the observed and simulated values is

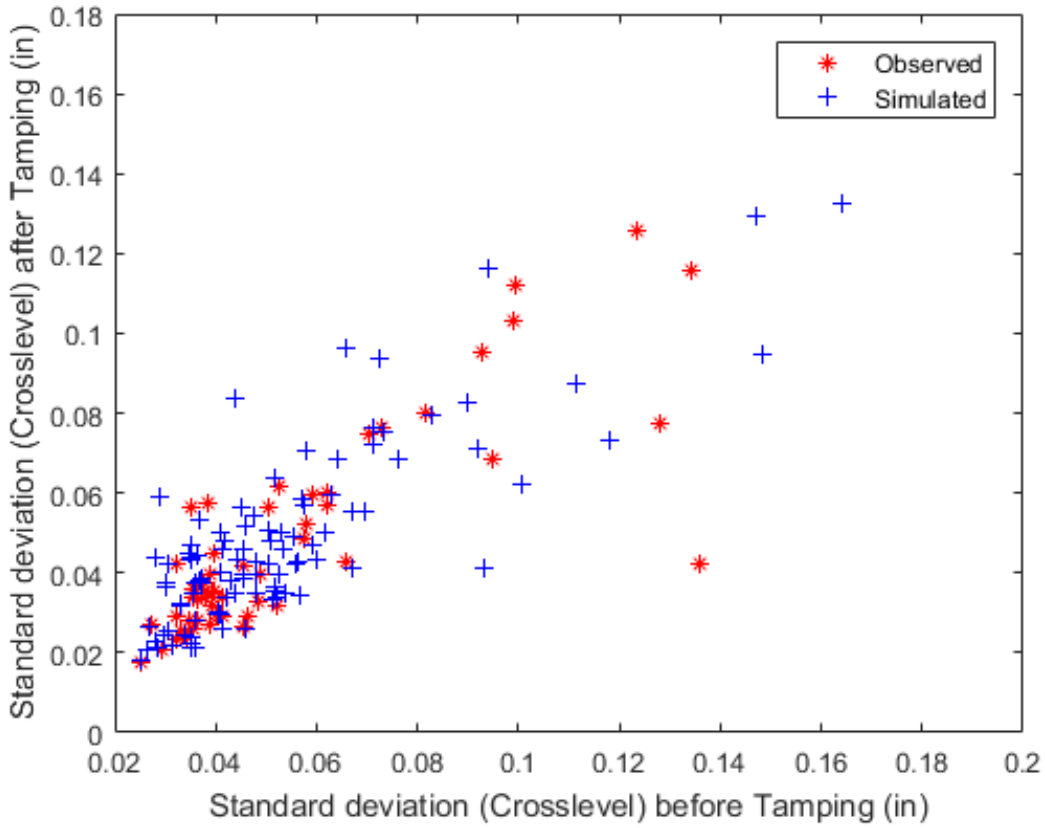


Figure 4.6: Comparison between real and simulated values for SD Crosslevel given 3-parameter lognormal marginals (Before tamping and after tamping) and Joe-Clayton (BB7) copula.

shown in figure 4.9.

The Student-t copula was found to offer the best fit of the underlying dependence between SD gage values before tamping and SD gage values after tamping. The selection of the t-copula suggests radially symmetric dependence with equal upper and lower tail dependence. Simulated values were produced given 2-parameter lognormal marginal (SD gage before tamping), 3-parameter log-logistic marginal (recovery value) and Student t-copula. This is illustrated in figure 4.10.

Table 4.26: Results for the fitted bivariate copula between SD values before tamping and SD recovery values after tamping for Gage

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
BB8	$\theta = 5.99$	$\delta = 0.53$	-15.19	-11.25	0.39
Gumbel	1.54	-	-12.91	-10.94	0.35
Joe	$\theta = 1.8$	-	-11.34	-9.36	0.31
Student-t copula	$\rho = 0.54$	$\nu = 30$	-11.27	-7.33	0.37
BB6	$\theta = 1$	$\delta = 1.54$	-10.91	-6.97	0.35
BB1	$\theta = 0$	$\delta = 1.54$	-10.91	-6.97	0.35
Frank	$\theta = 3.89$	-	-10.51	-6.54	0.38
Gaussian/Normal	$\rho = 0.55$	-	-9.85	-5.88	0.37
BB7	$\theta = 1.8$	$\delta = 0$	-9.33	-5.39	0.31
Clayton	$\theta = 0.66$	-	-6.26	-4.29	0.25

Table 4.27: Results for the fitted bivariate copula between values before tamping and after tamping for SD Gage

Copula	Parameter 1	Parameter 2	AIC	BIC	Kendall's Tau
Student-t copula	$\rho = 0.83$	$\nu = 2$	-54.82	-50.88	0.63
BB7	$\theta = 2.77$	$\delta = 1.26$	-48.61	-44.67	0.6
Gumbel	$\theta = 2.51$	-	-48.19	-46.22	0.6
BB1	$\theta = 0.26$	$\delta = 2.26$	-46.84	-42.90	0.61
BB6	$\theta = 1$	$\delta = 2.5$	-46.19	-42.25	0.6
Joe	$\theta = 3.16$	-	-44.00	-42.03	0.54
BB8	$\theta = 3.16$	$\delta = 1$	-42.00	-38.06	0.54
Gaussian/Normal	$\rho = 0.76$	-	-38.44	-36.47	0.55
Frank	$\theta = 7.21$	-	-38.10	-36.13	0.57
Clayton	$\theta = 1.87$	-	-33.19	-31.22	0.48

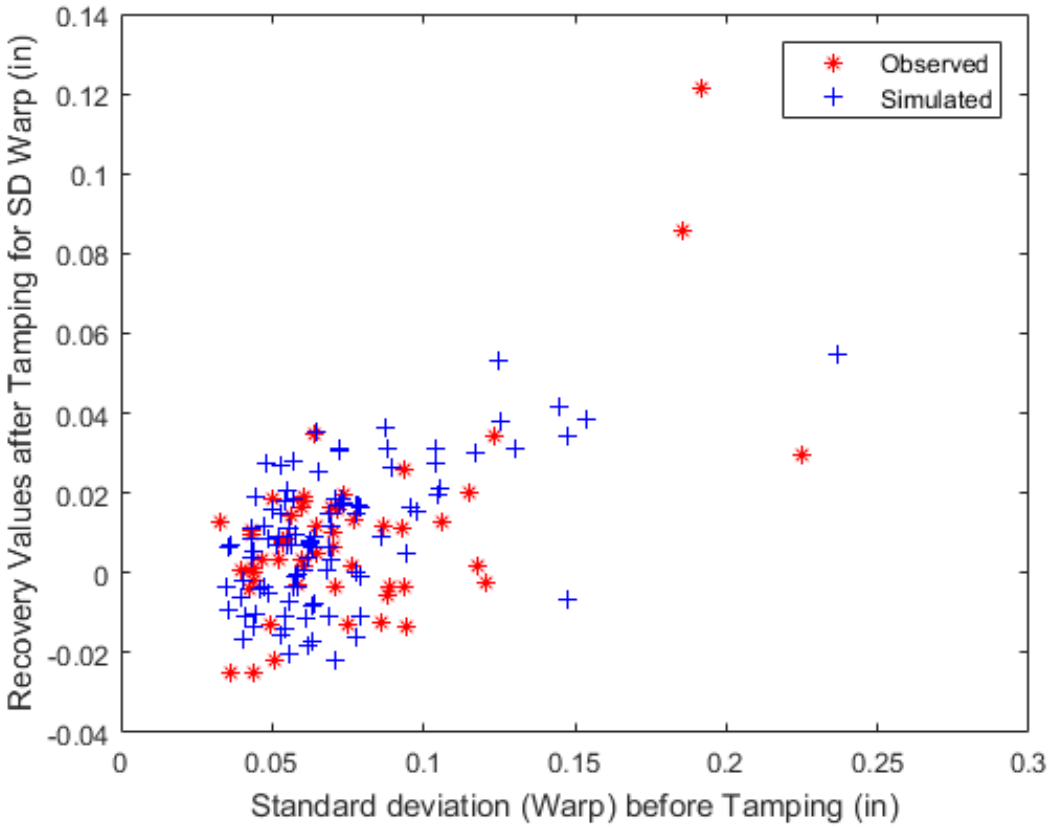


Figure 4.7: Comparison between real and simulated values for SD Warp given 3-parameter Lognormal marginal (Before tamping) 3-parameter loglogistic marginal (Recovery values) and Joe copula.

4.6.9 Correlation Analysis of Recovery Values of Geometry Parameters

Correlation analysis was conducted to measure the dependence between the tamping recoveries of the various track geometry parameter namely surface profile, alignment, cross level, warp and gage. The correlation measures employed include Pearson's correlation coefficient and concordance (or rank correlation) measures such as Kendall's Tau and Spearman's Rho. Pearson's correlation coefficient measures the linear dependence between random variables and assumes that the variables of interest are normal. Thus, the widely-used Pearson's coefficient is not suitable for evaluating non-linear dependence or dependence between non-normal distributions.

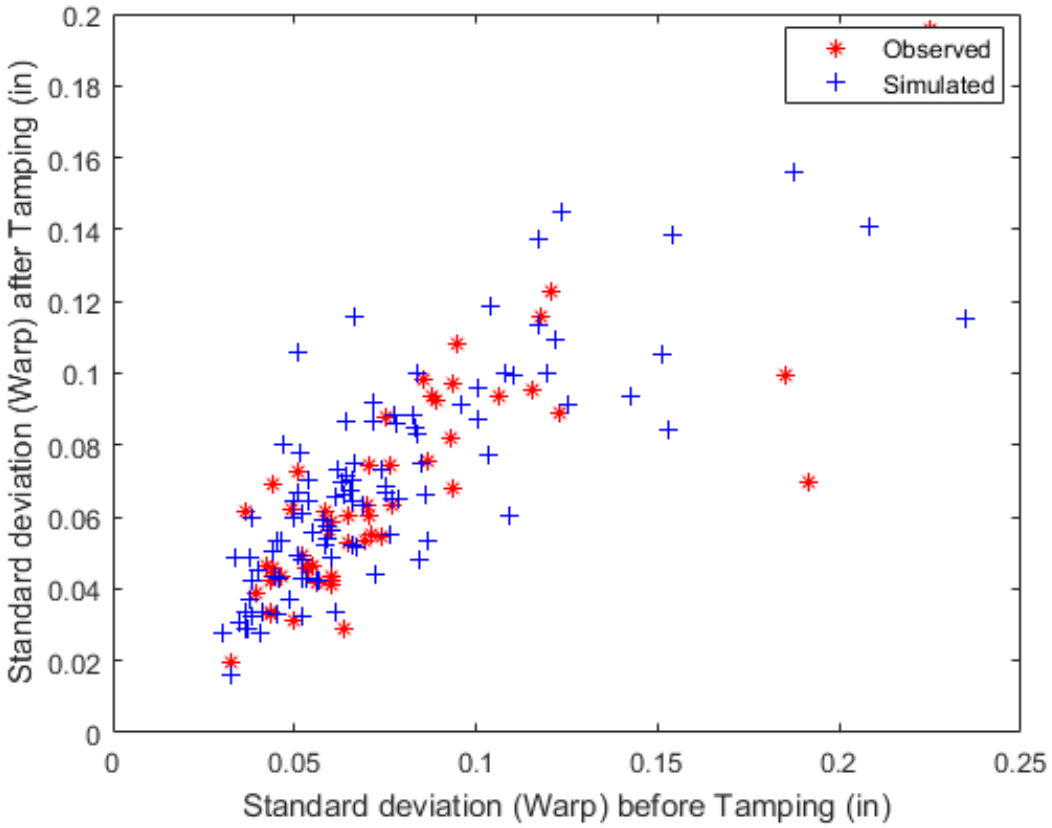


Figure 4.8: Comparison between real and simulated values for SD Warp given 3-parameter lognormal marginals (Before tamping and after tamping) and Gumbel copula.

The results of the linear correlation analysis are shown in Table 4.28. The highest dependence was found between the recoveries of SD warp and SD cross level. The fact that warp is a measure of the cross level variation offers some support to the high dependence observed. On the other hand, the lowest dependence was observed between the recovery values of SD gage and SD surface. Gage is a transverse horizontal parameter whereas surface is a vertical longitudinal parameter. In fact, generally gage was found to have relatively weak dependences between the other parameters with its highest dependence observed with alignment which is a horizontal longitudinal parameter unlike the others. The surface profile was found to have moderate dependence with

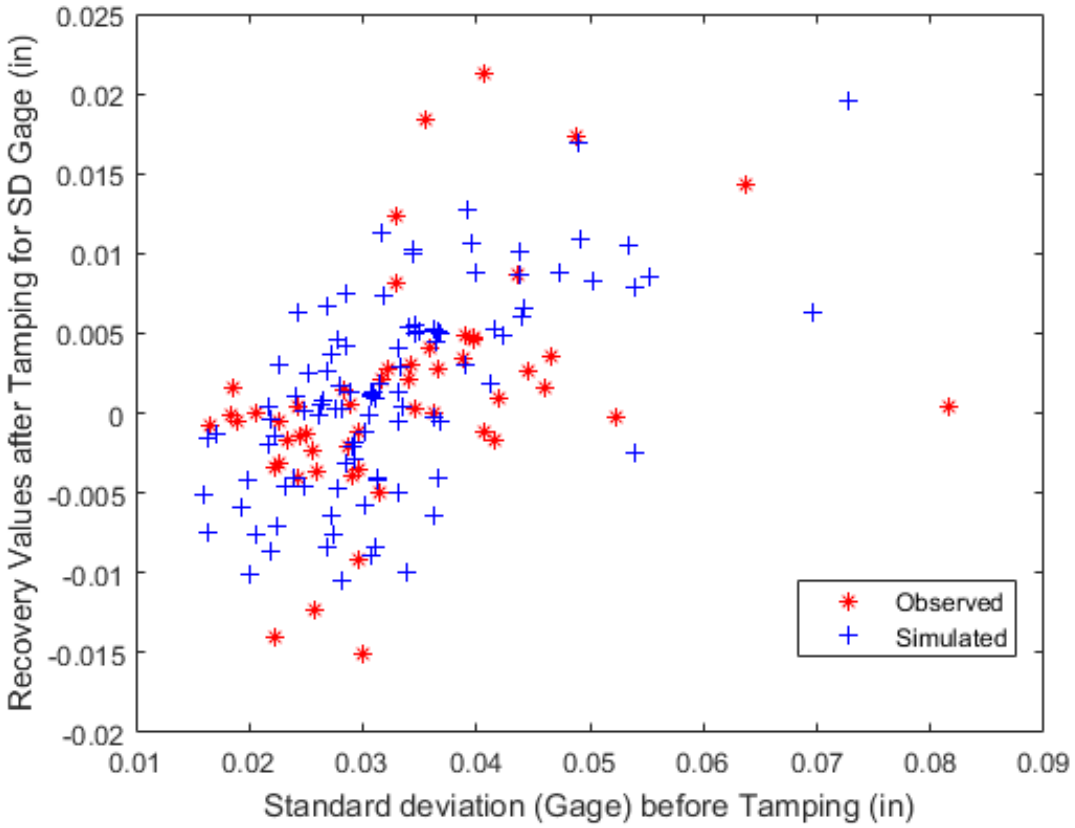


Figure 4.9: Comparison between real and simulated values for SD Gage given 2-parameter Lognormal marginal (Before tamping), 3-parameter log-logistic marginal (recovery value) and Joe-Frank (BB8) copula.

both cross level and warp which are also vertical parameters. Alignment were found to have moderate correlations with vertical parameters such as surface, cross level and warp parameters. Of these three parameters, surface profile was the parameter with the highest dependence with alignment which suggests that tamping affects the surface profile in more similar way to alignment in comparison with the others. Surface and alignment are both longitudinal parameters.

However, a comparison of the linear correlation results with the concordance dependence results shows a general reduction in the observed dependence between the recoveries of the various parameters. These are shown in concordance dependence

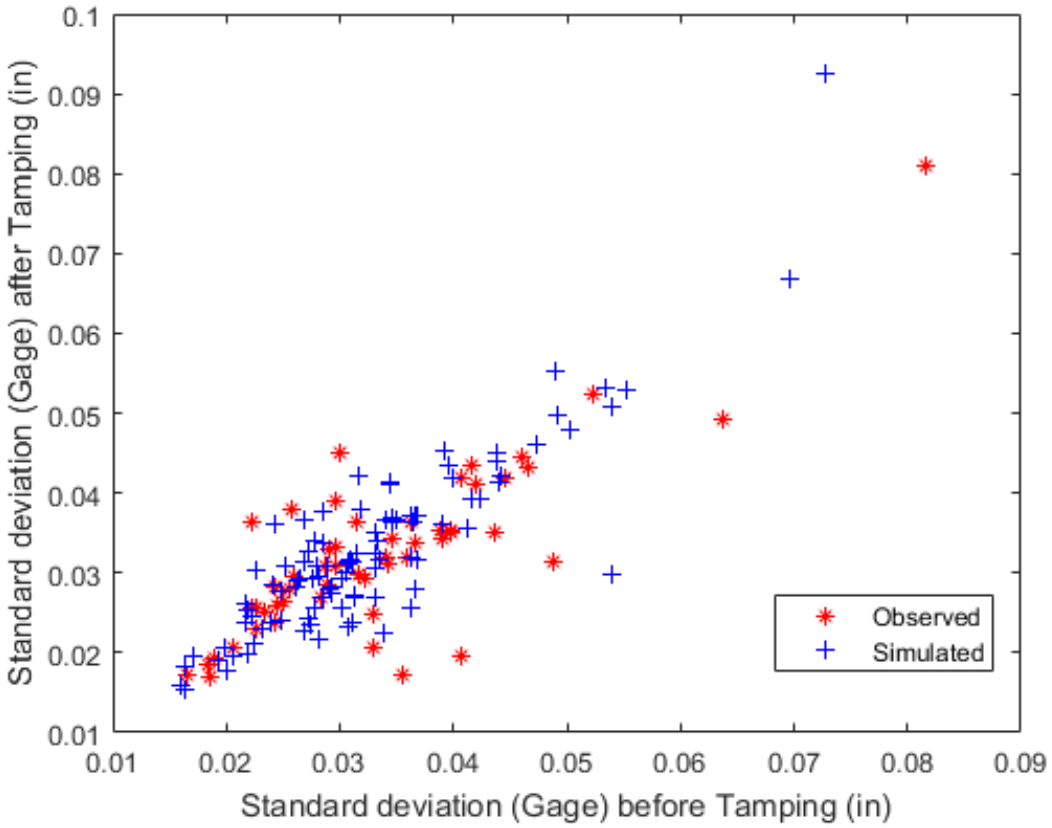


Figure 4.10: Comparison between real and simulated values for SD Gage given 2-parameter Lognormal marginal (Before tamping), 3-parameter log-logistic marginal (after tamping) and Student-t copula.

results such as the Kendall's Tau and Spearman's Rho correlation matrices in Tables 4.29 and 4.30 respectively. For instance, the linear dependence of 0.7919 was found to reduce to 0.3498 and 0.4721 by employing the Kendall's Tau and Spearman's Rho dependence measures which do not assume linear dependence or assume normality of the random variables. As a matter of fact the recoveries of warp and crosslevel and warp were found to assume 3-parameter lognormal distribution and 3-parameter log-logistic distribution as shown in Tables 4.14 and 4.18 respectively. Furthermore, the normal distribution was found to not fit the data as shown in the aforementioned tables. Additionally, an examination of the underlying dependence suggest a Student-t

Table 4.28: Pearson's correlation matrix of recovery values of geometry parameters

Parameter	Surface	Alignment	Cross level	Gage	Warp
Surface	1.00	0.50	0.30	-0.02	0.37
Alignment	0.50	1.00	0.53	0.19	0.61
Cross level	0.30	0.53	1.00	0.07	0.80
Gage	-0.02	0.19	0.07	1.00	0.13
Warp	0.37	0.61	0.79	0.13	1.00

Table 4.29: Kendall's tau correlation matrix of recovery values of geometry parameters

Parameters	Surface	Alignment	Cross level	Gage	Warp
Surface	1.00	0.17	0.07	0.02	0.15
Alignment	0.13	1.00	0.06	0.24	0.31
Cross level	0.07	0.06	1.0	0.12	0.35
Gage	0.02	0.24	0.12	1.00	0.12
Warp	0.15	0.31	0.35	0.11	1.00

copula. Thus, it may be quite misleading to employ linear correlation coefficient not only in modelling the tamping recovery of various parameters but also in analyzing the dependences of the various recoveries of these parameters.

4.6.10 Concluding Remarks

The effect of tamping on various parameters namely surface, alignment, cross level, warp and gage were evaluated by analyzing the recovery of these geometry parameters after tamping. Tamping recovery has been found to be predominantly dependent

Table 4.30: Spearman's rho correlation matrix of recovery values of geometry parameters

Parameters	Surface	Alignment	Cross level	Gage	Warp
Surface	1.00	0.18	0.08	0.04	0.22
Alignment	0.18	1.00	0.08	0.35	0.46
Cross level	0.08	0.08	1.00	0.16	0.47
Gage	0.04	0.35	0.16	1.00	0.17
Warp	0.21	0.46	0.47	0.17	1.00

on the track geometry condition before tamping. It has largely been modeled using deterministic techniques such as linear regression which assumes multivariate normal distribution and linear relationship between the variables. However, non-normality in most cases transpires in various forms: non-normality of marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables. Furthermore, deterministic techniques are not suitable given high degrees of uncertainty which happens to be observed in the recovery values of track geometry measures in majority of cases. Thus, probabilistic techniques are increasingly being employed which take into consideration the high variation in the restoration values after tamping even for similar track geometry condition. Majority of studies do not take into consideration the underlying dependence between the variables of interest. Thus, the authors employ a copula-based approach to model the tamping recovery phenomenon by combining arbitrary marginal distributions to form a joint distribution with the underlying dependence.

From marginal fitting results, the recoveries of the various parameters were found to be non-normal and were found to either fit a 3-parameter lognormal distribution (in the case of surface, alignment and warp) or 3-parameter log-logistic distribution (in the case of cross level and gage). Similarly, non-normal distributions were observed for the track quality condition (standard deviation of track geometry parameters) before and after tamping. Various copulas were fitted in order to find the copula which best describe the underlying dependence between the variables. The selection of copulas such as Gumbel, Joe and Joe-Clayton copulas (BB7) suggest the presence of asymmetric and tail dependence which cannot be appropriately captured using the widely-used linear regression. Thus, conventional correlation analysis appears not to be suitable for analyzing the dependences between the recovery values and tamping condition before tamping.

Correlation analysis of the recovery of various geometry parameters show that the use of Pearson's correlation coefficient which assumes normality of the variables

and linear dependence led to relatively high dependence values observed. However, the use of concordance measures such as Kendall's Tau and Spearman's Rho resulted in a general reduction in the observed dependences. These concordance measures are scale-invariant and are suitable for evaluating non-linear dependence and measure dependence irrespective of assumed distribution. Thus, the widely-used Pearson's correlation coefficient does not appear to be appropriate for analyzing the correlation between the recoveries of the various track geometry parameters. From the correlation analysis results, the strongest correlation was observed between warp and cross level recoveries with the weakest correlation observed between the surface and gage recoveries with varying levels in-between. This infers and gives credence to previous research that tamping affects the various track geometry parameters differently thus it is imperative to examine all the track geometry parameters and not focus on one or two parameters.

The copula-based approach was employed by considering only the predominant factor which is the track geometry condition or quality before tamping. However, this methodology can be extended to incorporate and examine other factors such as operational speed, tamping procedure, age of track components and number of previous tamping operations. To analyze the effect of various covariates on track geometry recoveries of several parameters, copula-based regression models can be employed taking into consideration the dependence between the recovery variables. In order to analyze the dependences between more than two variables, vine copulas are suggested which are more flexible than regular multivariate copulas. Vine copulas employ arbitrary bivariate copulas as building blocks for the construction of higher-dimensional multivariate distributions.

The copula-based tamping recovery model can be incorporated into track geometry maintenance scheduling models with track geometry degradation models and recovery models being the main components of these models. Degradation models that can be considered include linear and exponential regression models, polynomial models, multi-stage linear models, neural networks, grey models, path analysis, data mining,

models with random coefficient, Markov models, time series models and stochastic processes. There is the need to select an appropriate track geometry deterioration model that takes into consideration both the time and spatial variation of the track geometry degradation process (Soleimanmeigouni et al., 2016b). The combination of such a model with a copula-based approach that models the tamping recovery phenomena considering the underlying dependence will lead to better track geometry condition estimation for maintenance activity planning. The combination of such models will also result in a greater comprehension of track geometry maintenance modelling. This proposed methodology will be considered in a future case study. In order to integrate such degradation models and copula-based recovery models in track scheduling models, probabilistic optimization models need to be considered.

REFERENCES

Akaike, Hirotugu. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 15582523. doi: 10.1109/TAC.1974.1100705.

Atique, Farzana and Attoh-Okine, Nii. Using copula method for pipe data analysis. *Construction and Building Materials*, 106:140–148, 2016. ISSN 09500618. doi: 10.1016/j.conbuildmat.2015.12.027.

Attoh-Okine, Nii O. Pair-copulas in infrastructure multivariate dependence modeling. *Construction and Building Materials*, 49:903–911, 2013. ISSN 09500618. doi: 10.1016/j.conbuildmat.2013.06.055.

Audley, M. and Andrews, J. The effects of tamping on railway track geometry degradation. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 227(4):376–391, 2013. ISSN 0954-4097. doi: 10.1177/0954409713480439.

Ayuso, Mercedes; Bermúdez, Lluís, and Santolino, Miguel. Copula-based regression modeling of bivariate severity of temporary disability and permanent motor injuries. *Accident Analysis and Prevention*, 89:142–150, 2016. ISSN 00014575. doi: 10.1016/j.aap.2016.01.008.

Bedford, Tim and Cooke, Roger M. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268, 2001. ISSN 1573-7470. doi: 10.1023/A:1016725902970.

Berg, Daniel. Copula goodness-of-fit testing: An overview and power comparison. *European Journal of Finance*, 15(7-8):675–701, 2009. ISSN 1351847X. doi: 10.1080/13518470802697428.

Bhat, Chandra R. and Eluru, Naveen. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7):749–765, 2009. ISSN 01912615. doi: 10.1016/j.trb.2009.02.001.

Biau, Gérard and Wegkamp, Marten. A note on minimum distance estimation of copula densities. *Statistics and Probability Letters*, 73(2):105–114, 2005. ISSN 01677152. doi: 10.1016/j.spl.2005.02.006.

Brechmann, E C and Schepsmeier, U. Modeling dependence with C-and D-vine copulas: The R-package CDVine. *Journal of Statistical Software*, 52(3):1–27, 2013.

Caetano, Luis Filipe and Teixeira, Paulo Fonseca. Predictive Maintenance Model for Ballast Tamping. *Journal of Transportation Engineering*, 142(4):4016006, 2016. ISSN 0733-947X. doi: 10.1061/(ASCE)TE.1943-5436.0000825.

Chen, Song Xi and Huang, Tzee-ming. Nonparametric estimation of copula functions for dependence modeling. *Canadian Journal of Statistics*, 35(2):1–18, 2007. ISSN 03195724. doi: 10.1002/cjs.5550350205.

Clarke, Kevin A. A simple distribution-free test for nonnested model selection. *Political Analysis*, 15(3):347–363, 2007. ISSN 10471987. doi: 10.1093/pan/mpm004.

Czado, C.; Schepsmeier, U., and Min, A. Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, 12(3):229–255, 2012. ISSN 1471-082X. doi: 10.1177/1471082X1101200302.

Dalla Valle, Luciana; De Giuli, Maria Elena; Tarantola, Claudia, and Manelli, Claudio. Default probability estimation via pair copula constructions. *European Journal of Operational Research*, 249(1):298–311, 2016. ISSN 03772217. doi: 10.1016/j.ejor.2015.08.026.

Deheuvels, Paul. La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. *Acad. Royale de Belgique, Bulletin de la classe des sciences*, 5(65):274–292, 1979.

Dissmann, J.; Brechmann, E. C.; Czado, C., and Kurowicka, D. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis*, 59(1):52–69, 2013. ISSN 01679473. doi: 10.1016/j.csda.2012.08.010.

Dorey, M and Joubert, Phil. Modelling copulas: an overview. *The Staple Inn Actuarial Society*, pages 1–27, 2005.

Eluru, N; Paleti, R; Pendyala, R M, and Bhat, C R. Modeling Injury Severity of Multiple Occupants of Vehicles Copula-Based Multivariate Approach. *Transportation Research Record*, (2165):1–11, 2010. ISSN 0361-1981. doi: Doi10.3141/2165-01.

Embrechts, Paul; Mcneil, Alexander, and Straumann, Daniel. Correlation And Dependence In Risk Management: Properties And Pitfalls. 2002.

Embrechts, Paul; Lindskog, Filip, and Mcneil, Alexander. Modelling Dependence with Copulas and Applications to Risk Management. In *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. Elsevier, 2003. ISBN 9780444508966. doi: 10.1016/B978-044450896-6.50010-8.

Esveld, Coenraad. *Modern Railway Track, 2nd Edition*. 2001. ISBN 9080032433.

Famurewa, S. M.; Xin, T.; Rantatalo, M., and Kumar, U. Optimisation of maintenance track possession time: A tamping case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(1):12–22, 2013. ISSN 0954-4097. doi: 10.1177/0954409713495667.

Famurewa, S. M.; Juntti, U.; Nissen, A., and Kumar, U. Augmented utilisation of possession time: Analysis for track geometry maintenance. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 230(4): 1118–1130, 2016. ISSN 0954-4097. doi: 10.1177/0954409715583890.

Fermanian, M H and Scaillet, O. Nonparametric Inference of copulas for time series. *Journal of Risk*, 5(April 2007):25–54, 2003.

Fisher, N. I. and Switzer, P. Chi-Plots for Assessing Dependence. *Biometrika*, 72(2): 253, aug 1985. ISSN 00063444. doi: 10.2307/2336078.

Fisher, N. I. and Switzer, P. Graphical Assessment of Dependence: Is a Picture Worth 100 Tests? *The American Statistician*, 55:233–239, 2001. doi: 10.2307/2685807.

Genest, Christian and Boies, Jean-Claude. Detecting Dependence with Kendall Plots. *The American Statistician*, 57:275–284, 2003. doi: 10.2307/30037296.

Genest, Christian and Favre, Anne-Catherine. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12 (4):347–368, 2007. ISSN 1084-0699. doi: 10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, Christian and Rivest, Louis-Paul. Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*, 88(423): 1034, 1993. ISSN 01621459. doi: 10.2307/2290796.

Genest, Christian; Ghoudi, K, and Rivest, Louis-Paul. A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika*, 82(3):543–552, 1995.

Genest, Christian; Rémillard, Bruno, and Beaudoin, David. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44 (2):199–213, 2009. ISSN 01676687. doi: 10.1016/j.insmatheco.2007.10.005.

Genest, Christian; Carabarín-aguirre, Alberto, and Harvey, Fanny. Copula parameter estimation using Blomqvist's beta. *Journal de la Société Française de Statistique*, 154(1), 2013.

Grimaldi, Salvatore and Serinaldi, Francesco. Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources*, 29(8):1155–1167, aug 2006. ISSN 03091708. doi: 10.1016/j.advwatres.2005.09.005.

Huang, Wanling and Prokhorov, Artem. A Goodness-of-fit Test for Copulas. *Econometric Reviews*, 33(7):751–771, 2014. ISSN 07474938. doi: 10.1080/07474938.2012.690692.

Joe, Harry and Xu, James Jianmeng. The Estimation Method of Inference Functions for Margins for Multivariate Models. *Technical Report no. 166, Department of Statistics, University of British Columbia*, pages 1–21, 1996. ISSN 1098-6596. doi: 10.14288/1.0225985.

Jovanovic, Stanislav. Railway track quality assessment and related decision making. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 6:5038–5043, 2004. ISSN 1062922X. doi: 10.1109/ICSMC.2004.1400992.

Khouy, Iman Arasteh. *Cost-Effective Maintenance of Railway Track Geometry*. PhD thesis, Lulea University of Technology, 2013.

Kim, Gunky; Silvapulle, Mervyn J., and Silvapulle, Paramsothy. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, 51(6):2836–2850, 2007. ISSN 01679473. doi: 10.1016/j.csda.2006.10.009.

Krämer, Nicole and Schepsmeier, Ulf. Introduction to vine copulas, 2011.

Lewandowski, Daniel. *High Dimensional Dependence: Copulae, Sensitivity, Sampling*. PhD thesis, Delft University of Technology, 2008.

Liu, Jia. *Extreme Value Theory and Copula Theory : A Risk Management Application with Energy Futures*. PhD thesis, University of Victoria, 2011.

Ma, Ming-wei; Ren, Li-liang; Song, Song-bai; Song, Jia-li, and Jiang, Shan-hu. Goodness-of-fit tests for multi-dimensional copulas: Expanding application to historical drought data. *Water Science and Engineering*, 6(1):18–30, 2013. ISSN 1674-2370. doi: http://dx.doi.org/10.3882/j.issn.1674-2370.2013.01.002.

Miwa, Masashi. Mathematical Programming Model Analysis for the Optimal T rack Track Maintenance Schedule. *Quart Rep RTRI*, 43(3), 2002.

Morettin, Pedro A.; Toloi, Clelia M.C.; Chiann, Chang, and de Miranda, José C.S. Wavelet Estimation of Copulas for Time Series. *Journal of Time Series Econometrics*, 3(3), 2011. ISSN 1941-1928. doi: 10.2202/1941-1928.1033.

Nashad, Tammam; Yasmin, Shamsunnahar; Eluru, Naveen; Lee, Jaeyoung, and Abdel-Aty, Mohamed A. Joint Modeling of Pedestrian and Bicycle Crashes Copula-Based Approach. *Transportation Research Record*, (2601):119–127, 2016. ISSN 0361-1981. doi: 10.3141/2601-14.

Nelsen, Roger B. *An Introduction to Copulas*. Springer New York, New York, NY, 2nd edition, 2006. ISBN 0387947256. doi: 10.1007/978-0-387-98135-2.

Nicoloutsopoulos, Dimitris. *Parametric and Bayesian non-parametric estimation of copulas*. PhD thesis, University College London, 2005.

Oh, Dong Hwan and Patton, Andrew J. Simulated method of moments estimation for copula-based multivariate models. *Journal of the American Statistical Association*, 108(502):689–700, 2013. ISSN 01621459. doi: 10.1080/01621459.2013.785952.

Patton, Andrew J. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012. ISSN 0047259X. doi: 10.1016/j.jmva.2012.02.021.

Quiroga, L. M. and Schnieder, E. Monte Carlo simulation of railway track geometry deterioration and restoration. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(3):274–282, 2012. ISSN 1748-006X. doi: 10.1177/1748006X11418422.

Quiroga, L. M.; Schnieder, E., and Antoni, M. Holistic long term optimization of maintenance strategies on ballasted railway track. In *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012*, 2012.

Salvadori, G. and De Michele, C. On the use of copulas in dependent competing risk theory. *Journal of Hydrologic Engineering*, 12(August):369–380, 1992.

Schepsmeier, Ulf. *Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different families*. PhD thesis, Technical University of Munich, 2010.

Schepsmeier, Ulf and Czado, Claudia. Dependence modelling with regular vine copula models: A case-study for car crash simulation data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 65(3):415–429, 2016. ISSN 14679876. doi: 10.1111/rssc.12125.

Schwarz, Gideon. Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461–464, 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136.

Sklar, A. Fonctions de Répartition à n Dimensions et Leurs Marges. *Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

Soleimanmeigouni, I.; Ahmadi, A.; Arasteh Khouy, I., and Letot, C. Evaluation of the effect of tamping on the track geometry condition: A case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:408–420, 2016a. ISSN 0954-4097. doi: 10.1177/0954409716671548.

Soleimanmeigouni, I.; Ahmadi, A., and Kumar, U. Track geometry degradation and maintenance modelling: A review. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232:73–102, 2016b. ISSN 0954-4097. doi: 10.1177/0954409716657849.

Srinivas, S.; Menon, Devdas, and Prasad, A. Meher. Multivariate simulation and multimodal dependence modeling of vehicle axle weights with copulas. *Journal of Transportation Engineering*, 132(12):945–955, dec 2006.

Tsukahara, Hideatsu. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375, 2005.

Vuong, Quang H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307, 1989. ISSN 00129682. doi: 10.2307/1912557.

Wei, Wei. *Copula-based High Dimensional Dependence Modelling*. PhD thesis, University of Technology, Sydney, Australia, 2014.

Weiß, Gregor. Copula parameter estimation by maximum-likelihood and minimum-distance estimators: A simulation study. *Computational Statistics*, 26(1):31–54, 2011. ISSN 09434062. doi: 10.1007/s00180-010-0203-7.

Yan, Jun. Multivariate Modeling with Copulas and Engineering Applications. In *Springer Handbook of Engineering Statistics*, pages 973–990. Springer London, London, 2006. doi: 10.1007/978-1-84628-288-1_51.

Zhang, Shulin; Okhrin, Ostap; Zhou, Qian M., and Song, Peter X.K. Goodness-of-fit test for specification of semiparametric copula dependence models. *Journal of Econometrics*, 193(1):215–233, 2016. ISSN 18726895. doi: 10.1016/j.jeconom.2016.02.017.

Zilko, Aurelius A.; Kurowicka, Dorota, and Goverde, Rob M P. Modeling railway disruption lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies*, 68:350–368, 2016. ISSN 0968090X. doi: 10.1016/j.trc.2016.04.018.

Chapter 5

COPULA-BASED REGRESSION MODELS

5.1 Introduction

This chapter discusses the basic concepts of copula-based regression models which combine several marginal regression models with a bivariate parametric copula which characterizes the underlying dependence between the response variables. The various types of marginal regression models including general linear models, generalized linear models, generalized additive models and generalized additive models for location, scale and shape are discussed. Past applications of copula-based regression models in several transportation fields including modeling automobile crash severity are reviewed. The model formulation of the mixed copula-based regression model including marginal and copula selection and statistical inference (parameter estimation) are also provided. The chapter concludes with a case study on the application of copula-based regression models to bivariate severity analysis of train derailments. A joint mixed copula-based model for derailed cars and monetary damage is presented for the joint analysis of their relationship with a set of covariates that might influence both outcomes.

5.2 Marginal Regression Models

Regression analysis is a statistical method used in estimating the relationship among variables. Regression analysis seeks to determine the strength of the relationship between a dependent variable (also known as outcome or response variables) and a set of changing independent variables (also known as explanatory variables, predictors or covariates). Regression models are widely used in prediction and forecasting of response variables. Regression models are referred to as simple regression models if only one covariate is used in the prediction of a response variable whereas multiple regression

models are regression models where more than one predictor is used in the prediction of an outcome.

The corresponding regression models of the marginal distributions of the various responses can be referred to as the marginal regression models. The marginal regression models of a copula-based regression model can be defined via the following models:

- General Linear Models
- Generalized Linear Models
- General Additive Models
- General Additive Models for Location, Scale and Shape

5.2.1 General Linear Models

The general (or multivariate) linear model is a generalization of multiple linear regression model to the case of more than one response variable. A multiple linear regression model can be expressed as:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \tag{5.1}$$

where the response variable $y_i, i = 1, ..., n$ is modeled as a linear function of the covariates $x_j, j = 1, ..., p$ and an error term $\epsilon_i, i = 1, ..., n$. The errors are assumed to be independent and identically distributed such that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

Given n observations on q-vector of responses y and p-vector of covariates x, the multivariate regression model is given as

$$Y = XB + E \tag{5.2}$$

where $Y = (y_1, ..., y_n)' \in \mathbb{R}^{n \times q}$, $X = (x_1, ..., x_n)' \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times q}$ represent the responses, predictors and regression coefficient matrices respectively and E is the noise matrix comprising iid normal random variables. The model assumes the response variable and the error terms from the fitted models are normally distributed. However, they are not appropriate if the range of response variable is restricted such as the

case of binary or count variables. They are also not suitable if the variance of the response variable is dependent on the mean. Transformation of response variable may enhance linearity and homogeneity of variance thus making a general linear model applicable. However, this approach has limitations including the change of response variable, need for simultaneous enhancement of linearity and homogeneity of variance and transformation may not be defined on the boundaries of the sample space.

5.2.2 Generalized Linear Models

The marginal regression models (marginals) of the copula-based regression model employed in this chapter are defined via generalized linear models (GLMs). Generalized linear models (GLMs) were proposed by Nelder and Wedderburn (1972) to extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions. GLMs address the aforementioned limitations of normal-response models by extending its framework by considering distributions from the exponential family. The exponential family of distributions can be expressed as

$$f(y|\theta) = h(y)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(y)\right) \tag{5.3}$$

where $h(y) \geq 0$ and $t_1(y), ..., t_k(y)$ are real-valued functions of the observation of the random variable y (which cannot be dependent on the parameter θ and $c(\theta) \geq 0$ and $w_1(\theta), ..., w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (which is not dependent on y) (Casella and Berger, 2002). Thus, an exponential family requires the parameters (θ) and the value of the random variable y. Common distributions belonging to this family include normal, gamma, beta and Poisson distributions. Other distributions are members only when certain parameters are fixed or known such as binomial (given fixed number of trials), multinomial (given fixed number of trials) and negative binomial (given fixed number of failures).

The distributions of exponential families can be defined by the natural parameter, a function of the mean, and the dispersion parameter, a function of the variance

that is needed to produce standard error of the point estimates. For distributions such as Poisson, the variance is related to its mean thus the dispersion parameter is equal to one whereas for others such as Gamma, the dispersion parameter is estimated separately from the mean (Quinn and Keough, 2002).

Given an independent sample data (x_i, y_i), $i = 1, ..., n$ where y_i is the response, n is the sample size and $x_i = (x_{i1}, ..., x_{ip})^T$ is a vector of p covariates, a generalized linear model (GLM) comprises three parts namely:

1. A random component $f(y; \mu)$ which specifies the conditional distribution of the response variable given the covariates where μ is the mean of the distribution.
2. A systematic component $\eta = \beta^T x$ known as the linear predictor which specifies the variations in the response variable accounted for by the known covariates.
3. An invertible link function $g(\mu) = \beta^T x$ that ties the two components. It links the expected value of the response variable to the covariates by the function

$$g(\mu(x_i)) = \eta_i = \beta^T x_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} \tag{5.4}$$

where $\beta_{i1}, ..., \beta_{ip}$ are the parameters to be estimated. The vector of regression parameters is usually estimated using maximum likelihood estimation using Newton-Raphson algorithm with the expected Hessian. The maximum likelihood estimates are obtained by solving the score equations using the Fisher's Method of Scoring algorithm which can be fitted by iteratively reweighted least squares (IRLS) algorithm (Yee, 2008).

The IRLS algorithm is a more special case of the Newton-Raphson algorithm. The algorithm computes iterative updates. In the m-th iteration, the new estimate $\hat{\beta}^{(m+1)}$ is obtained from the previous estimate by $\hat{\beta}^m$

$$\hat{\beta}^{(m+1)} = \hat{\beta}^m - \mathbf{H}^{-1}\mathbf{u} \tag{5.5}$$

where $\mathbf{u}$ is the score vector and $\mathbf{H}$ is the Hessian matrix (the first and second derivatives of the log-likelihood respectively), both of which are evaluated at $\hat{\beta}^m$. The updates can be written as

$$\hat{\beta}^{(m+1)} = (X^\top W^{(m)} X)^{-1} X^\top W^{(m)} z^{(m)} \tag{5.6}$$

which is the score equation for a weighted least squares regression of $z^{(m)}$ on $\mathbf{X}$ with weights $W^{(m)} = diag(w_i)$ where adjusted (working) responses

$$z_i^{(m)} = \eta_i^{(m)} + \left(y_i + \mu_i^{(m)}\right) g'\left(\mu_i^{(m)}\right)$$

and adjusted (working) weights

$$w_i^{(m)} = \frac{a_i}{Var\left(\mu_i^{(m)}\right)\left(g'\left(\mu_i^{(m)}\right)\right)^2}.$$

Hence the estimates can be obtained using the IRLS algorithm as follows:

1. Start with initial value $\hat{\beta}^{(0)}$
2. For $m = 0, 1, 2, ...,$
   (a) Compute working responses $z_i^{(m)}$ and working weights $w_i^{(m)}$ based on $\hat{\beta}^{(m)}$
   (b) Solve for $\hat{\beta}^{(m+1)}$
   (c) Check for convergence of $\hat{\beta}$ and stop if met

The variables of the marginal regression models can be selected using the hierarchical model selection technique based on the deviance (Agresti, 2007). The deviance of a statistical model is is a goodness-of-fit statistic which can be defined as the likelihood-ratio statistic for comparing the model to the saturated model (S). The saturated model can be defined as the most complex model possible that offers a perfect fit to the data with a unique parameter for each observation (Agresti, 2007). It is a model that explains all the variation in the data (Quinn and Keough, 2002).

The maximized log-likelihood of the saturated model (L_S) is greater or equal to the maximized log-likelihood of a simpler model (L_G) due to its additional parameters. Thus, deviance can be defined as the test statistic for the hypothesis that all parameters in the saturated model but not in the simpler model equal zero. The deviance of a

generalized linear model G can be mathematically expressed as:

$$Deviance(G) = -2[L_G - L_S] \tag{5.7}$$

For normal-response models, the F-test comparison of the statistical models decomposes a sum of squares representing the variability in the data. This analysis of variance for decomposing variability generalizes to an analysis of deviance for GLMs. Given two GLMs, A and B, such that A is a special case of B, the likelihood-ratio statistic for testing that the simpler model A holds given the complex model B holds is $-2[L_A - L_B]$. Furthermore, the models can be compared by their deviances since

$$-2[L_A - L_B] = -2[L_A - L_s] - \{-2[L_B - L_s]\} = Deviance(A) - Deviance(B)$$

A large test statistic indicates a poor fit of A in comparison with B. For large samples, the statistic has an approximate chi-squared distribution, with degrees of freedom (df) equal to the difference between the residual df values of the two models. This difference in df equals the number of additional parameters that are in B but not in A (Agresti, 2007). In order to select the simpler model A over model B, a rule of thumb for evaluating the statistic requires that the difference in deviance between the models should not be more than twice the difference in the number of parameters estimated (Kreft and de. Leeuw, 1998).

