2018-01-01

A Quality Control Performance-Based Methodology For Pavement Management Systems

Edgar Daniel Rodriguez Velasquez
University of Texas at El Paso, edgaredrv@gmail.com

A QUALITY CONTROL PERFORMANCE-BASED METHODOLOGY FOR PAVEMENT MANAGEMENT SYSTEMS

EDGAR DANIEL RODRIGUEZ VELASQUEZ

Master's Program in Civil Engineering

APPROVED:

Carlos M. Chang, Ph.D., Chair

Vivek Tandon, Ph.D.

Vladik Kreinovich, Ph.D.

Charles H. Ambler, Ph.D.
Dean of the Graduate School

Dedication

I would like to dedicate this thesis to my beloved Grandmother, Mariana, who is in Heaven protecting me and guiding all my steps. We miss you!

A QUALITY CONTROL PERFORMANCE-BASED METHODOLOGY FOR PAVEMENT MANAGEMENT SYSTEMS

by

EDGAR DANIEL RODRIGUEZ VELASQUEZ, B.E.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Civil Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2018

Acknowledgements

First, I wish to thank God, the Father Almighty, for providing me the strength and wisdom throughout my graduate studies and for always being there for me, especially in the moments where I needed Him the most. I would like to express my deepest gratitude to my parents, Edgar and Fanny, to my sisters, Ana and Fatima, and to my girlfriend, Carolina, for their unconditional love and valuable guidance.

Secondly, I would like to acknowledge my Alma Mater, the University of Piura, who has given me the opportunity and resources to study a graduate level program at the University of Texas at El Paso. I also would like to thank Dr. Chang, Dr. Tandon, and Dr. Kreinovich for their academic support and dedication in the conception and continuous improvement of my research.

Finally, I want to express my sincerest thanks to all my friends for their help and advice, and for making this journey an unbelievable experience. Without them, this achievement would not have been possible.

Abstract

Transportation Asset Management is a decision-making process, which allocates available resources for operating, maintaining, enhancing, and expanding transportation infrastructure while considering its entire life cycle. Transportation infrastructure includes different types of assets and pavements are one of the main assets due to its social, economic, and environmental impacts to society. Transportation agencies implement Pavement Management Systems to support the pavement management process. While implementing and operating a Pavement Management System, one of the costliest procedures is collecting pavement condition data from the field. Good quality for pavement condition data is required to select the right preservation treatments, estimate the associated costs, model the pavement performance, justify budget needs, and apply well-timed maintenance and rehabilitation strategies.

This thesis focuses on the development of a framework that incorporates a systematic quality control method in the pavement management process. The methodology includes quality control validation checks and statistical tests for data collection of the pavement inventory, condition assessment, and performance modeling. The results of this research contribute to the improvement of data quality used in the pavement management process by identifying poor quality data collected either manually or automatically. This methodology can be applied to training programs, certification programs, pre-collection sites, verification sites, control sites, and sample audits, among other quality control processes.

Table of Contents

Acknowledgements v

Abstract vi

Table of Contents vii

List of Tables ix

List of Figures xi

List of Equations xii

Chapter 1: Introduction 1

1.1 Background 1

1.2 Problem Statement 5

1.3 Research Objectives 5

1.4 Thesis Organization 5

Chapter 2: Literature Review 7

2.1 Transportation Asset Management Overview 7

2.2 Pavement Management Systems 13

2.3 Pavement Condition Data 16

2.4 Quality in Pavement Condition Data 18

2.5 Quality Management Techniques 24

2.6 Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management 28

2.7 Pavement Performance Prediction Models in Pavement Management Systems 51

Chapter 3: Framework to Incorporate Quality Control in Pavement Management Systems 61

3.1 Framework for Quality Control in Pavement Management Systems 61

3.2 Stage I. Policy Goals and Objectives 65

3.3 Stage II. Pavement Inventory 66

3.4 Stage III. Condition Assessment 68

3.5 Stage IV. Statistical Quality Control Tools 74

3.6 Stage V. Performance Modeling 89

3.7 Stage VI. Determination of Needed Work and Funds 90

3.8 Stage VII. Identification of Candidate Projects 91
3.9 Stage VIII. Determination of Impacts of Funding Alternatives 92
3.10 Stage IX. Budget Allocation 93
3.11 Stage X. Feedback 93
Chapter 4: Case Studies Analysis 96
4.1 Raters Comparison Case Study 96
4.2 Rating Protocol Update Case Study 120
Chapter 5: Summary of Research Findings, Conclusions, and Recommendations 131
5.1 Summary of Research Findings 131
5.2 Conclusions from the Case Studies 133
5.3 Recommendations for Future Research 136
References 138
Appendix A 148
Appendix B 152
Appendix C 172
Appendix D 176
Appendix E 178
Appendix F 183
Appendix G 186
Appendix H 189
Appendix I 192
Appendix J 211
Appendix K 221
Appendix L 231
Appendix M 235
Appendix N 238
Vita 241

List of Tables

Table 1: Interpretation of Kappa Values 29

Table 2: Statistical Model-Free Replacement Techniques 31

Table 3: Statistical Model-Based Replacement Techniques 32

Table 4: Summary of Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management 48

Table 5: Statistical Quality Control Tools 75

Table 6: PCI Values of the Ground Truth (GT) and Raters 97

Table 7: MTC Passing Criteria Results for Raters 99

Table 8: Descriptive Statistics of PCI Values 100

Table 9: Shapiro-Wilk Test Results 103

Table 10: Variances Comparison Based on the F-Test and the Levene's Test 104

Table 11: Means Comparison Based on the T-Test and the Man-Whitney Test 105

Table 12: Percent Agreement Values when comparing each Rater with the Ground Truth 106

Table 13: Cohen's Kappa Coefficients: Comparison of Raters versus the Ground Truth 107

Table 14: Weights Used for Weighted Cohen's Kappa Calculation 108

Table 15: Weighted Cohen's Kappa Coefficients: Raters versus Ground Truth Comparison 109