5.2.3 Generalized Additive Models

Generative Additive Models (GAM) proposed by Hastie and Tibshirani (1990) can be defined as a generalized linear model (GLM) in which the linear predictor is linearly dependent on a sum of smooth functions of the covariates. GAM models are semi-parametric regression models since the response variable follows a parametric distribution but modelling of the distribution parameters, as function of covariates, may include non-parametric smoothing functions. For GAM models, equation 5.4 can

be rewritten as

$$g(\mu(x_i)) = \eta_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) \tag{5.8}$$

where f_j are the smooth function of the covariates, x_k. GAM permits the flexible specification of the dependence of the response variable on the covariates. This is specified in relation to (in terms of) smooth function instead of detailed parametric relationships (Wood, 2006). Thus, GAMs extend the flexibility of GLMs by allowing a variety of non-parametric smoothing functions other than just linear relationships. GAMs converge slower than GLMs with increasing n but have less bias. GLMs can be viewed as a special case of GAMs.

5.2.4 Generalized Additive Models for Location, Scale and Shape

Generalized Additive Models for location, scale and shape (GAMLSS) are semi-parametric regression type models proposed by Rigby et al. (2005) which extend generalized additive models (GAM) by not restricting the response variable to only the exponential family. Thus, the distributions of response variables of GAMLSS models can incorporate highly skewed and/or kurtotic continuous and discrete distributions. Also, they are suitable for heterogeneous response variables where the scale and shape parameters vary with the covariates. Furthermore, all its parameters can be dependent on flexible (linear/non-linear or smooth) functions of covariates. This permits not only the modeling of the mean (or location) parameter but also other parameters of the response distribution as linear and/or non-linear, parametric and/or additive non-parametric functions of covariates and/or random effects (Stasinopoulos and Rigby, 2007).

5.3 Copula Regression Models

5.3.1 General

The simultaneous modeling of multiple response or outcomes conditional on several covariates are achievable using copulas. Copulas provide a suitable and computationally tractable framework to model multivariate responses in a regression context by taking into consideration the dependences between the response variables. Copulas also address endogeneity by taking into account similar unobserved or omitted factors that may affect the response variables.

Although copula-based methodologies have not been employed in derailment severity, it has been employed in various transportation applications including modeling automobile crash severity. Bhat and Eluru (2009) employed a copula-based approach to accommodate residential self-selection effects in travel behavior modeling between neo-urbanist and conventional neighborhoods. The authors showed that the copula-based approach showed a considerable level of residential self-selection which could not be detected using bivariate normal dependence structures. Eluru et al. (2010) investigated the injury severity of multiple occupants of vehicles by employing a copula-based multivariate approach which takes into consideration similar unobserved factors that may simultaneously affect severity levels of the occupants in the same crashed-involved vehicle. The Frank copula-based model was found to outperform than the ordered-probit-logit model of independence.

Rana et al. (2010) employed a copula-based approach to address endogeneity in severity models of traffic crash injuries applicable to two-vehicle crashes. This approach took into account endogeneity due to similar unobserved factors affects the response variables namely crash type and injury severity as well as endogeneity between injury severity of the drivers involved in crash. A copula-based joint ordered logit-ordered logit (ORL-ORL) model was developed to jointly model the injury severity levels of the drivers whereas a copula-based joint multinomial logit-ordered logit (MNL-ORL) model was developed to jointly model the response variables. The copula-based models were found to outperform the independent models. Yasmin et al. (2014) built on

the work by Rana et al. (2010) by taking account varying dependences between injury severity and collision type across different categories of collision types. This was based on the hypothesis that collision type basically changes the injury severity pattern under consideration. Thus, collision type was incorporated as a vehicle-level variable instead of a crash-level variable. The methodology was considered the potential heterogeneity (across drivers) in the dependency structure. Wang et al. (2015) employed a copula-based approach to simultaneously model injury severity and vehicle damage by accommodating their dependences between the two outcomes due to common observed and unobserved factors. The Gaussian copula-based model was found to be the best model with lowest BIC value. Ayuso et al. (2016) investigated the the bi-dimensional nature of personal injuries in order to gather more insight into the interaction between of temporary disability and permanent motor injuries. The authors proposed a bivariate copula-based regression model for the joint analysis of their relationship with a set of factors that might influence the two categories of injury.

A mixed copula-based copula regression model was considered in this study namely copula-based generalized linear models (GLMs). This approach follows the methodology proposed by Krämer et al. (2013) which employs bivariate copula models to describe the joint distribution of a pair of continuous and discrete random response variables. The marginals are defined via generalized linear models for the two marginal regression models which are combined with various parametric copulas. In this study, four uni-parametric copulas are considered namely Gaussian, Clayton, Frank and Gumbel copulas. These bivariate copulas are fitted to the bivariate joint distribution with the method of maximum likelihood estimation.

5.3.2 Model Formulation

Despite the lack of uniqueness in the case of discrete or mixed pair of variables, copulas are still suitable for characterizing the dependence between the variables (Genest and Neslehova, 2007; Krämer et al., 2013). Given a continuous random variable A

and a discrete random variable B, the joint distribution is given by

$$F_{A,B|\theta}(a,b|\theta) = P(A \leq a, B \leq b) = C(F_A(a), F_B(b)|\theta) \tag{5.9}$$

where $C(...|\theta)$ is a parametric copula dependent on a parameter θ.

The joint density/probability mass function of the two variables can be expressed as:

$$\begin{aligned} f_{A,B}(a,b) &= \frac{\partial}{\partial a} P(A \leq a, B = b) && (5.10)\\ f_{A,B}(a,b) &= \frac{\partial}{\partial a} P(A \leq a, B \leq b) - \frac{\partial}{\partial a} P(A \leq a, B \leq b-1) && (5.11)\\ f_{A,B}(a,b) &= f_a(a)\left[D_1(F_A(a), F_B(b)) - D_1(F_A(a), F_B(b-1))\right] && (5.12) \end{aligned}$$

where $D_1(u,v) = \frac{\partial}{\partial u} C(u,v|\theta)$ for $u, v \in [0,1]$.

A joint derailment severity regression model is constructed by linking a marginal Gamma GLM for the monetary damage and a marginal zero-truncated Poisson (ZTP) GLM model for the number of derailed cars with a bivariate copula. The monetary damage A is modeled via a Gamma distribution

$$f_A(a|\mu,\delta) = \frac{1}{a\Gamma(\frac{1}{\delta})}\left(\frac{a}{\mu\delta}\right)^{\frac{1}{\delta}} \exp\left(-\frac{a}{\mu\delta}\right) \tag{5.13}$$

where $a > 0$, the mean parameter $\mu > 0$ and the dispersion parameter $\delta > 0$. The variance can be expressed as $\mu^2\delta$. The number of derailed cars is a positive count variable and is modeled as a zero-truncated Poisson (ZTP) distributed variable

$$f_b(b|\lambda) = \frac{\lambda^b}{b!(1 - exp(-\lambda))} \exp(-\lambda) \tag{5.14}$$

where $y = 1, 2, ...$ and parameter $\lambda > 0$.

Assuming $A_i \in \mathbb{R}_+, i = 1, 2, .., n$ are independent continuous random variables

modeled given a covariate vector $\boldsymbol{r_i} \in \mathbb{R}^p$, the marginal regression model can be expressed as $A_i \sim Gamma(\mu_i, \delta)$ with $\ln(\mu_i) = \boldsymbol{r_i}^\top \boldsymbol{\alpha}$. Similarly, assuming $B_i \in \mathbb{N}_{\geq 0}$ are independent discrete random variables modeled given a covariate vector $\boldsymbol{s_i} \in \mathbb{R}^q$, the marginal regression model can be expressed as $B_i \sim ZTP(\lambda_i)$ with $\ln(\mu_i) = \ln(e_i) + \boldsymbol{s_i}^\top \boldsymbol{\beta}$ where e_i is the exposure (time).

Given the unknown parameter vector, $\boldsymbol{\nu} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \theta, \delta)^\top \in \mathbb{R}^{p+q+2}$, the loglikelihood of the copula regression model parameters based on n observation pairs (a_i, b_i) can be expressed as

$$\ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} \ln\left(f_{A,B}(a_i, b_i|\boldsymbol{\nu})\right) \tag{5.15}$$

where $\boldsymbol{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, ..., b_n)^\top \in \mathbb{R}^n$. The maximum likelihood estimates is obtained by maximizing the loglikelihood

$$\hat{\boldsymbol{\nu}} = \arg\max_{v} \ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b}) \tag{5.16}$$

The loglikelihood needs to be maximized numerically since there is no closed-form solution. This is achieved using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm, an iterative method for solving unconstrained nonlinear optimization problems belonging to the quasi-Newton methods. A transformation of the copula parameter $\theta \in \Theta$ is required since it is generally restricted in its range. This is achieved by transforming the parameter θ using a function $h : \Theta \to \mathbb{R}$ such that $h(\theta)$ is unrestricted. The loglikelihood is subsequently optimized in relation to the vector $(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, h(\theta), \delta)^\top$. The parameters for the copula-based regression model is shown in table 5.1.

Asymptotic confidence intervals are used to quantify the uncertainty of the maximum likelihood estimates (MLE). To construct approximate confidence intervals, the Fisher information matrix is employed which can be defined as

$$\boldsymbol{I}(\nu) = E\left[\frac{\partial \ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{\nu}} \cdot \left(\frac{\partial \ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{\nu}}\right)^\top\right] \in \mathbb{R}^{(p+q+2)\times(p+q+2)} \tag{5.17}$$

Table 5.1: Parameters for the copula-based regression model

	Monetary Damage, A	**Number of derailed cars, B**	**Copula family**
Distribution	Gamma	Zero-truncated Poisson	Gaussian, Frank, Clayton, Gumbel
Parameters	$\mu > 0, \delta > 0$	$\lambda > 0$	$\theta \in \Theta$
Mean	$E(A) = \mu$	$E(B) = \frac{\lambda}{1-e^{-\lambda}}$	-
Variance	$Var(A) = \mu^2\delta$	$Var(B) = \frac{\lambda(1-e^{-\lambda}(\lambda+1))}{(1-e^{-\lambda})^2}$	-

Under regularity conditions, the maximum likelihood estimator $\hat{\nu}$ is consistent and asymptotically efficient with limiting distribution

$$\sqrt{N}(\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}) \xrightarrow{D} \mathcal{N}_{p+q+2}[0, \boldsymbol{I}^{-1}(\boldsymbol{\nu})] \tag{5.18}$$

where $\mathcal{N}_d$ is a d-dimensional multivariate normal distribution. The Fisher information matrix can also be expressed as follows:

$$\boldsymbol{I}(\boldsymbol{\nu}) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b})}{\partial^2 \boldsymbol{\nu}}\right] \tag{5.19}$$

To estimate the Fisher information, the observed Fisher information matrix is employed. Up to second-order terms, the observed Fisher information matrix has been found to be the best estimator of the expected Fisher information which is the variance of the MLE (Lehmann and Casella, 1998). It can be expressed as the Hessian matrix of the loglikelihood function as follows:

$$\hat{\boldsymbol{I}}(\boldsymbol{\nu}) = -\frac{\partial^2 \ell(\boldsymbol{\nu}|\boldsymbol{a}, \boldsymbol{b})}{\partial^2 \boldsymbol{\nu}} \tag{5.20}$$

To estimate the standard deviations/errors for the regression coefficients, the BFGS optimization algorithm approximation of the Hessian matrix is employed which is achieved by means of numerical derivatives (by computing the second partial derivatives explicitly) (Krämer et al., 2013).

The best copula regression model is selected by comparing pairs of copula families using the likelihood-ratio test for non-nested hypotheses by Vuong (1989). The Vuong test is a likelihood-ratio based test for comparing non-nested models. The regression models are non-nested since one copula-based regression model for one cannot be determined via a restriction of another copula-based regression model and both models have the same degrees of freedom (number of parameters) (Krämer et al., 2013).

Given that $\boldsymbol{\ell}^{(j)}, \boldsymbol{\ell}^{(k)} \in \mathbb{R}^n$ are the vectors of pointwise loglikelihoods of the models with copula family j and k respectively, the differences of the pointwise loglikelihood can be computed as follows:

$$d_i = \ell_i^{(j)} - \ell_i^{(k)}, \quad i = 1, .., n. \tag{5.21}$$

The Vuong test statistic can be mathematically defined as

$$T_V = \frac{\sqrt{n} \cdot \bar{d}}{\sqrt{\sum_{i=1}^{n} (d_i - \bar{d})^2}} \tag{5.22}$$

where $\bar{d}$ is the mean of the differences. The test statistic is asymptotically normally distributed with zero mean and unit variance. Copula family j is preferred to copula family k at significance level α if $T_V > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ where Φ is the standard normal distribution function. However, if $T_V < \Phi^{-1}\left(\frac{\alpha}{2}\right)$ copula family k is preferred. Else, no decision between the pair of copula families is made.

5.4 Case Study (Bivariate Derailment Severity)

5.4.1 Introduction

Safety is crucial and of paramount importance for every rail system worldwide (Liu, 2015). Derailments is a convoluted issue which is dependent on various factors including track and equipment parameters, running and environmental convictions, signal systems and human error (Mohammadzadeh et al., 2011). Derailments are the most

frequent kind of Federal Railroad Administration (FRA)-reportable mainline train accident in the United States (Barkan et al., 2003; Liu et al., 2012; Liu, 2015) and constituted about three-quarters of freight-train accidents from 2001 to 2010 (Liu et al., 2013). Furthermore, derailments are one of the catastrophic accidents in railway operations and the consequences of derailments are dire despite their relative low frequency (Zhao et al., 2006; Jeong et al., 2007; Liu et al., 2013). Some of these ramifications include injury, loss of life and property, interruption of services and destruction of the environment. Evaluating the degree and variability of derailment severity is as essential as estimating the probability of derailment (Liu et al., 2013). Thus, it is imperative to carefully examine train derailment severity in order to minimize and mitigate these consequences.

The severity of train derailments are usually evaluated by metrics such as the number of derailed cars, monetary damage or casualties. Most research have focused on analyzing the severity of a single outcome mostly the number of derailed cars. However, it is important to examine the multivariate nature of derailment severity by taking into consideration the dependences between multiple consequences. In this chapter, two of the most common evaluation metrics are examined namely the number of derailed cars and monetary damage in order to comprehend the interrelationship between the two outcomes. Monetary damage is the monetary value of destruction caused to infrastructure and rolling stock during a derailment. A joint mixed copula-based model is presented for the analysis of the relationship with the set of factors that might influence both outcomes. A simple flowchart of the process is shown in figure 5.1.

A bivariate copula which characterizes the dependence between the two response variables is used to link their marginal generalized linear regression models. The copula also takes into account endogeneity due to similar omitted or unobserved variables that might affect both outcomes. There is the presence of correlations across error terms of different marginal regressions models due to common unobserved factors influencing the response variable (Rana et al., 2010). Covariates may influence multiple response

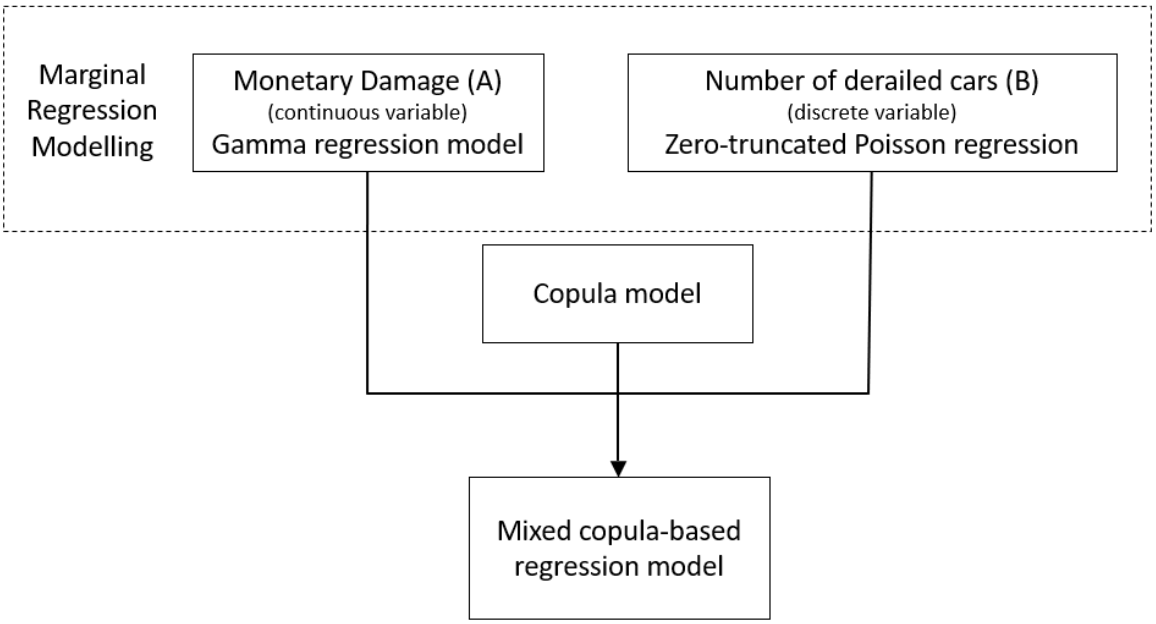


Figure 5.1: Mixed copula-based regression model.

variables, for instance rail friction and ground friction may influence the derailment severity with respect to the number of derailed cars, monetary damage and casualties.

5.4.2 Data

Data was obtained from the Rail Equipment Accident/Incident (REA) database maintained by the Federal Railroad Administration (FRA) of U.S. Department of Transportation (U.S. DOT). A "rail equipment accident/incident" is a collision, derailment, fire, explosion, act of God, or other event involving the operation of railroad on-track equipment (standing or moving). U.S. railroads are mandated to present detailed reports (Form 6180.54) to the FRA on all accidents or incidents whose damage costs exceed a specified monetary value. The damage incurred includes damage caused to the railroad track, signals, on-track equipment, track structures and roadbed as well as labor costs and the costs for acquiring new equipment and material. The reporting threshold is periodically altered to account for inflation and other adjustments and

has increased from \$5700 in 1990 to \$10,700 in 2017 (FRA, 2016). The relatively low threshold results in most accidents being reported to the FRA (Barkan et al., 2003).

The database contains detailed track accident information such as accident cause, number of derailed cars, total monetary damage, track type, track class, train length and derailment speed. 690 freight-train derailments occurring on Class I mainline track in the year 2005 were initially considered. The response variables considered were monetary damage and the number of derailed cars. The explanatory variables initially considered include derailment speed, residual train length, train power distribution and proportion of loaded railcars in the train (loading factor).

To cater for the effect (and variations) due to derailment cause, 124 derailments caused by broken rail were considered. Broken rails are the most frequent cause of freight-train derailments on Class I mainlines in the United States. Broken rails also result in a higher derailment severity in comparison with other causes such as bearing failure with the former causing twice as many derailed cars on average as that of the latter (Barkan et al., 2003). Due to their high frequency and severity, broken rails are more likely to present higher risk than other causes. All the derailments involving broken rails during the period were found to be non-distributed-power trains. Thus, train power distribution was subsequently removed from the analysis.

5.4.3 Analysis and Results

Table 5.2 presents the descriptive statistics of the variables of interest for broken-rail caused freight-train derailments. The average number of derailed cars involved was 10.64 with only 3 derailments resulting in only 1 derailed car. The average total monetary damage incurred during derailments was \$370,139.30. On average, a train derailed due to a broken rail was found to have a speed of 21.9 mph at the time of derailment, a loading factor of 0.74 and a residual train length of 47.03. The empirical distribution function for the monetary damage and the number of derailed cars are shown in figure 5.2. The Spearman's rho correlation coefficient between the variables of interest is shown in table 5.3. Speed, residual train length and loading factor were

found to have positive correlation with both response variables (derailed cars and total monetary damage). The dependence correlation coefficient of the two severity variables as well as the bivariate plot (Figure 5.3) confirmed our prior intuition that the two outcomes may be dependent.

Table 5.2: Descriptive statistics of variables for broken-rail-caused freight-train derailments

Variables	**Mean**	**Standard deviation**	**Minimum**	**Maximum**	**Type**
Monetary Damage (US $)	370139.3	518263.5	7948	2773500	Continuous
Derailed Cars	10.64	7.97	1	43	Count
Residual Train Length	47.03	30.42	2	134	Count
Derailment Speed (mph)	21.9	14.77	5	70	Continuous
Loading factor	0.74	0.32	0	1	Continuous

Table 5.3: Spearman's rho correlation coefficient between variables for broken-rail-caused freight-train derailments

Variables	Derailed Cars (D)	Monetary Damage (M)	Derailment Speed (S)	Residual Train Length (R)	Loading Factor (L)
Derailed Cars (D)	1	0.78	0.63	0.45	0.10
Monetary Damage (M)	0.78	1	0.62	0.44	0.22
Derailment Speed (S)	0.63	0.62	1	0.20	0.03
Residual Train Length (R)	0.45	0.44	0.22	1	-0.02
Loading Factor (L)	0.10	0.22	0.03	-0.02	1

The selection of the copula-based regression model for the bivariate derailment severity was conducted in two stages. In the first stage, the marginal regression models of the severity outcomes (monetary damage and number of derailed cars) were chosen.

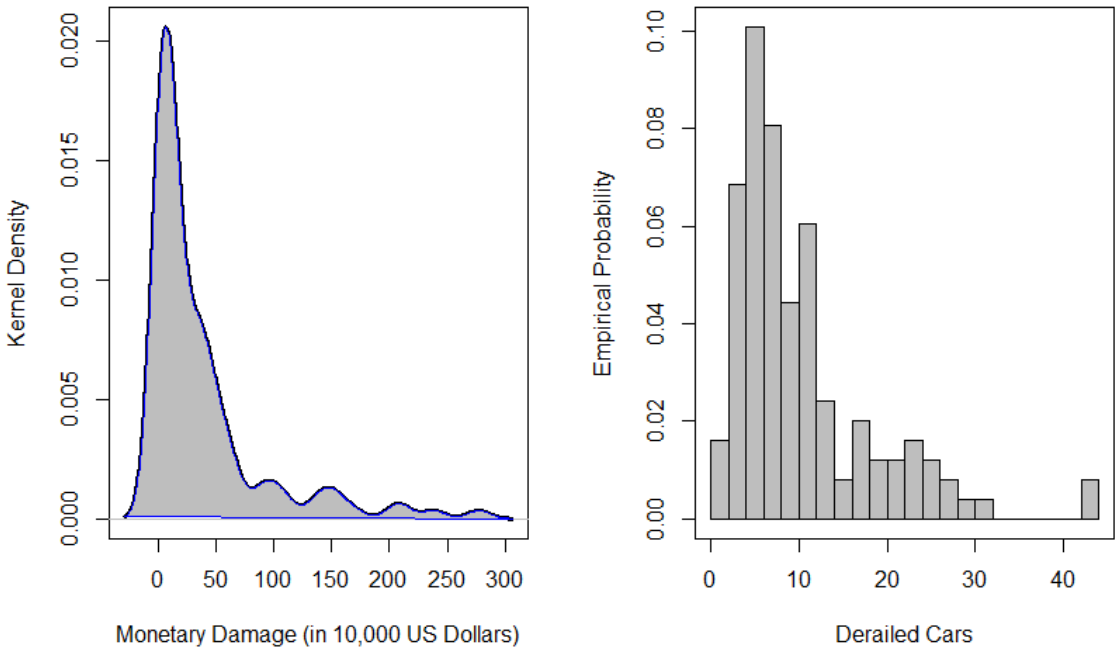


Figure 5.2: Empirical density functions of the monetary damage (left) and number of derailed cars (right).

Under the marginal regression models, the random components, systematic components and link functions are selected. In the second stage, the dependence structure between the response variables characterized by the bivariate copula is considered. The random components of the GLMs were selected based on the nature of the variables. Monetary damage is a continuous non-negative random variable and is skewed in nature. Furthermore, it is a common actuarial assumption that the size of loss or claim is Gamma distributed. Thus, gamma regression has historically been employed in the modelling of loss severity regression models (Burnecki et al., 2005; Gschlößl and Czado, 2007; Krämer et al., 2013). Based on AIC values, the Gamma marginal regression model was found to offer a better fit for monetary damage data compared to the normal or Gaussian marginal regression model. The Gamma model (1016.2) was

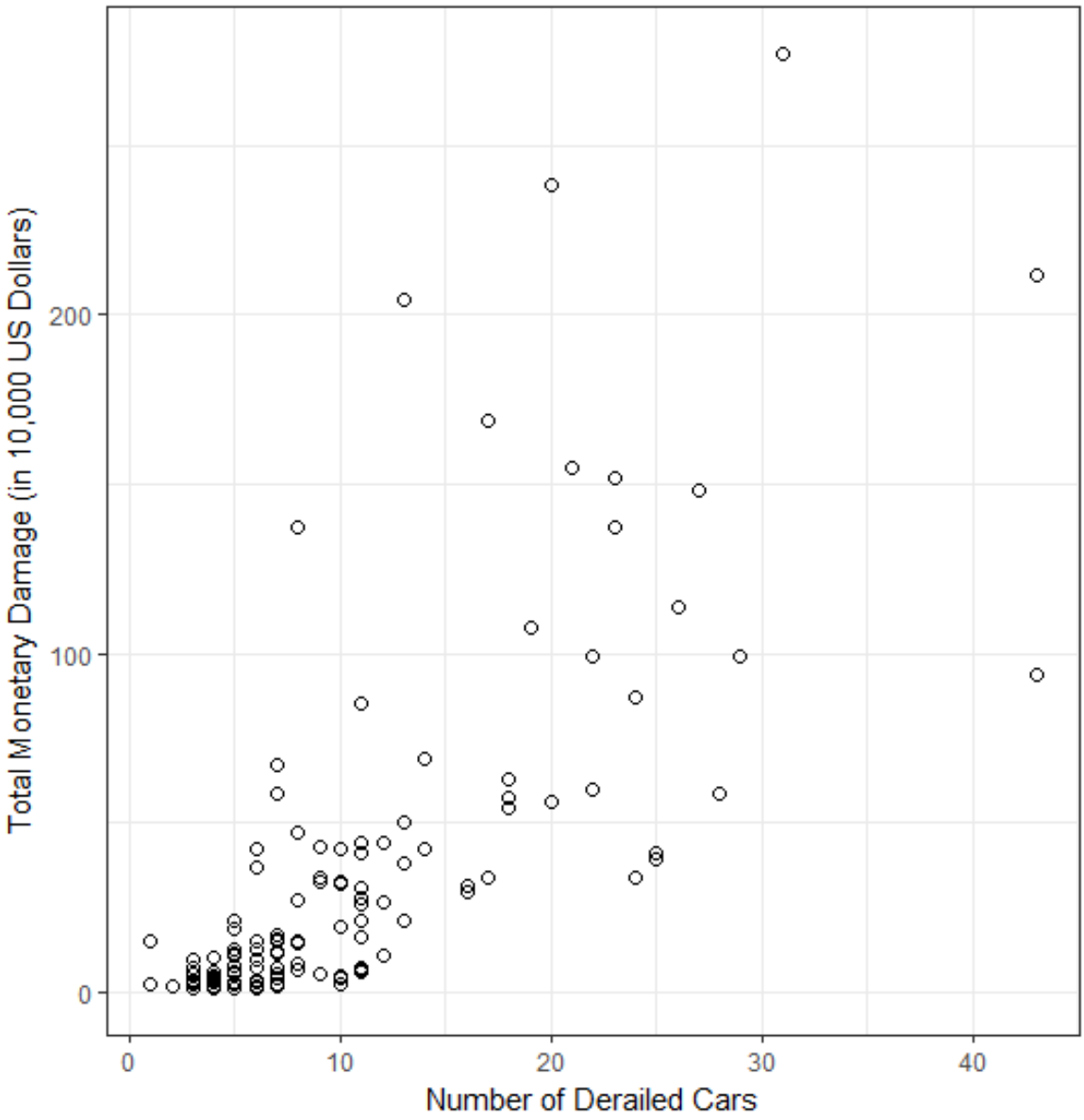


Figure 5.3: Bivariate plot of the number of derailed cars and overall monetary derailment damage.

found to have a lower AIC than the normal model (1268.3).

On the other hand, the number of derailed cars is a discrete positive random variable. The Poisson distribution is suitable for modeling count data and Poisson regression is suitable in predicting a count response variable from a set of covariates. However, zero-truncated Poisson models are more suitable for modeling count data without zeros. Since derailments always result in at least one derailed car, the number of derailed cars is assumed to follow a zero-truncated Poisson distribution. The empirical distribution function for the number of derailed cars is shown in figure 5.2 (right). The zero-truncated Poisson GLM is a GLM using the Poisson distribution conditional on the count variable being greater or equal to one (Czado et al., 2012).

Subsequently the systematic component of the marginal regression models were considered. Derailment severity has been found to increase exponentially with derailment speed and residual train length. Thus, logarithm transformation of these variables have been found to offer a better fit (Saccomanno and El-Hage, 1989, 1991; Liu et al., 2013). This was confirmed and adopted during the analysis. Similar to previous studies, the regression model initially took into consideration the main effect, higher-order component and interaction terms of the covariates. Liu et al. (2013) introduced interaction terms due to interaction between the covariates (derailment speed, residual train length, loading factor) which may be attributable to energy accumulation during derailments (Liu et al., 2013). However, incorporation of these terms were found to significantly increase the risk of overfitting and were subsequently eliminated. In order to achieve parsimonious model and avoid collinearity, variables selection of the multivariate marginal regression models was implemented using the hierarchical model selection technique based on the deviance (Agresti, 2007). Model selection was confirmed using Akaike's Information Criteria (AIC). The detailed model selection process for the multivariate marginal regression model is shown in table 5.4. Some higher-order components were eliminated during this process. The selected multivariate marginal

regression models are as follows:

$$Z_i = \beta_0 + \beta_1 \times S + \beta_2 \times R + \beta_3 \times L + \beta_4 \times R^2 + \beta_5 \times L^2 \tag{5.23}$$

where Z_i is the response variable (number of derailed cars or monetary damage in \$10,000s), S is the logarithmic derailment speed, R is the logarithmic residual train length, L is the loading factor and $\beta_0, ..., \beta_n$ are the regression coefficients to be estimated. The estimated parameters of the multivariate marginal regression model can be found in table 5.6. The link function employed for both the gamma and ZTP regression models was the log link $g(\mu) = \log(\mu)$ which models the log of the mean.

Table 5.4: Model selection of multivariate marginal regression model

Model	Predictors	Log likelihood	Number of parameters	Model comparison	Deviance Difference	AIC
1	S, R, L	-863.6	9	-	-	1745.2
2	S, R, L, S^2, R^2, L^2	-853.9	15	(1)-(2)	19.4 (df=6)	1737.9
3a	S, R, L, S^2, R^2	-860.3	13	(3a)-(2)	12.8 (df=2)	1746.7
3b	S, R, L, S^2, L^2	-856.1	13	(3b)-(2)	4.3 (df=2)	1738.1
3c	S, R, L, R^2, L^2	-854.6	13	(3c)-(2)	1.3 (df=2)	1735.2
4a	S, R, L, R^2	-861.2	11	(4a)-(3c)	13.2 (df=2)	1744.3
4b	S, R, L, L^2	-856.8	11	(4b)-(3c)	4.5 (df=2)	1735.7

Subsequently, the bivariate copula characterizing the dependency between the two outcomes variables was considered. Prior to selection of the bivariate copula, the Genest and Favre bivariate asymptotic independence test based on Kendall's Tau was performed to determine the independence of the pair of variables (Genest and Favre, 2007). The null hypothesis states that the variables are independent and the alternative hypothesis states that the variables are not independent. The independence copula is selected for the pair of variables if the p-value of the test is higher than 5% meaning the null hypothesis is accepted. Otherwise, the null hypothesis is rejected. The test resulted in p-value of 0 meaning the variables are dependent. The copulas considered for selection include Gaussian, Frank, Gumbel and Clayton copulas.

Given the selected marginal regression models, copula-based regression models were fitted for each bivariate copula. The Vuong test was conducted for each pair of copula-based regression models. Results of the pairwise Vuong test are shown in table 5.5. For each pair of copula families, a copula family is selected given a significance level of 5%. Thus, model 1 is chosen if the Vuong test statistic (value) is greater than 2 but model 2 is chosen if the value is less than -2. Otherwise, no decision among the pair of copula families is made. It was not possible to choose between Frank and Gaussian copulas thus Akaike Information Criterion (AIC) (Akaike, 1974) was used. The Gaussian copula model (1714) was found to have a lower AIC than the Frank copula model (1715) and was thus selected as the best fit. The final copula selection is a Gaussian copula regression model with gamma and zero-truncated Poisson marginals. The Gaussian copula indicates that the dependence between the outcomes (given the covariates) is radially symmetric with weak tail dependencies.

The Gaussian copula-based regression model was subsequently compared with an independent model which assumed no dependence between the two response variables. Results including parameter estimates, loglikelihood and AIC of two models can be found on table 5.6. The Gaussian copula-based regression model was found to have a lower AIC than the independent multivariate regression model (1735). The log likelihood ratio statistic (difference in deviance) between the independent and the

copula-based regression model is 23.18, which is greater than the critical chi-square value for a degree of freedom of 1 (as a result of the additional copula parameter) at any level of significance. Thus, this demonstrates the superior statistical fit of the copula regression model over the independent model and indicates that multivariate derailment severity analysis should be modeled in a joint manner. The copula parameter estimate $\hat{\theta} = 0.32$ corresponding to a theoretical Kendall's tau of 0.21. Although caution is needed during the interpretation of degree of dependence in the presence of discrete data (Genest and Neslehova, 2007), the log likelihood ratio statistic as well as the AIC and Kendall's tau of the copula regression model seems to confirm a positive dependence between the severities of total monetary derailment damage and the number of derailed cars.

The positive correlation also indicates that the unobserved factors that increase the derailment severity in terms of the number of derailed cars are positively correlated with the factors that contribute to higher overall monetary damage incurred during a derailment. Such correlations may arise due to the presence of several common unobserved but influential factors that may both severity outcomes during a derailment. These factors are usually not recorded during a derailment (such as ground and rail friction) or are not directly incorporated into the regression equation (such as derailment cause). Similar results have been observed in driver injury severity analysis during crashes (Rana et al., 2010). Failure to account for the dependence between the outcomes may lead to biased or distorted coefficient estimates in multivariate derailment severity models. This can be observed by comparing the estimates of the independence and copula-based regression models in table 5.6.

Greater differences were observed when comparing the dispersion estimates of the two models in comparison with their estimated coefficient values of the regressors. The incorporation of the dependence structure between the response variables has been found to have a greater influence on the variance estimates of the severity outcomes than the point estimates. In the case of Ayuso et al. (2016) who analyzed the bivariate severity of temporary disability and permanent motor injuries, the independence

assumption was found to result in the underestimation of the variance estimates of the severity outcomes. In this case, independence assumption led to an overestimation of the variance estimates of the covariates of the monetary damage severity outcome. The use of negative binomial distribution which incorporates a dispersion parameter is expected to have an impact on the variance estimates of the number of derailed cars. Thus, an assumption of independence introduces the risk of underestimating or overestimating the variance of the severity outcomes. For the aforementioned reasons provided, it seems inappropriate to assume independence between the two derailment severity outcomes.

Table 5.5: Values of Vuong test for each pair of copula-based regression models given monetary damage (gamma marginal model) and number of derailed cars (zero-truncated Poisson marginal model)

Model	**Model 2**				
	Copulas	Gaussian	Frank	Clayton	Gumbel
Model 1	Gaussian	-	0.0175	2.5625	2.5626
	Frank	-0.0175	-	2.4120	2.4138
	Clayton	-2.5625	-2.4120	-	1.2984
	Gumbel	-2.5626	-2.4138	-1.2984	-

This model can be used to simultaneously predict the expected monetary damage and expected number of derailed cars during a broken-rail caused freight train derailment. Due to the well-known variation in derailment severity due to accident cause, this model cannot be applied to other causes. However, the methodology can be applied to other cases. Sensitivity analysis was conducted to estimate the effect and relative importance of each covariate on the monetary damage and number of derailed cars. Changes in the model outcomes given ±10% variation in derailment speed, residual train length and loading factor were analyzed. Sensitivity analysis of the effect of these covariates on the expected monetary damage and number of derailed cars illustrated in the form of tornado diagrams are shown on figures 5.4 and 5.5 respectively.

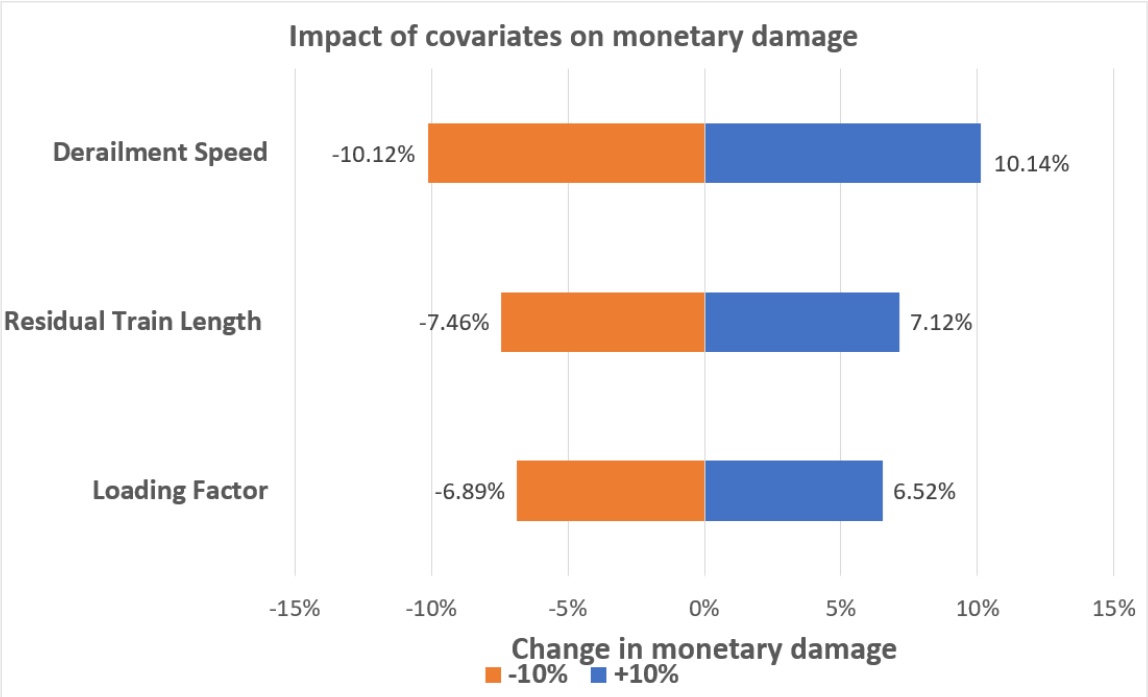


Figure 5.4: Tornado diagram showing the effect of various parameters on monetary damage.

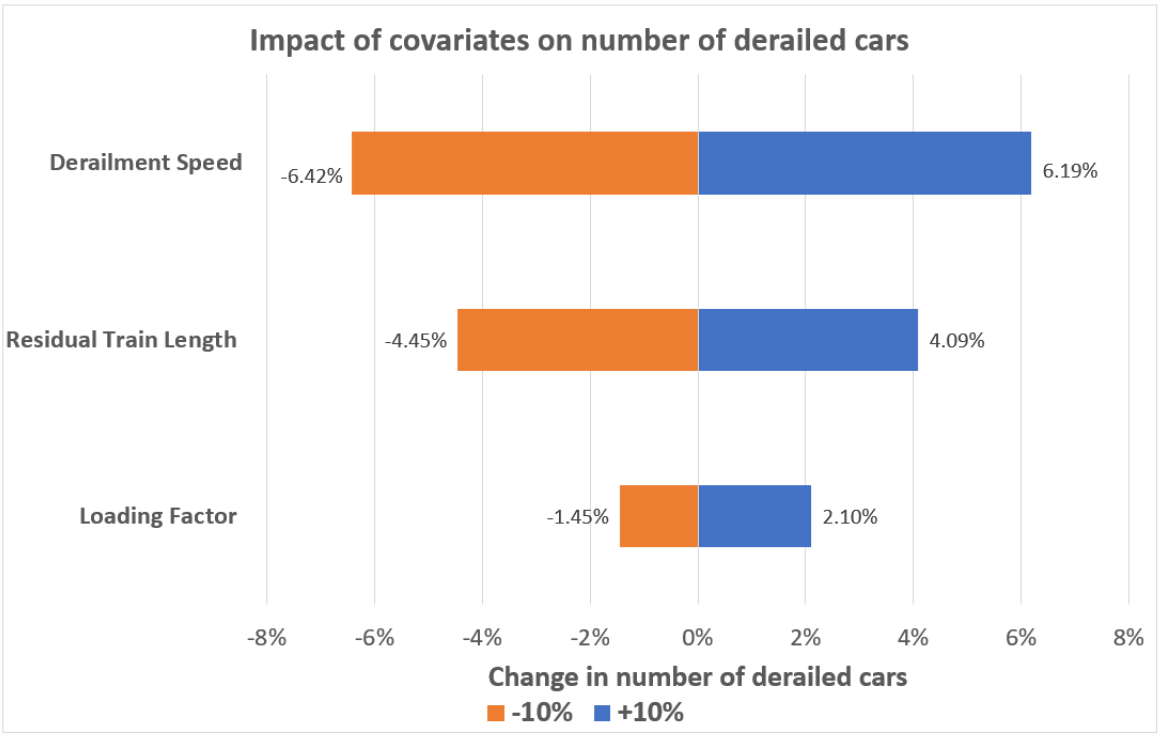


Figure 5.5: Tornado diagram showing the effect of various parameters on number of derailed cars.

Derailment speed was found to have the greatest effect on both the expected number of derailed cars and monetary damage. However, it was found to have a greater effect on the monetary damage outcome than the number of derailed cars. Residual train length was found to have the second most pronounced effect on both outcomes followed by loading factor with the least effect. Similar to derailment speed, both covariates were found to have a greater effect on the monetary damage outcome than the number of derailed cars. These results enable objective comparison of different train safety approaches that could be used to inform decision making by government and industry. For instance, one can argue for the reduction of freight train speeds in favor of a reduction in the number of cars in a train consist.

5.4.4 Concluding Remarks

The analysis of the severity of train derailments is critical due to their catastrophic nature. Most of the existing models have failed to consider the multivariate nature of derailment severity but have instead focused mainly on only one severity outcome namely the number of derailed cars. A mixed copula-based regression model of the number of derailed cars and monetary damage is proposed to jointly analyze their relationship with a set of explanatory variables which might influence both outcomes taking into consideration the dependence between the two responses. The joint model incorporates the discrete nature of the number of derailed cars and continuous nature of the monetary damage. The copula which describes the dependence between the response variables also takes into account endogeneity due to common unobserved or omitted factors. Results show that the Gaussian copula-based regression model is more appropriate than an independent multivariate regression model. Failure to account for the dependence between the outcomes may lead to biased coefficient estimates in multivariate derailment severity models. The log likelihood ratio statistic between the independent and the copula-based regression model was found to be greater than the critical chi-square value for a degree of freedom of 1 (as a result of the additional copula parameter) at any level of significance. This demonstrated the superior statistical fit

of the copula regression model over the independent model and indicates that multivariate derailment severity analysis should be modeled in a joint manner. Derailment speed was found to have the most pronounced effect on both the monetary damage and number of derailed cars. However, it was found to have a greater impact on monetary damage than the number of derailed cars.

The methodology can be extended to cater for bi-parametric copulas such as student t-copula, Clayton-Gumbel (BB1), Joe-Gumbel (BB6), Joe-Clayton (BB7) and Joe-Frank (BB8) copulas as well as marginal distributions to consider other distributions (such as lognormal distribution for monetary damage and zero-truncated negative binomial for the number of derailed cars).Furthermore, other derailment outcomes or severities such as casualties can be incorporated into a copula-based regression model. In order to analyze the dependences between more than two response variables, vine copulas are suggested which are more flexible than regular multivariate copulas. Vine copulas employ arbitrary bivariate copulas as building blocks for the construction of higher-dimensional multivariate distributions. Combining the marginal regression models of the two derailment severity outcomes with the underlying dependence facilitates a better comprehension of the train derailment severity distribution. The multidimensional approach to analyzing derailment severity consequences provides interesting challenges that are worth investigating and is intended to offer insights regarding the development and implementation of cost-effective safety improvement strategies.

Table 5.6: Parameter estimates of the Gaussian copula based regression model for monetary damage (Gamma regression marginal model) and number of derailed cars (Poisson regression marginal model) compared with the independence assumption.

Label	**Independence Model**				**Gaussian copula regression model**			
	Monetary Damage		Derailed Cars		Monetary Damage		Derailed Cars	
	α	SD	β	SD	α	SD	β	SD
Constant	-3.80	0.89***	-1.91	0.44***	-3.88	0.71***	-1.77	0.43***
Derailment Speed (S)	1.01	0.12***	0.65	0.04***	1.01	0.11***	0.63	0.04***
Residual Train Length (R)	1.08	0.50*	0.78	0.24**	1.1	0.43*	0.73	0.23**
Loading Factor (L)	2.34	1.16*	1.63	0.44***	2.41	0.95*	1.51	0.43***
R^2	-0.07	0.08	-0.07	0.03*	-0.07	0.07	-0.06	0.03*
L^2	-0.92	0.97	-1.25	0.35***	-1.02	0.81*	-1.15	0.35***
Dispersion Parameter	0.78		-		0.61		-	
Copula parameter	-				0.32			
Kendall's Tau	-				0.21			
Log-likelihood	-854.59				-843.00			
AIC	1735				1714			

REFERENCES

Agresti, Alan. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2007. ISBN 9780471226185.

Akaike, Hirotugu. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 15582523. doi: 10.1109/TAC.1974.1100705.

Ayuso, Mercedes; Bermúdez, Lluís, and Santolino, Miguel. Copula-based regression modeling of bivariate severity of temporary disability and permanent motor injuries. *Accident Analysis and Prevention*, 89:142–150, 2016. ISSN 00014575. doi: 10.1016/j.aap.2016.01.008.