Table 16: Fleiss' Kappa Results for All the Raters 110

Table 17: Interclass Correlation Coefficients for each Rater and Overall Agreement 111

Table 18: Kendall's Coefficient of Concordance for Each Rater and Overall Agreement 113

Table 19: Estimated Bias, Outliers, and Proportional Bias from the Bland-Altman Plot 116

Table 20: Recommended Acceptance Values for Statistical Quality Control Tools 117

Table 21: Summary of the Quality Control Tools Results 119

Table 22: Descriptive Statistics of PCI_7 and PCI_8 Values Including all the Data 122

Table 23: Descriptive Statistics of PCI_7 and PCI_8 Disregarding Outliers 123

Table 24: Descriptive Statistics of PCI_7 and PCI_8 for Each Agency 123

Table 25: PCI_8 versus PCI_7 Values for Individual Sections 124

List of Figures

Figure 1: Life Cycle of Infrastructure Assets 7

Figure 2: Total Auto and Truck VMT (trillions) and GDP (trillions of $2005). 10

Figure 3: VMT Tendency for the United States 1970-2014 11

Figure 4: VMT per Capita and GDP (based on estimates or forecasts) 11

Figure 5: Relationship of VMT and GDP, 1960-2017 12

Figure 6: Pavement Condition Data Quality Management Framework 21

Figure 7: Big Data Concept Applied to Pavement Management 38

Figure 8: Parts per Million Defective for ±3σ and ±6σ Quality. 42

Figure 9: Generic Asset Management System Components 62

Figure 10: An Asset Management Approach to Resource Allocation and Project Delivery 63

Figure 11: Framework for Quality Control in Pavement Management Systems 64

Figure 12: Flowchart of the Statistical Quality Control Tools 83

Figure 13: Mean, Range, and Standard Deviation Comparison 101

Figure 14: Distribution with a positive skewness 102

Figure 15: QQ plots for Rater 6 and Rater 7 103

Figure 16: Bland-Altman Diagrams for Rater 3 and Rater 17. 115

Figure 17: PCI_7 and PCI_8 Values of Pavement Sections Including all the Data 121

Figure 18: PCI_7 and PCI_8 Values of Pavement Sections Disregarding Outliers 122

Figure 19: Transition Condition Matrices for each Agency 125

Figure 20: Bland-Altman Diagrams for each Agency 127

List of Equations

Equation1: Mathematical Formulation of Bayesian Theorem. 35

Equation 2: Bayesian Theorem for Continuous Variables. 36

Equation 3: Bayesian Theorem for Discrete Variables..... 36

Chapter 1: Introduction

1.1 Background

Transportation Asset Management (TAM) is a systematic and strategic process focused on effectively operating, maintaining, enhancing, and expanding transportation infrastructure while considering its entire life cycle. Business and engineering knowledge and practices are used to improve the decision-making process based on good quality data and clear-defined objectives. In this way, available resources designated for infrastructure management would be adequately allocated and utilized (FHWA, 2007).

TAM includes the managing of the numerous transportation assets and their performance, such as pavements, bridges, tunnels, railroads, culverts, and ports, among others. Pavements are considered as one of the major transportation assets that contribute to the nation's sustainable development. Communication among communities, access to services, movement of freight and commodities, low vehicle operating costs, low fuel consumption, and low CO_2 emissions, are some of the social, economic, and environmental impacts that pavements, in acceptable condition, can provide. During the pavement management process, transportation agencies perform engineering and economic analysis with the purpose of adequately managing the pavement network. This process is supported through the implementation of a Pavement Management System (PMS). The American Association of State Highway and Transportation Officials (AASHTO) defines a Pavement Management System as a "set of tools or methods that assists decision-makers in finding the optimum strategies for providing, evaluating, and maintaining pavements in a serviceable condition over a period of time" (Huang, 2004). The purpose is to preserve these transportation assets at a level of service that will positively influence social, economic, and environmental aspects of society, while improving the quality of life.

Approximately 2.6 million miles of paved public roads in the United States are being managed through PMSs by transportation agencies. One of the costliest procedures of implementing and operating a PMS is gathering pavement condition data. It is critical to collect

pavement condition data that truly reflects real in situ conditions. Poor quality data costs American businesses $600 billion annually (Data Warehousing Institute, 2002). A critical component is the quality of pavement condition data collected from field surveys. Transportation agencies' decision-making process for asset management has its basis on the quality of pavement condition data (Flintsch & McGhee, 2009).

Good quality for pavement condition data is required to justify the costs associated with its acquisition. According to AASHTO (2001b), "a properly planned and implemented data collection program will significantly increase credibility, cost-effectiveness, and overall utility of a PMS". Good quality pavement condition data is needed for the performance models to accurately predict pavement condition and for the support of managerial decisions related to properly timed interventions. Poor quality data (erroneous data that does not represent real in situ pavement conditions) would generate erroneous pavement condition forecasts and, as a consequence, inadequate decisions regarding maintenance and rehabilitation strategies.

Quality in pavement condition data is crucial for an effective implementation of pavement management decisions, such as the selection of the most cost-effective pavement treatment and its optimal timing of application. Therefore, transportation agencies need to define procedures, techniques, tools, and alternatives for improving the quality of pavement data gathered from the field.

Quality can be evidenced in the variability of pavement condition data, which leads to a lack of data consistency. Variability can be measured by comparing the condition data of certain pavement segments against the correct values for those segment's condition. These correct values are known as ground truth values or reference values (Morian, Stoffels & Frith, 2002). Before and during data collection, these type of comparisons might identify any variability on the pavement condition data.

Other examples of variability would be the difference in rut depth between two wheelpaths, errors in the distress' extent and severity, variation in roughness measurements for the same pavement section using different equipment, etc. Random causes of variability cannot be

identified, but assignable causes of variability would be identifiable and avoided by equipment calibrations or additional training for raters (Montgomery, 2012).

Variability has a negative effect on pavement performance forecasting and consequently on treatment recommendations. If the prediction of pavement deterioration is not accurate due to high variability, pavement treatment recommendations might not reflect real needs nor conditions on the site. For instance, a pavement segment based on poor quality data is predicted to last 20 years, rather than 25 years. This situation can generate higher budget needs because treatments might be applied earlier than optimal during the infrastructure´s life. Therefore, decisions related to planning and programming strongly depend on the quality of pavement condition data.

More examples of the effect of variability in pavement condition data can be found in the calculation of indicators, such as present serviceability rating (PSR), present serviceability index (PSI), or pavement condition index (PCI). Pavement condition index is a numerical rating of the pavement condition that depends on the type, extension or density, and severity of pavement's distresses present in a sample of the network (ASTM D6433, 2018). Variability in the data gathered during the visual inspections may have a negative impact on pavement condition. Just one percent difference in the density of low-severity alligator cracking can make an 8-point difference in the 100-point PCI calculation. This difference might generate inadequate treatment recommendation and therefore might have economic consequences. If lower types of severity levels are considered for the distresses, the PCI variability would be reduced (Ponniah, Sharma & Kazmierowski, 2001). Variability on field survey data must be reduced to assure that the results reflect real pavement condition.

Variability in pavement condition data will have a considerable impact on the treatment strategies and budget needs. These errors should be eliminated by implementing an adequate quality management system. A case in Virginia showed that with the correction of errors in pavement condition data, the number of pavement sections requiring maintenance were reduced 83% and the number of pavement sections requiring no maintenance were increased in 22%. The final impact was a saving of $18 million in maintenance (Shekharan, Frith, Chowdhury, Larson &

Morian, 2007). Another research showed that a standard error of ±10 points based on a 0 to 100 scale for a condition index could result in a 2 to 6% error in the estimation of the number pavement segments that needed maintenance or rehabilitation (Saliminejad & Gharaibeh, 2014).

Tan and Cheng (2014), stated that errors in pavement condition data may have an impact on the current pavement state, deterioration rate, projection of future condition, maintenance and rehabilitation needs, and on the cost associated with repairing pavement sections. The selection of maintenance and rehabilitation treatments and budgets estimation are affected by poor data quality.

Background Summary

Transportation Asset Management (TAM) is a decision-making process that allocate available resources for operating, maintaining, enhancing, and expanding transportation infrastructure throughout its entire life cycle. Since pavements can be considered as core transportation assets for sustainability development, its effective management is necessary. Pavement management is effectively developed when a Pavement Management System (PMS) is defined and implemented. While implementing and operating a PMS, one of the costliest procedures is collecting pavement condition data from the field. Good quality for pavement condition data is required to select the right preservation treatments, estimate the associated costs, model the pavement performance, adequately justify budget needs, and apply well-timed maintenance and rehabilitation strategies. On the contrary, poor quality pavement condition data has a negative impact on pavement management practices leading to investment decisions that rely on inaccurate information.

Good quality in pavement condition data is required for an appropriate decision-making process. Economic losses due to the incorrect application of maintenance, rehabilitation, or reconstruction works during the pavement life are some examples of the problems encountered when poor quality data is used in the decision-making process.

1.2 Problem Statement

The problem to be addressed in this thesis is the need of a systematic quality control methodology integrated to pavement management practices. A comprehensive quality control framework is required to assist an agency through the entire pavement management process. This framework should include a quality control methodology that incorporates statistical tools to identify poor quality data, take corrective actions, and improve data quality.

1.3 Research Objectives

The research objectives defined in the thesis are:

1. To develop a framework that incorporates a systematic quality control method in the pavement management process to identify poor quality data and make corrective actions to improve its quality.
2. To propose a methodology that includes quality control validation checks and statistical tests of field data collected for the pavement inventory, condition assessment, and the development of pavement performance models. The methodology should be able to identify errors and outliers for corrections.
3. To apply the quality control methodology into two case studies. In the first case study, pavement condition field data from 18 raters for condition assessment are analyzed. In the second case study, a larger dataset is analyzed to evaluate the differences observed in pavement condition assessment using two versions of a rating distress manual.

1.4 Thesis Organization

This thesis is divided in five chapters.

Chapter 1 describes the importance of pavement condition data as part of a Pavement Management System and the negative impact of poor quality data in pavement managerial decisions.

A literature review is provided in Chapter 2, where the data quality management process is defined and the techniques and tools related to quality control, quality acceptance, and independent assurance are described. Statistical techniques, mathematical models, data analysis techniques, quality approaches, and softwares applicable to quality management are also covered in the same chapter as well as the impact of pavement condition data over pavement performance models.

A framework to incorporate a quality control in pavement management decision-making process, is described in Chapter 3. It includes a methodology to incorporate quality control into pavement management practices. The methodology focuses on pavement inventory, condition assessment, and performance modeling. Statistical quality control tools are proposed in a flowchart to evaluate the quality of pavement condition data by comparing it with reference values.

In Chapter 4, the methodology is implemented in two case studies: “Raters Comparison Case Study” and “Rating Protocol Update Case Study”. Real pavement condition data collected from the field is utilized in each case study. The results are interpreted and discussed in the same chapter.

Finally, Chapter 5 describes the conclusions and recommendations for future research.

Chapter 2: Literature Review

2.1 Transportation Asset Management Overview

Transportation Asset Management (TAM) is a decision-making process which includes strategic, systematic, and coordinated planning and programming of investments or expenditures in order to operate, maintain, upgrade, and expand physical transportation assets effectively throughout their entire life cycle (AASHTO, 2011). The purpose of TAM is to ensure that the assets provide a level of service acceptable to society, which will enhance quality of life with a positive impact in terms of sustainable development (see Figure 1). The life cycle of a transportation asset may typically include planning, design, construction, operation, maintenance, rehabilitation, and decommission. Each phase of the cycle should account for social, economic, and environmental impacts to achieve sustainability.