Barkan, Christopher P. L.; Dick, C. Tyler, and Anderson, Robert. T. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 1825(9):64–74, 2003. ISSN 03611981.

Bhat, Chandra R. and Eluru, Naveen. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7):749–765, 2009. ISSN 01912615. doi: 10.1016/j.trb.2009.02.001.

Burnecki, Krzysztof; Misiorek, Adam, and Weron, Rafał. Loss Distributions. In *Statistical Tools for Finance and Insurance*, pages 289–317. Springer-Verlag, Berlin/Heidelberg, 2005. doi: 10.1007/3-540-27395-6_13.

Casella, George. and Berger, Roger L. *Statistical inference*. Duxbury Press/Thomson Learning, Pacific Grove, CA, 2nd edition, 2002. ISBN 0534243126.

Czado, Claudia; Kastenmeier, Rainer; Brechmann, Eike Christian, and Min, Aleksey. A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305, 2012. ISSN 0346-1238. doi: 10.1080/03461238.2010.546147.

Eluru, N; Paleti, R; Pendyala, R M, and Bhat, C R. Modeling Injury Severity of Multiple Occupants of Vehicles Copula-Based Multivariate Approach. *Transportation Research Record*, (2165):1–11, 2010. ISSN 0361-1981. doi: Doi10.3141/2165-01.

FRA, . Monetary Threshold for Reporting Rail Equipment Accidents/Incidents for Calendar Year 2017. *Federal Registry*, 81(247):57–60, 2016.

Genest, Christian and Favre, Anne-Catherine. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12 (4):347–368, 2007. ISSN 1084-0699. doi: 10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, Christian and Neslehova, Johanna. A Primer on Copulas for Count Data. *ASTIN Bulletin*, 37(2):475–515, 2007. ISSN 0515-0361. doi: 10.2143/AST.37.2.2024077.

Gschlößl, Susanne and Czado, Claudia. Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225, 2007. ISSN 16512030. doi: 10.1080/03461230701414764.

Hastie, Trevor. and Tibshirani, Robert. *Generalized additive models*. Chapman and Hall, 1990. ISBN 9780412343902.

Jeong, D.Y.; Lyons, M.L.; Orringer, O, and Perlman, A.B. Equations of motion for train derailment dynamics. *Proceedings of the 2007 ASME Rail Transportation Division Fall Technical Conference, September 11-12, 2007 Chicago, IL*, RTDF2007-4: 1–7, 2007. ISSN 10788883. doi: 10.1115/RTDF2007-46009.

Krämer, Nicole; Brechmann, Eike C.; Silvestrini, Daniel, and Czado, Claudia. Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3):829–839, 2013. ISSN 01676687. doi: 10.1016/j.insmatheco.2013.09.003.

Kreft, Ita G. G. and de. Leeuw, Jan. *Introducing multilevel modeling*. Sage, 1998. ISBN 0761951415.

Lehmann, E L and Casella, G. *Theory of Point Estimation*. Number 3. Springer Texts in Statistics, New York, 2nd edition, 1998. ISBN 0387985026. doi: 10.2307/1270597.

Liu, Xiang. Statistical Temporal Analysis of Freight-Train Derailment Rates in the United States : 2000 to 2012. 2476(1):119–125, 2015.

Liu, Xiang; Saat, M., and Barkan, Christopher. Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*, 2289(2289):154–163, 2012. ISSN 0361-1981. doi: 10.3141/2289-20.

Liu, Xiang; Saat, M. Rapik; Qin, Xiao, and Barkan, Christopher P L. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. *Accident Analysis and Prevention*, 59:87–93, 2013. ISSN 00014575. doi: 10.1016/j.aap.2013.04.039.

Mohammadzadeh, Saeed; Sangtarashha, Manie, and Molatefi, Habibollah. A novel method to estimate derailment probability due to track geometric irregularities using reliability techniques and advanced simulation methods. *Archive of Applied Mechanics*, 81(11):1621–1637, 2011. ISSN 09391533. doi: 10.1007/s00419-011-0506-3.

Nelder, J A and Wedderburn, R W M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) Journal of the Royal Statistical Society. Series A (General J. R. Statist. Soc. A*, 13517213(3):370–384, 1972. ISSN <null>. doi: 10.2307/2344614.

Quinn, Gerald Peter and Keough, Michael J. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002. ISBN 9780521009768.

Rana, Tejsingh; Sikder, Sujan, and Pinjari, Abdul. Copula-Based Method for Addressing Endogeneity in Models of Severity of Traffic Crash Injuries. *Transportation Research Record: Journal of the Transportation Research Board*, 2147(2147):75–87, 2010. ISSN 0361-1981. doi: 10.3141/2147-10.

Rigby, R. A.; Stasinopoulos, D. M., and Lane, Peter W. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(3):507–554, 2005. ISSN 00359254. doi: 10.1111/j.1467-9876.2005.00510.x.

Saccomanno, F. F. and El-Hage, S. M. Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. *Transportation Research Record*, (1245):34–51, 1989.

Saccomanno, F. F. and El-Hage, S. M. Establishing derailment profiles by position for corridor shipments of dangerous goods. *Canadian Journal of Civil Engineering*, 18 (1):67–75, 1991. ISSN 0315-1468. doi: 10.1139/l91-009.

Stasinopoulos, D Mikis and Rigby, Robert A. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of statistical software*, 23(7):1–46, 2007.

Vuong, Quang H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307, 1989. ISSN 00129682. doi: 10.2307/1912557.

Wang, Kai; Yasmin, Shamsunnahar; Konduri, Karthik C.; Eluru, Naveen, and Ivan, John N. Copula-Based Joint Model of Injury Severity and Vehicle Damage in Two-Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2514:158–166, 2015. ISSN 0361-1981. doi: 10.3141/2514-17.

Wood, Simon N. Generalized Additive Models: An Introduction with R. *Journal of the Royal Statistical Society A*, 170(1):388, 2006. ISSN 09641998. doi: 10.1111/j.1467-985X.2006.00455_15.x.

Yasmin, Shamsunnahar; Eluru, Naveen; Pinjari, Abdul R., and Tay, Richard. Examining driver injury severity in two vehicle crashes - A copula based approach. *Accident Analysis and Prevention*, 66:120–135, 2014. ISSN 00014575. doi: 10.1016/j.aap.2014.01.018.

Yee, Thomas W. VGAM Family Functions for Generalized Linear and Additive Models. pages 1–22, 2008.

Zhao, Jianmin; Chan, Andrew H C, and Stirling, Alan B. Risk analysis of derailment induced by rail breaks-a probabilistic approach. *Annual Reliability and Maintainability Symposium*, 00(C):486–491, 2006.

Chapter 6

VINE COPULA MODELS

This chapter provides a detailed overview of vine copula models. The relative inflexibility problem of classical multivariate extensions of bivariate elliptical and archimedean copulas is considered as the starting point for introducing the concept of vine copulas. Vine copulas are multivariate copulas constructed hierarchically from bivariate copulas as building blocks. The theory of pair-copula construction upon which vine copulas are developed is explained and the graphical representation of vine copulas known as regular vines is also discussed. The various vine structure selection methods, parameter estimation techniques and pair-copula families selection procedures of vine copulas are also reviewed. This chapter concludes with a case study in which high-dimension dependence of derailment severity data is modeled using vine copulas. Vine copula methodology is subsequently applied to simulation modeling of multivariate derailment severity data.

6.1 Multivariate Elliptical Copulas

Majority of the extant literature concentrates on bivariate copulas. In the multivariate case, multivariate elliptical copulas such as multivariate normal and multivariate t-copulas are popular. Multivariate normal distributions comprises of a normal (Gaussian) copula with normal margins. Thus, the n-variate Gaussian copula is the copula of the n-variate normal distribution with linear correlation matrix R can be mathematically expressed as

$$C(u_1, ..., u_n) = \Phi_R^n(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)) \tag{6.1}$$

where Φ_R^n is the joint distribution joint distribution function of the n-variate standard normal distribution function with linear correlation matrix R and Φ^{-1} is the inverse of the univariate standard normal distribution function.

On the other hand, a student-t copula with Student-t margins does not necessarily form a multivariate-t distribution since the distribution must have the same degree of freedom at all the margins. A t-copula with t-margins can have varying degrees of freedom for different margins thus providing a lot more flexibility in modeling multivariate heavy-tailed data (Yan, 2006). The n-variate student-t copula can be expressed as

$$C_(u_1, ..., u_n) = t_{\nu,R}^n(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_n)) \tag{6.2}$$

where $R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$ for $i, j \in 1, ..., n$ and $t_{\nu,R}^n$ is the distribution function of $\frac{\sqrt{\nu}\mathbf{Y}}{\sqrt{S}}$ where $S \sim \chi_\nu^2$ and $\mathbf{Y} \sim \mathcal{N}_n(0, R)$ are independent. Despite their wide usage, multivariate elliptical copulas are relatively inflexible in representing the entire dependence structure of the multivariate data.

6.2 Multivariate Archimedean Copulas

The use of other multidimensional copulas apart from multivariate elliptical copulas are relatively limited as a result of construction complexities, computational and theoretical limitations. Multivariate extensions of Archimedean copulas were subsequently suggested such as partially nested Archimedean copulas, hierarchical Archimedean copulas and hierarchical Archimedean copulas. However, their flexibility are restricted due to the additional restraints on the parameters caused by the extensions (Dalla Valle et al., 2016).

The most popular Archimedean multivariate extension is the exchangeable multivariate Archimedean copula (EAC). EAC can be mathematically expressed as

$$C(u_1, ..., u_n) = \varphi^{-1}\left(\varphi(u_1), ..., \varphi(u_n)\right) \tag{6.3}$$

where φ is generator of the copula (continuous strictly decreasing convex function)

and φ^{-1} is its pseudo-inverse. EACs are however highly restrictive since they permit the specification of only one distribution parameter, regardless of dimension (all d-dimensional marginal distributions ($d < n$) are identical) (Wei, 2014). More flexible multivariate Archimedean extensions using the nested copula approach were subsequently proposed by Joe (1997). Types of nested Archimedean copulas include fully nested Archimedean construction (FNAC), partially nested Archimedean construction (PNAC) and hierarchically nested Archimedean construction (HNAC).

The construction of fully nested Archimedean construction (FNAC) copulas are quite simple but notationally cumbersome. The scheme of FNAC is to add dimensions stepwise. An n-dimensional FNAC permits the specification of $(n-1)$ bivariate copulas and corresponding distributional parameters whereas the $\frac{(d-1)(d-2)}{2}$ copulas and parameters are implicitly given through the construction (Berg, 2009). The 4-dimensional FNAC copula can be expressed as

$$C(u_1, u_2, u_3, u_4) = C_{31}(u_4, C_{21}(u_3, C_{11}(u_1, u_2))) \tag{6.4}$$

$$= \varphi_{31}^{-1}\{\varphi_{31}(u_4) + \varphi_{31}(\varphi_{21}^{-1}\{\varphi_{21}(u_3) + \varphi_{21}(\varphi_{11}^{-1}\{\varphi_{11}(u_1) + \varphi_{11}(u_2)\})\})\} \tag{6.5}$$

Figure 6.1 illustrates the 4-dimensional FNAC copula. Two pairs (u_1, u_3) and (u_2, u_3) both have copula C_{21} with dependence parameter θ_{21} whereas three pairs (u_1, u_4), (u_2, u_4) and (u_3, u_4) all have copula C_{31} with dependence parameter θ_{31}. The FNAC is a construction of partial exchangeability and thus certain conditions need to be fulfilled in order to achieve a proper n-dimensional copula. These conditions (such as generators having to be strict with completely monotone inverses) put restrictions on copula parameters (Berg, 2009).

The partially nested Archimedean construction (PNAC) initially suggested by Joe (1997) is a hybrid of EAC and FNAC. The 4-dimensional PNAC is expressed as follows

$$C(u_1, u_2, u_3, u_4) = C_{21}(C_{11}(u_1, u_2), C_{12}(u_3, u_4)) \tag{6.6}$$

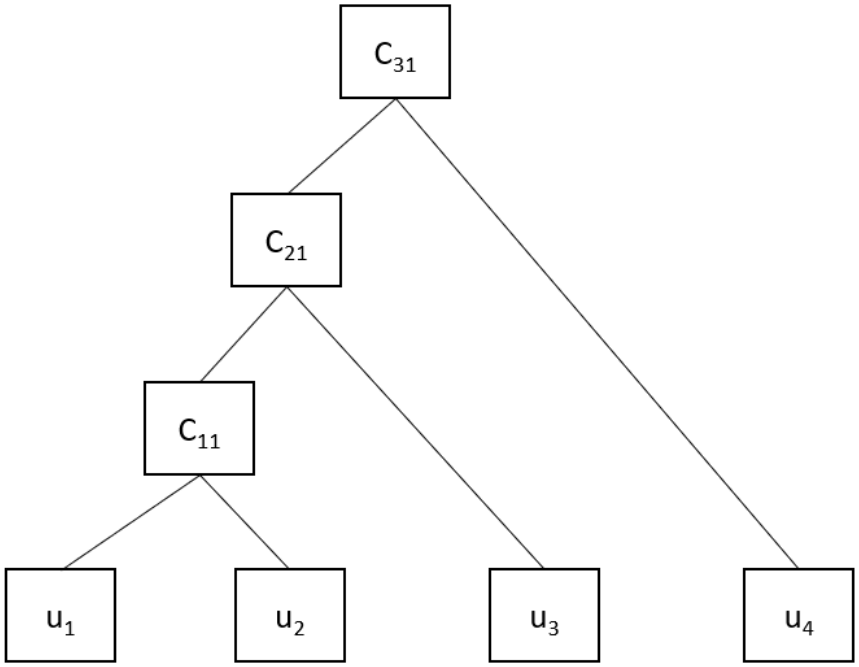


Figure 6.1: Fully nested Archimedean construction.

Figure 6.2 illustrates the 4-dimensional PNAC. Two pairs (u_1, u_2) and (u_3, u_4) are initially coupled with copulas C_{11} and C_{12} with generators φ_{11} and φ_{12} respectively. The two copulas are subsequently coupled with copula C_{21}. The resulting copula is exchangeable between u_1 and u_2 as well as between u_3 and u_4. Thus, PNAC can be viewed a combination of EAC and FNAC.

The hierarchical nested Archimedean construction (HNAC) was initially proposed by Joe (1997) and elaborated upon by Savu and Trede (2010). A hierarchy of Archimedean copulas is constructed based on arbitrary nesting. The copula at a given level in the hierarchy does not have to bivariate. Figure 6.3 illustrates a 10-dimensional HNAC copula. This copula can be given as

$$C(u_1, ..., u_{10}) = C_{21}(C_{11}(u_1, u_2, u_3), C_{12}(u_4, u_5, u_6, u_7), C_{13}(u_8, u_9, u_{10})) \tag{6.7}$$

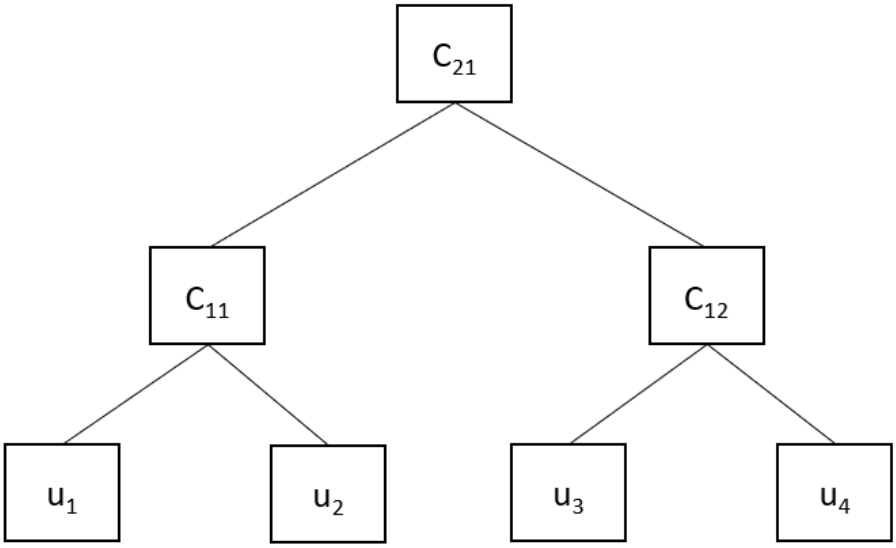


Figure 6.2: Partially nested Archimedean construction.

At the first level, there are three copulas. Copula C_{11} is a 3-dimensional EAC coupling variables u_1, u_2 and u_3. Copula C_{12} is a 4-dimensional EAC coupling variables u_4, u_5, u_6 and u_7 whereas copula C_{13} is a 3-dimensional EAC coupling variables u_8, u_9 and u_10. In the second level of hierarchy, the three copulas from the previous tier are joined by copula C_{21} which is a 3-dimensional EAC. Thus, it is a partially exchangeable copula.

In general, multivariate copulas extended from classical bivariate parametric copulas lack higher dimensional flexibility and cannot accommodate different tail dependencies for different pairs of variables. For this reason, pair copula construction addresses this flexibility limitation. structure (Czado et al., 2012; Schepsmeier and Czado, 2016; Dalla Valle et al., 2016).

6.3 Pair Copula Construction

The decomposition of higher-dimension multivariate copula into bivariate copulas is referred to as pair copula construction (PCC). A pair copula construction (PCC)

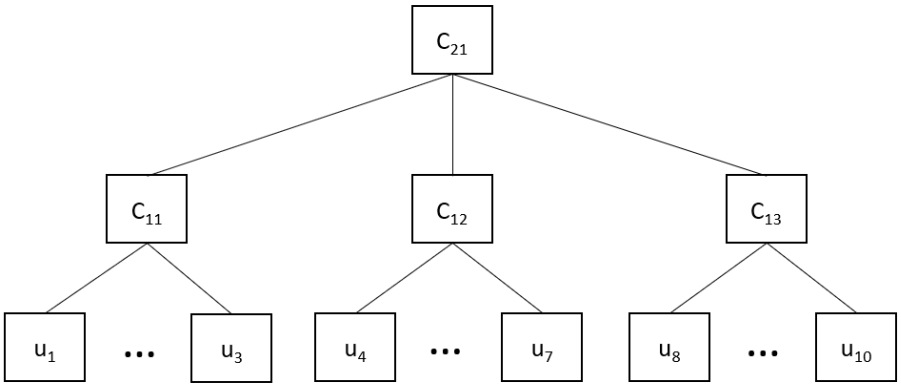


Figure 6.3: Hierarchical nested Archimedean construction.

represents complex multivariate dependence structures through the construction of flexible high-dimensional copulas via a cascade of bivariate copulas as building blocks which are highly flexible in expressing the underlying dependence and tail dependence structure (Dalla Valle et al., 2016). Joe (1996) initially employed the use of pair copula constructions to represent complex multivariate dependence structures which was based on Sklar's theorem utilizing cumulative distribution functions. The methodology was subsequently examined and organized in a graphical and systematical manner called regular vines by Bedford and Cooke (2001, 2002) offering expressions for the joint density. Their methodology was however limited to only Gaussian pair-copulas. Aas et al. (2009) extended this methodology to consider arbitrary pair-copulas and developed standard maximum likelihood estimation (MLE) for special cases of vine copulas, where the arduous task was to offer a good initial point for the needed high dimensional optimization. This was achieved using sequential estimation.

Pair copula construction can be explained by decomposing a trivariate copula. Via recursive conditioning, the three-dimensional joint density can be given as follows

$$f(x_1, x_2, x_3) = f_3(x_3) \cdot f(x_2|x_3) \cdot f(x_1|x_2, x_3) \tag{6.8}$$

According to Sklar's theorem

$$f(x_1, x_2, x_3) = c_{123}(F_1(x_1), F_2(x_2), F_3(x_3)) \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \tag{6.9}$$

where c_{123} is the trivariate copula density. In the bivariate case,

$$f(x_2, x_3) = c_{23}(F_2(x_2), F_3(x_3)) \cdot f_2(x_2) \cdot f_3(x_3) \tag{6.10}$$

for a bivariate copula c_{23}. Thus,

$$f(x_2|x_3) = \frac{f(x_2, x_3)}{f_3(x_3)} = c_{23}(F_2(x_2), F_3(x_3)) \cdot f_2(x_2) \tag{6.11}$$

Similarly, for the trivariate case, it can be expressed as

$$f(x_1|x_2, x_3) = \frac{f(x_1, x_3|x_2)}{f(x_3|x_2)} = c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f(x_1|x_2) \tag{6.12}$$

where $c_{13|2}$ is the copula density for $f(x_1|x_2, x_3)$ with margins $F_{1|2}$ and $F_{3|2}$. Similar to equation 6.11, $f(x_1|x_2)$ can be expressed as

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \tag{6.13}$$

Hence,

$$f(x_1|x_2, x_3) = c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \tag{6.14}$$

Finally, the construction of the trivariate copula density via a cascade of bivariate copulas is given as

$$\begin{aligned} c_{123}(F_1(x_1), F_2(x_2), F_3(x_3)) &= c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \\ &\times c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \end{aligned} \tag{6.15}$$

This decomposition is not unique and the variables can be arranged in $3! = 6$ ways (Wei, 2014). In order to make PCC tractable for inference and model selection purposes, the construction is simplified by assuming that the pair-copula are independent of the conditional variables (Haff et al., 2010). Thus, the pair-copula $c_{13|2}$ is independent of the conditioning variable X_2.

The joint density of an n-dimensional random vector $X = (X_1, ..., X_n)$ can be decomposed as

$$f(x_1, ..., x_n) = f_n(x_n)f(x_{n-1}|x_n)f(x_{n-2}|x_{n-1}, x_n)...f(x_1|x_2, ..., x_n) \tag{6.16}$$

This can be subsequently decomposed into marginal densities and bivariate copulas

$$f(x|\nu) = c_{x\nu_j|\nu_{-j}}(F(x|\nu_{-j}), F(\nu_j|\nu_{-j}))f(x|\nu_{-j}) \tag{6.17}$$

where ν is a k-dimensional vector, ν_j is an arbitrary element of ν and ν_{-j} is the $(m-1)$-dimensional vector ν excluding ν_j. The pair-copula can be applied to transformed variables, which are marginal conditional distribution of $F(x|\nu)$ (Wei, 2014). $F(x|\nu)$ can be generally expressed as

$$f(x|\nu) = \frac{\partial C_{x\nu_j|\nu_{-j}}(F(x|\nu_{-j}), F(\nu_j|\nu_{-j}))}{\partial F(\nu_j|\nu_{-j})} \tag{6.18}$$

where $C_{x\nu_j|\nu_{-j}}$ is a bivariate copula distribution function. The PCC of n-dimensional random variables is not unique with the number of decompositions growing considerably with increase in dimension. For instance, there are 240 possible decompositions for a 5-dimensional distribution/density. Thus, there is the need to represent/describe and organize the multitude of possible pair-copula compositions graphically through the use of regular vines.

6.4 Regular Vines

Vine copulas are the most researched copulas arising from pair-copula construction (Wei, 2014). Vine copulas have two main advantages over other copula models: they are computationally efficient for discrete variables (computationally advantageous expression for the likelihood function) rendering maximum likelihood inference for higher dimensional datasets and combinations of different asymmetric copula families result in very flexible higher dimensional distributions (Stöber et al., 2015).

Vine copulas employ a graphical representation known as regular vine (R-vine) which comprises of a series of trees (undirected acyclic graphs). These trees which consist of nodes and edges are known as dependence trees since they describe dependence structures in high-dimensional distributions (Schepsmeier, 2010). An R-Vine is a special case of vine (nested series of connected trees), $V = T_1, ..., T_{n-1}$ in which two edges in tree T_j are linked by an edge in T_{j+1} only if these edges in T_j share a common node. Each edge of the R-vine is related to a particular pair-copula in a given PCC and the edges of a tree T_j form the nodes for tree T_{j+1} where $j = 1, ..., n-1$.

There are two special cases of regular vine copulas which were proposed by Aas et al. (2009) namely Canonical vines (C-Vines) and Drawable vines (D-vines). C-vines have star structures in their tree sequence with a unique node called root node or node of maximal degree that links all other nodes for each tree. The specification of a four-dimensional C-Vine in the form of a nested set of trees is illustrated in Figure 6.4. This structure comprises of 3 trees, $T_j, j = 1, ..., 3$ with each tree T_j consisting of $5 - j$ nodes and $4 - j$ edges. Each tree has a root node that is linked to $d - j$ edges. Each edge label corresponds to the subscript of the pair-copula density. For instance, the edge $34|12$ in Tree 3 corresponds to the copula density $c_{34|12}(\cdot)$.

D-vines on the other hand have path structures where each node has a degree of not more than 2 (i.e. each node is linked to not more than two other nodes). The specification of a four-dimensional D-Vine is illustrated in Figure 6.5. D-vines may be preferred to C-vines when one does not want to assume the existence of a particular node that dominates the dependencies. D-vine models have been more widely used

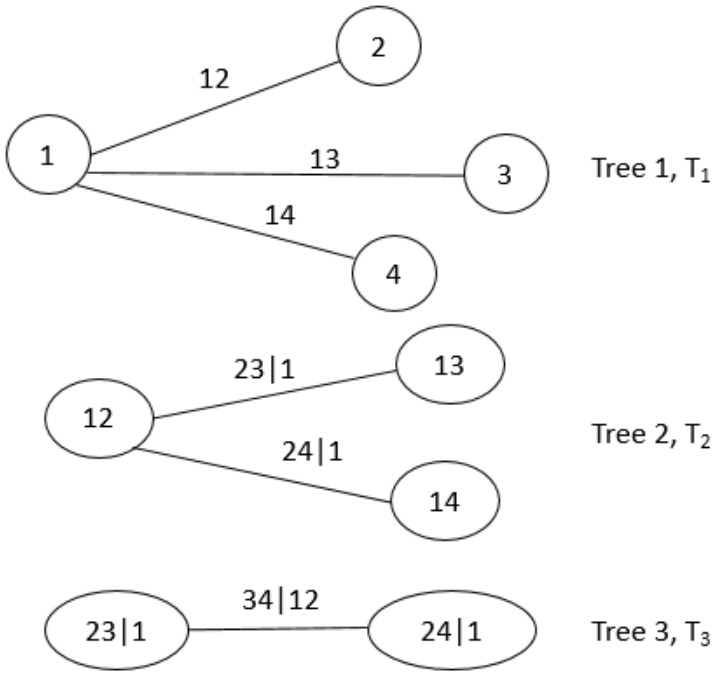


Figure 6.4: Four dimensional C-Vine Structure.

than C-vine models. These special types of R-vines are convenient to use since the initial tree (in the case of D-vines) and the order of the root nodes (in the case of C-vines) determine their structure entirely. However, these special cases of R-vines are restrictive cases with arbitrary R-vines copulas more flexible in modelling of complex dependences in higher-order dimensions (Dissmann et al., 2013; Stöber et al., 2015; Schepsmeier and Czado, 2016).

The d-dimensional density $f(x_1, ..., x_d)$ corresponding to a C-vine is given as:

$$\prod_{k=1}^{d} f(x_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+1|1,...,j-1} \{F(x_j|x_1, ..., x_{j-1}), F(x_{j+1}|x_1, ..., x_{j-1})\} \tag{6.19}$$

The d-dimensional density corresponding to a D-vine can be expressed as:

$$\prod_{k=1}^{d} f(x_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,...,i+j-1} \{F(x_i|x_i + 1, ..., x_{i+j-1}), F(x_{i+j}|x_{i+1}, ..., x_{i+j-1})\} \tag{6.20}$$

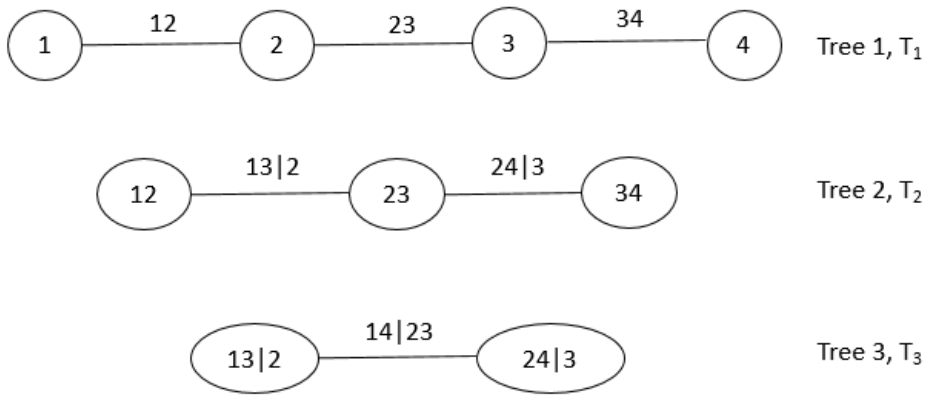


Figure 6.5: Four dimensional D-Vine Structure.

where f_k, $k = 1, .., d$ denotes the marginal densities, and $c_{(...)}$ denotes bivariate copula densities with index j identifying the trees, with index i running over the edges for each tree.

The four-dimensional C-Vine structure can be mathematically expressed as:

$$f_{1234} = f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot c_{12} \cdot c_{13} \cdot c_{14} \cdot c_{23|1} \cdot c_{24|1} \cdot c_{34|12} \tag{6.21}$$

where f_1, f_2, f_3, f_4 are the nodes in Tree 1; c_{12}, c_{13}, c_{14} are the nodes in Tree 2 and edges in Tree 1; $c_{23|1}, c_{24|1}$ are the nodes in Tree 3 and edges in Tree 2; and $c_{34|12}$ is the edge in Tree 3.

The four-dimensional D-Vine structure can be expressed as:

$$f_{1234} = f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{14|23} \tag{6.22}$$

where f_1, f_2, f_3, f_4 are the nodes in Tree 1; c_{12}, c_{23}, c_{34} are the nodes in Tree 2 and edges in Tree 1; $c_{13|2}, c_{24|3}$ are the nodes in Tree 3 and edges in Tree 2; and $c_{14|23}$ is the edge in Tree 3.

There are two stages of model estimation of vine copulas: graph theory which

establishes the dependency structure of the data and statistical inference techniques (such as sequential estimation, maximum likelihood, Bayesian estimation) which are employed in order to fit the bivariate copulas. Vine Copula models comprise of three elements namely the vine tree structure, copula families (for each edge) and copula parameters.

6.5 Vine Structure Selection Methods

The problems which arise during model selection include the huge number of possible vine structure during structure selection and $\frac{d(d-1)}{2}$ number of pair-copulas during copula selection and parameter estimation. For this reason, structure selection is performed tree-wise using approaches such as optimal C-vines structure selection proposed by Czado et al. (2012), Traveling Salesman Problem for D-vines, Maximum Spanning Tree Algorithm for arbitrary R-vines proposed by Dissmann et al. (2013) and Bayesian approaches such as Reversible Jump Markov Chain Monte Carlo(MCMC) (Krämer and Schepsmeier, 2011).

6.5.1 Maximal Spanning Tree Algorithm

Maximum Spanning Tree Algorithm proposed by (Dissmann et al., 2013) is conducted tree-by-tree using a top-bottom approach beginning with the selection of the first tree to the last tree. Edge weights are selected appropriately to demonstrate large dependencies with the assumption that higher weights provide a better fit of the chosen characteristics. These weights are estimated using a sequential estimation approach. With the given weights, Prim's Algorithm can be applied to choose the tree structure that maximizes the sum of edge weights in each tree. Possible edge weights include the following (Krämer and Schepsmeier, 2011; Wei, 2014):

1. Concordance measures such as Kendall's tau and Spearman's rho
2. Information criteria such as AIC and BIC of the pair copula
3. P-values of formal goodness-of-fit tests and variants
4. Distances

For concordance measure weights, the aim is to capture the strongest pairwise dependences in the data. The strongest pairwise dependencies are selected for the first tree of the R-vine. Kendall's τ is usually employed because of its suitability for evaluating non-linear dependence and its invariance under monotone transformations of the margins. The variables that maximize the sum absolute value of Kendall's τ among all pairs make up the tree. Since the true Kendall's τ is not known, empirical estimates are employed.

When it is beneficial to determine a good fitting of the R-vine to the data, the choice of copula family can be conducted in a manner that fits the corresponding observations well. The most popular goodness-of-fit measure is the Akaike Information Criterion (AIC). The choice of pair-copula from an array of bivariate copula families is separated in terms of the parameters for each pair of variables. The corresponding AIC is computed and the copula family with the smallest value is chosen. However, AIC has its limitations. It does not permit the evaluation of statistical significance unlike statistical goodness-of-fit (GoF) tests which can produce results of their p-value. Thus, the p-values of these formal tests can be used to tackle this problem. Since the performance of sequential estimation is dependent on the choice of pair-copula for the corresponding pair of pseudo-data values. Thus, formal GoF tests can be considered during selection.

6.5.2 Sequential Bayesian Tree Selection

Min and Czado (2010) proposed a Bayesian analysis of pair-copula constructions based on Reversible Jump Markov Chain Monte Carlo(MCMC). Reversible Jump MCMC initally proposed by Green (1995) is an extension of the standard MCMC that permits the simulation of the posterior distribution on spaces of varying dimensions and sampling from discrete-continuous posterior distributions. The approach by Min and Czado (2010) is to obtain a sequential estimate of the posterior distribution of the R-vine tree structure, the bivariate copula families and their corresponding parameters. It allows for easier computation of credible intervals of parameter estimates which are

difficult to attain using maximum likelihood estimation (MLE). Modeling the prior density function which favors sparse models can help prevent the selection of models with runaway complexity. On the other hand, the use of non-informative flat priors permits tree-by-tree MLE of the R-vine tree structure, pair-copula families and the corresponding parameters (Wei, 2014).

6.6 Parameter Estimation

Vine copula models can be estimated either sequentially or by joint maximum likelihood estimation (MLE) (Czado et al., 2012; Brechmann and Schepsmeier, 2013). Aas et al. (2009) employed a sequential estimation approach since joint MLE of regular vine copula parameters can be computationally intensive. Joint MLE can be conducted using two-stage estimation approach. Marginal parameters are initially estimated parametrically or non-parametrically and then the copula parameters are subsequently estimated. The parametric approach is known as Inference for Margins and the semi-parametric appraoch is known as Maximum Pseudo Likelihood estimation. These have been previously discussed in sections 4.4.1.2 and 4.4.2 respectively.

Sequential estimation is carried out tree-wise starting with the first tree. The parameters of unconditional copulas are initially estimated which are subsequently used to estimate parameters of the pair-copulas with single conditioning variable. The resulting estimates are in turn used to estimate pair-copula parameters with two-conditioning variables. This is continued sequentially till all parameters are estimated (Czado et al., 2012). Sequential estimation offers a much quicker way of estimating copula parameters than joint MLE since it only estimates bivariate copulas. Resulting values from sequential estimation can be subsequently employed as initial values for numerical high dimensional optimization of the log-likelihood to obtain joint maximum likelihood estimates.

6.7 Copula Families Selection

For a given R-vine structure, there is the need to consider how to select the appropriate pair-copula from a set of families. Copula selection can be performed using goodness-of-fit tests such as tests based on Rosenblatt's probability integral transform, Genest and Favre bivariate asymptotic independence test, information criteria such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) and graphical tools such as contour plots, scatter plots, lambda function, Kendall's plots (K-plots) and Chi-plots (Krämer and Schepsmeier, 2011; Brechmann and Schepsmeier, 2013). These have been discussed at length in section 4.5. However, the pair-copulas are usually selected independently using AIC.

For general R-vine copulas, the choice of pair-copula is dependent on the selection of copulas in a preceding tree. Due to impracticability of a joint selection of copula families, copula families selection is conducted tree-wise as proposed in the sequential estimation. The copula families of the first tree are selected initially and subsequently estimated. Given the selected pair-copulas of the previous tree and their corresponding parameters, the bivariate copula families of the next tree are chosen. The copula selection approach normally coincides with most vine tree selection procedures. However, the sequential copula selection approach accumulates in the selection (Wei, 2014). Thus, there is the need to check and compare the resulting model with alternate models. This can be achieved using comparison tests for non-nested models such as the likelihood ratio tests based on Vuong (1989) and Clarke (2007) tests which have been previously discussed in section 4.5.3.

6.8 Limitations of Vine Copulas

The vine copula methodology has some shortcomings. In order to make PCC tractable for inference and model selection purposes, the construction is simplified by assuming that the pair-copula are independent of the conditional variables, except through the conditional distributions (Haff et al., 2010). Thus, the copulas corresponding to conditional distribution are constant irrespective of the values of variables that

they are conditioned (Stöber et al., 2013). In other words, the conditional pair copulas are assumed to depend on the conditional variables only indirectly through the conditional margins (Acar et al., 2012). Haff et al. (2010) demonstrated that a simplified PCC is a good approximation even when the simplifying assumption is far from being fulfilled by the actual model. However, Acar et al. (2012) suggested that this view is too optimistic and subsequently showed that an uncritical use of the simplifying assumption may be misleading. Additionally, the flexibility of vine copulas can be disadvantageous due to the plethora of different vine structures to select from and it is a priori not clear which structure to employ. Thus, PCCs do not have a unique solution (Geidosch and Fischer, 2016; Monstvilaite, 2016). The manner of the manner of finding the best-fitting model can be done heuristically (at hand) or by employing highly time-consuming Markov chain Monte Carlo methods Schepsmeier and Czado (2016).

Furthermore, vine copula methodology for continuous data is employed in this dissertation despite the mixed nature of the variables of interest. This is due to the computation complexity of handing discrete variables. There are two common techniques for copula modelling of discrete variables (Stöber et al., 2015):

1. For copulas functions available in closed form, calculation of the probability mass function by taking finite differences of the copula function for the discrete margins. This results in the exponential growth in the number of evaluations of the copula function with the number of discrete variables. The limitation of this approach is the significant increase in computational complexity with dimension and sample size.

2. Introduction of latent continuous variables (as an alternative to direct application of a copula to discrete data). In this method, the dependence structure of the latent variables is modelled instead. The ability to apply popular dependence models and the avoidance of the technicalities when dealing with discrete copulas

adds to the appeal of this method. However, the latent variables make inference computationally hard.

To tackle these issues, Stöber et al. (2015) developed a vine copula approach which considers applications of pair copula constructions for mixed data. The implementation of this methodology may change the results of this paper.

6.9 Case Study (Modeling High-Dimensional Dependence of Derailment Severity)

6.9.1 Introduction

Despite the relatively low frequency of train derailments, they have been a major concern due to their high consequence justifying the need to critically examine the severity of train derailments in order to minimize and mitigate the resulting damage (Jeong et al., 2007; Liu et al., 2013). Derailments may result in loss of life and property, interruption of services and destruction of the environment (Liu et al., 2013), and are the most frequent kind of Federal Railroad Administration (FRA)-reportable mainline train accident in the United States (Barkan et al., 2003; Liu et al., 2012; Liu, 2015). Derailments made up about three-quarters of freight-train accidents in the United States from 2001 to 2010. Therefore, analyzing the magnitude and variability of derailment severity is as important as estimating the likelihood of derailment (Liu et al., 2013).

Derailment severity may be influenced by factors like car mass, derailment speed, residual train length (number of cars after the point of derailment), derailment cause, ground friction, rail friction, derailment cause, proportion of loaded railcars in the train (loading factor) and train power distribution. Estimation of these variables is often established through exact estimation or the determination of statistical distributions and time history of the examined factors (Mohammadzadeh and Ghahremani, 2010).

It is important to know the interrelationships or dependencies between these variables in order to better understand how to reduce the severity and consequences of train derailments. The most widely used statistical dependence model is that of

the multivariate Gaussian distribution (Dissmann et al., 2013). Multivariate normal distributions are usually used for multivariate data which assumes linear dependence structure and no asymmetric or tail dependence (Schepsmeier and Czado, 2016). However, non-normality transpires in various forms: nonnormality of marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables (Yan, 2006; Attoh-Okine, 2013). Thus, conventional correlation analysis is not suitable for analyzing data with non-normality, tail dependency and skewness. The objective of this case study is to model the underlying high-dimensional dependences between the variables by taking into account these nonlinearities.

6.9.2 Data

Data was obtained from the Rail Equipment Accident/Incident (REA) database maintained by the Federal Railroad Administration (FRA) of U.S. Department of Transportation (U.S. DOT). A "rail equipment accident/incident" is a collision, derailment, fire, explosion, act of God, or other event involving the operation of railroad on-track equipment (standing or moving). U.S. railroads are required to present detailed reports (Form 6180.54) to the FRA on all accidents or incidents whose damage costs exceed a specified monetary value. The damage incurred includes damage caused to the railroad track, signals, on-track equipment, track structures and roadbed as well as labor costs and the costs for acquiring new equipment and material. The reporting threshold is periodically changed to account for inflation and other adjustments and has increased from \$5700 in 1990 to $\$10,700$ in 2017 (FRA, 2016). The relatively low threshold results in most accidents being reported to the FRA (Barkan et al., 2003).

The database contains detailed track accident information such as accident cause, number of derailed cars, total monetary damage, track type, track class, train length and derailment speed. 690 freight-train derailments occurring on Class I mainline track in the year 2005 were initially considered. The variables considered include

the number of derailed cars, monetary damage, derailment speed, residual train length and proportion of loaded railcars in the train (loading factor).

To cater for the effect (and variations) due to derailment cause, 124 derailments caused by broken rail were considered. Broken rails are the most frequent cause of freight-train derailment on Class I mainlines in the United States. Broken rails also result in a higher derailment severity in comparison with other causes such as bearing failure with the former causing twice as many derailed cars on average as that of the latter (Barkan et al., 2003). Due to their high frequency and severity, broken rails are more likely to present higher risk than other causes. Initial exploratory data analysis indicated non-normality of the marginal distributions of the variables under examination as shown in Figures 6.6, 6.7 and 6.8. Copula analysis was employed since copula models address the limitations of conventional correlation analysis by taking into account non-normality, tail dependence, skewness and other nonlinearities.

Pseudo-observations (or copula data) were computed by transforming the dataset into marginally uniform data using empirical probability integral transformation (empirical distribution functions) and an asymptotically negligible scaling factor $\frac{n}{n+1}$ where n is the number of observations. This factor is employed to compel the variates to fall within the open unit hypercube $(0,1)^4$ in order to avoid issues with density evaluation at the boundaries of closed unit hypercube $[0,1]^4$ (Schepsmeier and Czado, 2016). Transformation can also be achieved by means of parametric probability integral transformation. Figure 6.9 shows pairs plots of the transformed derailment data set with scatter plots above the diagonal and contour plots with standard normal margins below the diagonal.

Detailed exploratory data analysis was conducted on the pair of variables Derailment Speed and Derailed Cars. Illustrative tools such as Kendall's plot (K-plot), Chi-plot and lambda function can be employed for detecting dependence of the pair of variables (see (Fisher and Switzer, 1985, 2001; Genest and Rivest, 1993; Genest and Boies, 2003; Genest and Favre, 2007). Figure 6.10 shows the K-plot, chi-plot and empirical lambda function (black line), theoretical lambda function for Gumbel copula

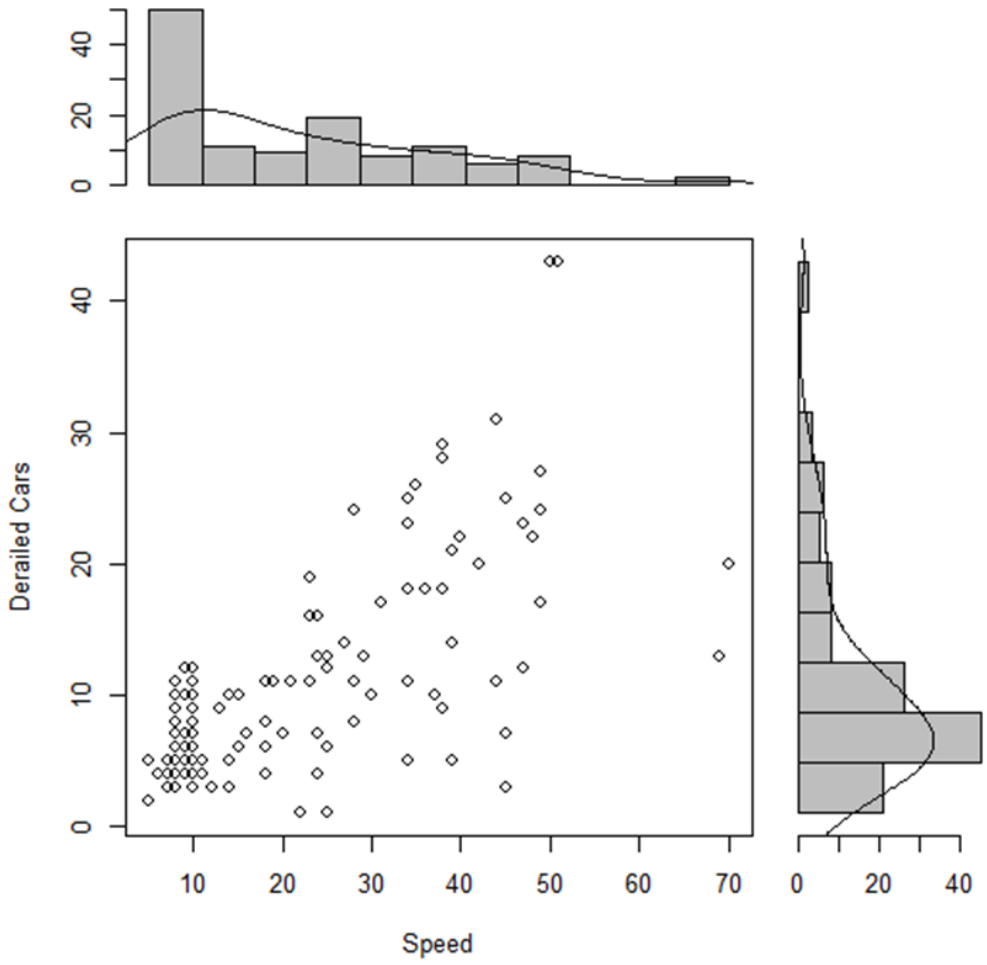


Figure 6.6: Scatter histogram of Derailed Cars against Derailment Speed.

distribution as well as independence and comonotocity limits (dashed line). The contour plot in row 2 column 1 of figure 6.9 as well as the curve located above the main diagonal (or line $y = x$) in the Kendall's plot and majority (in this case all) of the pair of observations in the Chi-plot being positive values of χ in Figure 6.10 show that variables are positively dependent.