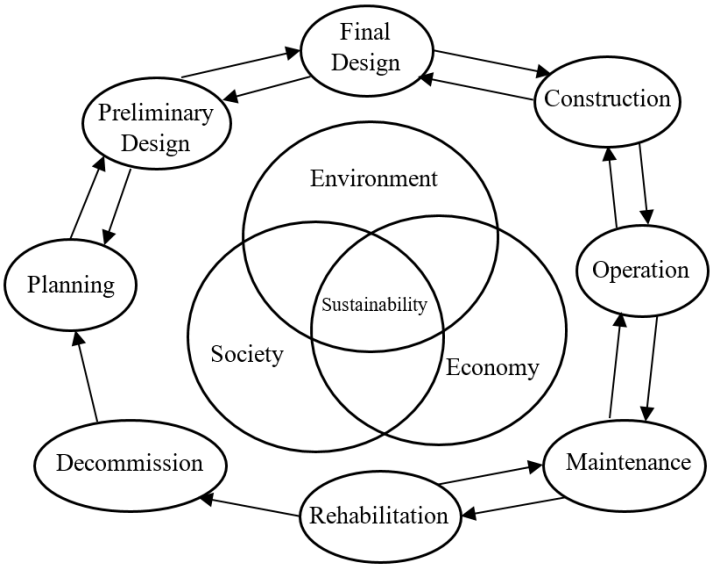


Figure 1: Life Cycle of Infrastructure Assets.
Source: adapted from Chang (2016).

TAM comprises managing various types of transportation assets. According to Uddin, Hudson, and Haas (2013), transportation infrastructure can include different types of assets. Ground transportation facilities (e.g. pavement networks, bridges, tunnels, and railroads), air transportation facilities (e.g. airports, heliports, and air-traffic control services), waterways and

ports (inland waterways, shipping channels, and terminals), mass transit facilities (e.g. subways, bus transit, light rail, and monorails), and pipelines (e.g. natural gas ducts, and crude oil ducts) are some examples of transportation assets.

One of the transportation assets that can be considered as fundamental for a nation's sustainable development are the pavement networks. Pavement networks are ground transportation facilities that provide mobility and access to the user. These types of assets are important for the nation's economic progress and for communication among communities. Low number of car accidents, low maintenance costs, and reduced Greenhouse Gas (GHG) emissions are some examples of the social, economic, and environmental impacts that pavements in acceptable conditions can provide. These conditions are defined by transportation agencies that manage the pavements to obtain maximum benefits with available funds.

Pavement management has been used by federal, state, and local transportation agencies to assess pavement condition, estimate needed funds, identify pavement preservation recommendations, and justify funding alternatives, among others (AASHTO, 2012). AASHTO (1993) defines pavement management as "a set of tools or methods that assist decision-makers in finding optimum strategies for providing, evaluating, and maintaining pavements in a serviceable condition over a period of time". Pavement management is policy driven and involves all management levels: strategic, network, project-selection, and project level.

Strategic pavement management decisions are related to general pavement management problems. These challenges are mainly associated with investment analysis and funding allocation. Some strategic decisions are related to the determination of funds for transportation facilities regarding maintenance, rehabilitation, and construction works. Other decisions at this management level are associated with the definition of funding alternatives, justification of funds, policies definition, and the appropriate communication to funding authorities.

Pavement management decisions at the network level are related to the development of a prioritization program for treatment lists of candidate pavement sections, taking into account maintenance and rehabilitation needs, funding needs, impact of funding alternatives, and budget

constraints. Network level decisions are associated with the budget process and can be used to identify pavement maintenance, rehabilitation, and reconstruction needs; determine funds necessary to fulfill these needs; identify viable funding alternatives and strategies to be analyzed; estimate the impact of funding options; and develop optimal pavement budget recommendations.

The project-selection level centers around a specific location and includes prioritization of projects, scheduling, and the identification of physical and financial constraints not previously considered. At the project-selection level, pavement management decisions have an influence on the completion of prioritization and optimization processes of pavement segments programmed for work. Furthermore, the improvement of cost estimates for the selected pavement segments are also covered at this management level.

Project level decisions are related to the assessment of needs for maintenance, rehabilitation, construction or cause of deterioration. Other managerial decisions at this level are the definition of viable strategies regarding design, maintenance, rehabilitation, and reconstruction; the completion of final project designs; the planning of construction schedules; the cost-effectiveness analysis of strategies; and the definition of constraints related to safety, time and economic limitations.

As previously stated, pavement networks are considered crucial transportation assets due to their positive impact on a nation's economic development. Figure 2 shows the comparison between United States Vehicle-Miles of Travel (VMT) with the Gross Domestic Product (GDP) ranging from 1936 to 2011. It is noticeable that both indicators have grown mostly in parallel since 1936, excluding the years in which World War II occurred (Ecola & Wachs, 2012).

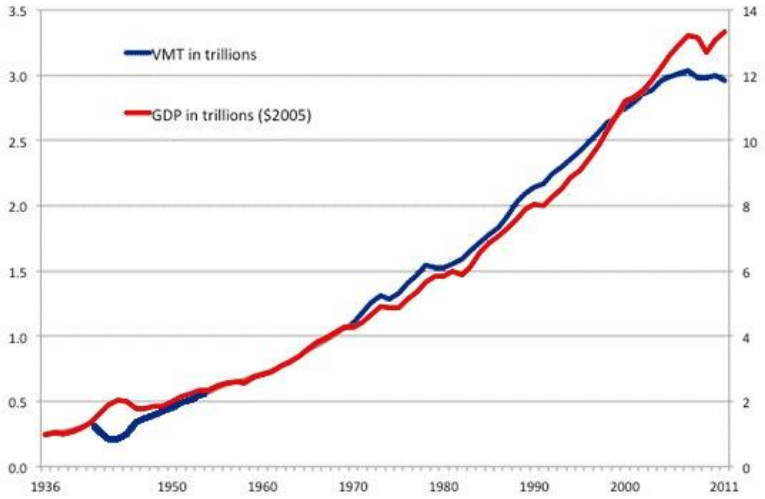


Figure 2: Total Auto and Truck VMT (trillions) and GDP (trillions of $2005).
Source: Ecola and Wachs (2012).