6.9.3 Vine Copula Analysis and Results

Four-dimensional C-vine and D-vine copula models were applied to the transformed derailment severity data. The analysis includes the selection of C-vine and

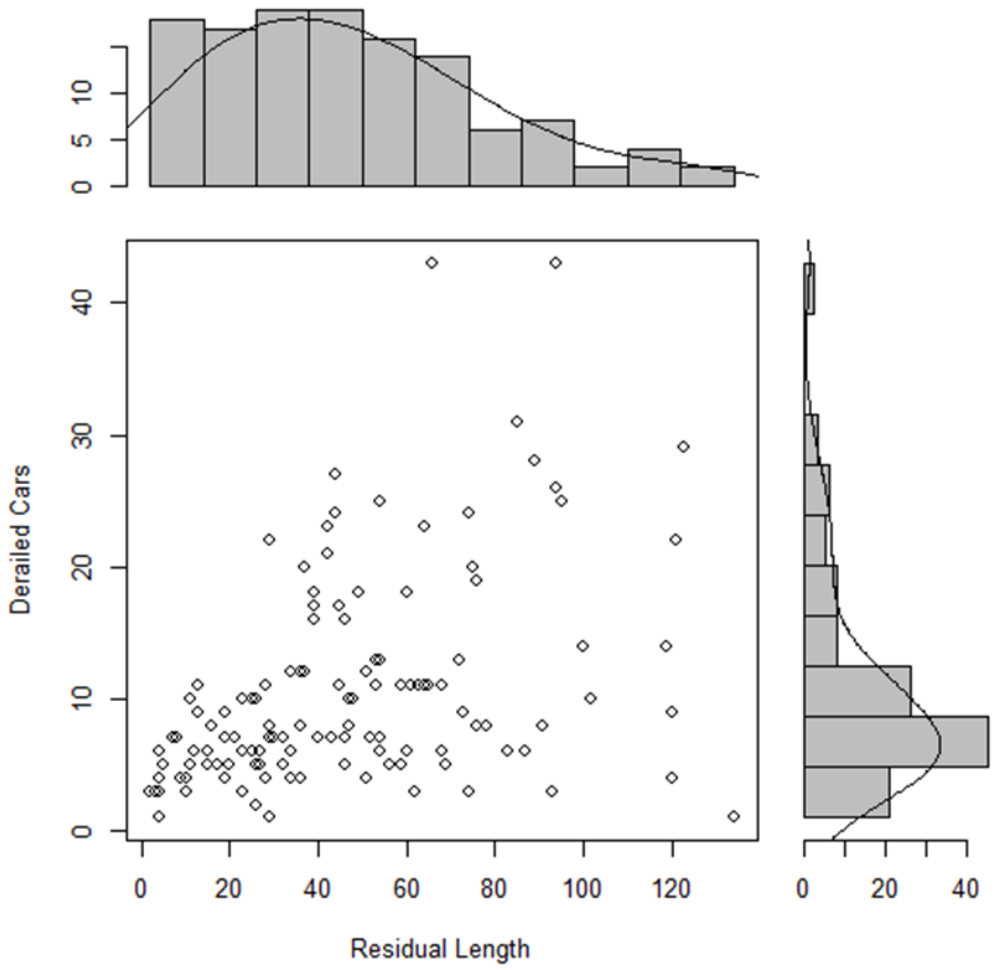


Figure 6.7: Scatter histogram of Derailed Cars against Residual Train Length.

D-Vine structures, set of pair-copula families and their corresponding parameters and evaluation of the alternative models. The selection of the tree structures of the C- and D-vines was obtained given the data. The structures can also be obtained via manual selection or through expert knowledge. The order of the root nodes and the first tree completely determine the structures of the C-vine and D-vine copula models respectively. The root node of each tree of the C-vine was determined by establishing the node with the strongest dependencies with other nodes. This is achieved by finding the node with the maximum row sum of the absolute values in the empirical Kendall's tau matrix. As shown in Table 6.1, Derailed Cars was identified as the first root node.

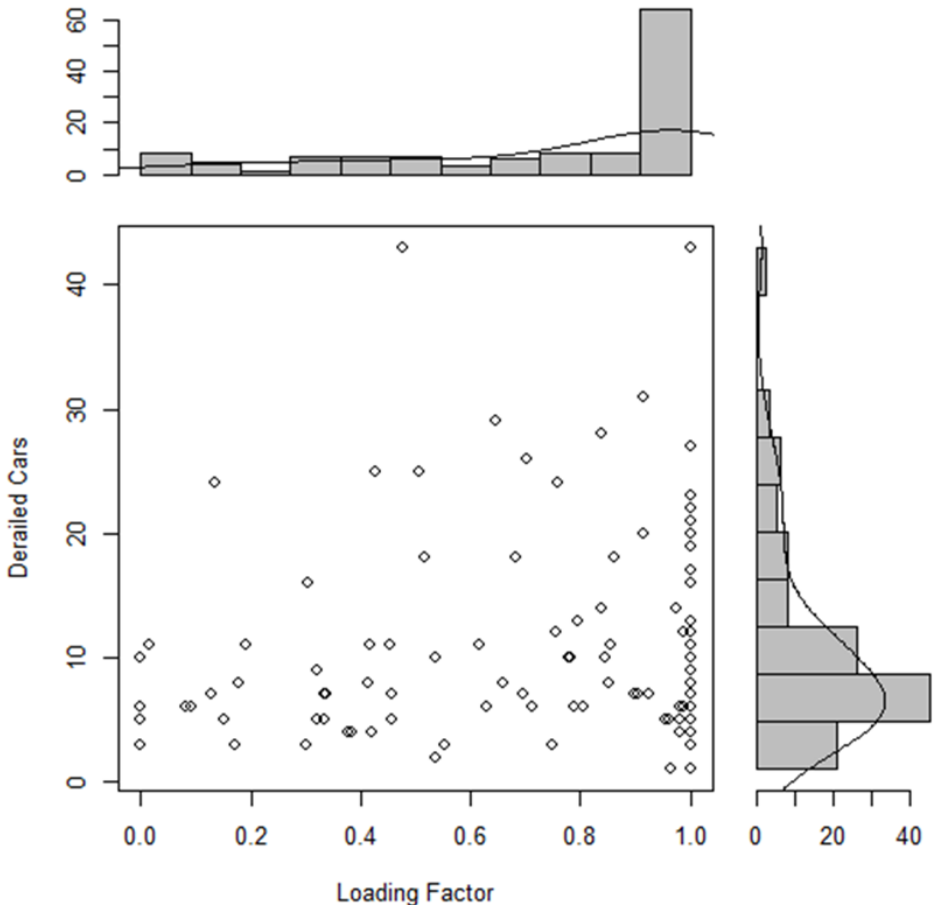


Figure 6.8: Scatter histogram of Derailed Cars against Loading Factor.

Subsequently, given the initial root node and the sequential C-vine identification procedure outlined in Czado et al. (2012), the next root node was identified to be Speed (as shown in Table 6.2) followed by Residual Train Length and finally Loading Factor.

The structure of the D-vine is determined by establishing the order of the first tree. This can be determined by finding the path which maximizes the pairwise dependences (Kendall's τ) of the variables of interest (Dalla Valle et al., 2016). This can be transformed into a traveling salesman problem where the shortest Hamiltonian path is determined in terms of weights $1 - |\tau|$ (Dissmann et al., 2013; Dalla Valle et al., 2016; Schepsmeier and Czado, 2016). The order of the D-vine obtained was Speed, Derailed Cars, Residual Length and Loading Factor.

Table 6.1: Empirical Kendall's τ matrix and the sum over the absolute entries of each row for the Derailment data set

	Derailed Cars	Speed	Residual Length	Loading Factor	Sum of Absolute τ
Derailed Cars	1	0.482	0.330	0.061	**1.873**
Speed	0.482	1	0.140	0.027	1.649
Residual Length	0.330	0.140	1	-0.011	1.481
Loading Factor	0.061	0.027	-0.011	1	1.089

Table 6.2: Empirical Kendall's tau matrix and the sum over the absolute entries of each row for the Derailment data set given derailed cars (D) as first root

	D, S	**D, R**	**D, L**	**Sum of Absolute τ**
D,S	1	-0.11	-0.15	**1.18**
D,R	-0.11	1	-0.05	1.16
D,L	-0.07	-0.05	1	1.12

The family set of pair-copulas to choose from must include at least one bivariate copula family that permits positive dependence and at least another one that permits negative dependence. The pair-copula families considered during the analysis were the independent copula, elliptical bivariate Gaussian (Normal) and Student t-copulas, the single parameter Archimedean copulas such as bivariate Clayton, Gumbel, Frank, Joe as well as two-parameter Archimedean copulas such as Clayton-Gumbel (BB1) and Joe-Clayton (BB7) which permit different non-zero lower and upper tail dependence coefficients. Rotated versions (90^0 and 270^0) of these Archimedean copulas were considered to fit negative dependences (with the exception of Frank copula which has no rotated versions). This catalogue for the implementation of copula family choice address a vast range of dependence behavior. Properties of these copulas are found in Table 4.3. The best fitting pair copula for each pair of variables was selected using the Akaike Information Criterion (Akaike, 1973) which corrects the log likelihood of a copula for the number of parameters. AIC was chosen for bivariate copula selection ahead of other alternative criteria such as Vuong (1989) and Clarke (2007) goodness-of-fit tests and Bayesian Information Criteria (Schwarz, 1978) as a result of its high

performance in simulation analysis and its greater reliability (Dissmann et al., 2013; Dalla Valle et al., 2016).

Independence copula was included in the selection by computing the independence of each pair of variables of the R-vines prior to bivariate copula selection. This achieved by conducting the Genest and Favre bivariate asymptotic independence test based on Kendall's Tau. The observance of conditional independence between variables leads to a reduction in the number of levels of the pair copula decomposition leading to the simplification of the construction. The independence copula for this pair-copula term was selected if the p-value of the test was higher than 0.05. The pair-copula parameter estimation was conducted based on the sequential estimation approach proposed by Aas et al. (2009). The resulting values from sequential estimation were subsequently employed as initial values for joint maximum likelihood estimation. Sequential and maximum likelihood estimates and Kendall's tau values for C-vine and D-vine copula models can be found in Tables 6.3 and 6.4 respectively.

Figures 6.11 and 6.12 show the C- and D-vine copula models with family and Kendall's tau values in each tree. It shows 3 trees for 4 variables where G - Gumbel Copula, t - Student's t copula, C90 - rotated Clayton (90^0) copula and I - Independence Copula. In tree 1 of the C-Vine (Figure 6.11), moderate positive correlation was observed between the number of derailed cars and derailment speed with relatively weak positive correlations observed between derailed cars and residual train length. On the other hand, no (very weak) dependence between derailed cars and loading factor. The selection of the Gumbel copula suggests the underlying dependence between the number of derailed cars and derailment speed is asymmetric and exhibits positive upper tail dependence. The number of derailed cars during a derailment was found to be highly correlated positively at high values of derailment speed but poorly (positive) correlated at low derailment speeds. The selection of the student-t copula suggest the underlying dependence between the number of derailed cars and residual train length is symmetric and exhibits both upper and lower tail dependence.

In tree 2 of the C-Vine, the conditional dependence between derailment speed

Table 6.3: Sequential and maximum likelihood parameter estimates and Kendall's tau values for C-vine copula model

Variable Pair	**Copula Type**	**Sequential Estimation**		**MLE**		τ
		Parameter 1	**Parameter 2**	**Parameter 1**	**Parameter 2**	**Tau**
Derailed Cars (D), Speed (S)	Gumbel	1.8213561	0	1.8196274	0	0.45
Derailed Cars (D), Residual Length (R)	Student-t	0.4780607	4.135542	0.4790877	4.310644	0.32
Derailed Cars (D), Loading Factor (L)	Independent	0	0	0	0	0
$(S, R)\|D$	Rotated (90^0) Clayton	-0.2519488	0	-0.2535085	0	-0.11
$(S, L)\|D$	Independent	0	0	0	0	0
$(R, L)\|(D, S)$	Independent	0	0	0	0	0

Table 6.4: Sequential and maximum likelihood parameter estimates and empirical tau values for D-vine copula model

Variable Pair	**Copula Type**	**Sequential Estimation**		**MLE**		**Emp. Tau**
		Parameter 1	**Parameter 2**	**Parameter 1**	**Parameter 2**	
Speed (S), Derailed Cars (D)	Gumbel	1.8213561	0	1.8196274	0	0.45
Derailed Cars (D), Residual Length (R)	Student-t	0.4780607	4.135542	0.4790877	4.310644	0.32
Residual Length (R), Loading Factor (L)	Independent	0	0	0	0	0
$(S, R)\|D$	Rotated (90^0) Clayton	-0.2519488	0	-0.2535085	0	-0.11
$(D, L)\|R$	Independent	0	0	0	0	0
$(S, L)\|(R, D)$	Independent	0	0	0	0	0

and residual train length given the number of derailed cars was found to be negative and relatively weak and can be characterized by a 90^0 rotated Clayton copula which also indicates the presence of tail dependence. On the other hand, it was observed that derailment speed and loading factor are conditionally independent given derailed cars.

In tree 1 of the D-Vine (Figure 6.12) moderate positive correlation was observed between the number of derailed cars and derailment speed with relatively weak positive correlations observed between derailed cars and residual train length. This is similar to the C-Vine with the same copulas and dependencies identified. However, no (very weak) correlation was found between derailment speed and loading factor. In tree 2 of the D-Vine, the derailment speed and loading factor were found to be conditionally independent given the number of derailed cars. Similar to the C-Vine, negative conditional dependence between derailment speed and residual train length given the number of derailed cars was observed.

The Akaike Information Criterion (AIC), Bayesian Information Criterion and log-likelihood were used to measure which R-vine structure models the data better. The log-likelihood is a popular measure of goodness of fit. AIC and BIC on the other hand are classical model comparison measures, taking the model complexity into account (Schepsmeier and Czado, 2016). The C- and D-Vine models were found to have similar AIC, BIC and log-likelihood values as shown in table 6.5. The C- and D-Vine copulas were subsequently compared with the multivariate Gaussian copula approach. The most widely used statistical dependence model is the multivariate Gaussian distribution (Dissmann et al., 2013). The multivariate Gaussian copula is obtained from the multivariate normal or Gaussian distribution and is the dependence structure for linear correlation (Dorey and Joubert, 2005). Multivariate normal distributions (or multivariate Gaussian copulas) are usually used for multivariate data which assumes linear dependence structure and no tail dependence. The multivariate Gaussian copula can be expressed as an R-vine with Gaussian pair copulas where the parameters are established by the associated partial correlations (Schepsmeier and Czado, 2016). The Vine Copulas were found to have greater log-likelihood and lower AIC and BICs.

This demonstrates that the use of the more popular multivariate Gaussian copula which assumes normality of the marginal and joint distributions is not appropriate. This demonstrates the importance of characterizing tail dependence, skewness, and non-normality within the data. The Log-likelihoods, number of parameters, AIC and BIC for the vine copulas and multivariate Gaussian copula using maximum likelihood estimation and/or sequential estimation can be found in Table 6.5.

Table 6.5: Log-likelihood, number of parameters, AIC and BIC for Vine copulas and Multivariate Gaussian copula using maximum likelihood estimation (MLE) or sequential estimation (SE)

	D-vine copula model	**C-vine copula model**	**Multivariate Gaussian copula**
Log-likelihood (SE)	53.7602	53.7602	*
Log-likelihood (MLE)	53.7636	53.7639	43.8882
Number of parameters	3	3	6
AIC (SE)	-99.5204	-99.5204	*
AIC (MLE)	-99.5279	-99.5279	-75.7763
BIC (SE)	-88.2393	-88.2393	*
BIC (MLE)	-88.2468	-88.2468	-58.8546

Similarly, 4-dimensional C- and D-Vine models involving monetary damage instead of derailed cars were analyzed. The 4-dimensional monetary damage severity C- and D-Vine models are shown in figures 6.13 and 6.14. Similar to derailed cars, the dependence between monetary damage and derailment speed was found to be moderate with upper tail dependence observed between the pair (characterized by the Gumbel copula). However, unlike derailed cars which appeared to be independent of loading factor, monetary damage was found to be dependent on loading factor with the Gaussian copula indicating radial symmetric dependence between the pair with no tail dependence. Thus, loading factor may be appropriate in predicting the total monetary damage incurred during derailments however caution is needed when being used to predict the number of derailed cars.

For the monetary damage models, the D-Vine was found to have a lower AIC value. This was confirmed using the log-likelihood and Bayesian Information Criterion

(BIC) with the D-Vine having a greater log-likelihood and a lower BIC value as shown in table 6.6. The likelihood ratio tests based on Vuong (1989) and Clarke (2007) suitable for non-nested model comparison were also performed with Schwarz correction. The Vuong test (log-likelihood ratio for non-nested models) which is based on the difference in the log-likelihoods was used to examine whether differences in the log-likelihood and the AIC of the two vine-copula models were statistically significant (Ayuso et al., 2016). The results of both tests concluded that these differences were not statistically significant and failed to reject the null hypothesis of statistical indistinguishability of the two vine copula models as shown in Table 6.7).

Table 6.6: Log-likelihood, number of parameters, AIC and BIC for "monetary damage severity" Vine copulas and Multivariate Gaussian copula using maximum likelihood estimation (MLE) or sequential estimation (SE)

	D-vine copula model	**C-vine copula model**	**Multivariate Gaussian copula**
Log-likelihood (SE)	54.97051	51.66251	*
Log-likelihood (MLE)	54.97459	51.66251	48.07538
Number of parameters	4	3	6
AIC (SE)	-101.941	-97.32502	*
AIC (MLE)	-101.9492	-97.32502	-84.15077
BIC (SE)	-90.6599	-88.86418	*
BIC (MLE)	-90.66805	-88.86418	-67.22908

Table 6.7: Pairwise non-nested model comparison using Vuong and Clarke tests with Schwarz correction

Null Hypothesis	**Method**		**Alternative C-vine copula model**
D-vine Copula Model	Vuong	Statistics	0.3101814
		p-value	0.756423
		Decision	D-Vine = C-Vine
	Clarke	Statistics	68
		p-value	0.3232476
		Decision	D-Vine = C-Vine

Vine copula method can be applied to simulation modeling of multivariate derailment severity data and has already been applied in areas such as pipeline infrastructure (Atique and Attoh-Okine, 2016) and pavement infrastructure (Attoh-Okine, 2013). 150 pseudo-observations were generated given the bivariate copulas and parameters of the C-Vine copula model in table 6.3. Figure 6.15 shows the scatter plot of the simulated data which reproduces the general pattern in the original data.

6.9.4 Concluding Remarks

Exploratory data analysis showed that the marginal distributions of the variables (such as derailment speed, residual length, loading factor, derailed cars) were not normal as well as the joint distribution of these variables. Conventional correlation analysis which assumes multivariate normality is generally not suitable for analyzing the dependencies between variables with non-normality, tail dependence, skewness and other nonlinearities. Copula models however address the limitations of conventional correlation analysis were deemed to be more appropriate for analyzing the dependences within the derailment data. Special cases of vine copulas: Canonical vines (C-Vines) and Drawable vines (D-Vines) were used to model the dependences within the derailment data. Some of the pairwise dependences were found to show asymmetric and tail dependence violating the multivariate normality assumption. These vine copulas models were found to be better at modeling the derailment data in comparison with multivariate Gaussian copulas which assume multivariate normality. It was found that loading factor may be appropriate in predicting the total monetary damage incurred during derailments however caution is required when being used to predict the number of derailed cars. Vine copula methodology was applied to simulation modeling of multivariate derailment severity data. Insights gained from dependence modeling will improve railroad safety decision making by facilitating deeper comprehension of train derailment severity distribution which will guide future safety research in the railroad industry. In conclusion, this approach is applicable in railroad industry based on its satisfactory results.

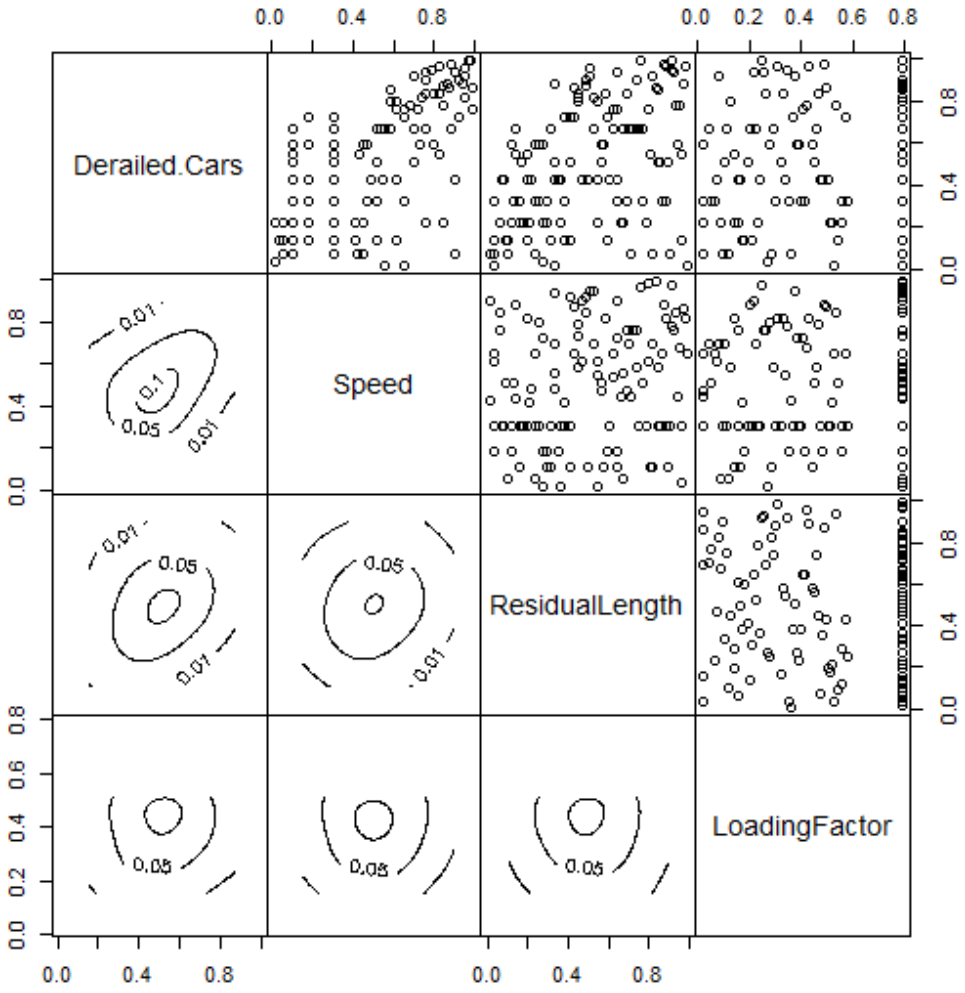


Figure 6.9: Pairs plot of transformed derailment data set with scatter plots above and contour plots with standard normal margins below the diagonal.

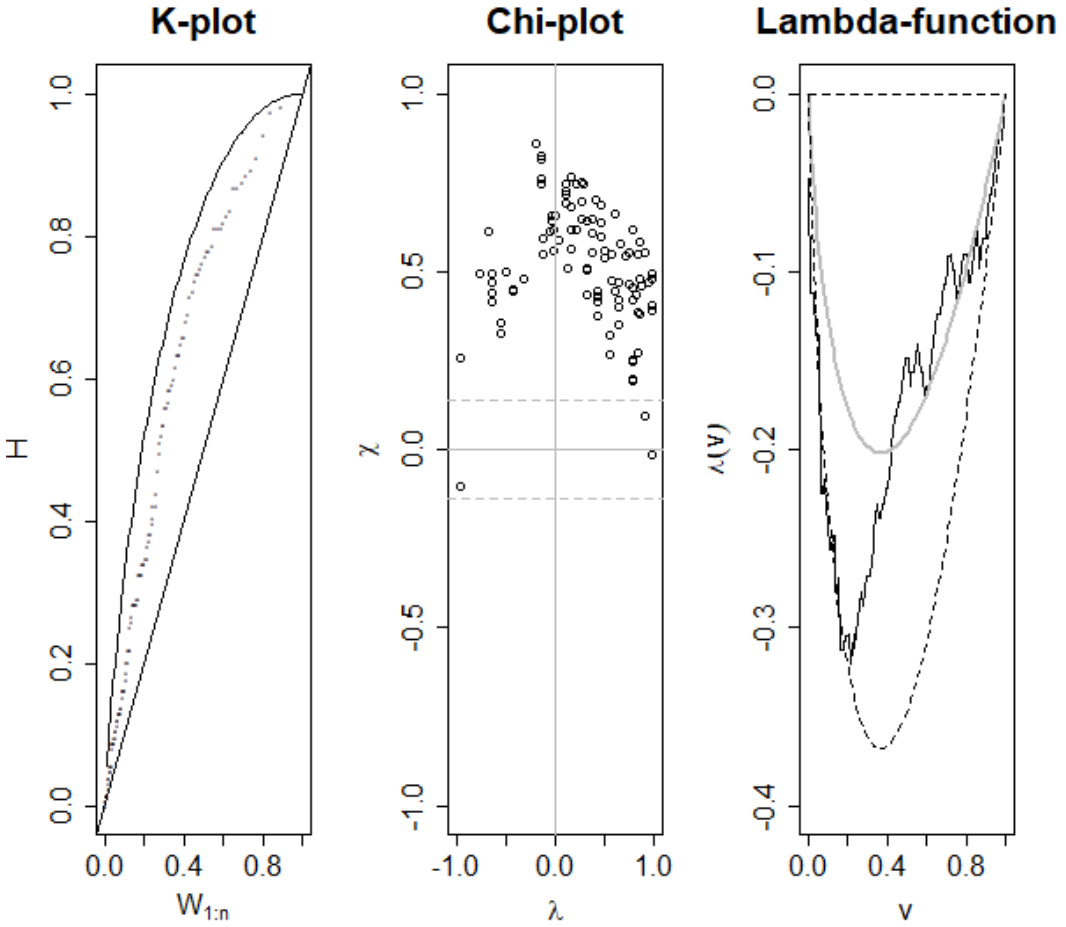


Figure 6.10: Left panel: K-plot. Middle panel: chi-plot. Right panel: empirical lambda-function (black line), theoretical lambda-function of Gumbel copula (grey line) as well as independence and comonotonicity limits (dashed lines).

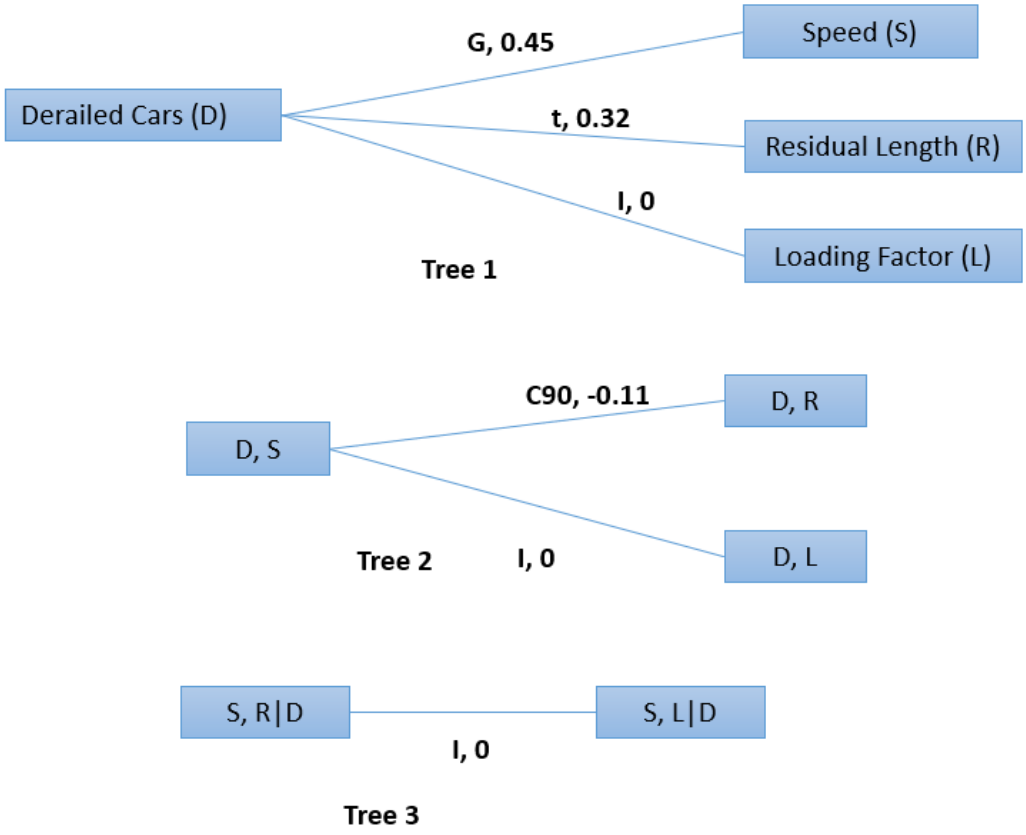


Figure 6.11: Four dimensional C-vine, where G - Gumbel Copula, t - Student's t copula, C90 - rotated Clayton (90^0) copula, I - Independence Copula with corresponding tau values shown on the links with the copula family.

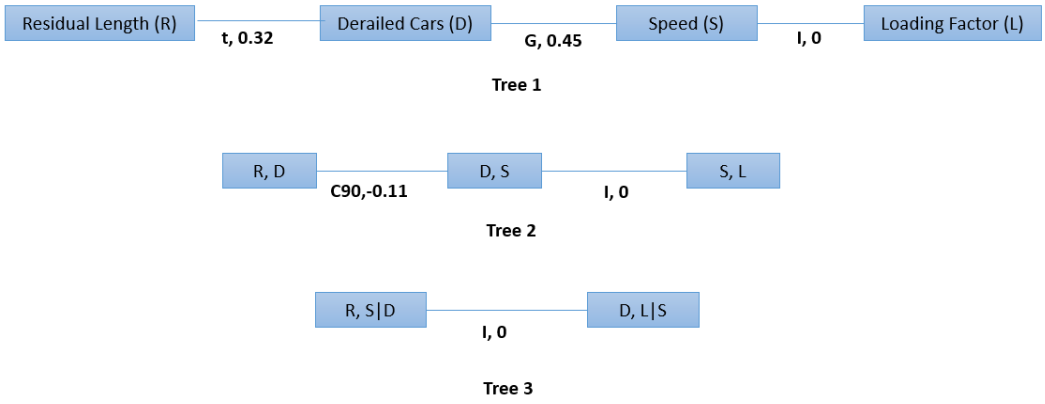


Figure 6.12: Four dimensional D-vine, where G - Gumbel Copula, t - Student's t copula, C90 - rotated Clayton (90^0) copula, I - Independence Copula with corresponding tau values shown on the links with the copula family.

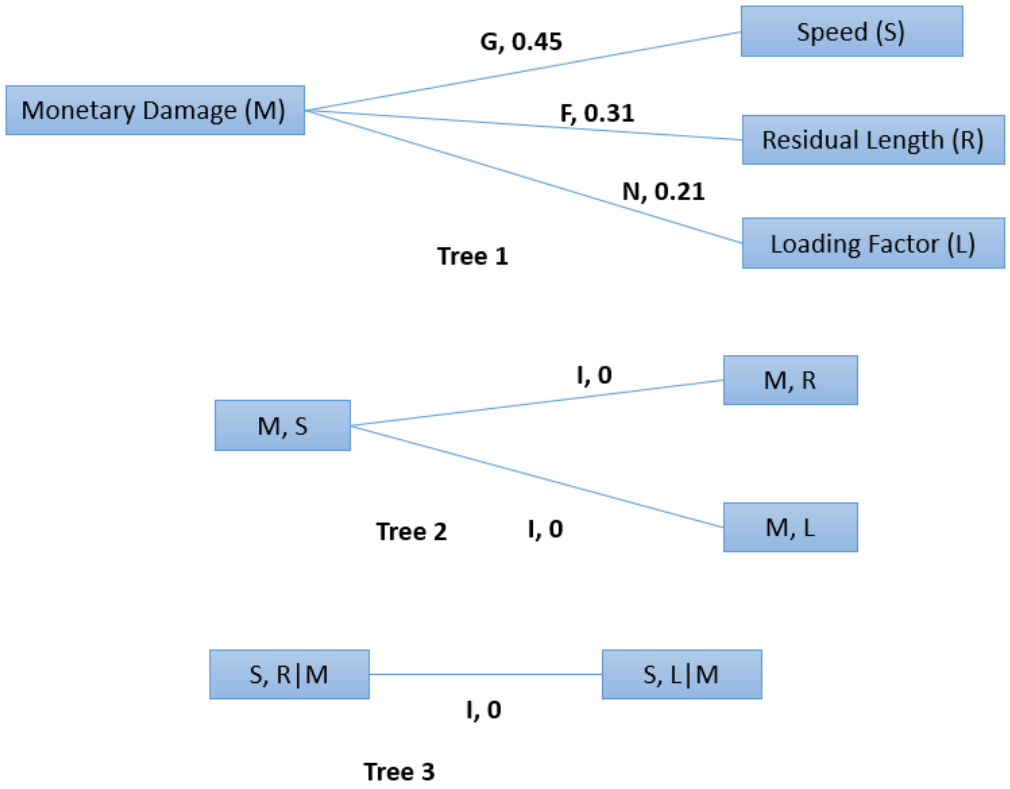


Figure 6.13: Four dimensional C-vine, where G - Gumbel Copula, F - Frank copula, N - Normal/Gaussian copula and I - Independence Copula with corresponding tau values shown on the links with the copula family.

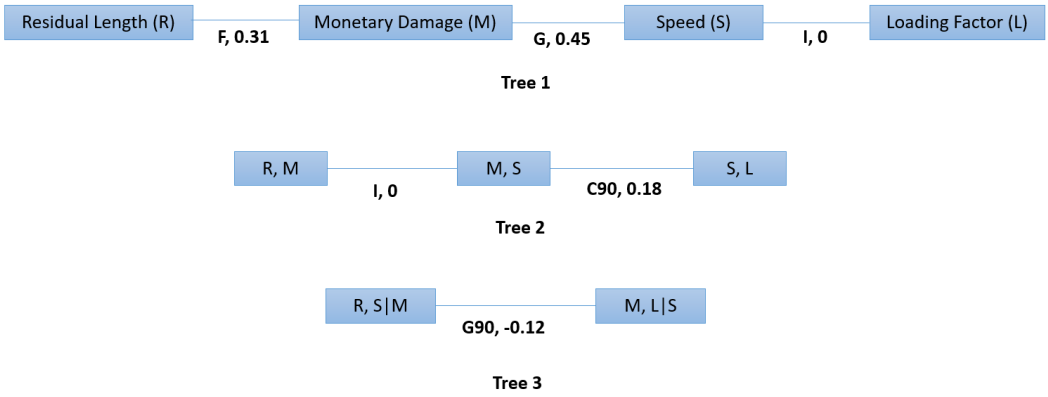


Figure 6.14: Four dimensional D-vine, where F - Frank Copula G - Gumbel Copula, C90 - rotated Clayton (90^0) copula, G90 - rotated Gumbel (90^0) copula, I - Independence Copula with corresponding tau values shown on the links with the copula family.

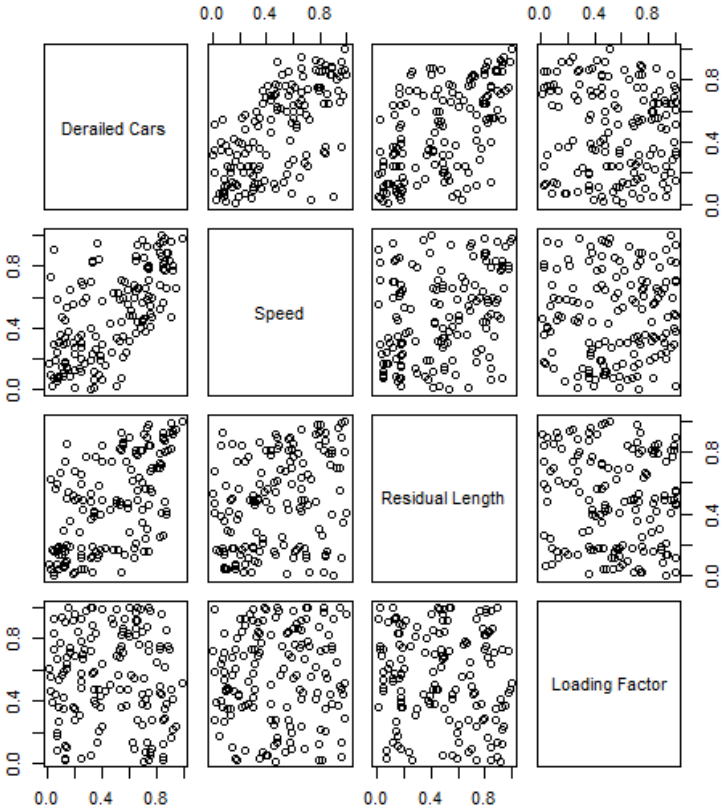


Figure 6.15: Simulated derailment severity data using C-Vine copula model

REFERENCES

Aas, Kjersti; Czado, Claudia; Frigessi, Arnoldo, and Bakken, Henrik. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44 (2):182–198, 2009. ISSN 01676687. doi: 10.1016/j.insmatheco.2007.02.001.

Acar, Elif F.; Genest, Christian, and Neslehova, Johanna. Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90, 2012. ISSN 0047259X. doi: 10.1016/j.jmva.2012.02.001.

Atique, Farzana and Attoh-Okine, Nii. Using copula method for pipe data analysis. *Construction and Building Materials*, 106:140–148, 2016. ISSN 09500618. doi: 10.1016/j.conbuildmat.2015.12.027.

Attoh-Okine, Nii O. Pair-copulas in infrastructure multivariate dependence modeling. *Construction and Building Materials*, 49:903–911, 2013. ISSN 09500618. doi: 10.1016/j.conbuildmat.2013.06.055.

Ayuso, Mercedes; Bermúdez, Lluís, and Santolino, Miguel. Copula-based regression modeling of bivariate severity of temporary disability and permanent motor injuries. *Accident Analysis and Prevention*, 89:142–150, 2016. ISSN 00014575. doi: 10.1016/j.aap.2016.01.008.

Barkan, Christopher P. L.; Dick, C. Tyler, and Anderson, Robert. T. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 1825(9):64–74, 2003. ISSN 03611981.

Bedford, Tim and Cooke, Roger M. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268, 2001. ISSN 1573-7470. doi: 10.1023/A:1016725902970.

Bedford, Tim and Cooke, Roger M. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002. ISSN 2168-8966.

Berg, Daniel. Copula goodness-of-fit testing: An overview and power comparison. *European Journal of Finance*, 15(7-8):675–701, 2009. ISSN 1351847X. doi: 10.1080/13518470802697428.

Brechmann, E C and Schepsmeier, U. Modeling dependence with C-and D-vine copulas: The R-package CDVine. *Journal of Statistical Software*, 52(3):1–27, 2013.

Clarke, Kevin A. A simple distribution-free test for nonnested model selection. *Political Analysis*, 15(3):347–363, 2007. ISSN 10471987. doi: 10.1093/pan/mpm004.

Czado, C.; Schepsmeier, U., and Min, A. Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, 12(3):229–255, 2012. ISSN 1471-082X. doi: 10.1177/1471082X1101200302.

Dalla Valle, Luciana; De Giuli, Maria Elena; Tarantola, Claudia, and Manelli, Claudio. Default probability estimation via pair copula constructions. *European Journal of Operational Research*, 249(1):298–311, 2016. ISSN 03772217. doi: 10.1016/j.ejor.2015.08.026.

Dissmann, J.; Brechmann, E. C.; Czado, C., and Kurowicka, D. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis*, 59(1):52–69, 2013. ISSN 01679473. doi: 10.1016/j.csda.2012.08.010.

Dorey, M and Joubert, Phil. Modelling copulas: an overview. *The Staple Inn Actuarial Society*, pages 1–27, 2005.

Fisher, N. I. and Switzer, P. Chi-Plots for Assessing Dependence. *Biometrika*, 72(2):253, aug 1985. ISSN 00063444. doi: 10.2307/2336078.

Fisher, N. I. and Switzer, P. Graphical Assessment of Dependence: Is a Picture Worth 100 Tests? *The American Statistician*, 55:233–239, 2001. doi: 10.2307/2685807.

FRA, . Monetary Threshold for Reporting Rail Equipment Accidents/Incidents for Calendar Year 2017. *Federal Registry*, 81(247):57–60, 2016.

Geidosch, Marco and Fischer, Matthias. Application of Vine Copulas to Credit Portfolio Risk Modeling. *Journal of Risk and Financial Management*, 9(2):4, 2016. ISSN 1911-8074. doi: 10.3390/jrfm9020004.

Genest, Christian and Boies, Jean-Claude. Detecting Dependence with Kendall Plots. *The American Statistician*, 57:275–284, 2003. doi: 10.2307/30037296.

Genest, Christian and Favre, Anne-Catherine. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007. ISSN 1084-0699. doi: 10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, Christian and Rivest, Louis-Paul. Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*, 88(423):1034, 1993. ISSN 01621459. doi: 10.2307/2290796.

Green, Peter J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711.

Haff, I H; Aas, K, and Frigessi, A. On the Simplified Pair-Copula ConstructionSimply Useful or Too Simplistic? *Journal of Multivariate Analysis*, 2010. doi: 10.1016/j.jmva.2009.12.001.

Jeong, D.Y.; Lyons, M.L.; Orringer, O, and Perlman, A.B. Equations of motion for train derailment dynamics. *Proceedings of the 2007 ASME Rail Transportation Division Fall Technical Conference, September 11-12, 2007 Chicago, IL*, RTDF2007-4: 1–7, 2007. ISSN 10788883. doi: 10.1115/RTDF2007-46009.

Joe, Harry. Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. pages 120–141. Institute of Mathematical Statistics, 1996. ISBN 0-940600-40-4. doi: 10.1214/lnms/1215452614.

Joe, Harry. *Multivariate models and dependence concepts*. Chapman & Hall, 1997. ISBN 9780412073311.

Krämer, Nicole and Schepsmeier, Ulf. Introduction to vine copulas, 2011.

Liu, Xiang. Statistical Temporal Analysis of Freight-Train Derailment Rates in the United States : 2000 to 2012. 2476(1):119–125, 2015.

Liu, Xiang; Saat, M., and Barkan, Christopher. Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*, 2289(2289):154–163, 2012. ISSN 0361-1981. doi: 10.3141/2289-20.

Liu, Xiang; Saat, M. Rapik; Qin, Xiao, and Barkan, Christopher P L. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. *Accident Analysis and Prevention*, 59:87–93, 2013. ISSN 00014575. doi: 10.1016/j.aap.2013.04.039.

Min, Aleksey and Czado, Claudia. Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4):511–546, 2010. ISSN 14798409. doi: 10.1093/jfinec/nbp031.

Mohammadzadeh, Saeed and Ghahremani, Soodabeh. Estimation of train derailment probability using rail profile alterations. *Structure and Infrastructure Engineering*, 2479(August 2013):1–20, 2010. ISSN 1573-2479. doi: 10.1080/15732479.2010.500670.

Monstvilaite, Monika. *Portfolio Value-at-Risk Using Regular Vine Copulas Matematiska institutionen*. PhD thesis, Stockholm University, Stockholm, Sweden, 2016.

Savu, Cornelia and Trede, Mark. Hierarchies of Archimedean copulas. *Quantitative Finance*, 10(3):295–304, 2010. ISSN 14697688. doi: 10.1080/14697680902821733.

Schepsmeier, Ulf. *Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different families*. PhD thesis, Technical University of Munich, 2010.

Schepsmeier, Ulf and Czado, Claudia. Dependence modelling with regular vine copula models: A case-study for car crash simulation data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 65(3):415–429, 2016. ISSN 14679876. doi: 10.1111/rssc.12125.

Schwarz, Gideon. Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461–464, 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136.

Stöber, Jakob; Joe, Harry, and Czado, Claudia. Simplified pair copula constructions-Limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118, 2013. ISSN 0047259X. doi: 10.1016/j.jmva.2013.04.014.

Stöber, Jakob; Hong, Hyokyoung Grace; Czado, Claudia, and Ghosh, Pulak. Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics and Data Analysis*, 88:28–39, 2015. ISSN 01679473. doi: 10.1016/j.csda.2015.02.001.

Vuong, Quang H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307, 1989. ISSN 00129682. doi: 10.2307/1912557.

Wei, Wei. *Copula-based High Dimensional Dependence Modelling*. PhD thesis, University of Technology, Sydney, Australia, 2014.

Yan, Jun. Multivariate Modeling with Copulas and Engineering Applications. In *Springer Handbook of Engineering Statistics*, pages 973–990. Springer London, London, 2006. doi: 10.1007/978-1-84628-288-1_51.

Chapter 7

CONCLUDING REMARKS

7.1 Introduction

In many infrastructure applications of data analysis including railroad applications such as track geometry recovery and derailment severity; non-normality of data transpires in various forms. These include non-normality of the marginal distribution of some variables and in some instances multivariate non-normality of the joint distribution of a group of variables despite normal marginal distributions of all the individual variables (Yan, 2006; Attoh-Okine, 2013). Copulas describe the dependence structure between variables are generally suitable for analyzing the dependence between variables with non-normality. Copulas allow for the separate modeling of arbitrary marginal distributions and the dependence structure. Copulas are suitable for modeling various forms of dependence including tail dependence and asymmetric dependence. Copulas (copula-based methodologies) can be applied as standalone models or in tandem with other alternate models.