Sundquist and McCahill (2015) made a similar analysis by comparing U.S. VMT in trillions with U.S. VMT per capita. VMT per capita is calculated by dividing the total annual miles of a vehicle travel by the total population. Figure 3 shows both indicators. It is noticeable that both curves, VMT per capita and total VMT, have a peak in year 2004 and 2007 respectively, followed by a decrease and finally a rise starting in 2012. The authors also determined the relations between the actual VMT per capita, the GDP based on estimates (1970-1995), and the GDP based on forecasts (1995-2014) using regression analysis. Figure 4 shows that the trend between these two curves are similar throughout all the analysis period, with more correlation from 1970 to 1995. After these years, the difference between the curves is considerable. From Figures 3 and 4, it can be concluded that total VMT has a positive correlation with the GDP-based estimate.

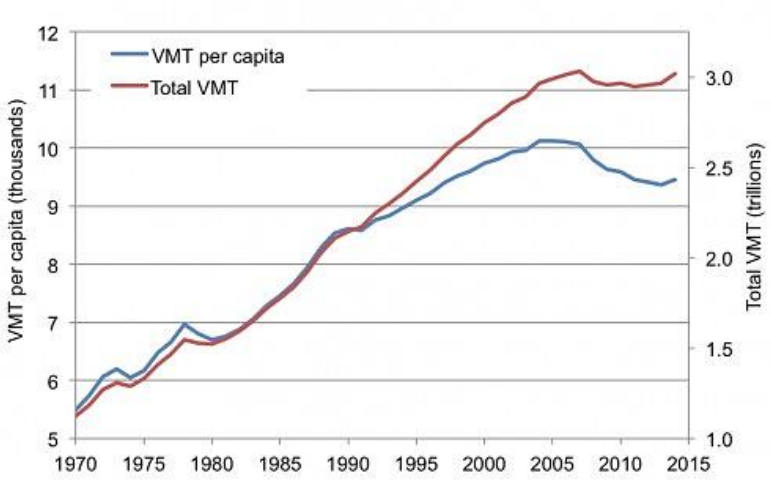


Figure 3: VMT Tendency for the United States 1970-2014.
Source: FHWA and Census Bureau (2015).

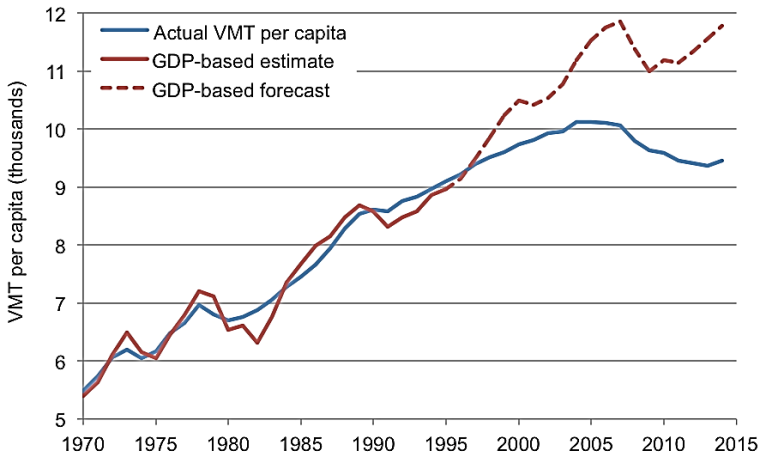


Figure 4: VMT per Capita and GDP (based on estimates or forecasts).
Source: FHWA, Census Bureau, and Bureau of Economic Analysis (2015).

Another graph that shows the relation between VMT and GDP is presented in Figure 5. According to the Office of Energy Efficiency and Renewable Energy (EERE) of the United States, the VMT and the GDP have grown at an average rate of 3.5% annually from 1960 to 1997. After those years, the growth of GDP (2.1%) has surpassed the growth of VMT (1%). The gap between them in 2017 is the largest since 1960. The factor that contributes to this behavior is the growth of those economic activities that do not involve an increase in travel. However, the contribution of VMT to the growth of the GDP is important and has been continuous throughout the years (EERE, 2018).

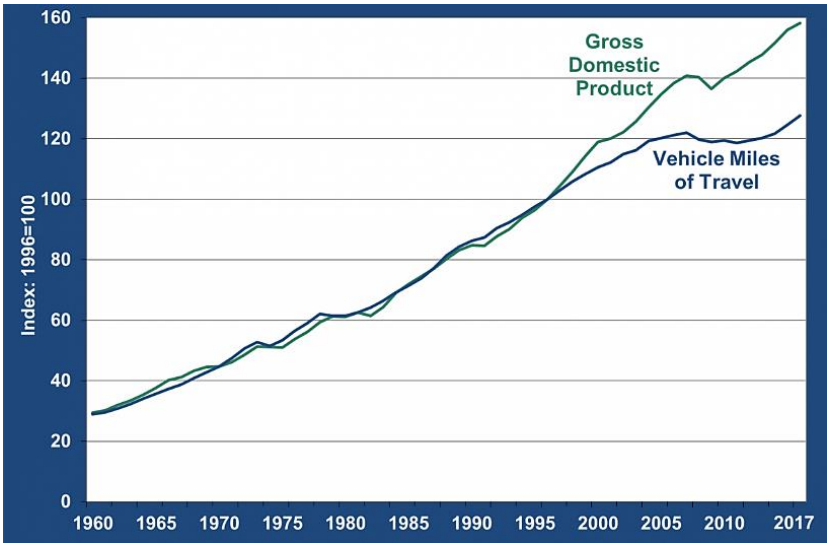


Figure 5: Relationship of VMT and GDP, 1960-2017.
Source: Office of Energy Efficiency and Renewable Energy (2018).

Since pavement networks are important in the contribution of this economic development, these assets must preserve an acceptable condition during their entire service life. A pavement network's current condition is the initial state of the asset that can be reported based on measurable indicators that represent specific characteristics including roughness, friction, serviceability, and distress extension, among others. Current condition is the starting point to evaluate the deterioration of pavement condition over time, make managerial decisions, and address all the problems that may affect its performance. Some of the challenges encountered by pavement networks are the deterioration of infrastructure over time, an increased demand of people using the infrastructure, and a limited availability of funds and resources.