This dissertation provides detailed copula analysis of various railroad maintenance and safety applications such as track geometry recovery and derailment severity. State-of-the-art literature review of existing track geometry (tamping) recovery models and derailment severity models is presented with a discussion of the gaps in the literature also provided. This dissertation has introduced copula-based approaches as a technique that can be used in describing and analyzing the underlying dependence between the variables of interest in various railroad engineering applications. Three copula-based methodologies namely bivariate copula models, copula-based regression models and vine copula models were applied to railroad (maintenance and safety) applications. A (bivariate) copula-based approach was developed to evaluate (estimate)

the tamping recovery of track geometry parameters such as surface (longitudinal level), alignment, cross level, gage, and warp. A joint mixed copula-based regression model for derailed cars and monetary damage is developed for the combined analysis of their relationship with a set of covariates that might affect both severity outcomes. Vine copulas, a cascade of bivariate copulas as building blocks, was used to model the high-dimension dependencies within the derailment severity data taking into account the non-linearities in the data.

7.2 Conclusions

The following conclusions were drawn from this dissertation and can divided into three main parts namely:

- Tamping Recovery of Track Geometry
- Bivariate Derailment Severity using Copula-based Regression Models
- Dependence Modeling of Derailment Severity using Vine Copulas

7.2.1 Tamping Recovery of Track Geometry

- Based on exploratory data analysis as well as confirmatory data analysis, the marginal and joint distributions of the variables of interest were found to be not normal.
- From the marginal fitting results, the recoveries of the various parameters were found to be non-normal and were found to either fit a three-parameter log-normal distribution (in the case of surface, alignment, and warp) or three-parameter log-logistic distribution (in the case of cross level and gage). Similarly, non-normal distributions were observed for the track quality condition (SD of track geometric parameters) before and after tamping.
- Tail and asymmetric dependences between the variables of interest were identified by the selected copulas. For instance, the track geometry recovery of the surface parameter was found to be poorly correlated at low values but highly correlated at high values.

- Correlation analysis of the recovery of various geometry parameters show that the use of Pearson's correlation coefficient which assumes normality of the variables and linear dependence led to relatively high dependence values observed. However, the use of concordance measures such as Kendall's Tau and Spearman's Rho resulted in a general reduction in the observed dependences. These concordance measures are scale-invariant and are suitable for evaluating non-linear dependence and measure dependence irrespective of assumed distribution. Thus, the widely-used Pearson's correlation coefficient does not appear to be appropriate for analyzing the correlation between the recoveries of the various track geometry parameters.
- From the correlation analysis results, the strongest correlation was observed between warp and cross level recoveries with the weakest correlation observed between the surface and gage recoveries with varying levels in-between. This infers and gives credence to previous research that tamping affects the various track geometry parameters differently thus it is imperative to examine all the track geometry parameters and not focus on one or two parameters.
- Copula modeling can be used in generating large volumes of data with similar dependence patterns as the original track geometry data set.
- In general, conventional correlation analysis does not appear to be suitable for analyzing the dependences between the variables of interest.
- Copulas appropriately model tamping recovery phenomenon taking into consideration the underlying dependence between the variables.

7.2.2 Bivariate Derailment Severity using Copula-based regression models

- Failure to consider for the dependence between the severity outcomes may lead to biased or distorted coefficient estimates in derailment severity models.
- The incorporation of copulas in derailment severity models have a greater influence on the dispersion/variance estimates than the point estimates.
- Derailment speed was found to have most pronounced effect on both monetary

damage and number of derailed cars. This was followed by residual train length and loading factor.

- In general, the covariates were found to have a greater effect on the monetary damage outcome than the number of derailed cars.
- Results enable objective comparison of different train safety approaches that could be used to inform decision making. An argument can be made for the reduction of freight train speeds in favor of a reduction in the number of cars in a train consist.
- Combining the marginal regression models of the two derailment severity outcomes with the underlying dependence facilitates a better comprehension of the train derailment severity distribution.

7.2.3 Dependence Modeling of Derailment Severity using Vine Copulas

- Some of the pairwise dependencies were found to show asymmetric and tail dependence violating the multivariate normality assumption.
- Both number of derailed cars and monetary damage were found to be poorly correlated with low values of derailment speed but highly correlated with much higher values.
- It was found that loading factor may be appropriate in predicting the total monetary damage incurred during derailments however caution is required when being used to predict the number of derailed cars.
- These vine copulas models were found to be better at modeling the derailment data in comparison with multivariate Gaussian copulas which assume multivariate normality.
- Vine copula methodology was used to generate large volumes of multivariate derailment severity data.

Copulas in general describe the dependence structure between the variables and thus are suitable for modeling various forms of dependence including linear dependence,

tail dependence, symmetric and asymmetric dependence. Furthermore, copula modeling is an emerging methodology which is suitable for analyzing the dependence between arbitrary marginal distributions. Thus, it is useful for analyzing the dependence between variables irrespective of the nature of the variables which might be non-normal, discrete, skewed, heavy-tailed or thin-tailed. This is particularly useful in the railroad industry since data take several forms (as shown in the exploratory data analysis in Chapter 3).

This dissertation offers major contributions for improving data analysis in railroad maintenance and safety by accounting for the underlying dependence between the variables of interest irrespective of the observed marginal distribution. The copula-based methodology allows the railroads to appropriately analyze the effectiveness of maintenance activities such as tamping at different track geometry quality or condition levels prior to maintenance. This is conducted without assuming constant (average) linear dependence between track geometry recovery and track quality before maintenance. For instance, upper tail dependence was observed between the track geometry recovery of parameters (such as surface and alignment) and the track condition before tamping. This suggests that hasty tamping or tamping at low standard deviation levels is not as effective as tamping at high standard deviation levels. This provides railroad maintenance managers further evidence for the need to optimally execute tamping by employing condition-based maintenance instead of time-based maintenance. Early tamping has been found to lead to shorter track lifecycle and the failure to attain track design capacity (Quiroga et al., 2012; Famurewa et al., 2013).

High dimensional copulas such as vine copulas provide analysts further insight into the complex interactions between various variables including the relationship between covariates which are often not considered critically. Bivariate or vine copula modeling is also useful for the generation of large volumes of data when it is not easy to obtain large data sets for analysis. This is particularly useful for railroad data analysts and researchers who can use the large data sets generated to analyze track geometry degradation, track geometry restoration and train derailment severity.

Copula models are also useful since they can be combined with other models. For instance, bivariate (or vine) copulas can be combined with generalized linear models to form copula-based regression models or with Bayesian Networks to form Copula Bayesian Networks. Copula-based regression models allow railroad safety regulators to simultaneously model and predict several derailment severity outcomes conditional on a set of covariates or factors while taking into consideration the underlying dependence between the outcomes which may be skewed or discrete in nature. This allows safety analysts to evaluate the effect of the individual covariates on the joint distribution of the derailment severity outcomes. For instance, derailment speed was found to have the greatest effect on both derailment severity outcomes followed by residual train length and finally loading factor. This permits the objective comparison of different train safety approaches from a severity standpoint that could be used to inform decision making made by railroad safety regulators from government and industry.

7.3 Future Research

Despite the major contributions of this dissertation, there are vast opportunities for future implementation of copula-based methodologies in railroad engineering particularly railroad maintenance and safety applications. Recommendations and future work can divided into two main sections namely

- Recovery and Degradation Modeling of Track Geometry
- Derailment Severity Modeling

7.3.1 Recovery and Degradation Modeling of Track Geometry

- Incorporate the copula-based tamping recovery model into track geometry maintenance scheduling model with the track geometry degradation models and recovery models being the main components of the model.
- Extend the copula-based methodology to incorporate others factors such as operational speed, tamping procedure, age of track components, and number of previous tamping operations. To analyze the effect of various covariates on track

geometry recoveries of several parameters, copula-based regression models can be employed taking into consideration the dependence between the response variables. In order to analyze the dependence between more than two variables, vine copulas are recommended which are more flexible than regular multivariate copulas.

- There is the need to select an appropriate track geometry deterioration model that takes into consideration both the time and spatial variation of the track geometry degradation process.
- Develop copula autoregressive models as an appropriate track geometry degradation model that takes into account both of the time and spatial variation of the track geometry degradation process. To employ copula autoregressive models, there is the need for substantial data points between tamping interventions in order to effectively model the degradation of track geometry.
- In order to integrate such degradation models and copula-based recovery models in track maintenance scheduling models, probabilistic optimization models need to be considered.

7.3.2 Derailment Severity Modeling

- Extend derailment severity copula-based regression methodology to consider bi-parametric copulas such as Student t-copula, Clayton-Gumbel (BB1), Joe-Gumbel (BB6), Joe-Clayton (BB7) and Joe-Frank (BB8) copulas.
- Extend derailment severity copula-based regression methodology to consider other marginal distributions (such as lognormal distribution for monetary damage and zero-truncated negative binomial and Poisson-inverse Gaussian for the number of derailed cars).
- Extend the methodology to consider other derailment severities such as casualties by developing vine-copula based regression models in order to simultaneously model more than two response variables given a set of covariates.
- Extend the methodology to investigate the effect of other covariates on train

derailment severity such as train power distribution, rail friction, ground friction and car mass.

- Develop copula additive models in order to consider various types of covariate effects other than linear relationships by allowing a variety of non-parametric smoothing functions. The other covariate effects that can be considered include non-linear, random and spatial effects. The (dependence and marginal) model parameters of copula are permitted to be dependent on additive predictors which incorporate these covariate effects (Marra and Radice, 2017).

In summary, results from this dissertation provide greater insight and comprehension of the train derailment severity and track geometry recovery phenomena considering various forms of dependence between the variables of interest. These results will aid decision making which would help reduce the consequences of train derailments as well as improve track maintenance strategies.

REFERENCES

Attoh-Okine, Nii O. Pair-copulas in infrastructure multivariate dependence modeling. *Construction and Building Materials*, 49:903–911, 2013. ISSN 09500618. doi: 10.1016/j.conbuildmat.2013.06.055.

Famurewa, S. M.; Xin, T.; Rantatalo, M., and Kumar, U. Optimisation of maintenance track possession time: A tamping case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(1):12–22, 2013. ISSN 0954-4097. doi: 10.1177/0954409713495667.

Marra, Giampiero and Radice, Rosalba. Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113, 2017. ISSN 01679473. doi: 10.1016/j.csda.2017.03.004.

Quiroga, L. M.; Schnieder, E., and Antoni, M. Holistic long term optimization of maintenance strategies on ballasted railway track. In *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012*, 2012.

Yan, Jun. Multivariate Modeling with Copulas and Engineering Applications. In *Springer Handbook of Engineering Statistics*, pages 973–990. Springer London, London, 2006. doi: 10.1007/978-1-84628-288-1_51.

Appendix A

TRACK GEOMETRY EXPLORATORY DATA ANALYSIS

A.1 Foot-by-foot measurements

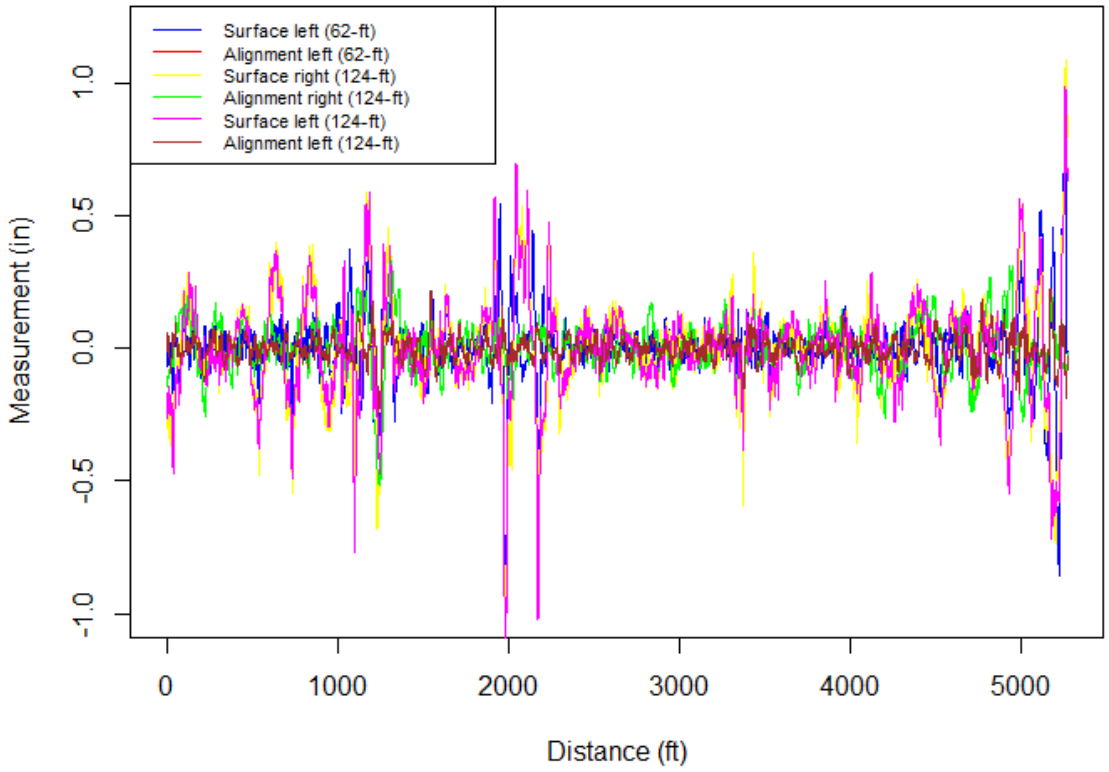


Figure A.1: Illustration of spatial variation of various track geometry parameters at a given inspection date

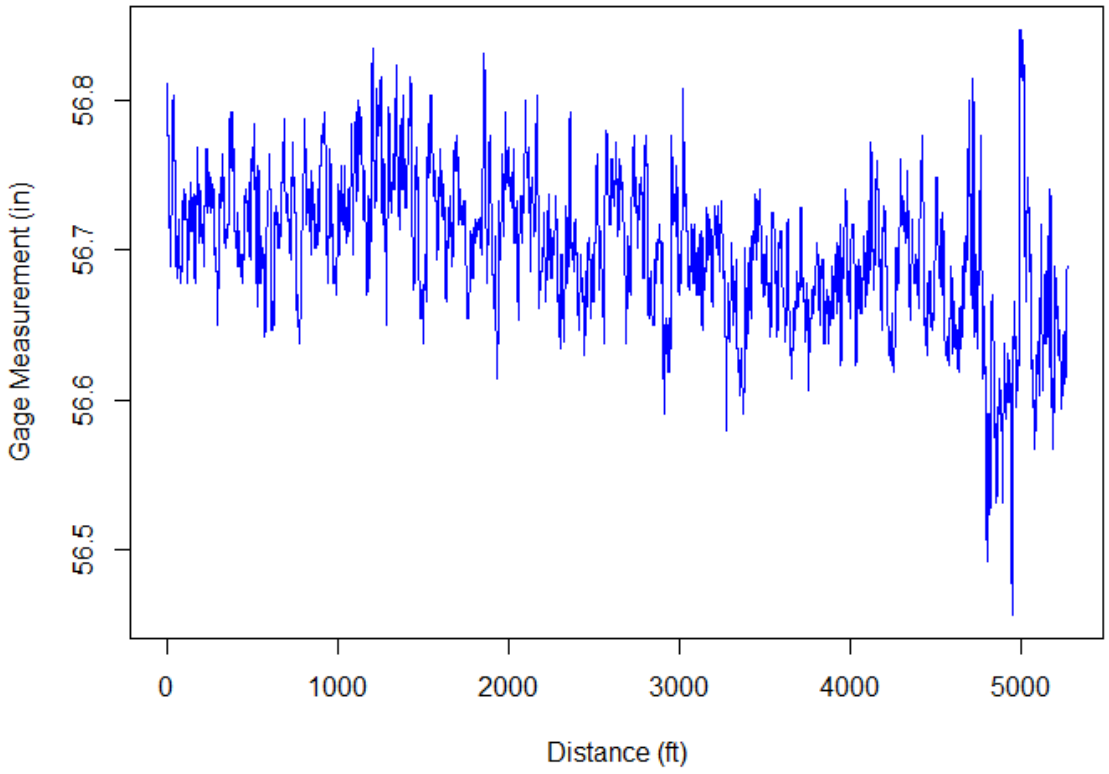


Figure A.2: Illustration of spatial variation of gage at a given inspection date

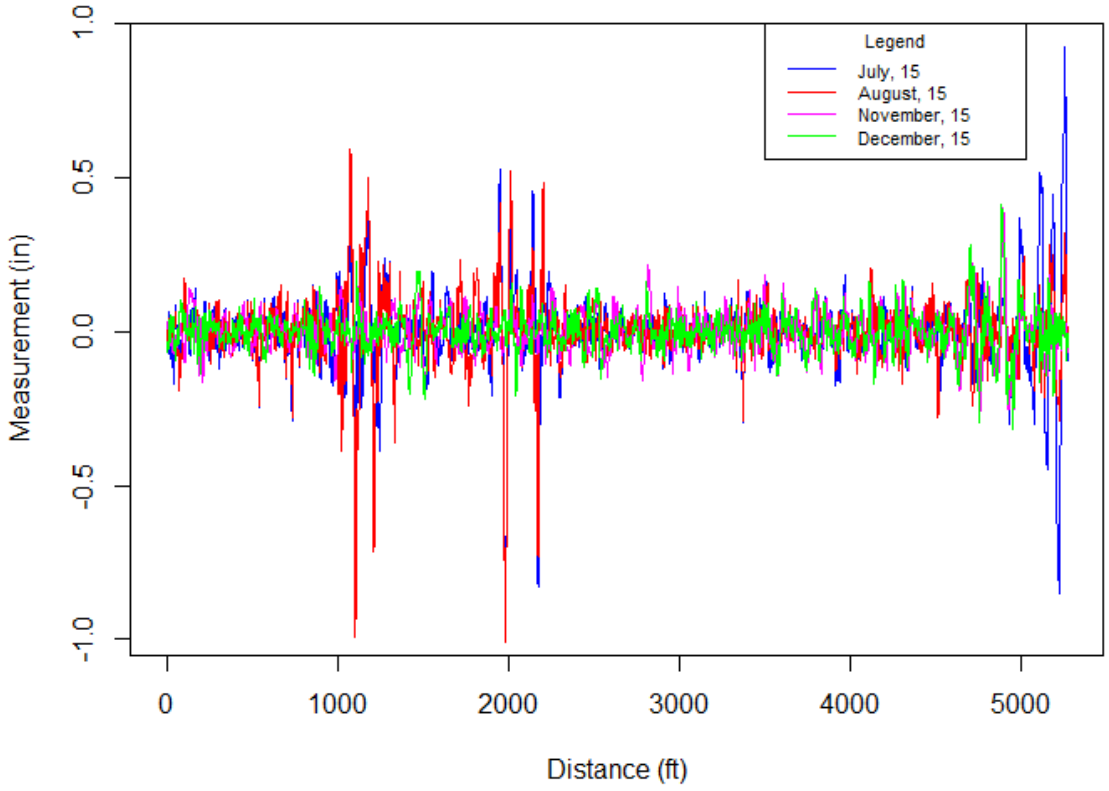


Figure A.3: Illustration of surface left (62-ft) track geometry parameter at multiple inspecton dates

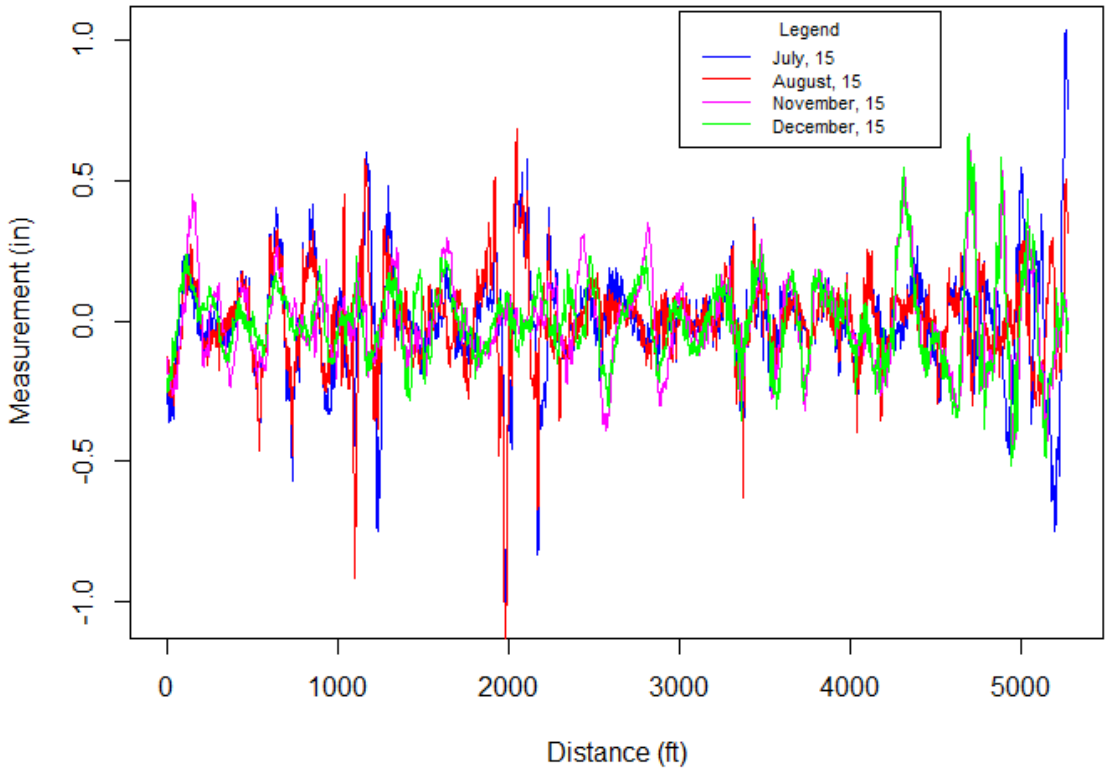


Figure A.4: Illustration of surface right (124-ft) track geometry parameter at multiple inspecton dates

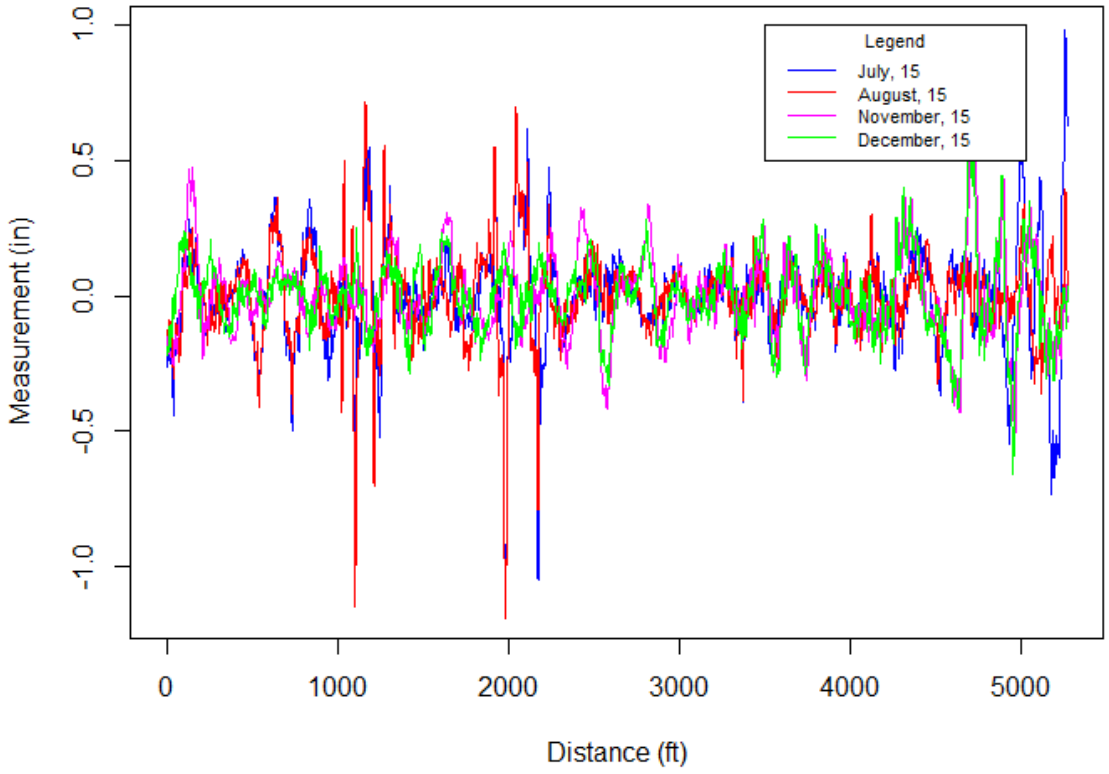


Figure A.5: Illustration of surface left (124-ft) track geometry parameter at multiple inspecton dates

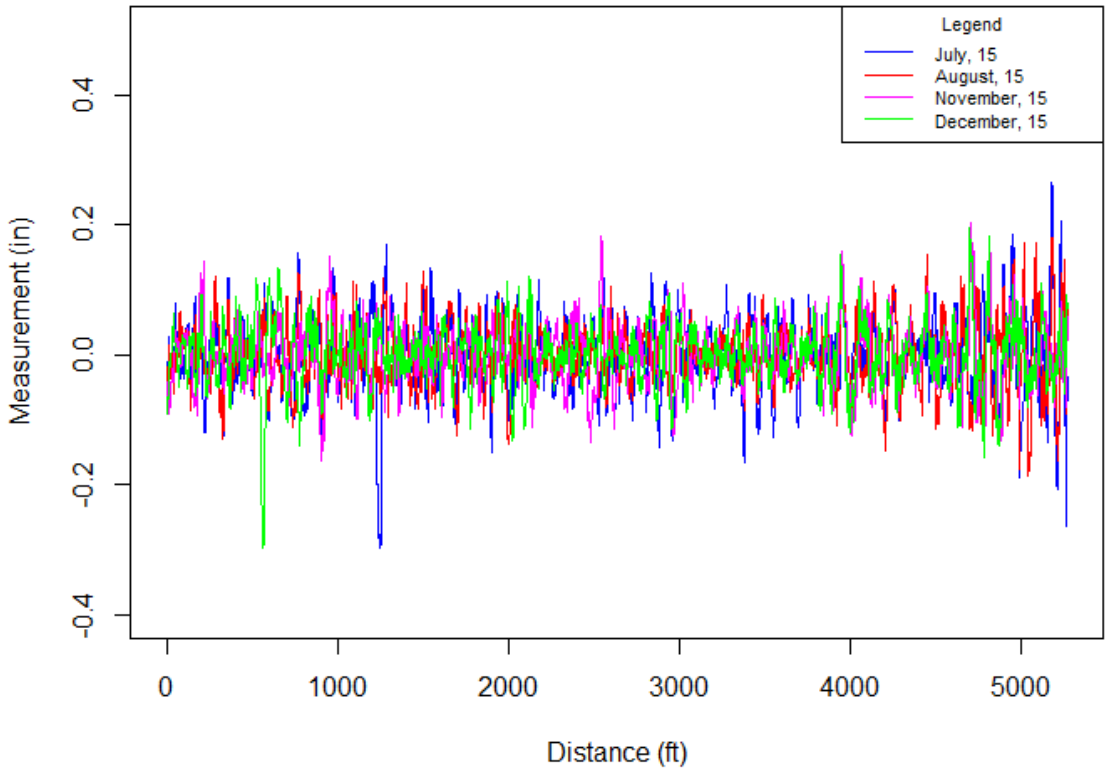


Figure A.6: Illustration of alignment right (62-ft) track geometry parameter at multiple inspecton dates

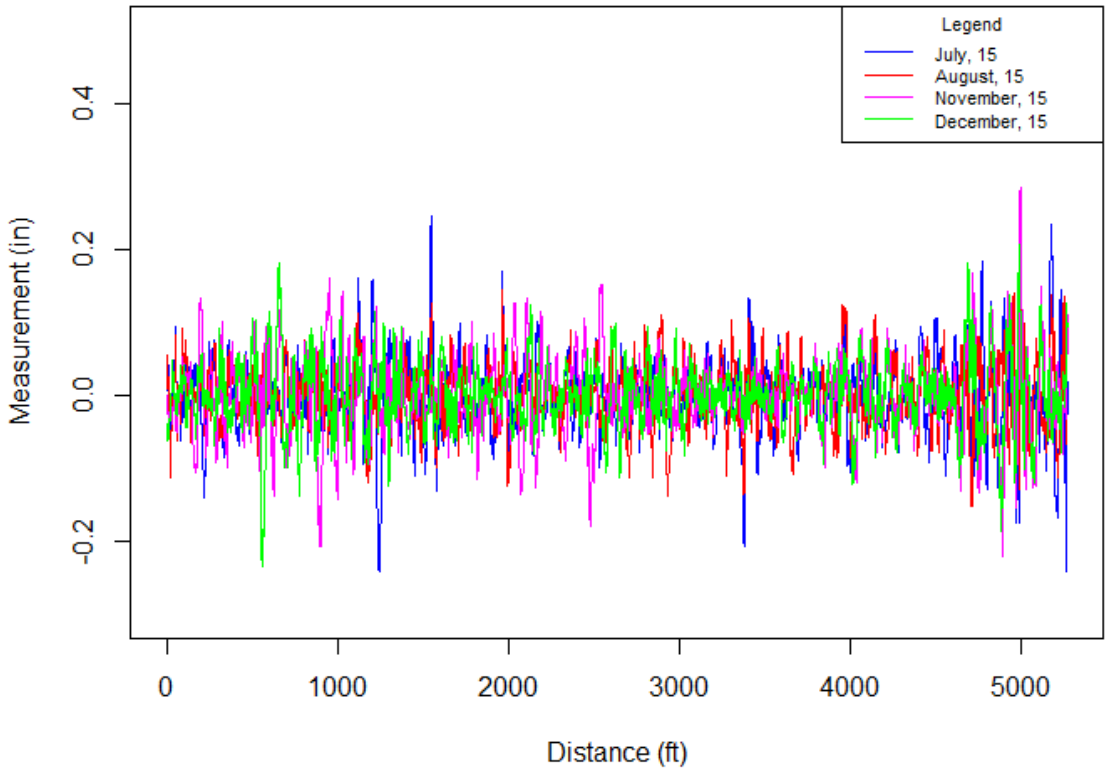


Figure A.7: Illustration of alignment left (62-ft) track geometry parameter at multiple inspecton dates

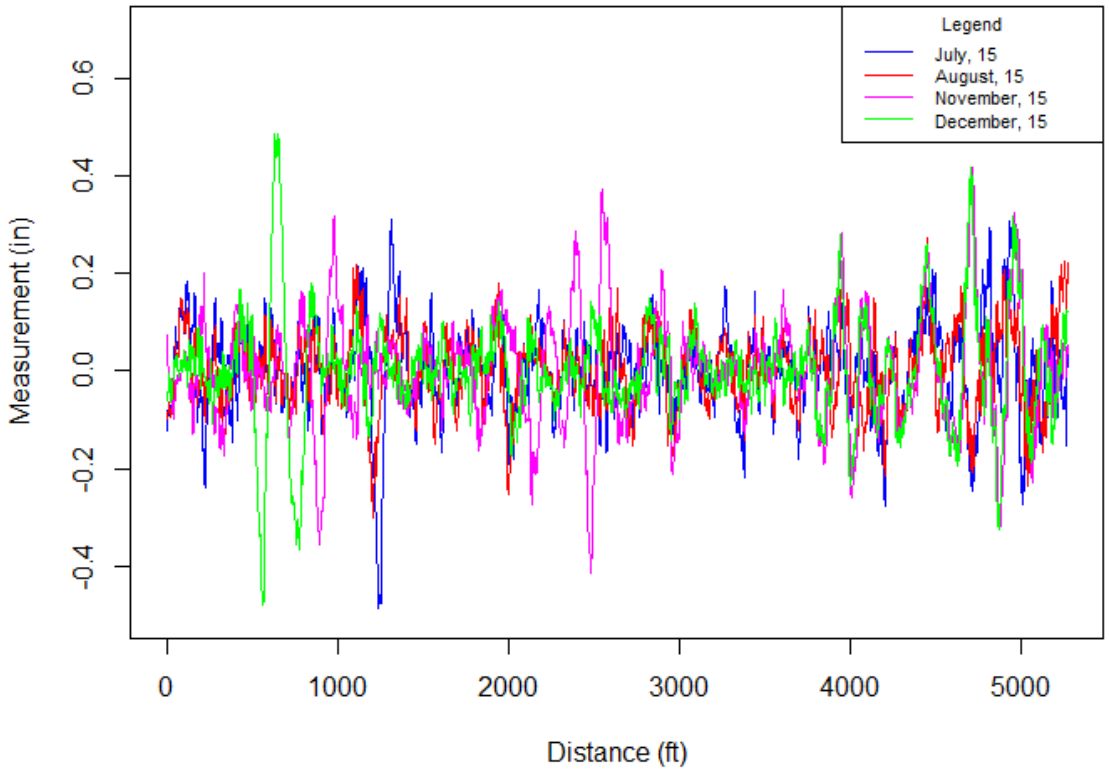


Figure A.8: Illustration of alignment right (124-ft) track geometry parameter at multiple inspecton dates

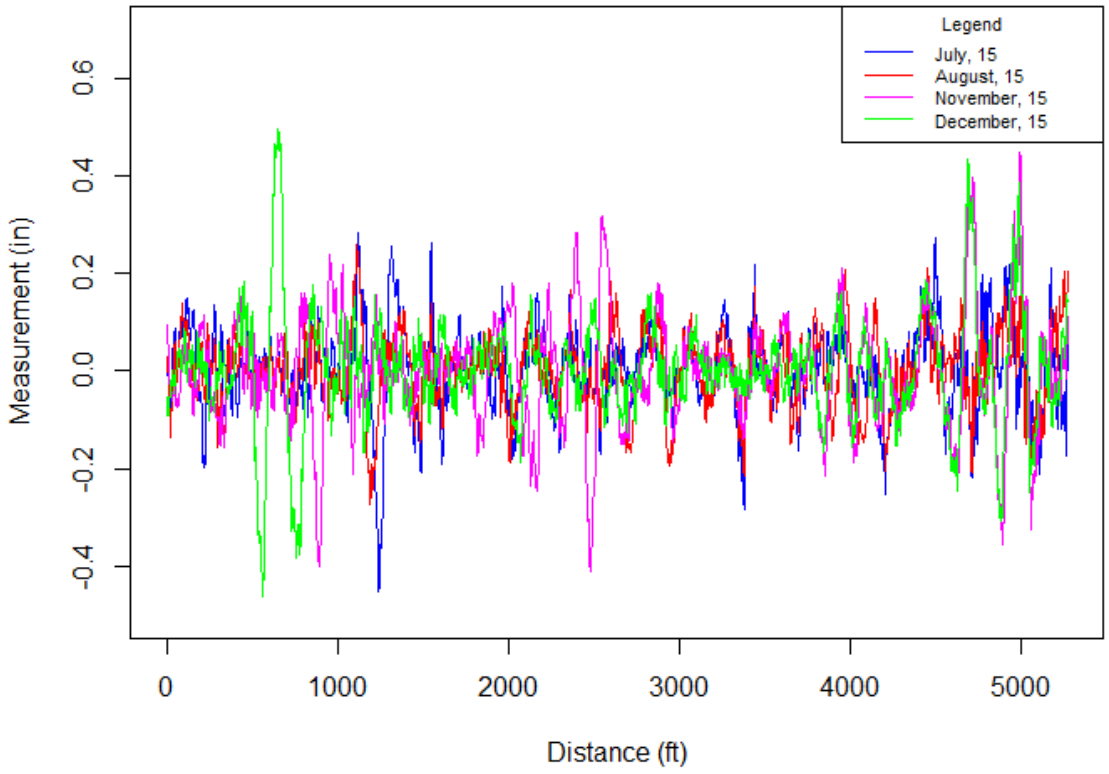


Figure A.9: Illustration of alignment left (124-ft) track geometry parameter at multiple inspecton dates

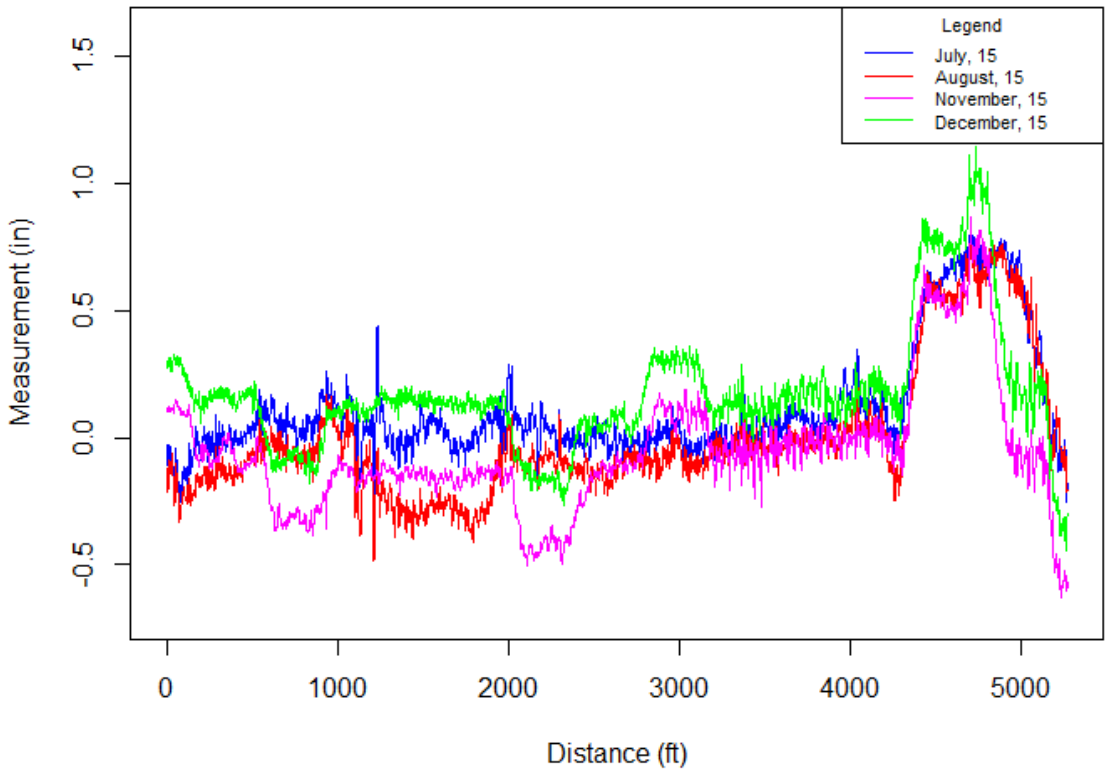


Figure A.10: Illustration of crosslevel track geometry parameter at multiple inspecton dates

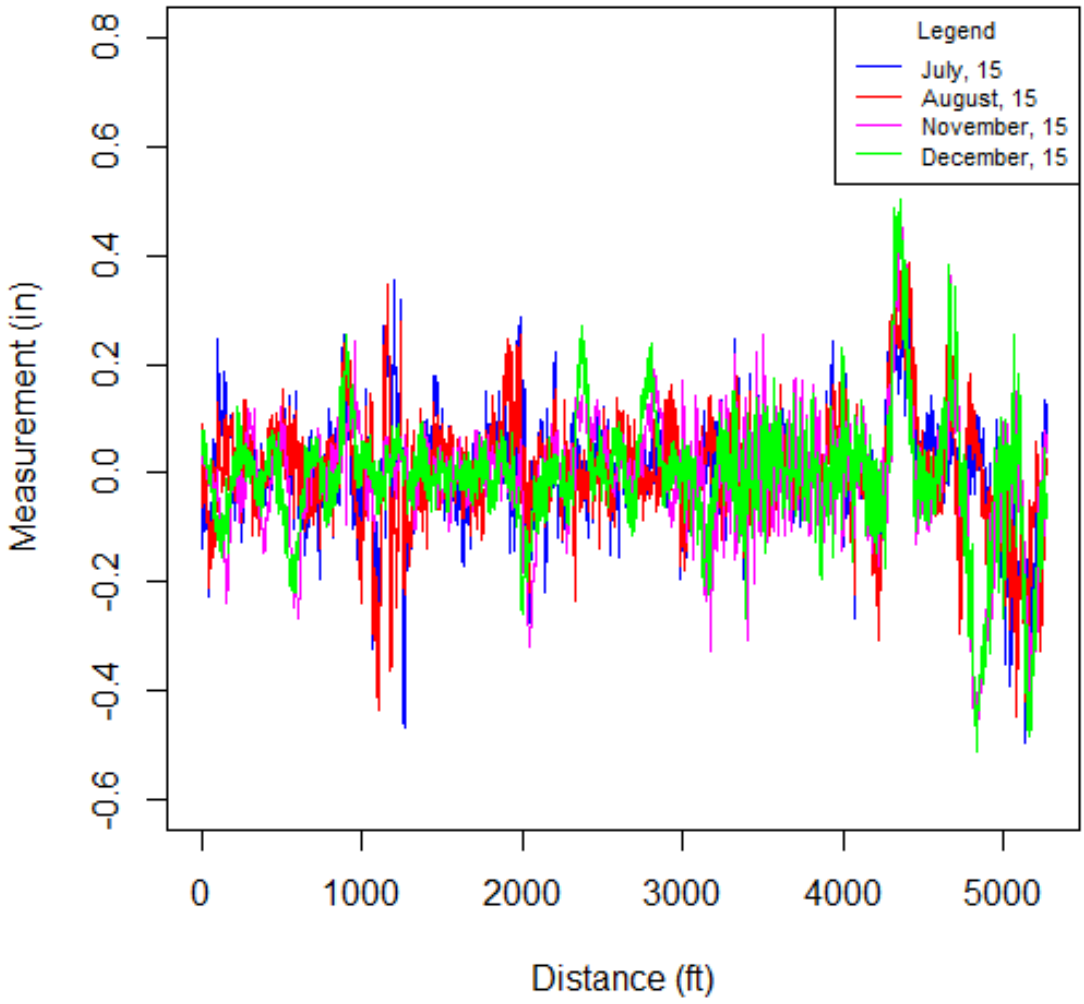


Figure A.11: Illustration of warp (62-ft) track geometry parameter at multiple inspecton dates

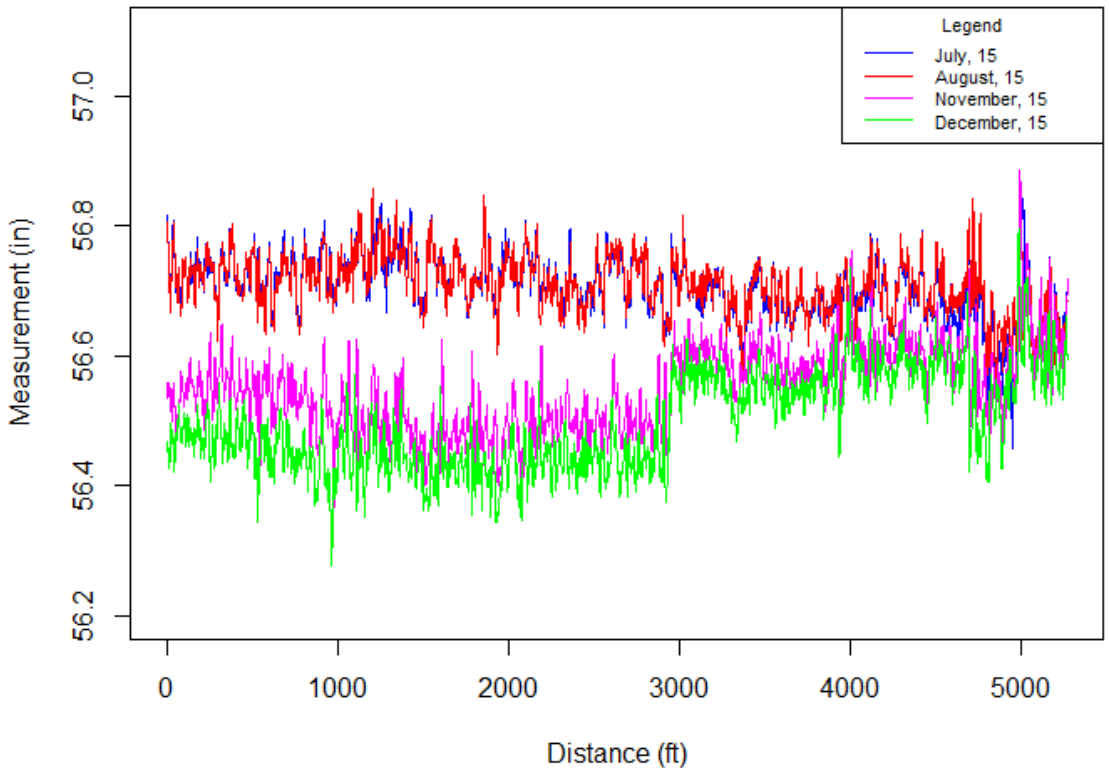


Figure A.12: Illustration of gage track geometry parameter at multiple inspecton dates