Satisfactory, cost-effective, and well-timed decisions are required by the highway agencies to address those problems and guarantee pavement networks to provide a level of service that will generate positive effects, taking into account social, economic, and environmental impacts. In order to achieve this, highway agencies in the United States have implemented Pavement Management Systems.

2.2 Pavement Management Systems

A Pavement Management System (PMS) is a group of decision-support tools and procedures that provide cost-effective strategies to transportation agencies in order to provide, evaluate, and maintain pavements in a serviceable condition over time (AASHTO, 1993). Pavement management includes planning, programming, and budgeting treatment needs related to maintenance or rehabilitation activities (Al-Zou'Bi, Chang, Nazarian & Kreinovich, 2015). The decision-support methods and pavement management tools that are part of a PMS are all interrelated. The purpose of a PMS is to provide, evaluate, and maintain pavement networks in a serviceable condition (AASHTO, 1990). PMSs can assist a transportation agency in the application of cost-effective treatments, allocation of funds, and quality improvement of the pavement network (AASHTO, 2001a).

Pavement management tools can be applied to all management levels. According to Haas, Hudson, and Zaniewski (1994), pavement management activities are principally conducted at two distinctive levels: the network level and the project level. The network level represents an overall aspect of pavement management and covers the general budget and planning activities. The project level is more specific and has a more centered perspective, focusing on a defined component of the entire network. At the project level, decision-makers are responsible for designing maintenance, rehabilitation, and reconstruction strategies, as well as analyzing funding allocation.

Network level activities are typically related to the budget process, which is the final product to achieve at this management level. The first activity is the identification of needs and funds for pavement maintenance, rehabilitation, and reconstruction. After this process, some feasible funding alternatives and strategies are compared to forecast and quantify their future influence on the performance of the pavement, including social, economic, and environmental impact to the user. At the end, pavement budget recommendations can be developed. PMSs can assist network level management during the planning, programming, budgeting, and analysis phases (Huang, 2004).

Project level activities aim to define the most optimal maintenance, rehabilitation, or reconstruction strategy with the highest cost-effectiveness and feasibility, taking into account economical and time-related constraints for the specific pavement section. The project level includes the needs assessment for construction or deterioration cause; the identification of feasible strategies related to maintenance, rehabilitation, and reconstruction; the evaluation of alternatives based on cost-effective criteria; and the selection of the most appropriate strategy taking into account imposed constraints. Technical staff usually use project level results for the design phase, where final plans and specifications are developed (Huang, 2004).

According to Muench, Mahoney, and Pierce (2003), there are five broad elements of PMSs: inventory data, pavement condition surveys, analysis scheme, decision criteria, and implementation procedures.

- Inventory data is the information that describes basic characteristics of pavement sections in the road network. Location, construction history, traffic, and the physical geometry of the cross section are some examples of inventory data.
- Pavement condition surveys comprise data collected from pavement sections, which express pavement performance. The data collected, such as skid resistance, ride quality, distresses, etc. are useful for pavement sections condition monitoring and for the management activities' effectiveness measurement.
- The analysis scheme includes the mathematical interpretation of data through algorithms to forecast future pavement performance, perform cost analysis, and estimate the impact of maintenance and rehabilitation strategies.
- The decision criteria contains guidelines or rules (e.g. decision trees and decision matrices) to guide decisions regarding pavement management, such as the selection of appropriate maintenance and rehabilitation techniques for roadway sections.
- The implementation procedures are methods to execute the management decisions to the pavement sections.

Each of the previously described elements have to be satisfactorily implemented for a PMS to be effective.

A PMS must be designed based on reliable data, clearly defined procedures, and calibrated models in order to quantify the consequences of the possible decisions that are being evaluated. The data needed in a PMS is very broad. It can include inventory information (e.g. pavement structure, geometry of the road, and cost per mile of the pavement section), road usage (e.g. volume of traffic, trucks axle configuration, and types of loads applied), pavement condition (e.g. smoothness, type, quantity and severity of distresses, skid resistance, and structural capacity), and pavement construction, maintenance, and rehabilitation activities throughout the life cycle.

Pavement condition data is a critical component of any Pavement Management System (Pierce, McGovern & Zimmerman, 2013). One of the costliest parts of operating a PMS is the collection of pavement distress data at the network and project level. For that reason, the quality of the pavement condition data needs to be accurate, complete, and consistent due to its influence on pavement management decisions regarding funding allocation to maintenance and rehabilitation needs as well as needs assessments for construction, maintenance or rehabilitation to refine final project designs considering safety, time, and economic constraints.

Highway agencies need pavement condition data quality management techniques to gather reliable data at the lowest level of detail, sufficient to make appropriate decisions (Bennett, Chammoro, Chen, Solminihac & Flintsch, 2005). Appropriate decisions will generate positive social, economic, and environmental impacts after they are made. For instance, timely cost-effective maintenance strategies applied to a pavement segment based on pavement condition data gathered from the field will provide users with comfortable road circulation due to the adequate performance of the asset.

According to Flintsch and McGhee (2009), the techniques used for pavement data quality management are: (1) calibration of the equipment and verification of the analysis criteria before data collection, (2) testing of known control or verification sites before and during data collection,

(3) software routines, (4) time-series data analysis, (5) independent data verification and validation, and (6) the use of blind site monitoring.

2.3 Pavement Condition Data

Pavement Management Systems depends on complete, accurate, and reliable pavement condition data. This type of data is the support for pavement performance modeling; maintenance, rehabilitation, and reconstruction planning, and program effectiveness evaluation.

At the network level, pavement condition data is collected in large quantities. The data collected at the network level is related to the smoothness and distresses of the pavement structure. Automated technologies are used to gather pavement condition data from road networks in short periods of time. These large volumes of data are usually converted into condition indices. Based on these indices, network level activities can be performed, such as the assessment of pavement conditions, definition of maintenance and rehabilitation strategies, application of budget allocation programs, and prioritization of pavement segments, among others (Flintsch & McGhee, 2009).