A.2 Histogram and Quantile-Quantile Plot

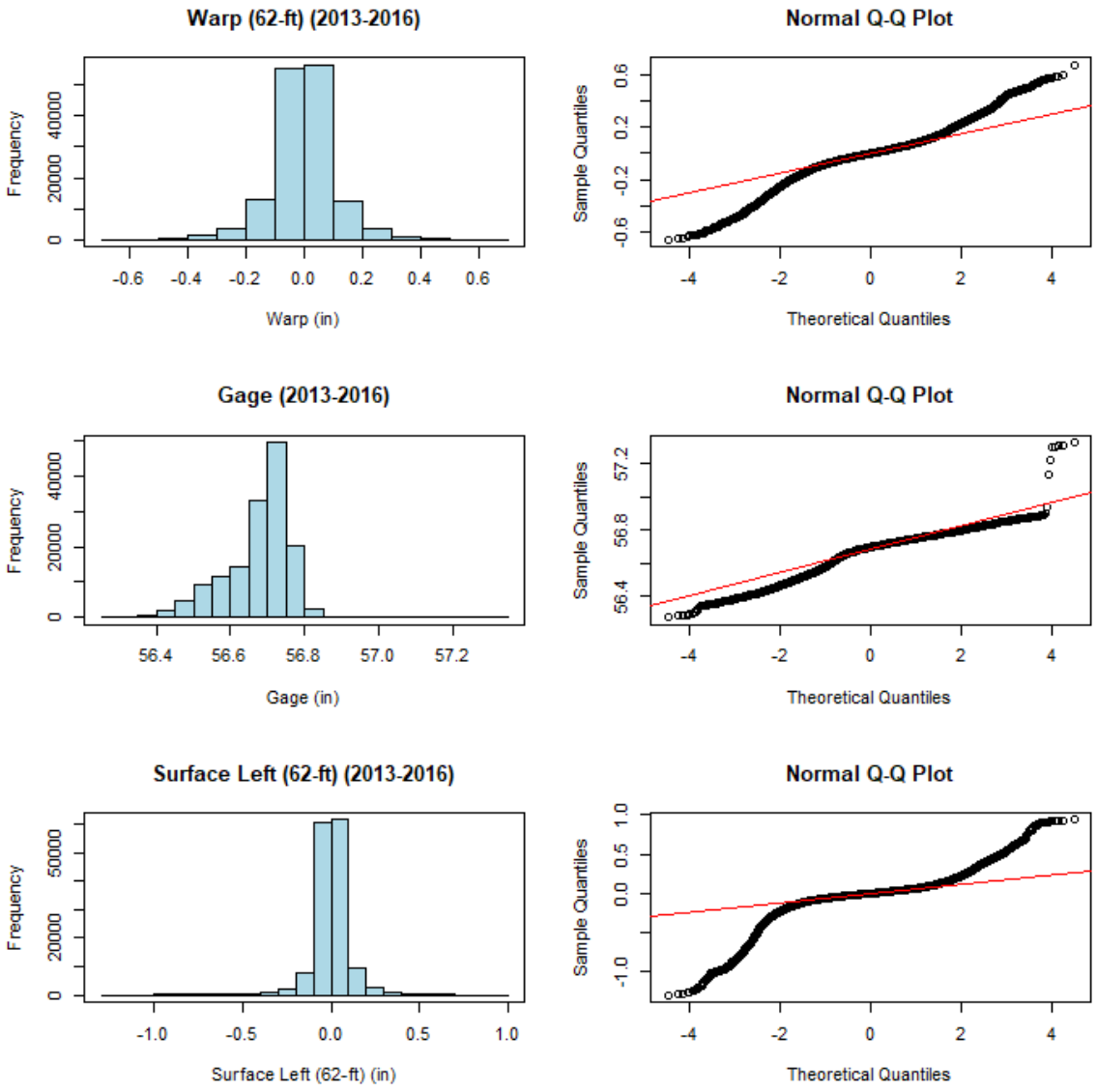


Figure A.13: Histograms and Q-Q plots for warp, gage and surface left (62-ft) data points from 2013 to 2016

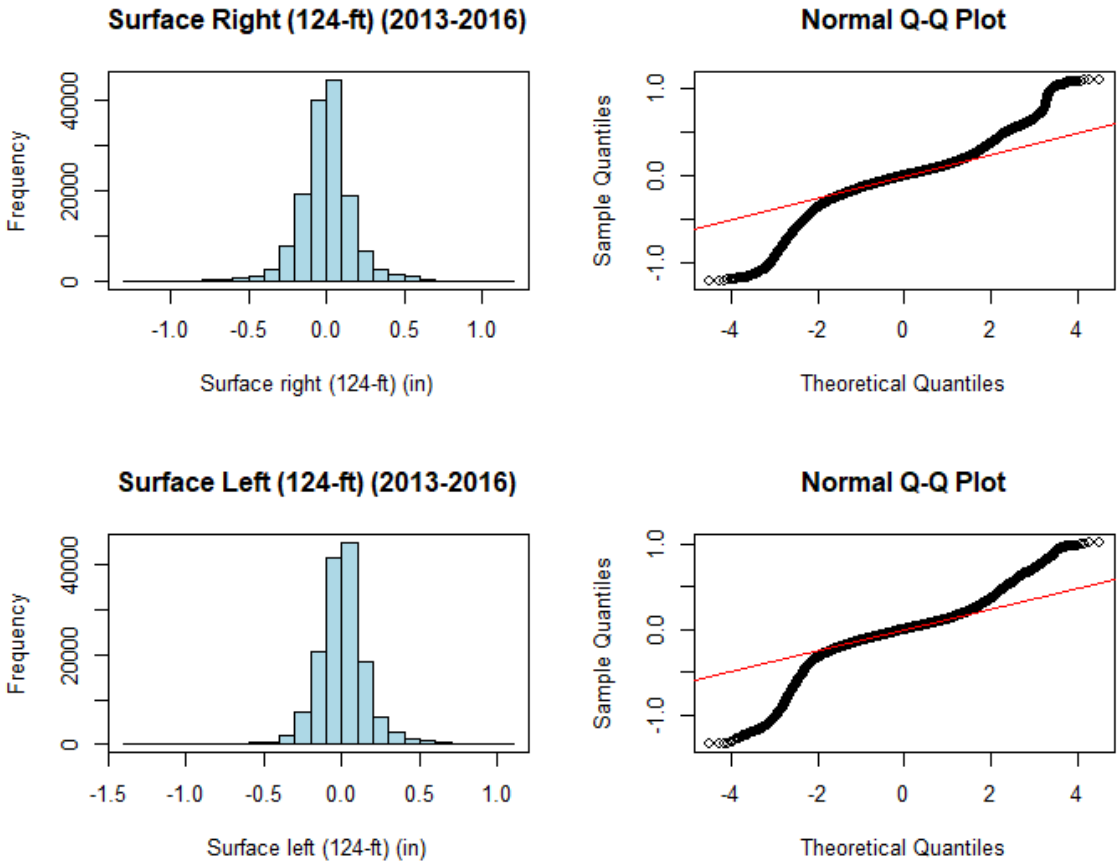


Figure A.14: Histograms and Q-Q plots for surface right (124-ft) and surface left (124-ft) data points from 2013 to 2016

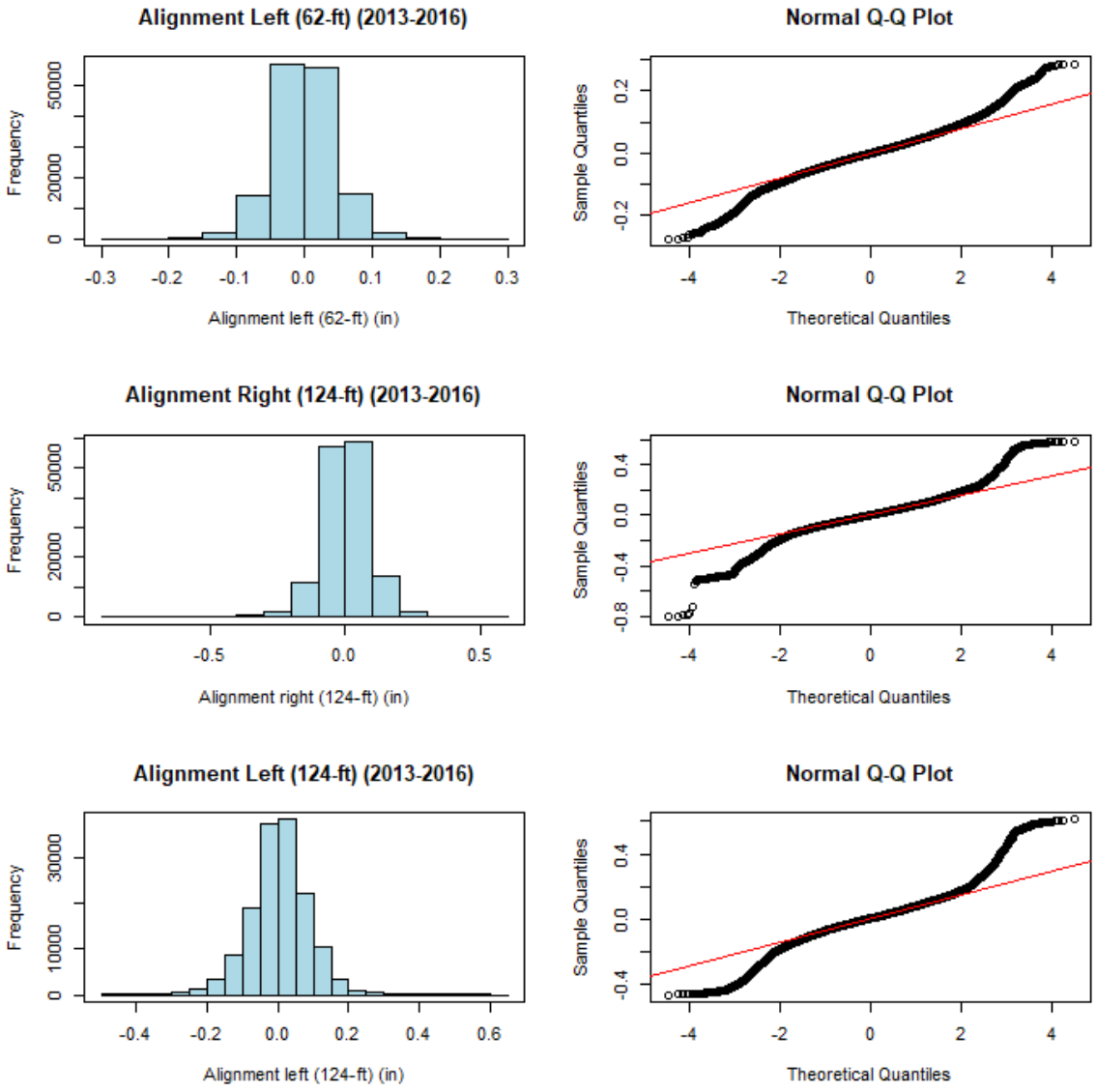


Figure A.15: Histograms and Q-Q plots for alignment left (62-ft), alignment right (124-ft) and alignment left (124-ft) data points from 2013 to 2016

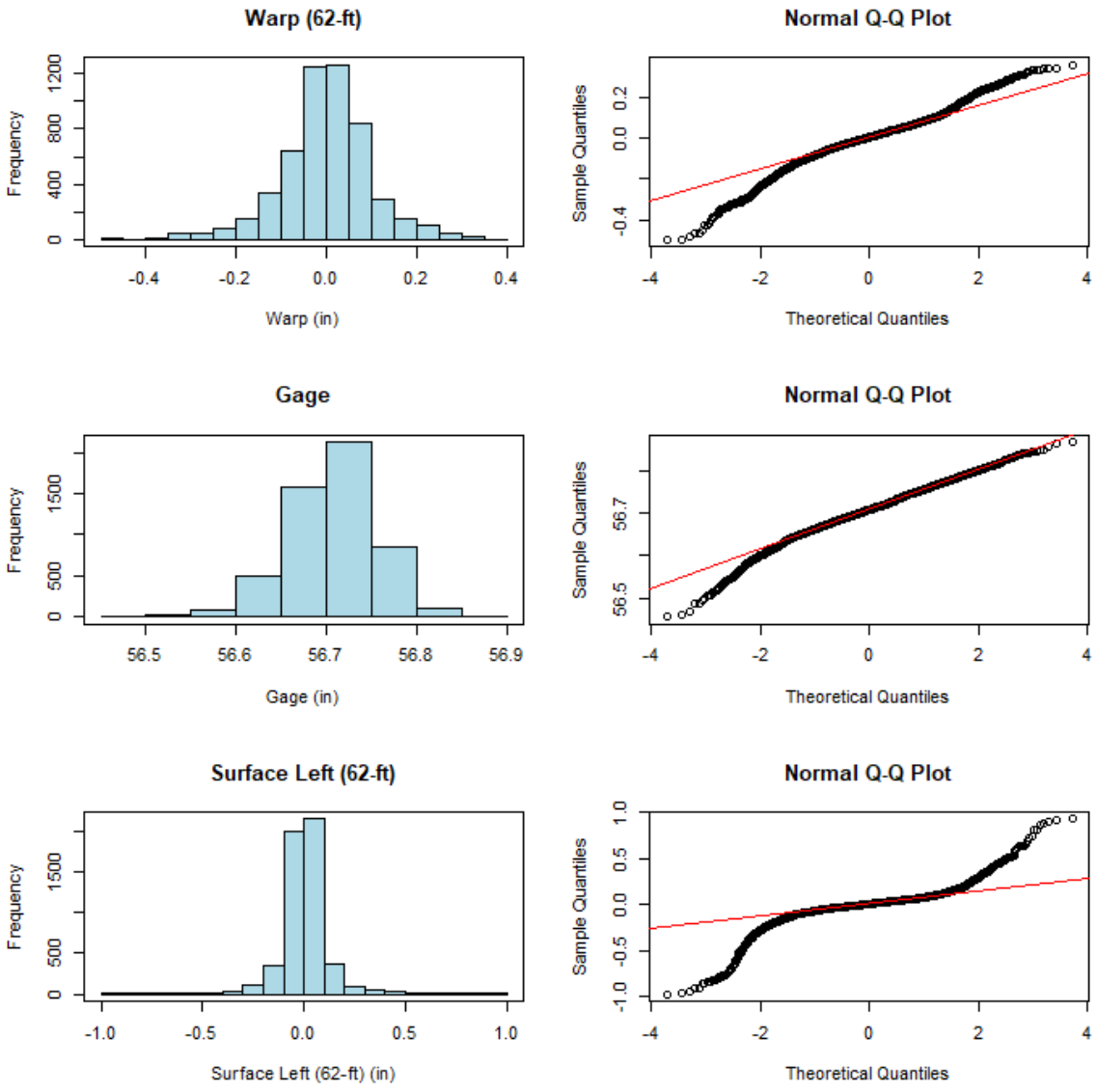


Figure A.16: Histograms and Q-Q plots for warp, gage and surface left (62-ft) data points for a given inspection date

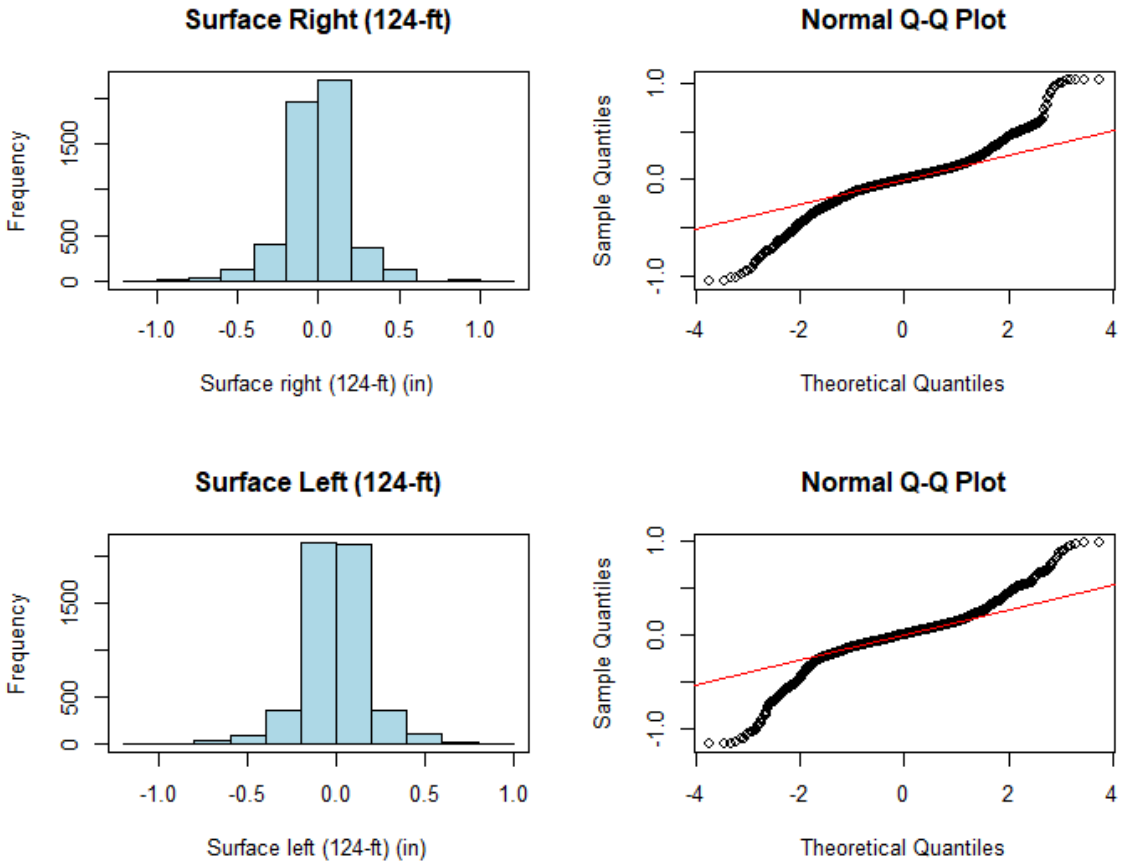


Figure A.17: Histograms and Q-Q plots for surface right (124-ft) and surface left (124-ft) data points for a given inspection date

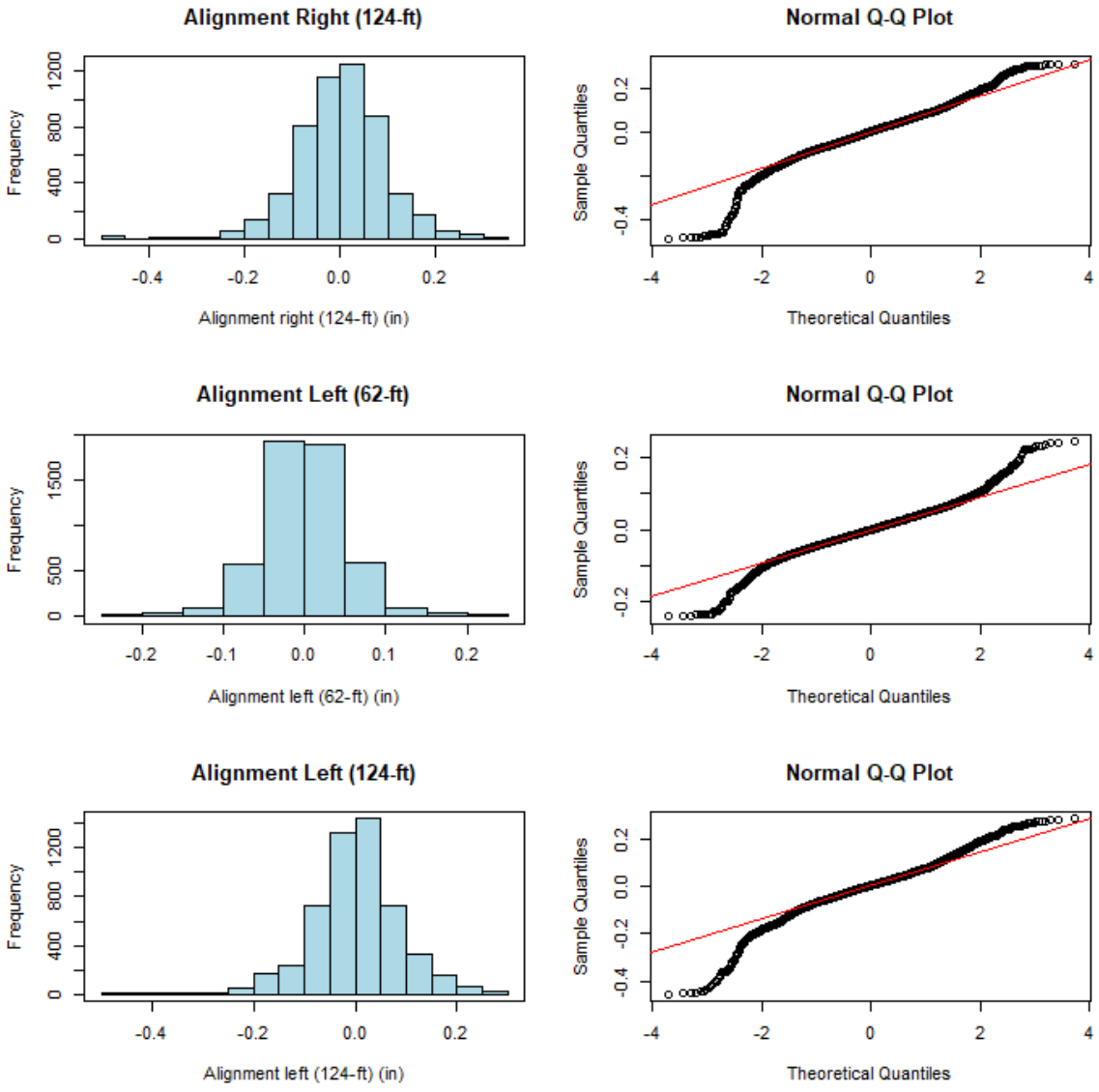


Figure A.18: Histograms and Q-Q plots for alignment left (62-ft), alignment right (124-ft) and alignment left (124-ft) data points for a given inspection date

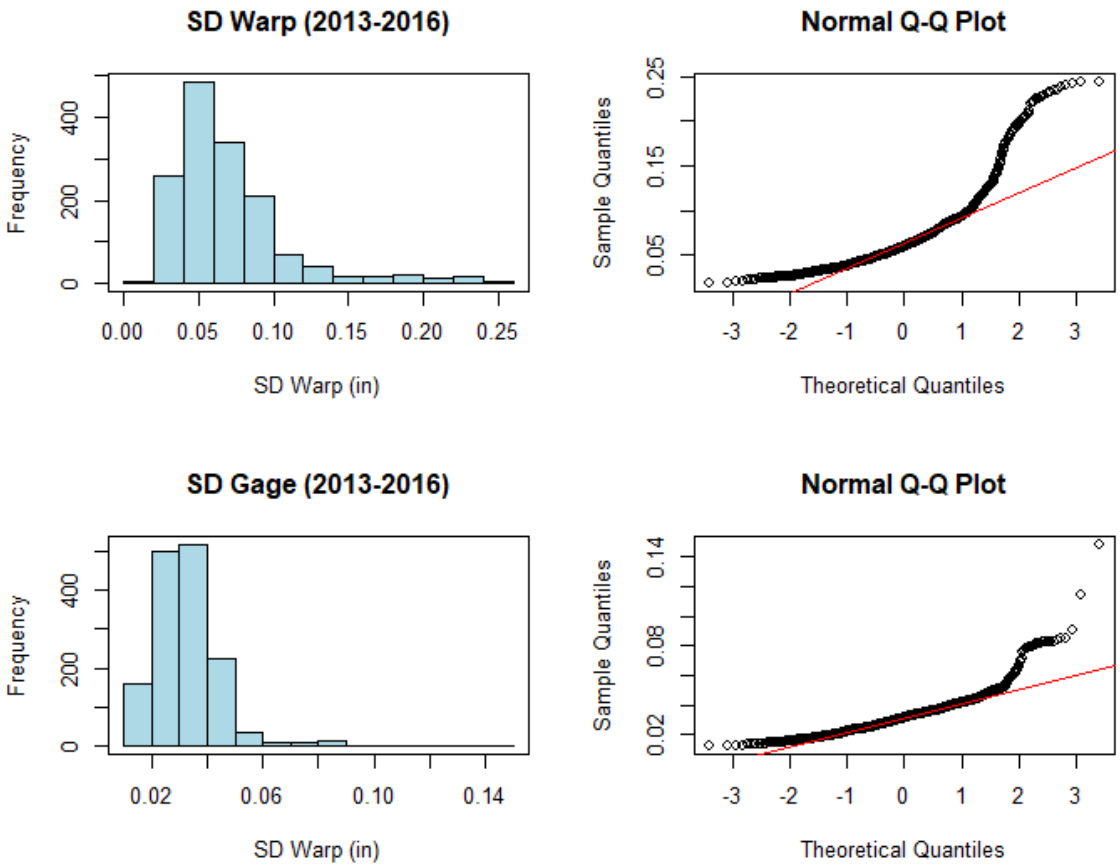


Figure A.19: Histograms and Q-Q plots for SD warp and SD gage data points from 2013 to 2016

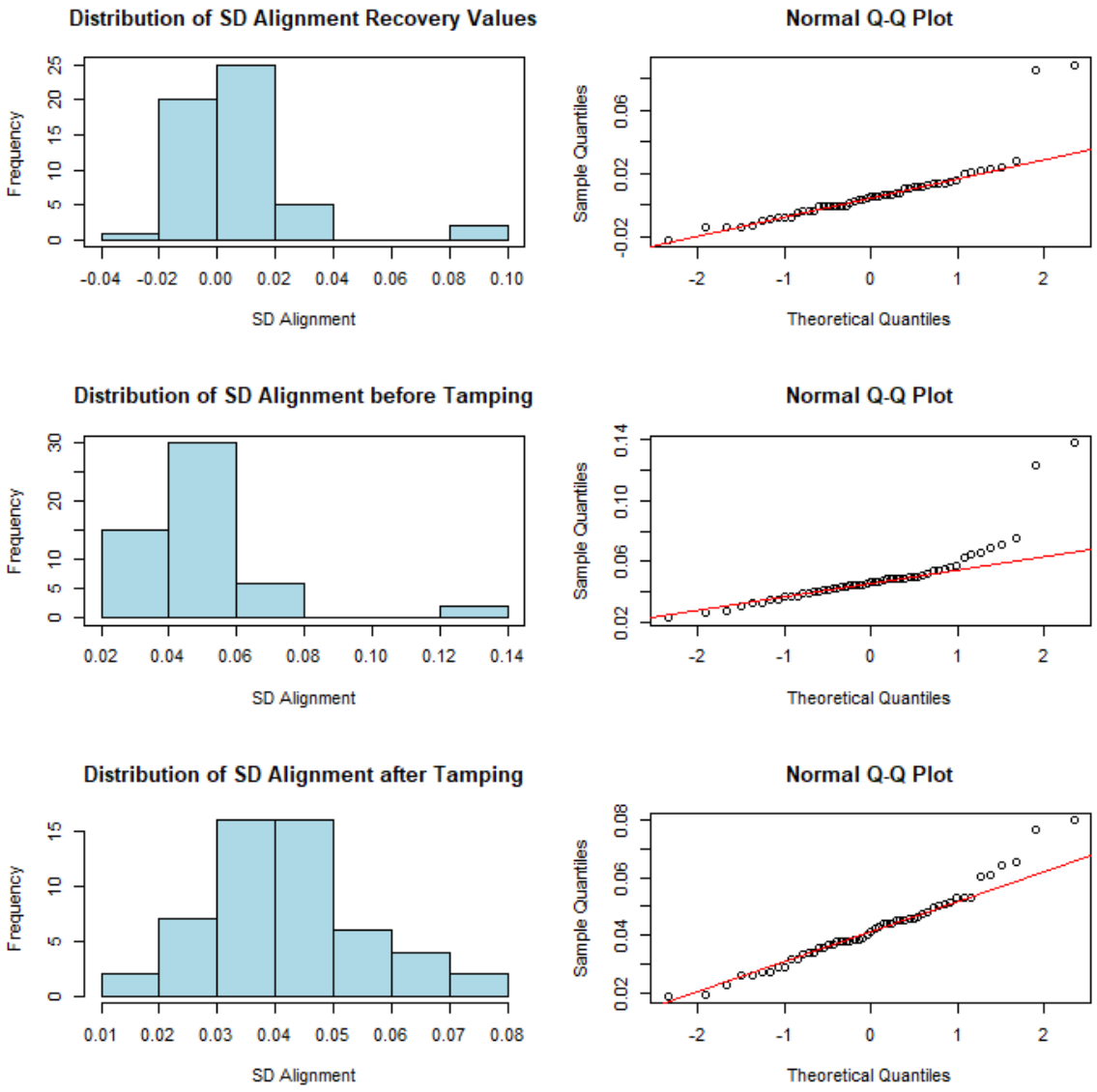


Figure A.20: Histograms and Q-Q plots for SD alignment recovery values, SD alignment before tamping and SD alignment after tamping

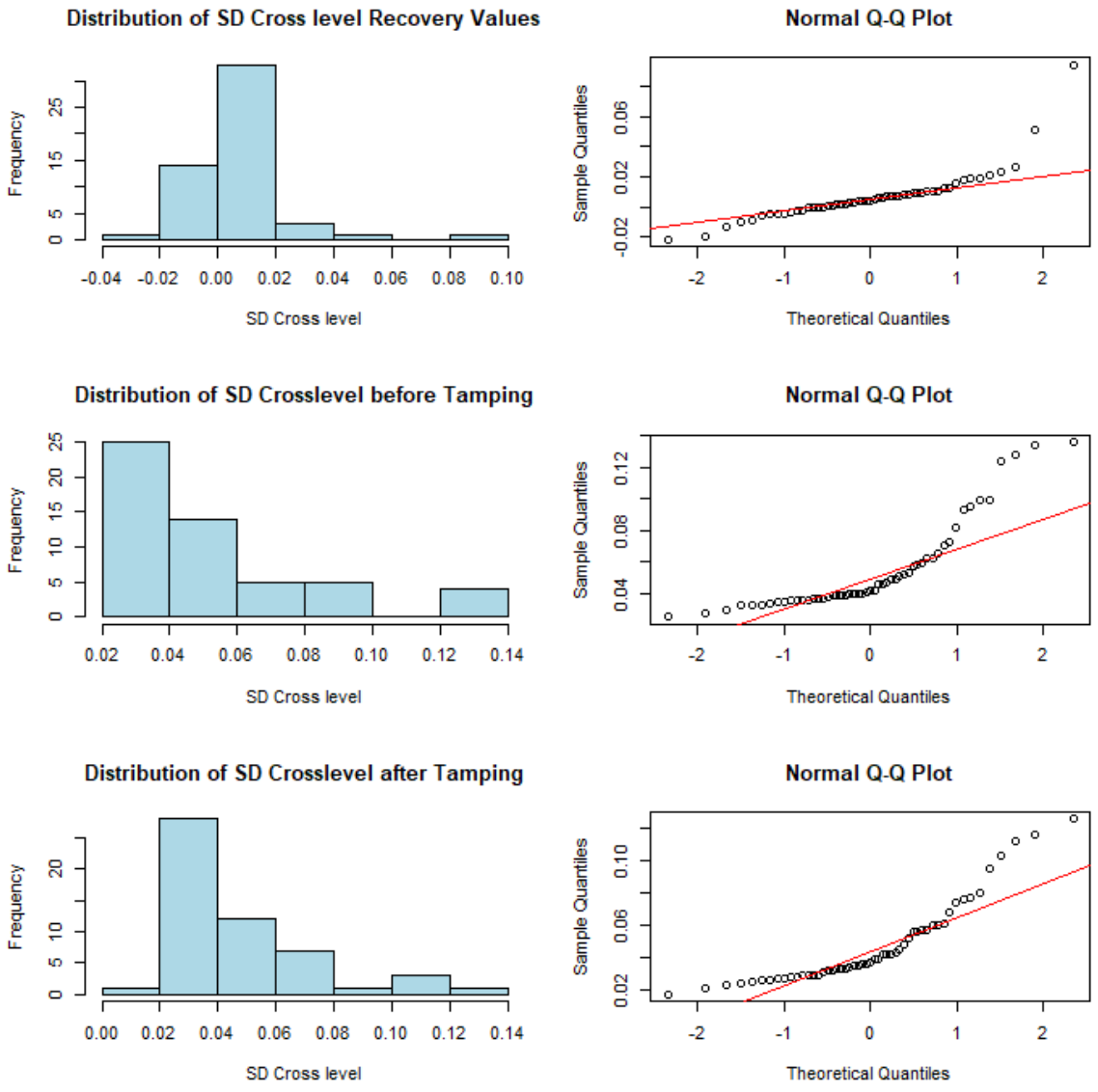


Figure A.21: Histograms and Q-Q plots for SD crosslevel recovery values, SD crosslevel before tamping and SD crosslevel after tamping

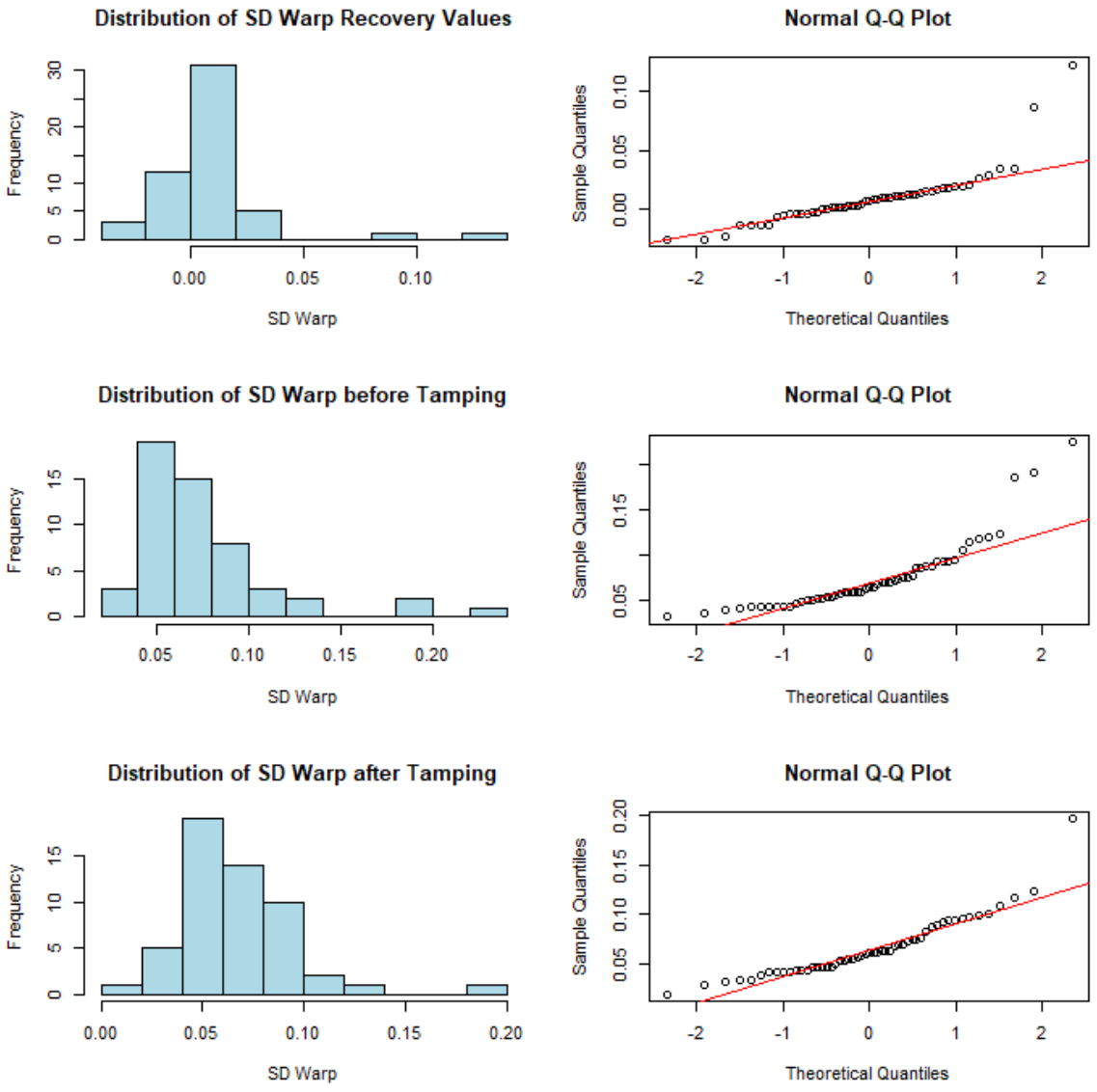


Figure A.22: Histograms and Q-Q plots for SD warp recovery values, SD warp before tamping and SD warp after tamping

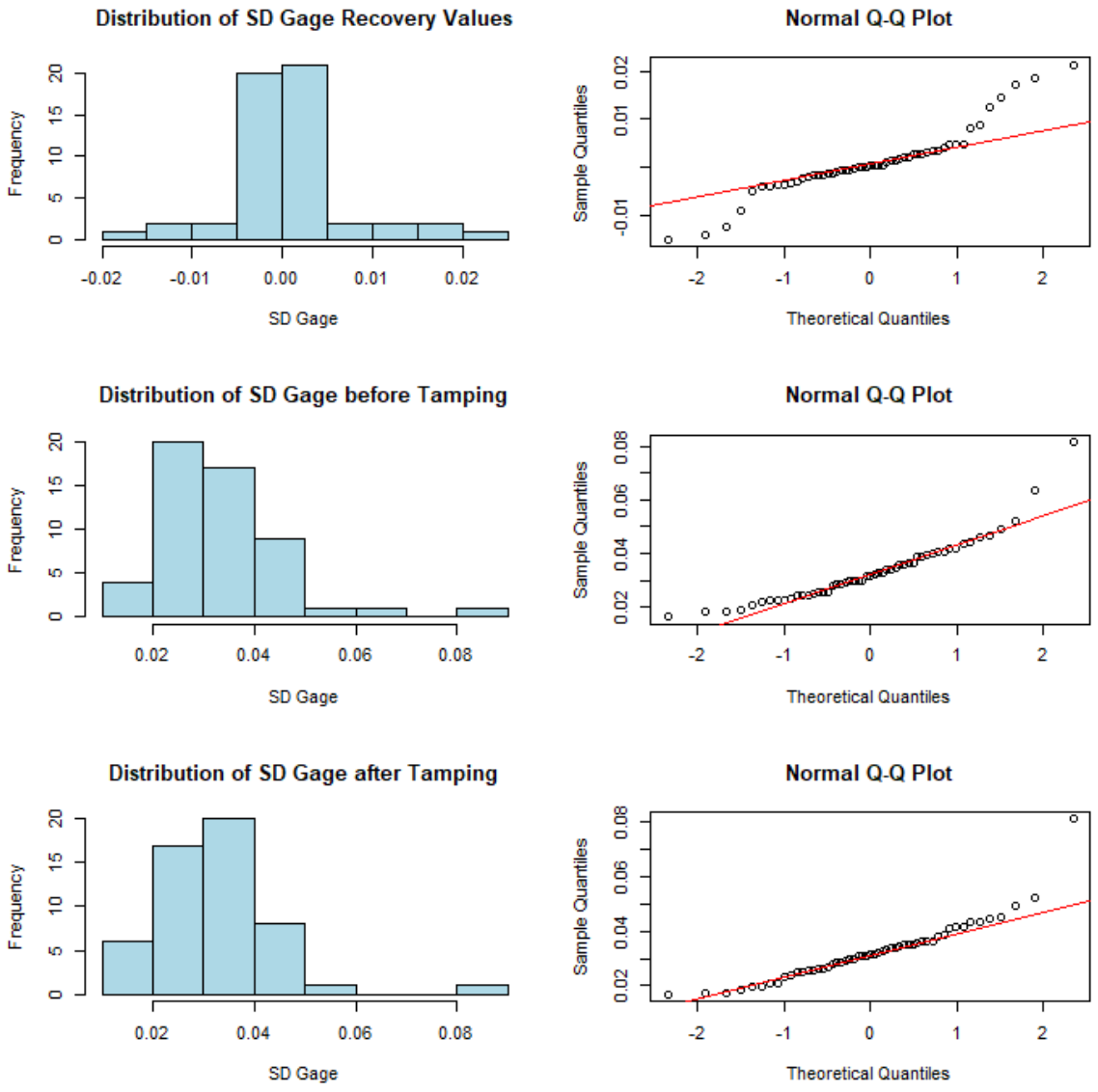


Figure A.23: Histograms and Q-Q plots for SD gage recovery values, SD gage before tamping and SD gage after tamping

A.3 Box and whisker Diagrams

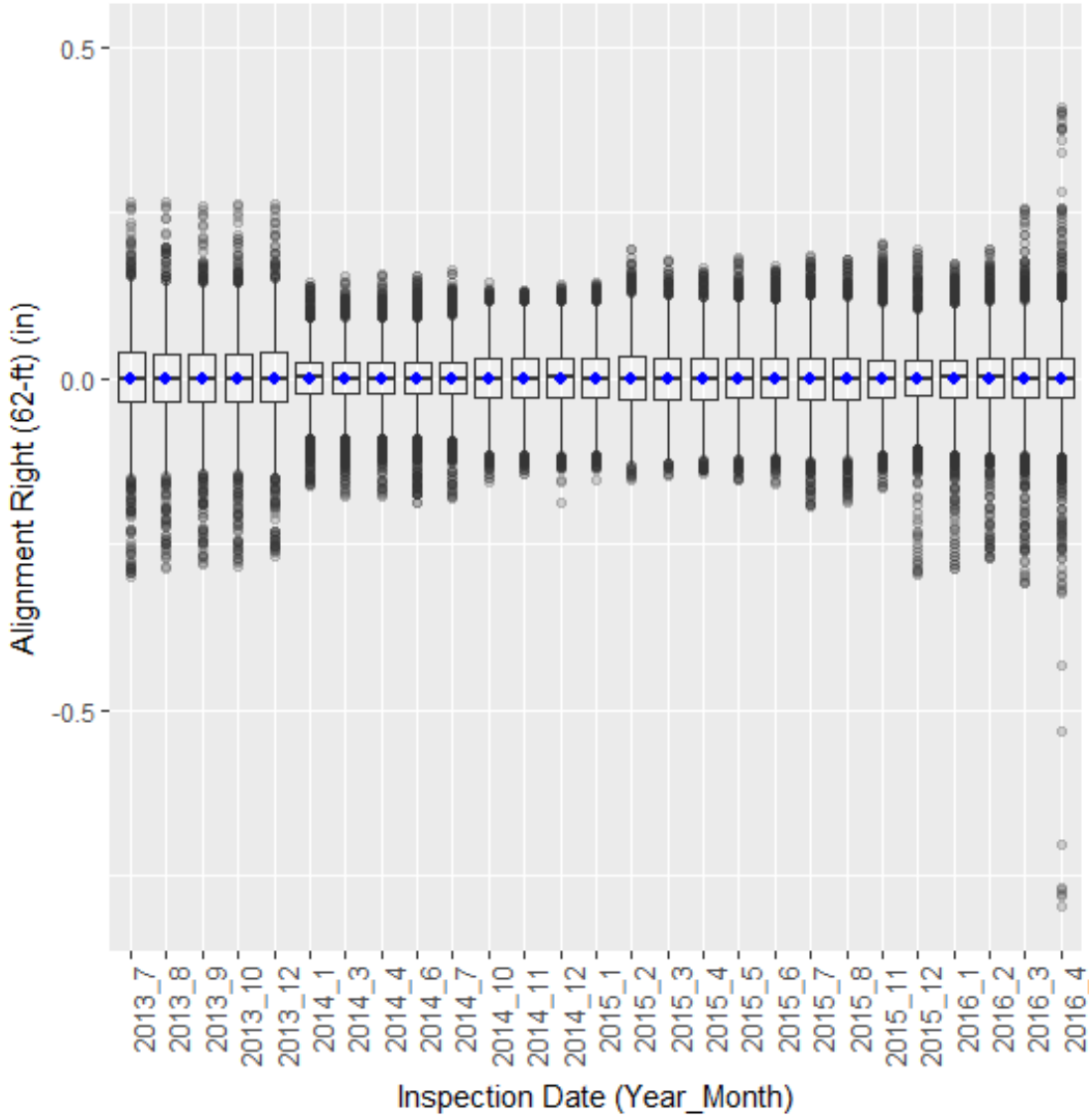


Figure A.24: Box plot of alignment right (62-ft) data points across all the inspection dates

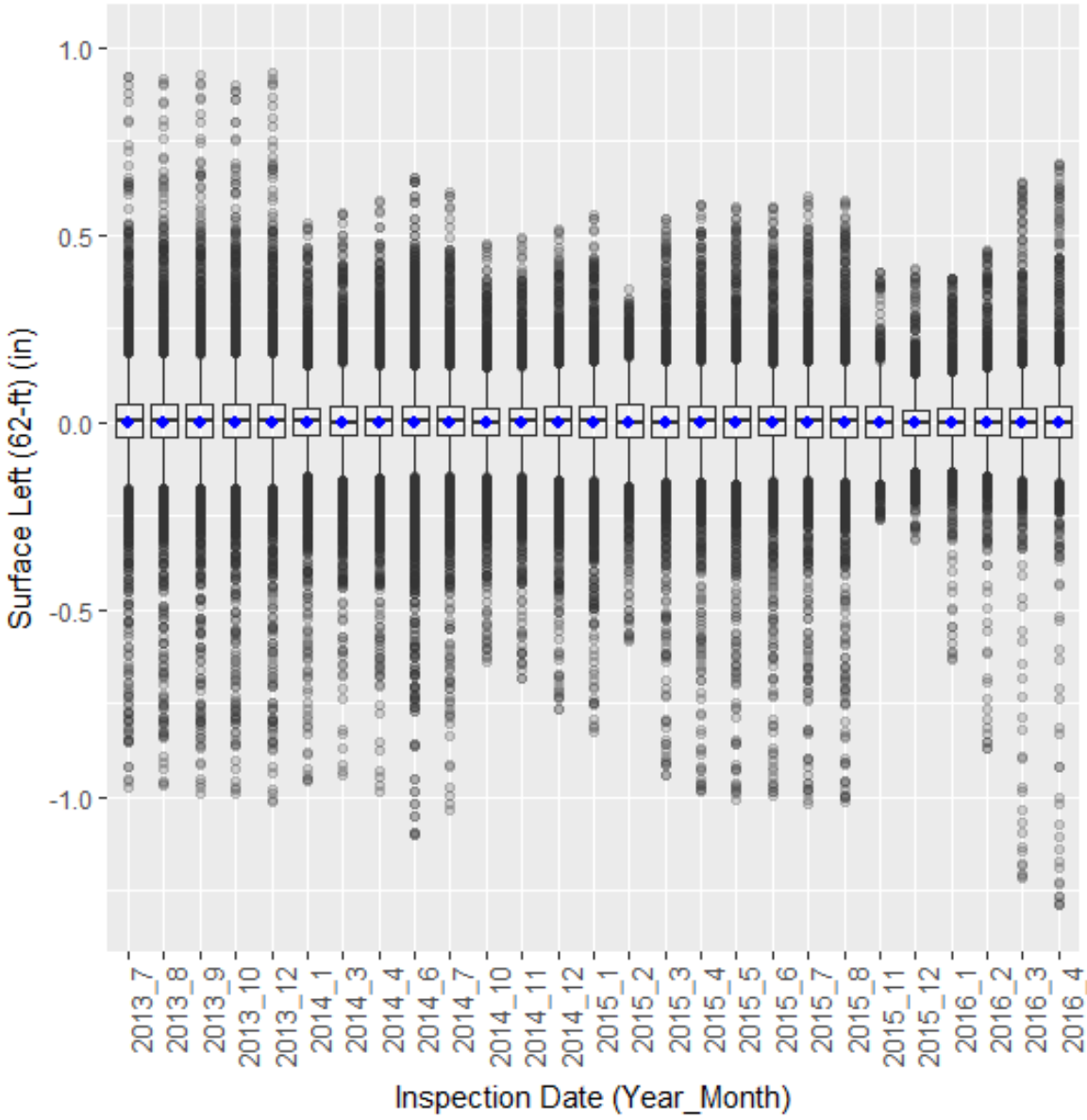


Figure A.25: Box plot of surface left (62-ft) data points across all the inspection dates