At the project level, more detailed pavement condition data is collected. The data collected at the project level is related to pavement distresses, distresses' severity, friction and pavement's structural capacity (deflections). Walking surveys are used to gather pavement condition data at the project level. Based on this data, project level activities can be performed, such as the definition of more specific maintenance and rehabilitation methods, determination of funding requirements for particular projects, treatment designs, and selection of treatments based on decision tress, among others (Flintsch & McGhee, 2009).

The most common methods for collecting pavement condition data for network level management and project level management, are manual and automated surveys. Automated surveys are divided in semi-automated and fully automated.

Manual surveys

Manual surveys are performed by traveling at slow speed or by walking and identifying the distresses on a pavement surface. Manual surveys can be applied to the entire length of a road or be performed to a sample. Distresses can be registered on paper, smartphones, tablets, computers or any other similar device.

Examples of standards associated with manual surveys are the Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys ASTM D6433-18 (ASTM, 2018) and the Distress Identification Manual for The Long-Term Pavement Performance Program (FHWA, 2014), published by ASTM and FHWA, respectively. Private and governmental agencies as the Metropolitan Transportation Commission (MTC) of California have also created their own data gathering procedures and distress rating protocols called Condition Index Distress Identification Manual for Flexible Pavements (MTC, 2016a), and Condition Index Distress Identification Manual for Rigid Pavements (MTC, 2016b). MTC is a public, governmental transportation agency responsible for planning, funding allocation, and managing streets, roads, highways, transit systems, airports, and other transportation assets for the nine counties located in the San Francisco Bay Area.

Automated surveys

Automated surveys are performed by traveling at high speed in vans, which are equipped with lasers, computers, and cameras designed for pavement data collection. Transverse, as well as longitudinal profiles of the road surface, are captured trough digital images. All the information acquired by automated surveys requires processing before using it in managerial decisions.

The processing for semi-automated surveys includes the visualization of distresses in the collected images. The personnel perform these inspections at workstations. Software is utilized to display the images and record the distresses. If sensors are used to acquire pavement condition data, processing is needed for determining indices as the International Roughness Index, or distresses as rut depth or faulting.

The processing for fully automated surveys does not include any personnel. A pattern recognition technology is used to identify the distresses based on the collected images. Video and laser technology is utilized to detect and determine the type of distresses. Processing when sensors are used has the same features when semi-automated surveys are applied (Pierce et al., 2013).

Examples of standards associated with automated surveys are the Standard Guide for Classification of Automated Pavement Condition Survey Equipment, Automated Data Collection for Pavement Condition Index Survey, and the Automated Pavement Distress Collection Techniques. These documents where published by the American Society for Testing and Materials (ASTM), the Transportation Research Board (TRB), and the National Cooperative Highway Research Program (NCHRP), respectively (ASTM, 2016; Gregory, 2003; McGhee 2004).

2.4 Quality in Pavement Condition Data

ISO 9000 (2005) defines quality as "the degree to which a set of inherent characteristics fulfill requirements". Under the pavement condition data quality management approach, requirements can be considered as features of the data collection process. These features can be specified in a contract or defined by the transportation agency.

Traditional data quality concepts and principles have changed over the last few decades (ISO, 2008). The traditional approach of data quality considers a unique reference value, also called true value or ground truth, which represents the correct value for pavement condition data. This value is determined by the most trained, certified, and experienced raters, who utilize proper methodologies and calibrated equipment. For surface deterioration ratings, the reference value is estimated by consensus due to the subjectivity of the distresses.

The measurements have to get close to the true value to be considered as quality data. Deviations are due to errors, which can be random or systematic. If the measurements are dispersed around the reference value, data contains random errors and has low precision. For instance, if during a visual inspection, a rater assigns a higher severity level to a structural distress (e.g.

alligator cracking), the final calculation of the condition indicator might be affected by this random error.

When the mean of the measurements is shifted away from the true value, data contains systematic errors and has low accuracy. For example, the Pavement Condition Index (PCI) is an indicator that represents pavement condition and can range from 0, being the worst condition, to 100, being the best condition (ASTM, 2018). For the determination of the PCI, specific charts are used to determine the deduct values for each distress. Then, based on the deduct values, the pavement condition index is calculated. If the chart corresponding to weathering is used to calculate the deduct values for rutting, the final PCI would be affected by a systematic error.

The combination of random and systematic errors has a higher negative impact on data quality. However, in large numbers of measurements, systematic errors have more influence on data quality than random errors (Flintsch & McGhee, 2009).

The current data quality approach incorporates the concepts of trueness and uncertainty instead of the old terminology of precision and accuracy. Trueness is related to the closeness between the mean of the measurements and the true value, and uncertainty describes the acceptable dispersion of the measured values. Standard deviations, confidence intervals, and other statistical indicators can represent the trueness or uncertainty of a set of measurements (Pierce et al., 2013).

The two data quality approaches identify cases when the collected data do not reflect the real pavement condition in the field. In order to reduce this situation, data quality management becomes relevant to achieve reliable and complete pavement condition data. Transportation agencies implement data quality management through the application of Quality Management Systems. This system defines the procedures for data acquisition, training, data processing, quality control, and quality acceptance. These activities are considered under a Quality Management Plan. This plan is a document that describes the planning, implementation, and assessment processes needed to evaluate the agency´s effectiveness of its pavement data collection, quality control, and acceptance.

Regulations in the United, such as the Moving Ahead for Progress in the 21st Century Act (MAP-21) and Fixing America's Surface Transportation Act (FAST ACT), have required the development of a Pavement Data Quality Management Plan for data collection and processing to evaluate pavement performance. These legislations provide long-term funding for surface transportation infrastructure planning and investment. Both acts have required transportation agencies to implement an effective Quality Management Plan for pavement condition data collection (FHWA, 2012a & FHWA, 2016).