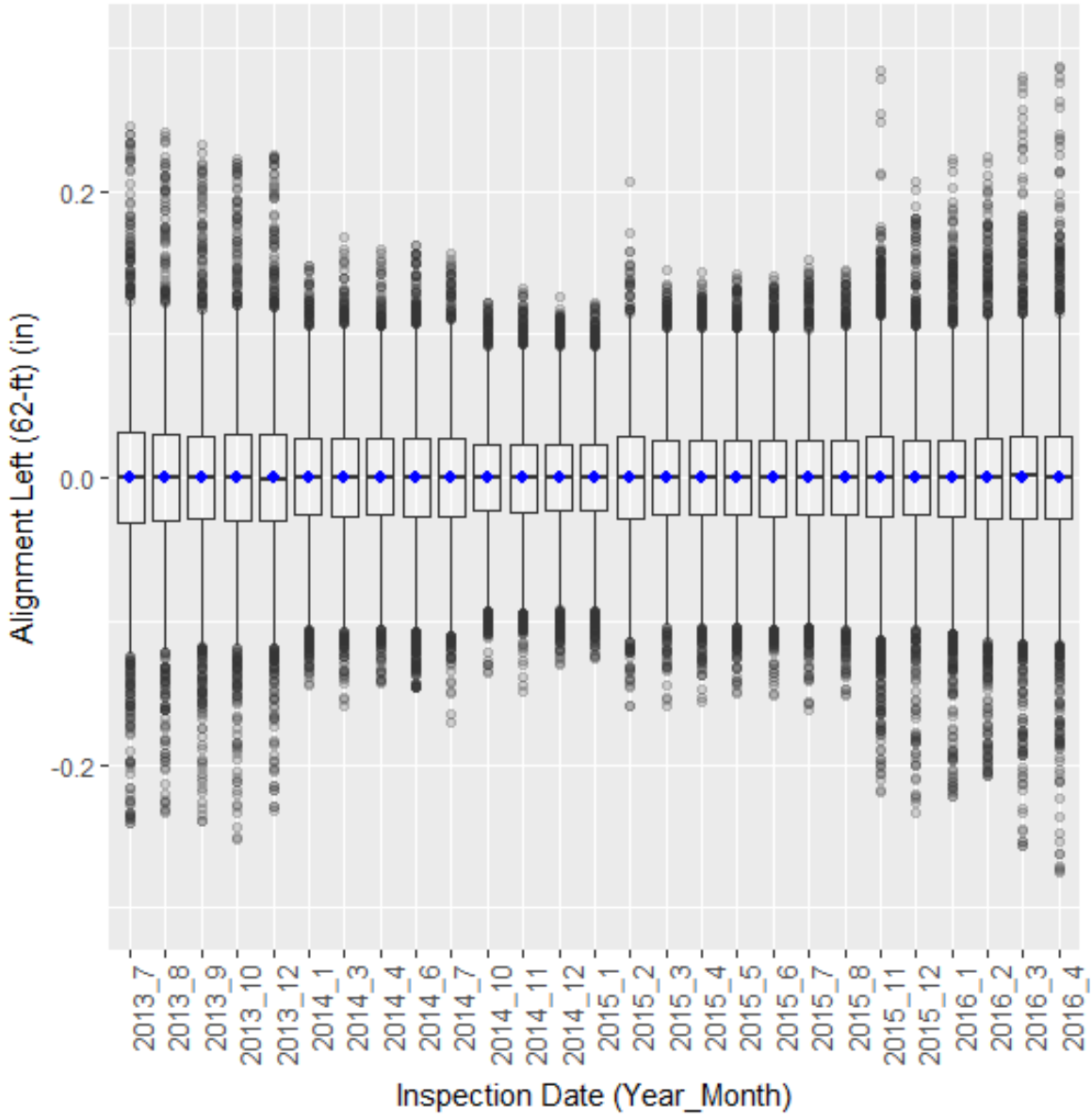


Figure A.26: Box plot of alignment left (62-ft) data points across all the inspection dates

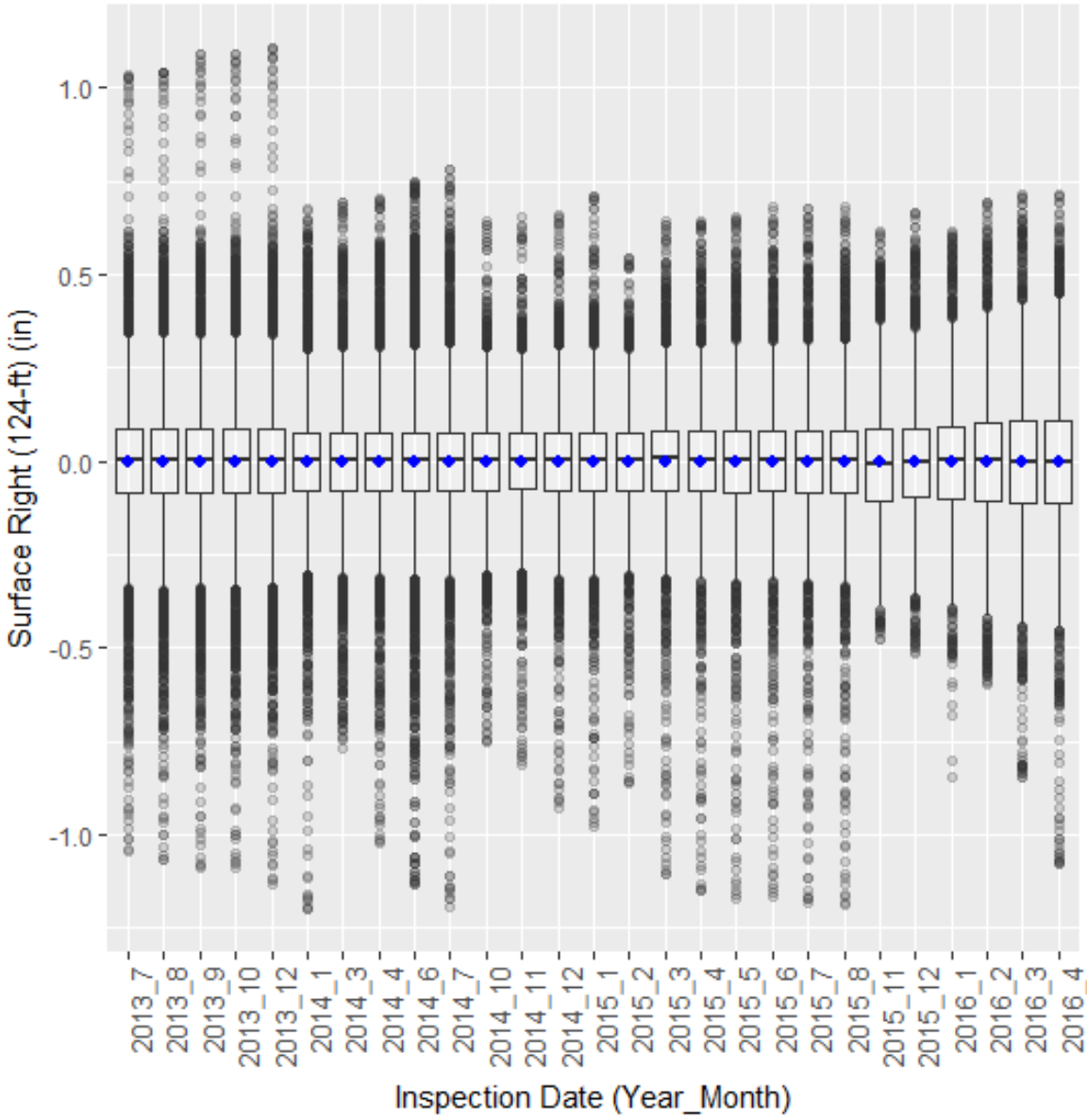


Figure A.27: Box plot of surface right (124-ft) data points across all the inspection dates

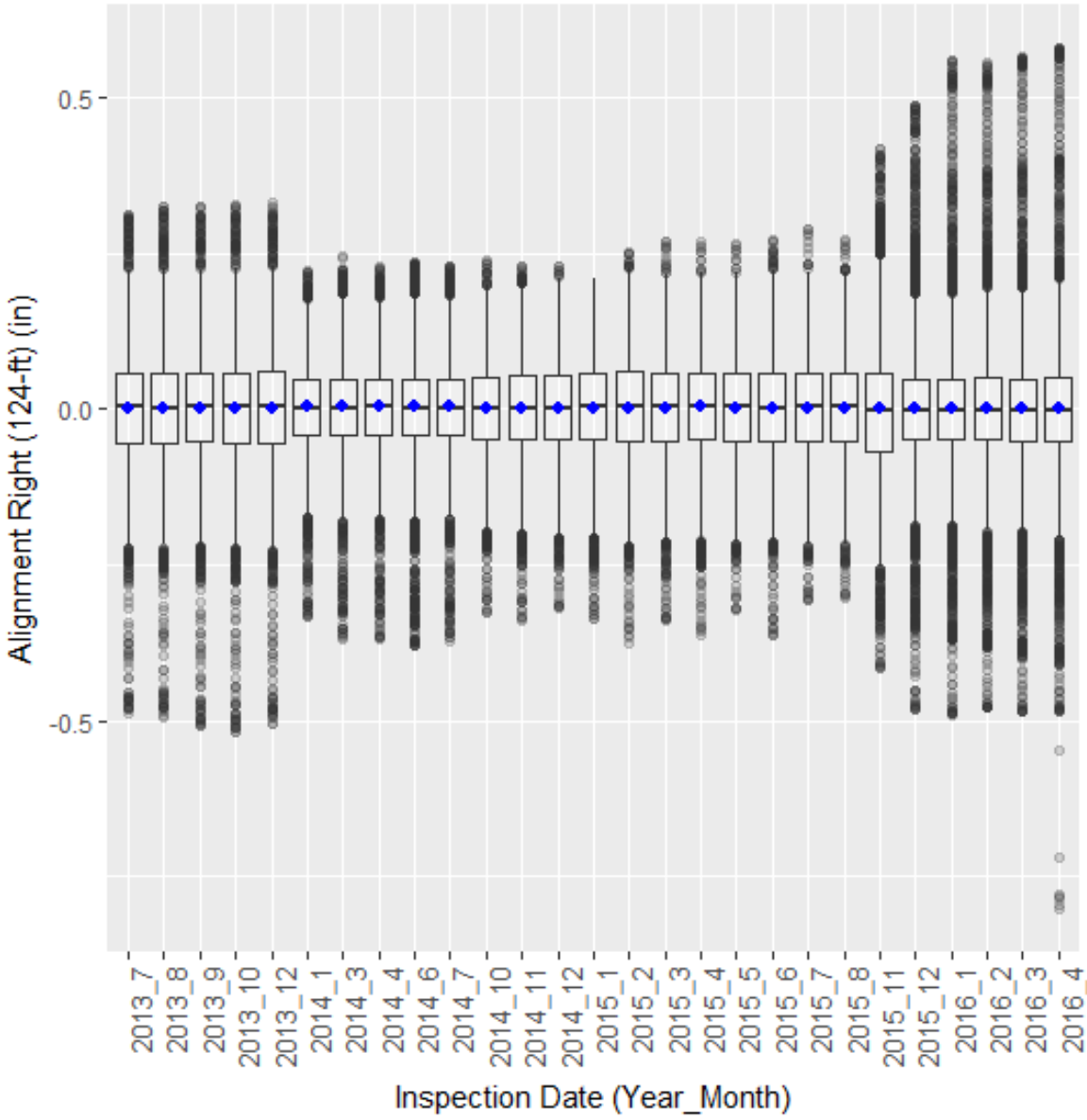


Figure A.28: Box plot of alignment right (124-ft) data points across all the inspection dates

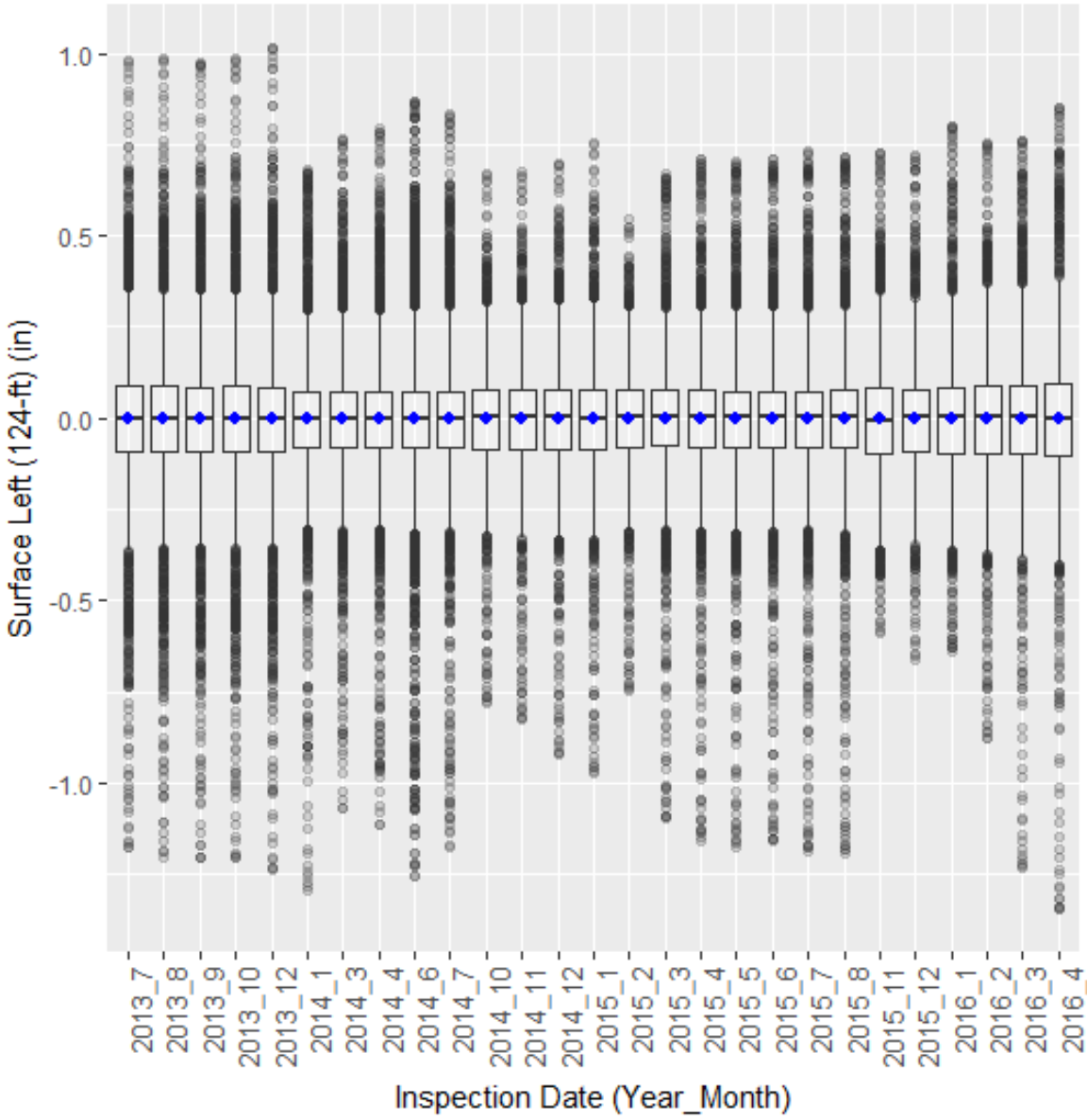


Figure A.29: Box plot of surface left (124-ft) data points across all the inspection dates

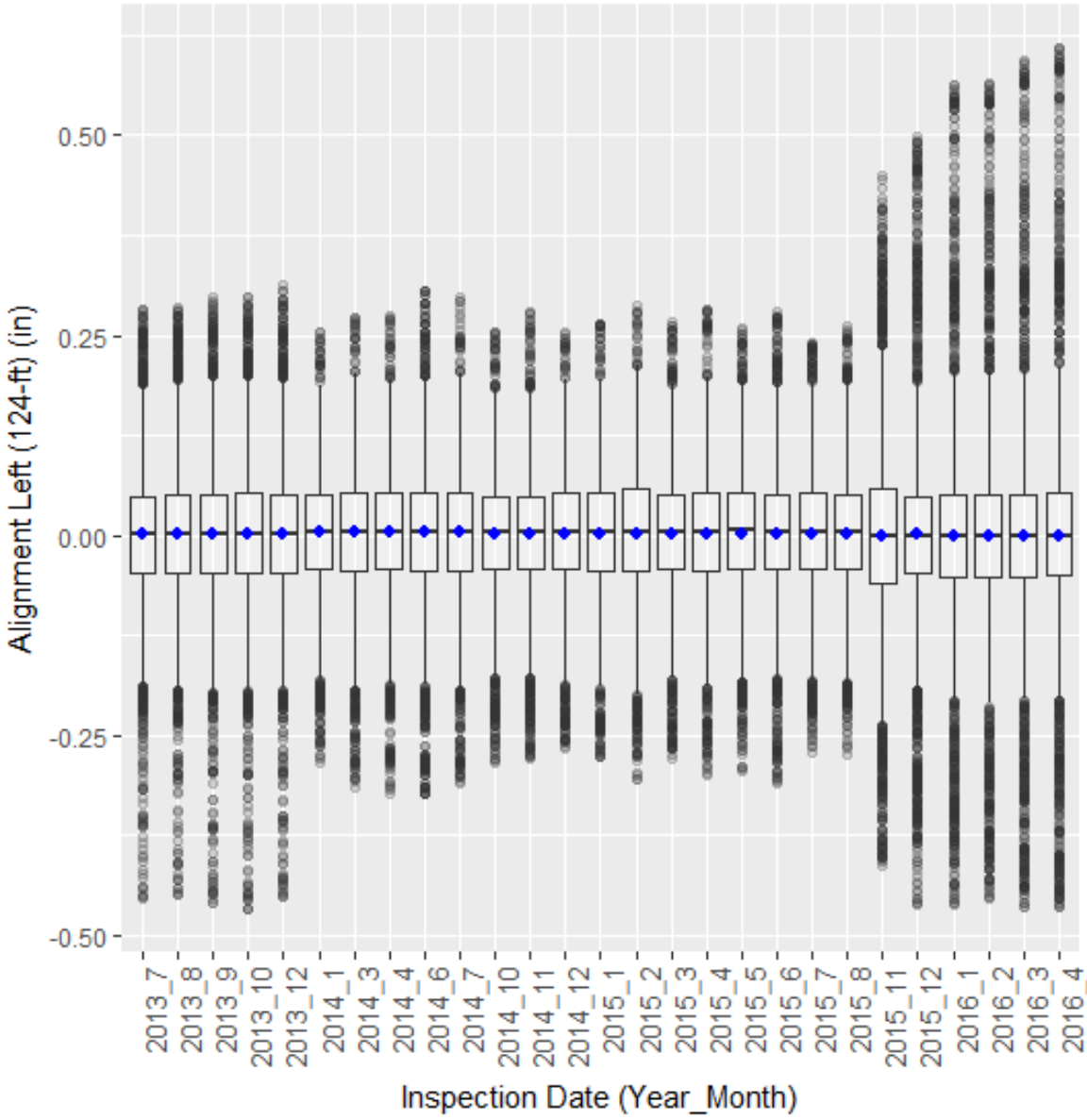


Figure A.30: Box plot of alignment left (124-ft) data points across all the inspection dates

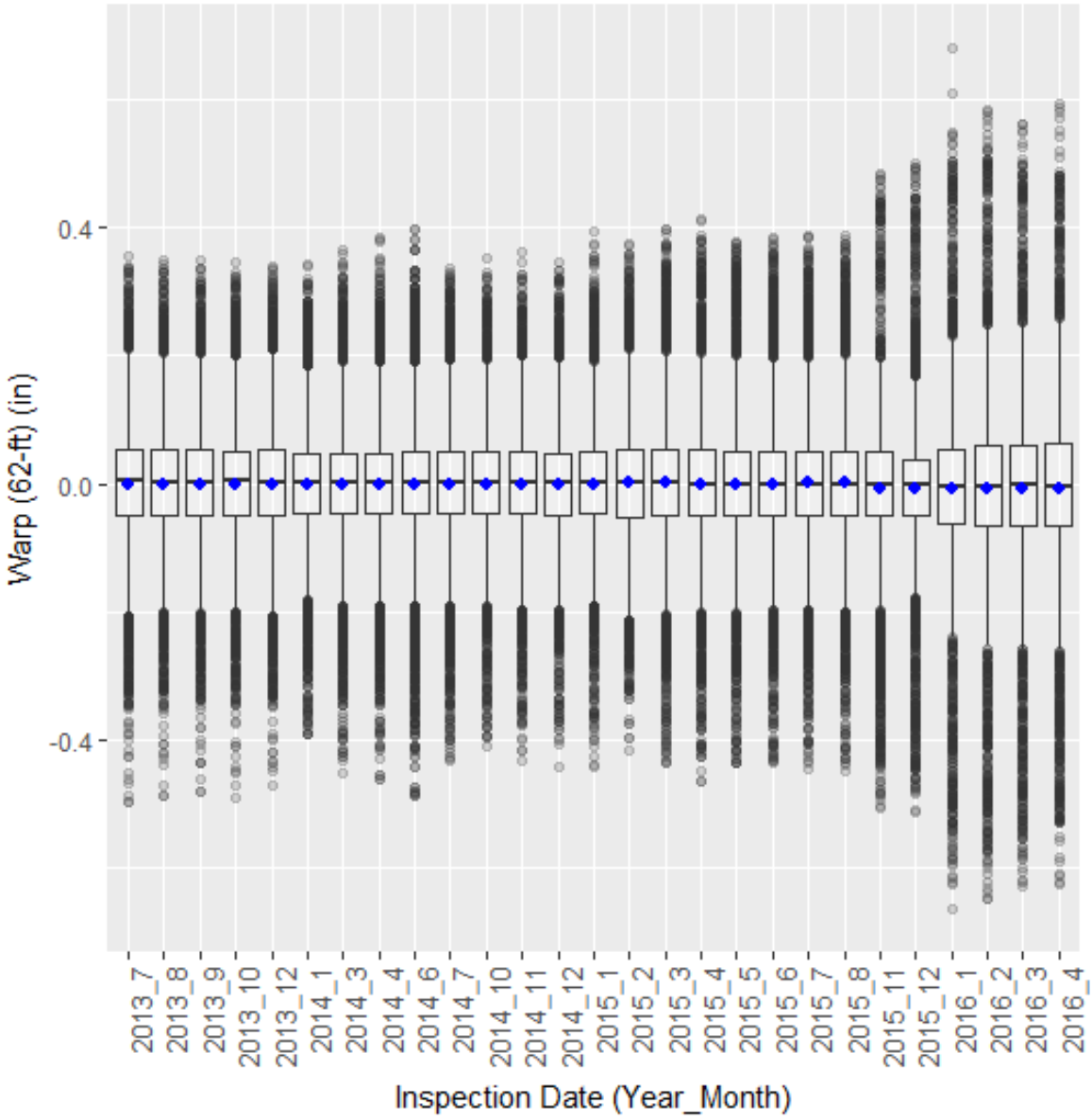


Figure A.31: Box plot of warp (62-ft) data points across all the inspection dates

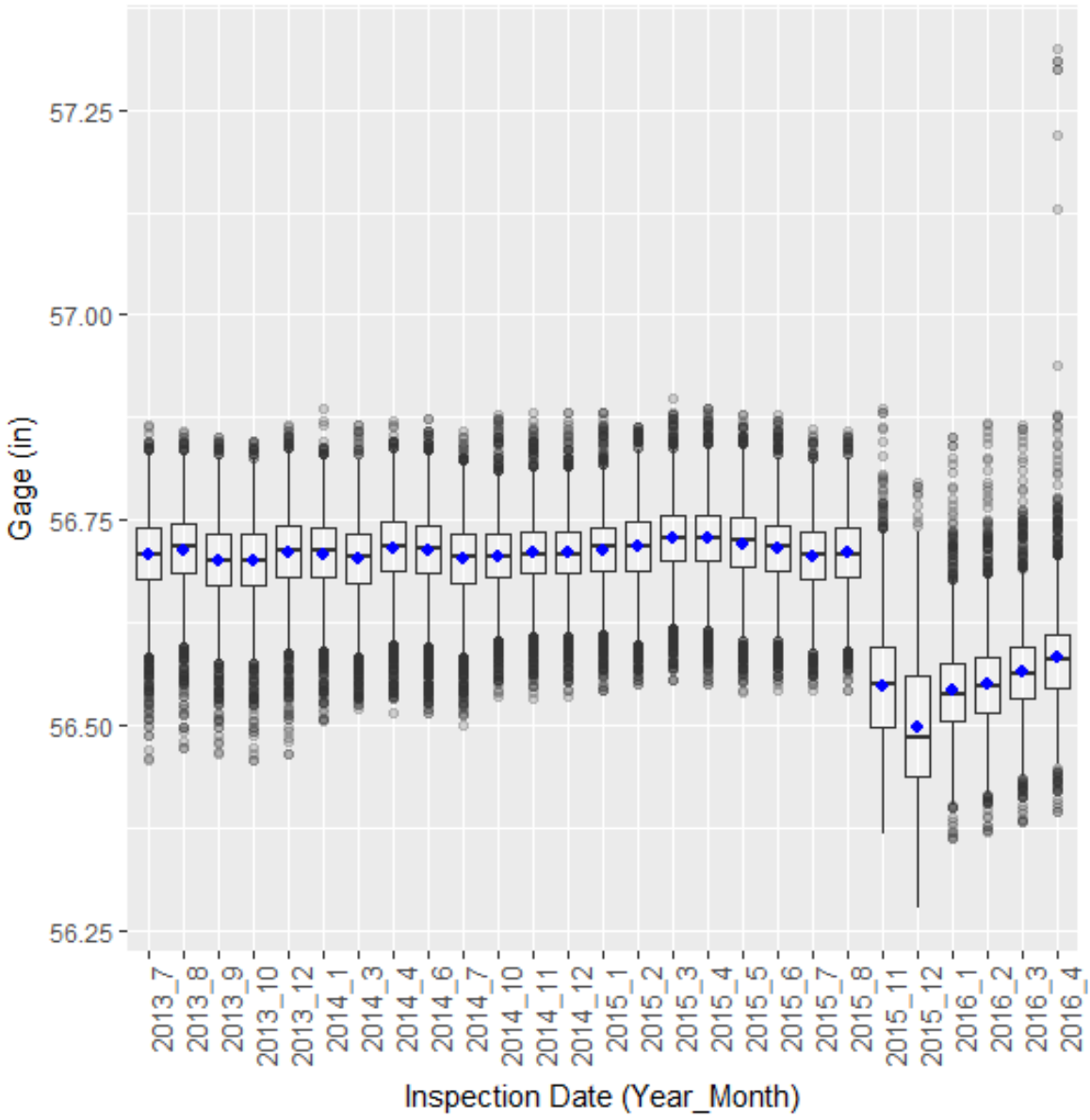


Figure A.32: Box plot of gage data points across all the inspection dates

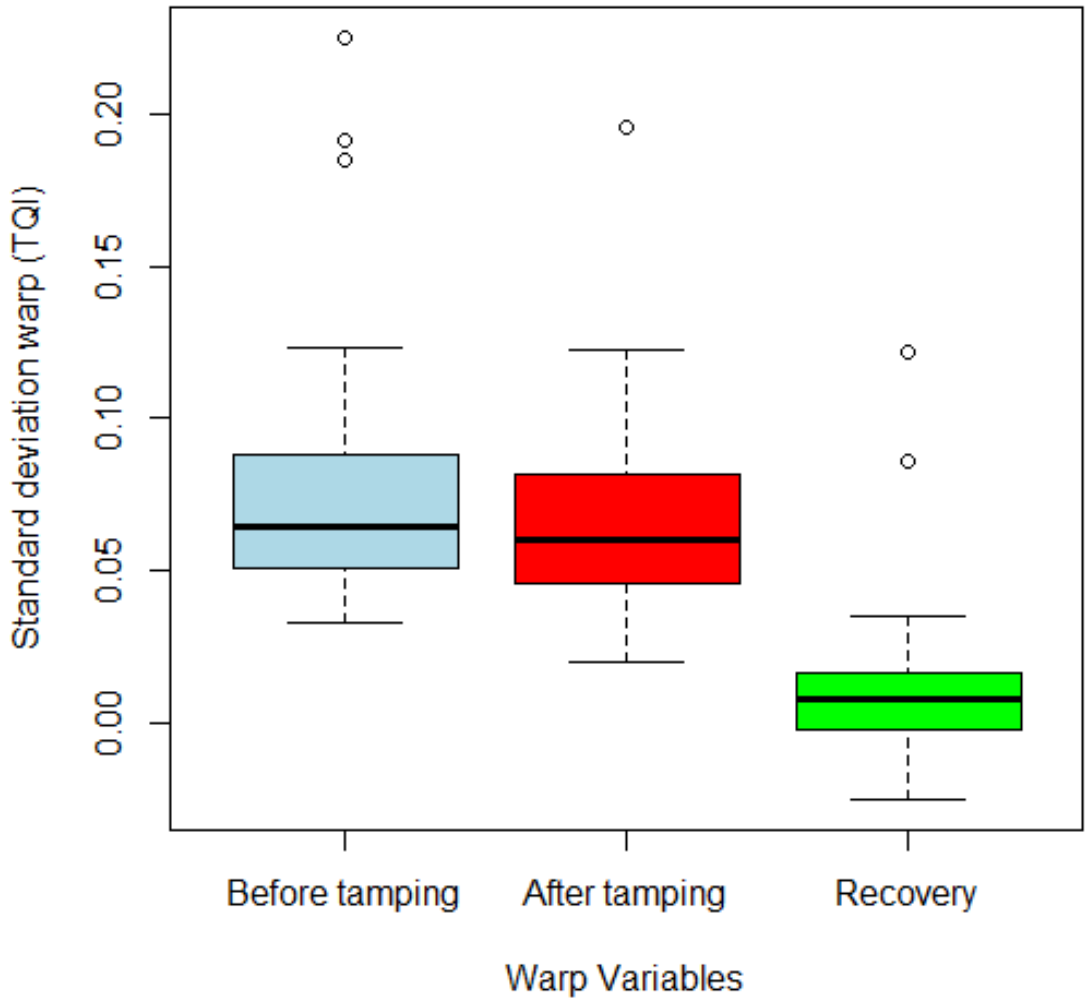


Figure A.33: Box plot of TQI before tamping, TQI after tamping and recovery values for SD warp

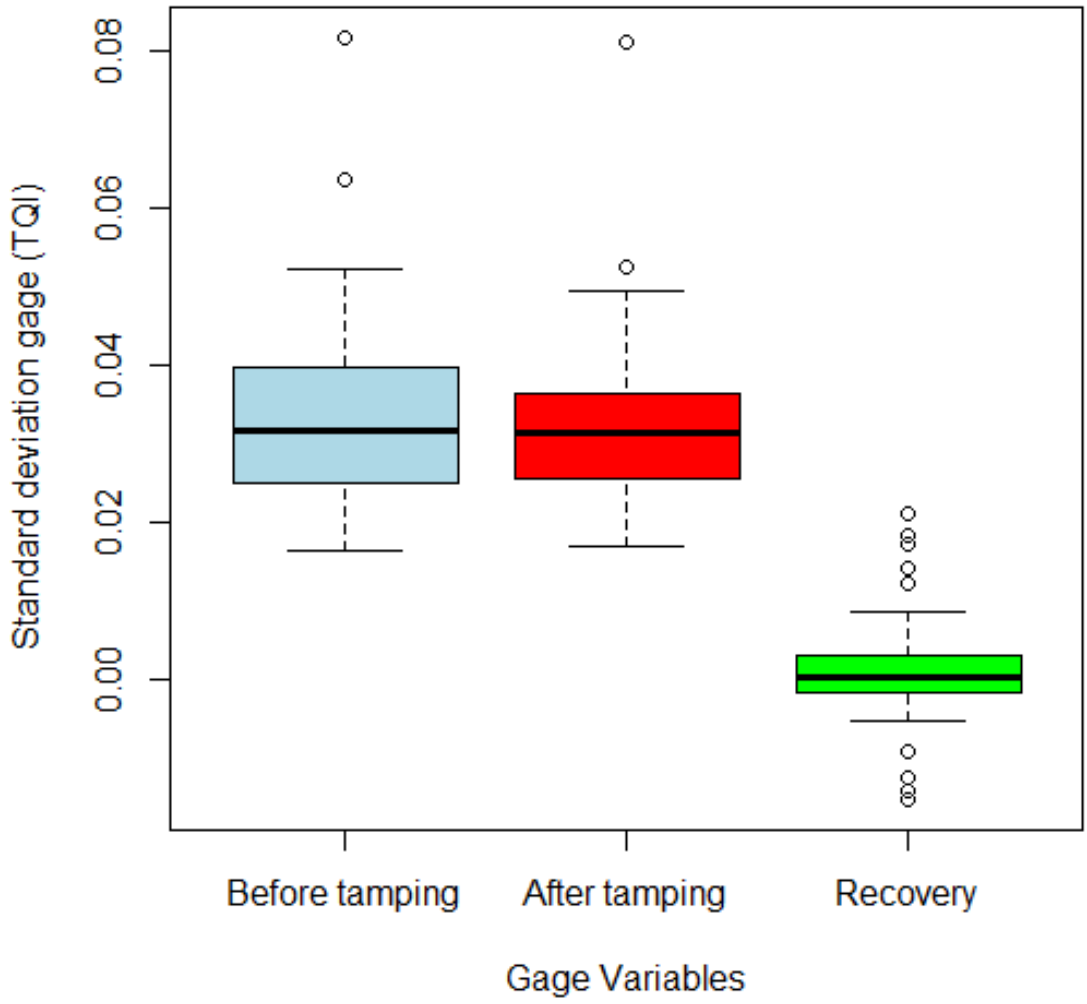


Figure A.34: Box plot of TQI before tamping, TQI after tamping and recovery values for SD gage

Appendix B

DERAILMENT SEVERITY EXPLORATORY DATA ANALYSIS

B.1 Dataset Description

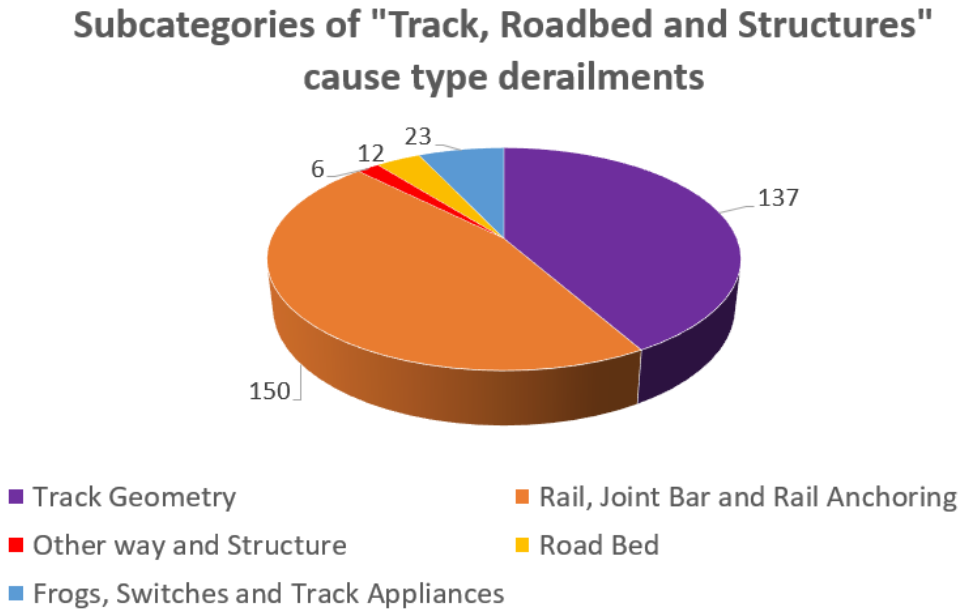


Figure B.1: Subcategory breakdown of "Track, Roadbed and Structures" cause type derailments

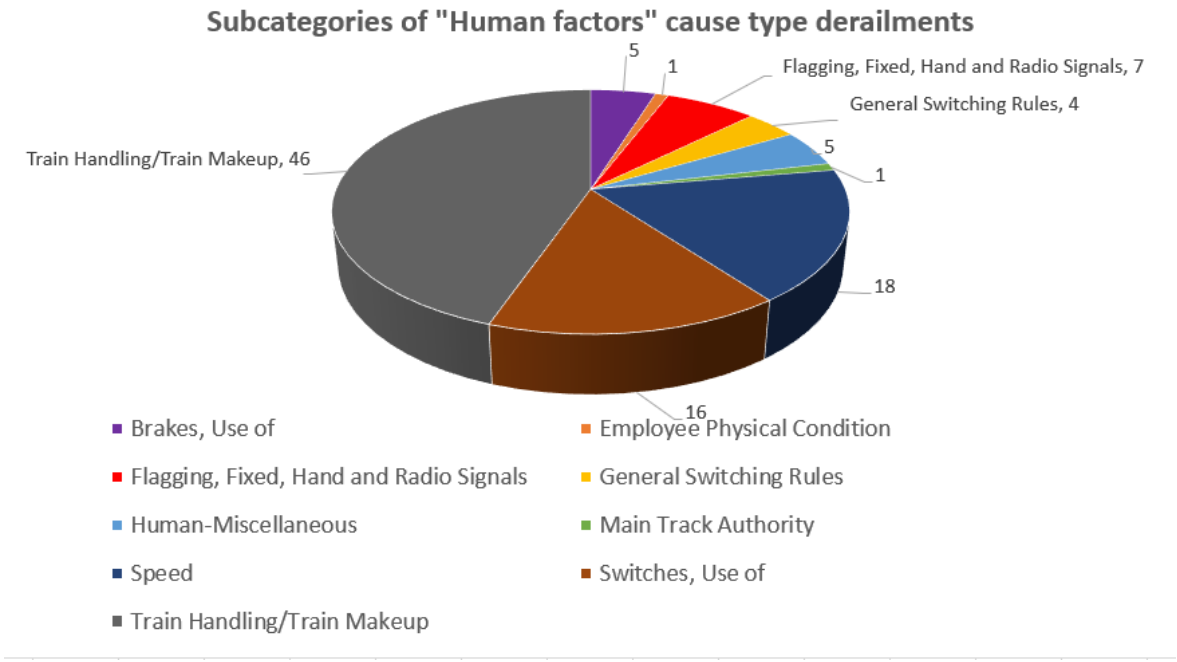


Figure B.2: Subcategory breakdown of "Human factors" cause type derailments

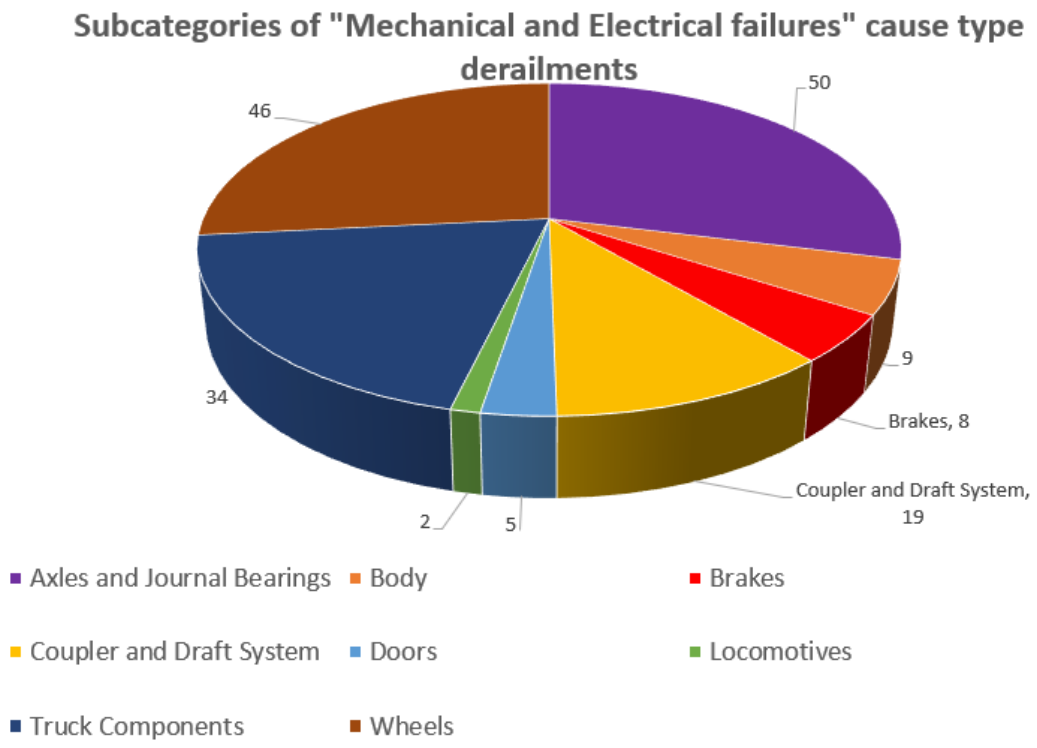


Figure B.3: Subcategory breakdown of "Mechanical and Electrical Failures" cause type derailments

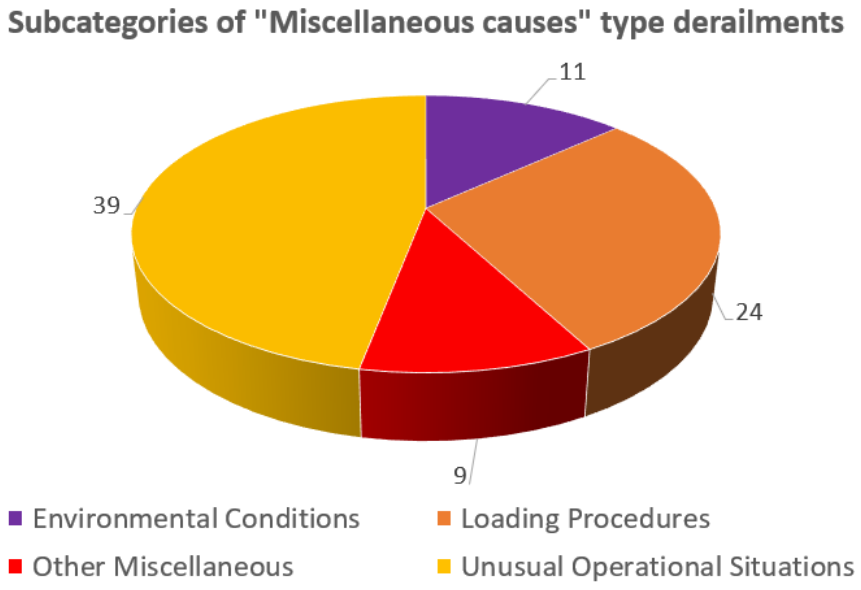


Figure B.4: Subcategory breakdown of "Miscellaneous Causes" type derailments

B.2 Histogram and Quantile-Quantile Plot

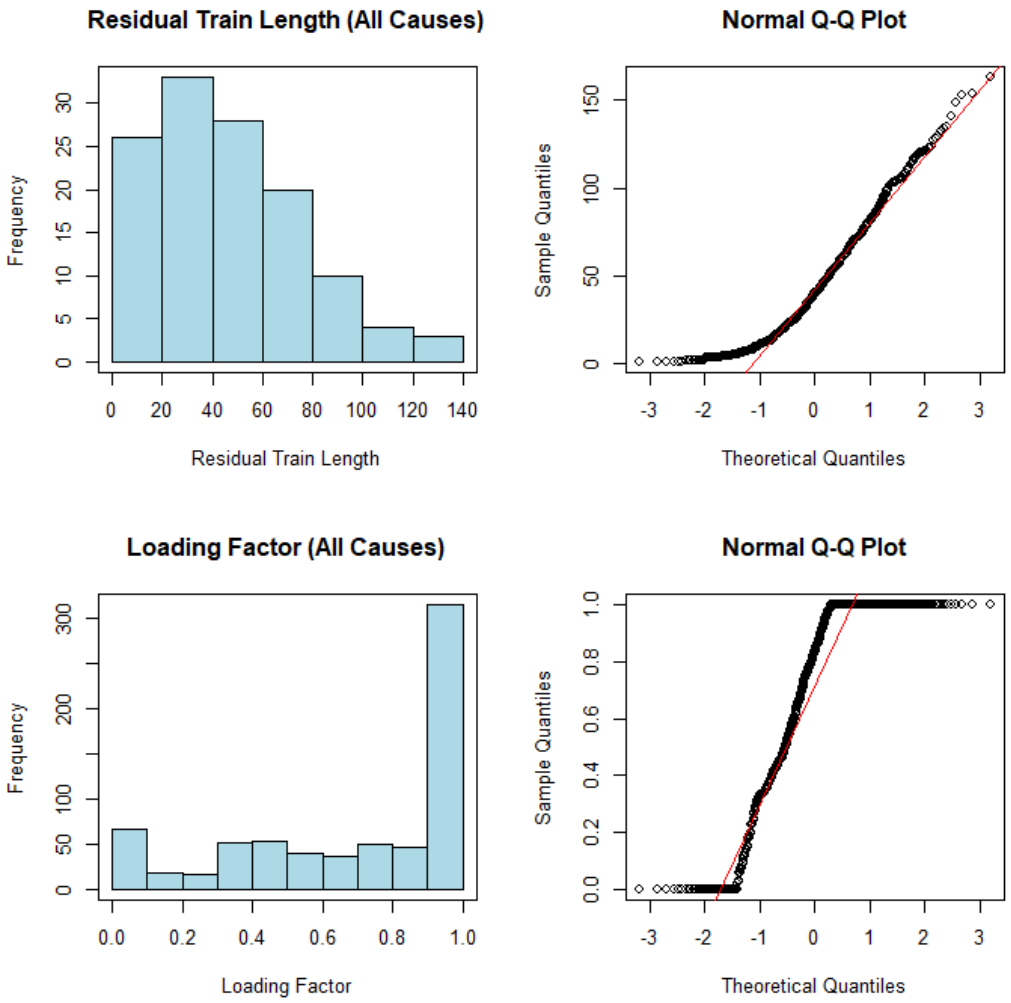


Figure B.5: Histograms and Q-Q plots for residual train length and loading factor for all freight-train derailments occurring on Class I mainline track

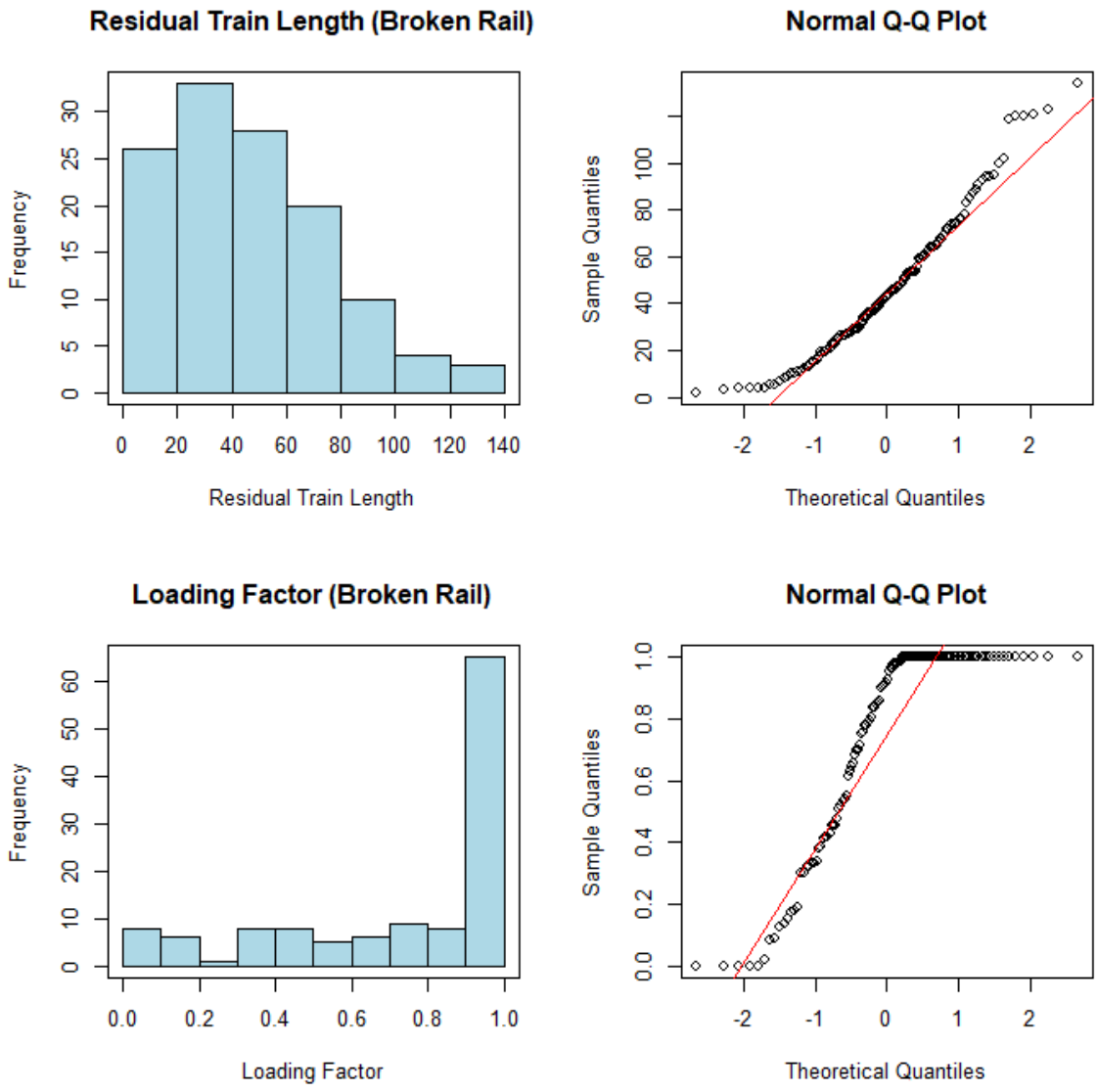


Figure B.6: Histograms and Q-Q plots for monetary damage, derailed cars and derailment speed for broken-rail caused freight-train derailments occurring on Class I mainline track

B.3 Box and whisker diagram

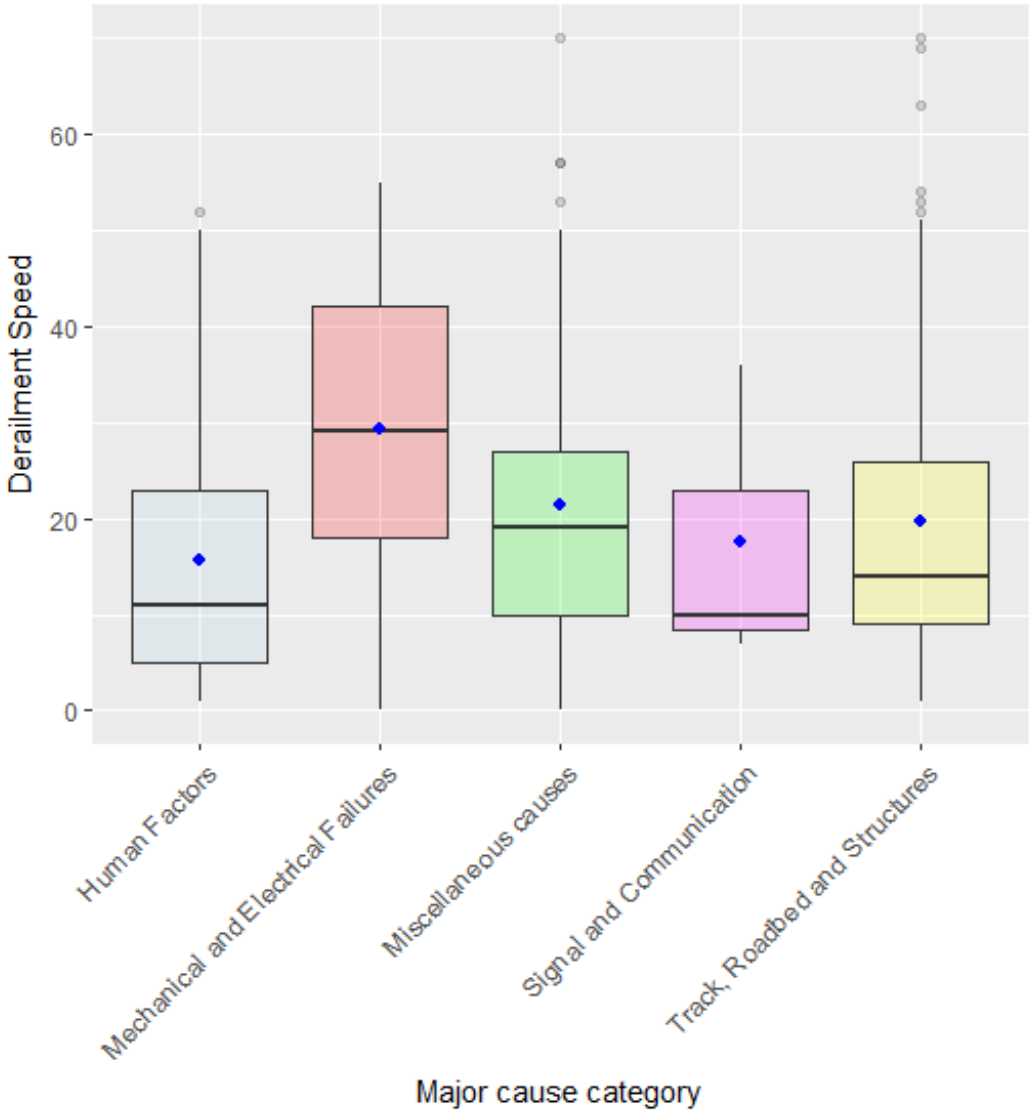


Figure B.7: Box plot illustrating distribution of derailment speed across all major accident cause category

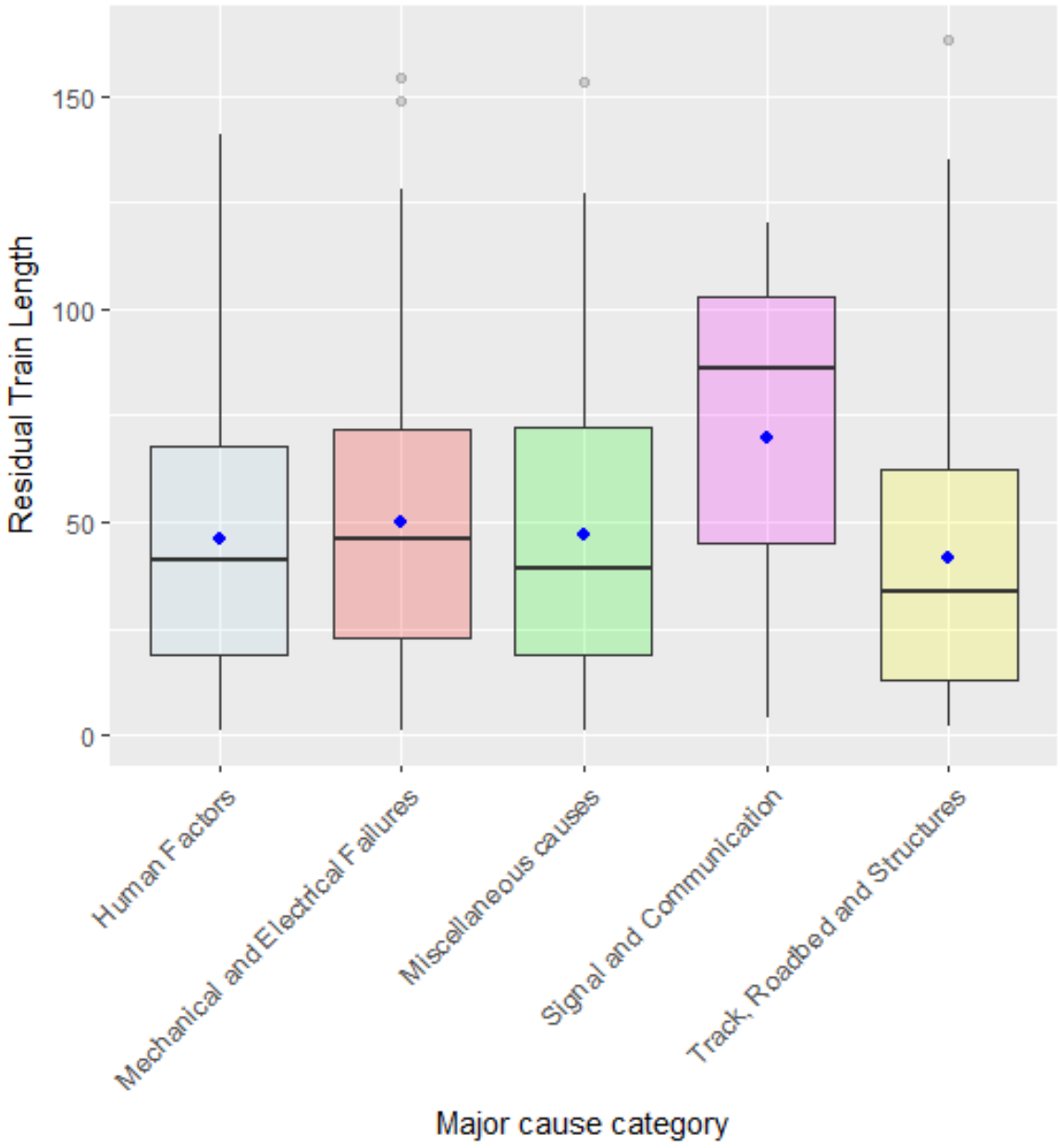


Figure B.8: Box plot illustrating distribution of residual train length across all major accident cause categories

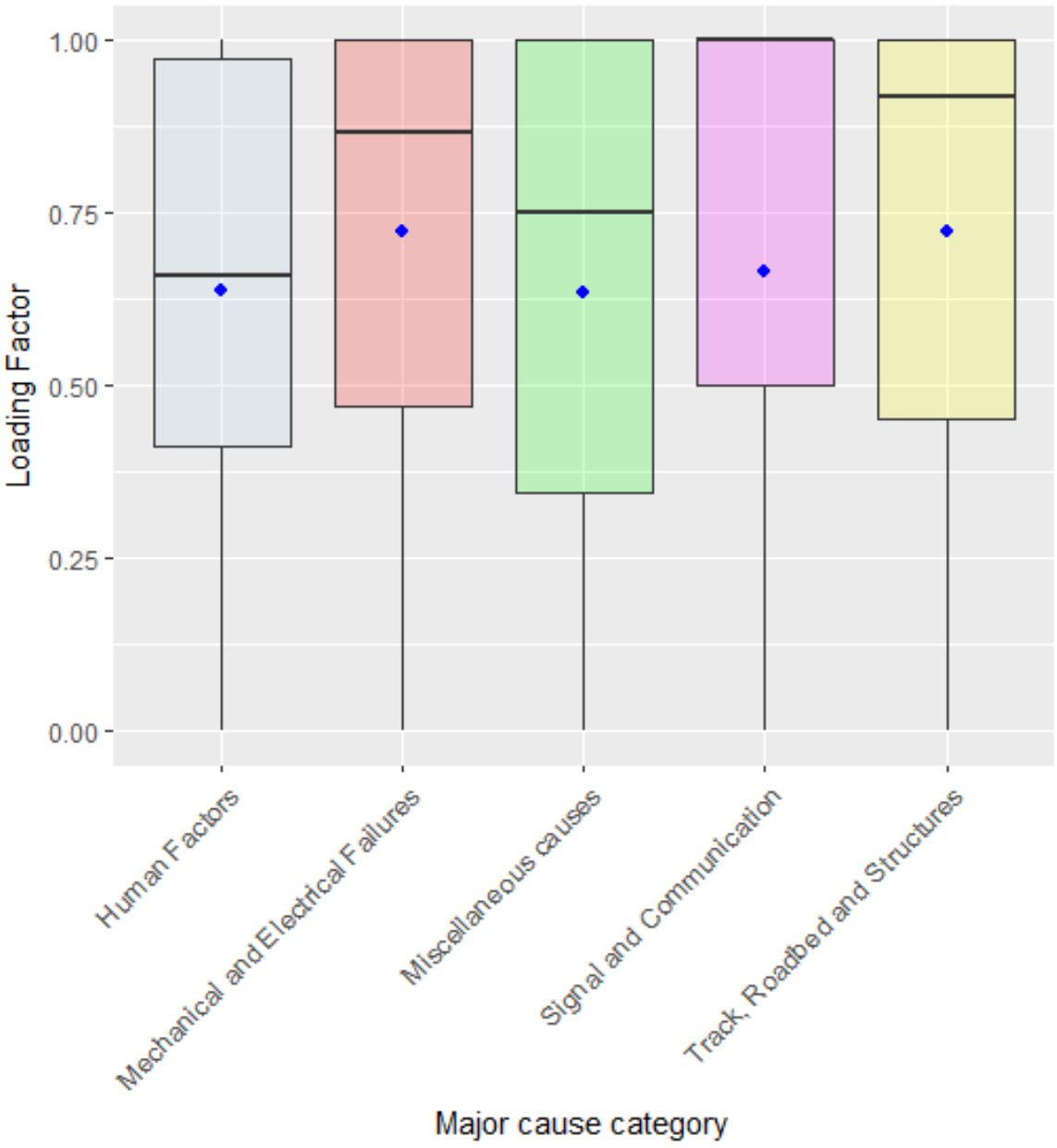


Figure B.9: Box plot illustrating distribution of loading factor across all major accident cause categories

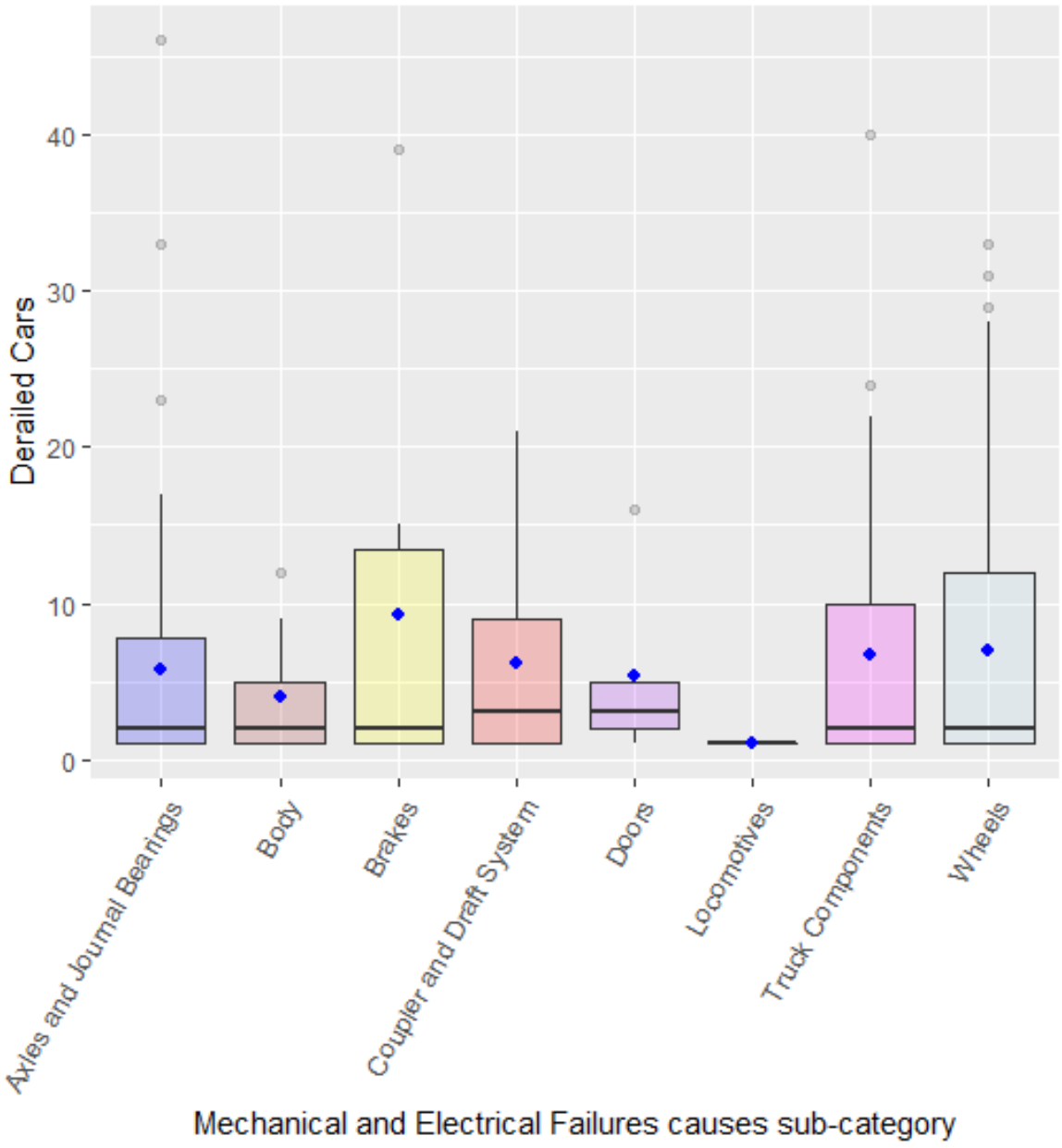


Figure B.10: Box plot illustrating distribution of derailed cars across Mechanical and Electrical failures causes sub-category

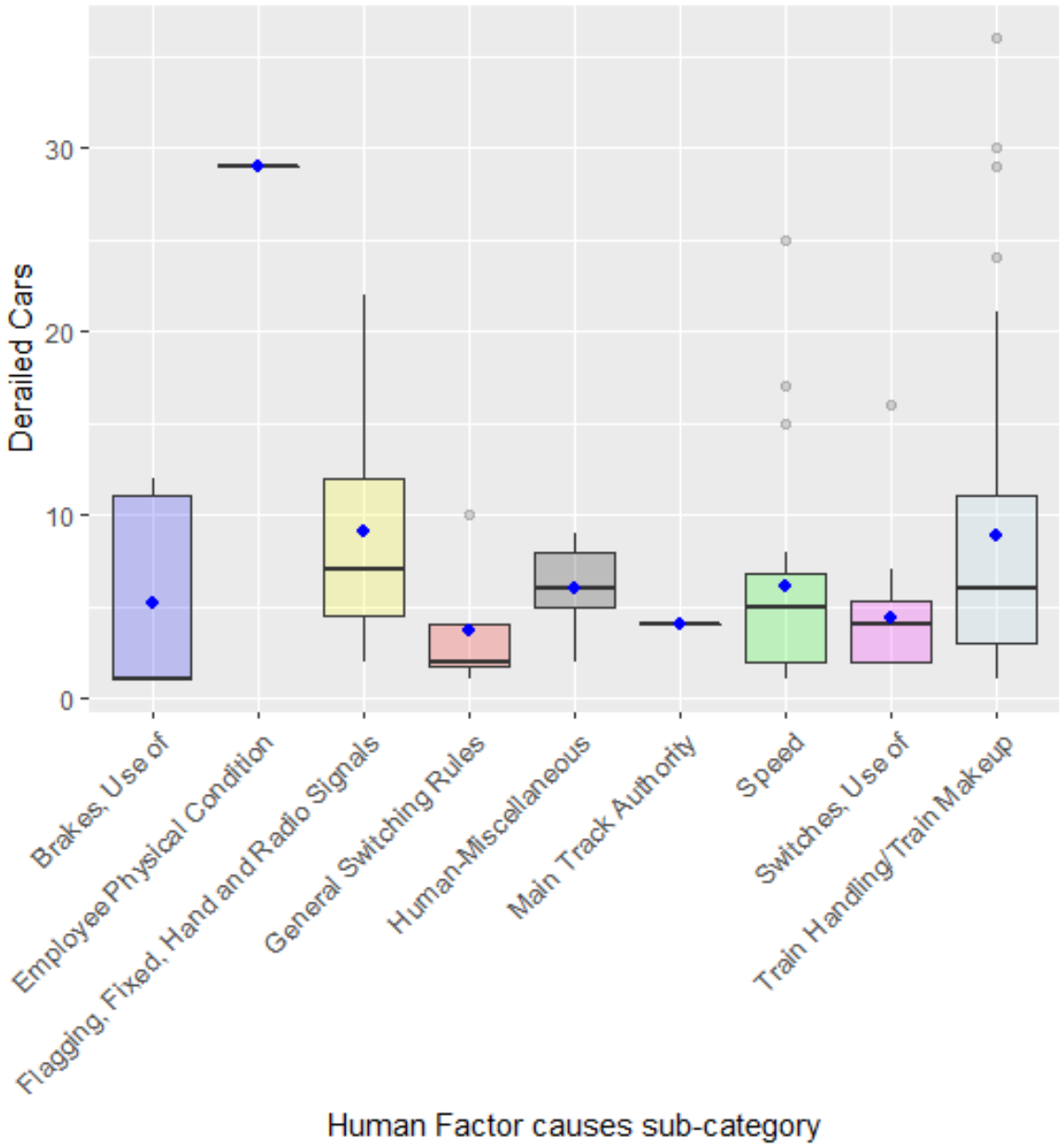


Figure B.11: Box plot illustrating distribution of derailed cars across Human Factors causes sub-category

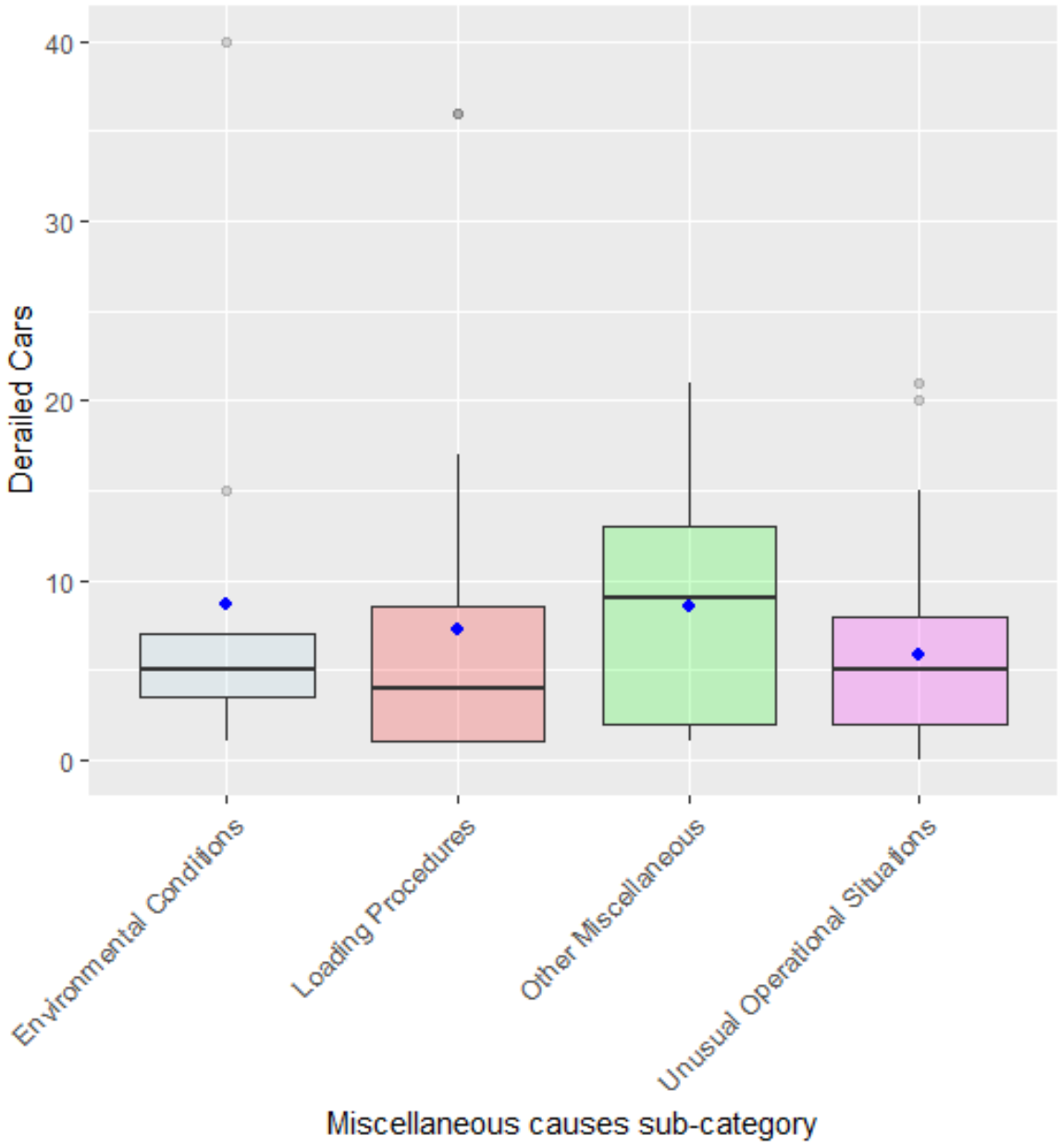


Figure B.12: Box plot illustrating distribution of derailed cars across Miscellaneous causes sub-category

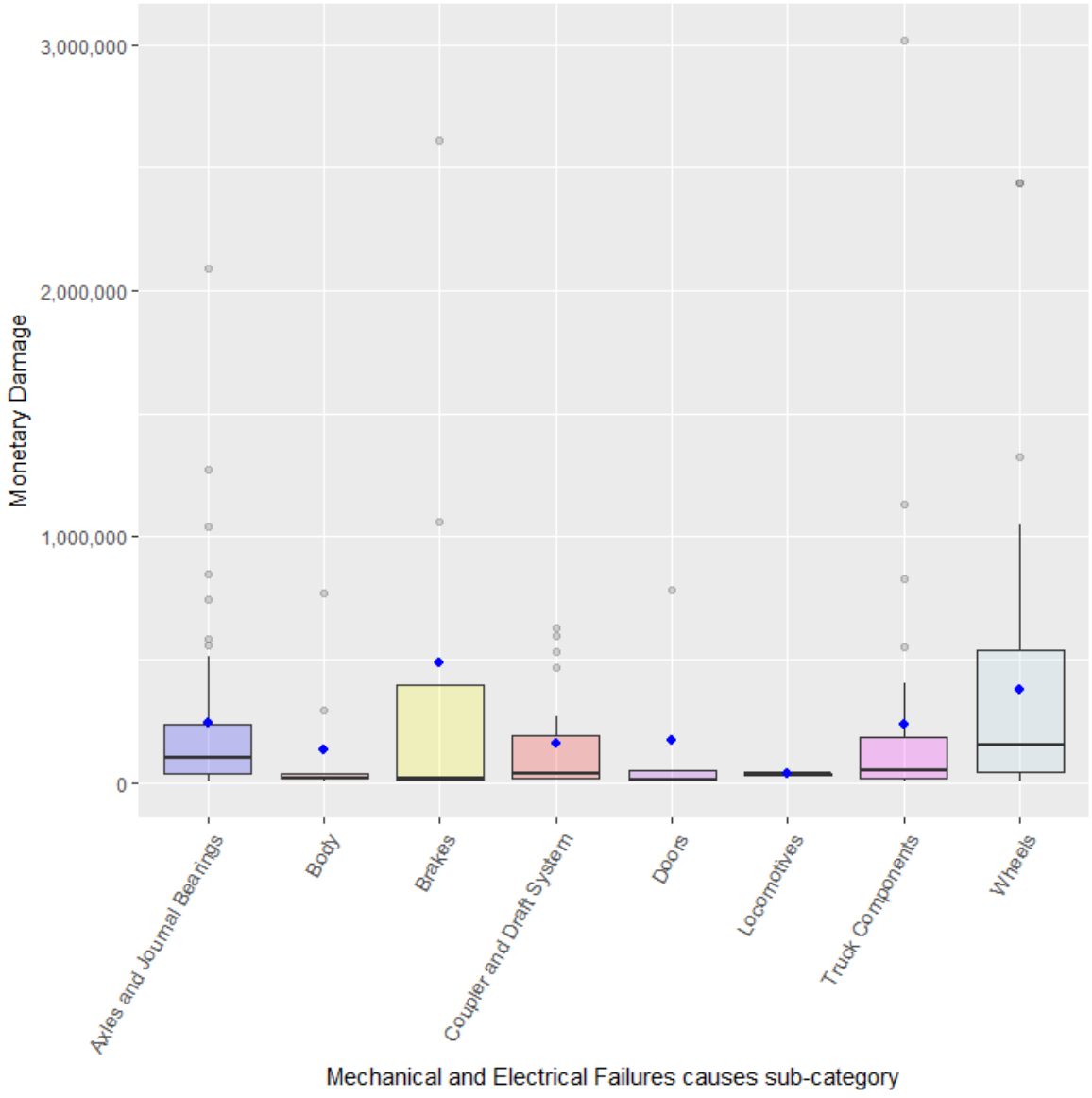


Figure B.13: Box plot illustrating distribution of monetary damage across Mechanical and Electrical failures causes sub-category

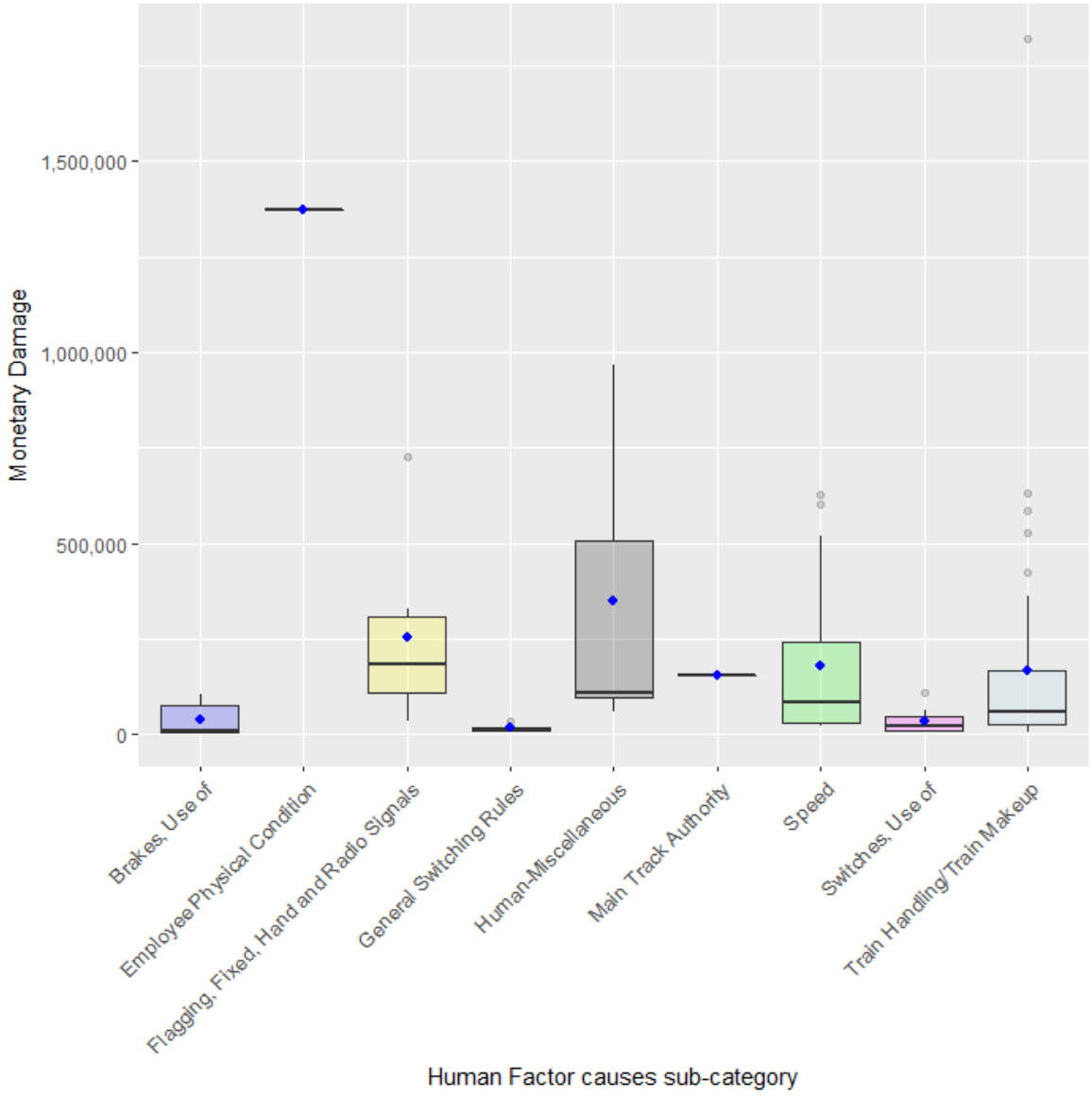


Figure B.14: Box plot illustrating distribution of monetary damage across Human Factors causes sub-category

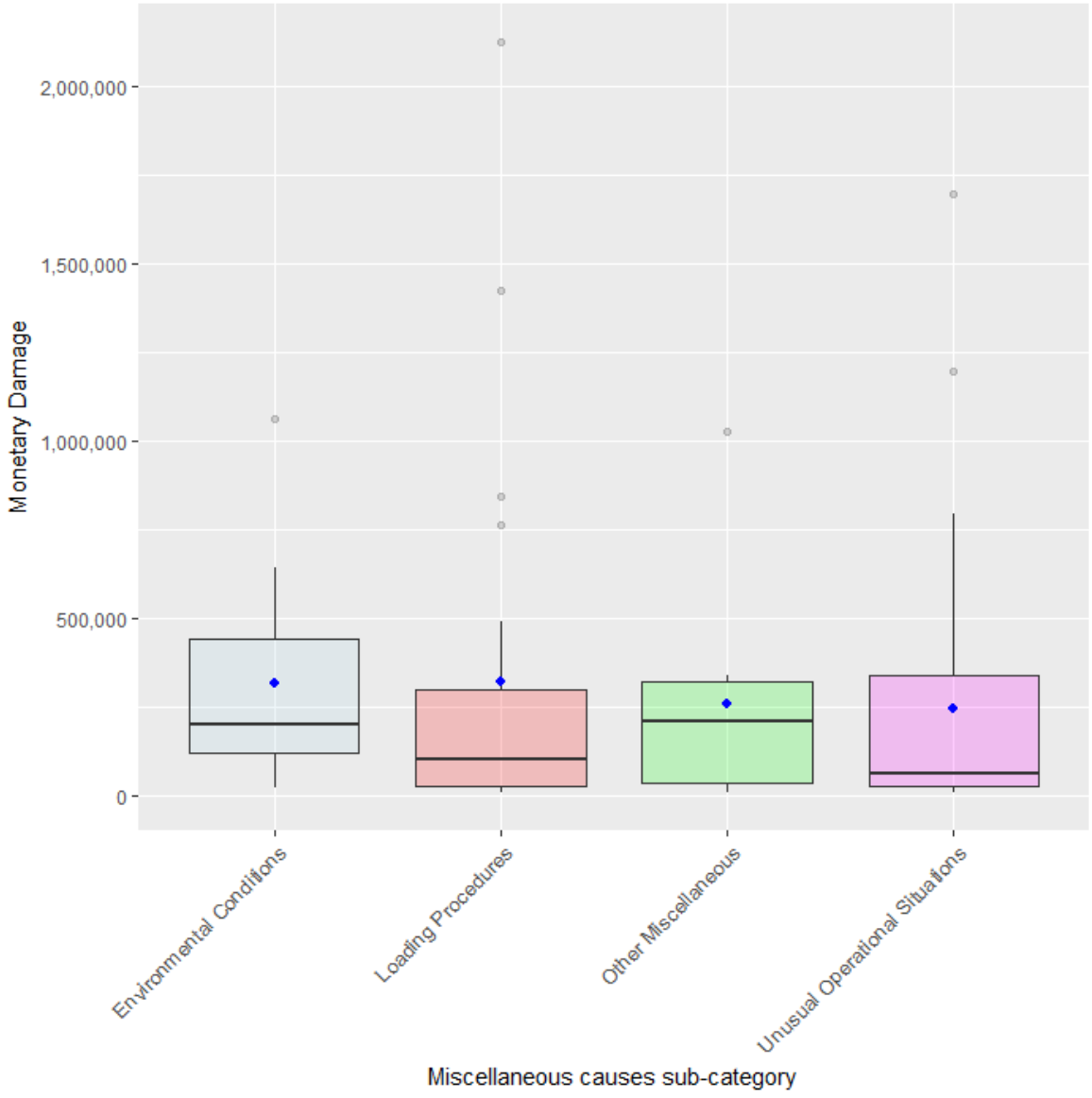


Figure B.15: Box plot illustrating distribution of monetary damage across Miscellaneous causes sub-category

Appendix C

PERMISSIONS

7/2/2018 RightsLink Printable License

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Jul 02, 2018

This Agreement between Emmanuel Martey ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4380820316871
License date	Jul 02, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Wiley Books
Licensed Content Title	Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering
Licensed Content Author	Nii O. Attoh-Okine
Licensed Content Date	May 1, 2017
Licensed Content Pages	1
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 1.1 Track structure components
Will you be translating?	No
Title of your thesis / dissertation	Copula-based Models in Railroad Maintenance and Safety Analysis
Expected completion date	Jul 2018
Expected size (number of pages)	325
Requestor Location	Emmanuel Martey 301 DuPont Hall NEWARK, DE 19716 United States Attn: Emmanuel N Martey
Publisher Tax ID	EU826007151
Total	0.00 USD

Figure C.1: Permission to use Figure 2.1

G Suite

Emmanuel Martey <enmartey@udel.edu>

Figure Permission for PhD Dissertation

Silvia Nunez <silgalnu@udel.edu> Mon, Jul 9, 2018 at 9:48 PM
To: Emmanuel Martey <enmartey@udel.edu>

Dear Emmanuel,

Thank you for contacting me. I would be very happy that my figure appears in your dissertation. I appreciate that you find my work useful.

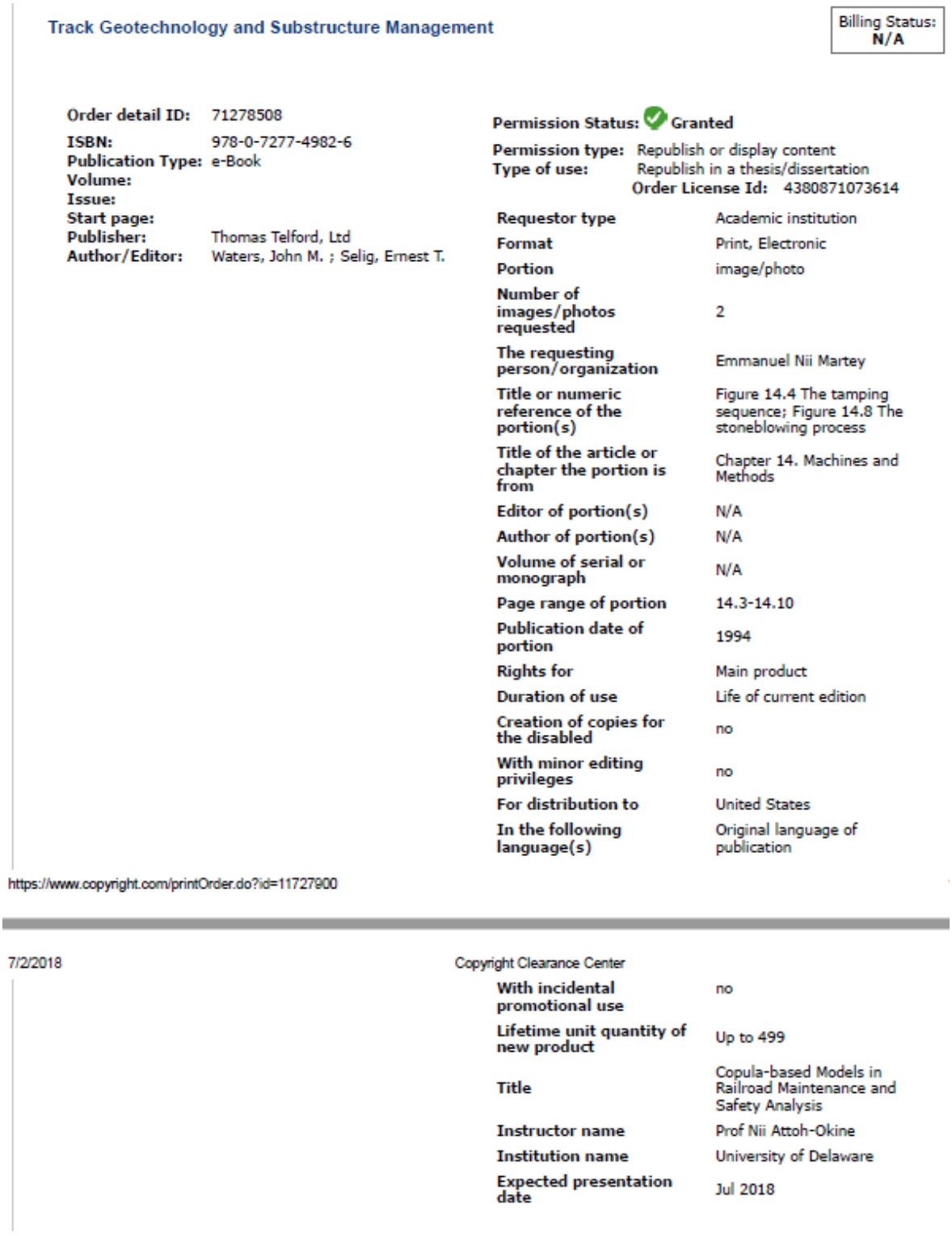
Best regards,

Silvia A. Galván Núñez
--

On Mon, Jul 9, 2018, 9:37 AM Emmanuel Martey <enmartey@udel.edu> wrote:

> Good morning Dr. Silvia Galvan-Nunez,
>
> My name is Emmanuel Nii Martey, a PhD candidate at the Department of Civil and Environmental Engineering at the University of Delaware. I am currently wrapping up work on my dissertation entitled "Copula-based models in Railroad Maintenance and Safety Analysis".
> I would like to ask permission to use the Track Geometry Parameters figure (Figure 2.4) from your PhD dissertation (Hybrid Bayesian-Wiener Process in Track Geometry Degradation Analysis) in my own dissertation. Please find attached the figure in question. Thank you for the consideration and hope to hear from you soon.
>
> Best regards,
>
> Emmanuel Nii Martey

Figure C.2: Permission to use Figure 2.2



Track Geotechnology and Substructure Management

Billing Status: N/A

Order detail ID: 71278508
ISBN: 978-0-7277-4982-6
Publication Type: e-Book
Volume:
Issue:
Start page:
Publisher: Thomas Telford, Ltd
Author/Editor: Waters, John M. ; Selig, Ernest T.

Permission Status: Granted
Permission type: Republish or display content
Type of use: Republish in a thesis/dissertation
Order License Id: 4380871073614

Requestor type	Academic institution
Format	Print, Electronic
Portion	image/photo
Number of images/photos requested	2
The requesting person/organization	Emmanuel Nii Martey
Title or numeric reference of the portion(s)	Figure 14.4 The tamping sequence; Figure 14.8 The stoneblowing process
Title of the article or chapter the portion is from	Chapter 14. Machines and Methods
Editor of portion(s)	N/A
Author of portion(s)	N/A
Volume of serial or monograph	N/A
Page range of portion	14.3-14.10
Publication date of portion	1994
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	United States
In the following language(s)	Original language of publication

https://www.copyright.com/printOrder.do?id=11727900

7/2/2018 Copyright Clearance Center

With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	Copula-based Models in Railroad Maintenance and Safety Analysis
Instructor name	Prof Nii Attoh-Okine
Institution name	University of Delaware
Expected presentation date	Jul 2018

Figure C.3: Permission to use Figures 2.3 and 2.4