An effective Quality Management Plan is focused on the definition of methods, standards, protocols, or guides that will be used to adequately gather pavement condition data on the field. These documents must clearly describe the types of pavement distresses, the severity levels associated with each distress, the measurement of the distresses, how frequently the inspections should be made, and the procedure to calculate the final condition value.

If low quality pavement condition data is detected, the Quality Management Plan should define corrective procedures (before the data gathering process is performed) as soon as possible. Corrective activities may include calibration of equipment, further rater training, the re-collection of data, or rerating of pavement segments.

An example of a data Quality Management Plan is the one developed by the Metropolitan Transportation Commission (MTC). The plan includes three components. The first component is a prequalification process for consultants called the rater certification program (StreetSaver Academy, 2018). Raters who are certified have the skills to determine pavement condition with a desired accuracy level. The second component is the quality control plan applied before, during, and after data collection. Finally, the third component is quality acceptance to check the effectiveness of the quality control process by the consultants (Tan & Cheng, 2014).

Shekharan, Frith, Chowdhury, Larson, and Morian (2006) analyzed the benefits that the implementation of a Quality Management Plan can provide to a transportation agency:

- Better fulfillment of external data requirements.
- Better credibility within the agency.

- Better integration with other transportation agencies.
- More appropriate maintenance and rehabilitation recommendations.
- Improved quality (accuracy and consistency) in pavement condition data.
- Improved transportation management decisions
- Improved deficient pavement identification and reporting
- Improved condition indices calculations.
- Better determination of budget needs.

Figure 6 shows a framework associated with pavement condition data quality management. The framework considers the three relevant quality management processes, which are quality control, quality acceptance, and independent assurance. For each process, quality management techniques are listed depending on its application: before, during, or after data collection.

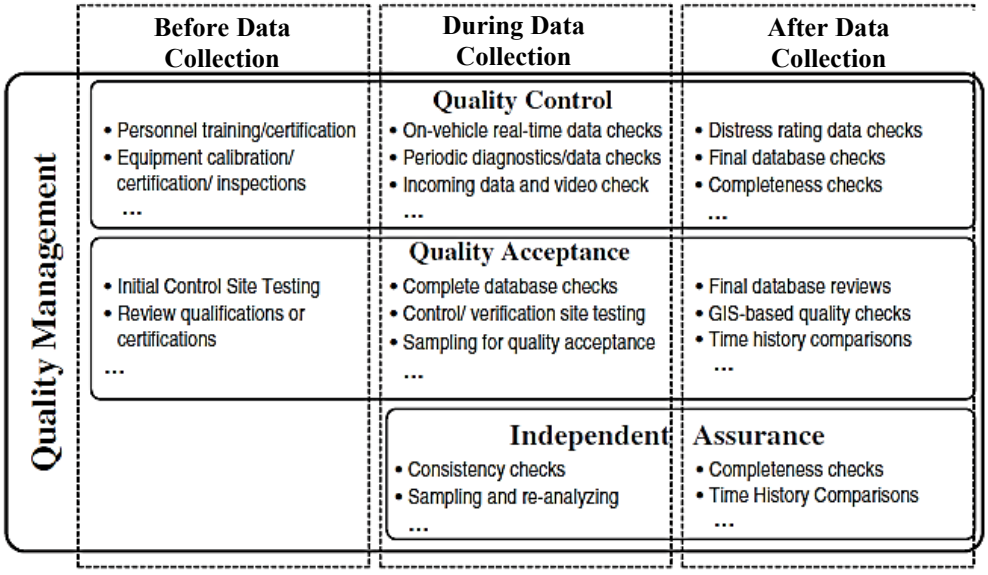


Figure 6: Pavement Condition Data Quality Management Framework.
Source: adapted from Flintsch and McGhee (2009).

2.4.1 Quality Control

The Federal Highway Administration defines quality control as the "actions taken to measure the quality of the data to identify its compliance with the required quality standard" (Simpson, Rada, Bryce, Serigos, Visintine & Groeger, 2018). This actions or techniques are able to assess, calibrate, validate, and verify data gathering processes to obtain quality pavement condition data.

Quality control aims to measure the variability in the data obtained during and after the data acquisition process. Once it is quantified, the variability´s causes are determined and controlled, if possible, to reduce or keep it within acceptable limits. The collection process is adjusted to minimize variability.

Causes of variability for pavement condition data collection can be related to the equipment used to collect the data, the operator´s experience while using the equipment, the rater´s skills if manual or automated inspections are performed, in-site environmental factors, the pavement condition, traffic, and the geometric characteristics of the road, among others. These potential causes have to be taken into account and controlled due to their influence on the quality of the data.

The impact of data variability also affects forecasting pavement deterioration, maintenance and rehabilitation timing, and allocating the budget due to treatment recommendations. In addition, the quality control process must identify any problems during data collection, as soon as possible. This will avoid the gathering of large amounts of low quality data.

As indicated in the pavement condition data quality management framework (Figure 6), quality control can be implemented not only during or after data collection, but also before. In that stage, quality control techniques can include: (1) an update of clearly documented manuals and procedures, (2) an analysis of the criteria for data collection, (3) a training and supervision of personnel in charge of the data collection, (4) an equipment calibration, testing of controlled segments, maintenance, and inspection procedures, and (5) a certification of raters and equipment (Flintsch & McGhee, 2009).

During data collection, quality control techniques can cover: real-time data checks, periodic data checks, and the testing of control sections. Finally, after data acquisition is performed, quality control can include distress rating data reviews, database reviews, and software routines to identify out of range data or missing road sections or elements.

2.4.2 Quality Acceptance

Once the pavement condition data has been collected, it is necessary to evaluate if the information gathered from the field is in conformity with the acceptance criteria. Quality acceptance techniques are then needed to conduct this verification process, before the data is used to support managerial decisions at any level of transportation management.

Quality acceptance processes require that all the data or an appropriate sample be validated to meet the acceptance criteria. Data precision, accuracy, and reliability are established to check if any part of the data needs corrections or needs to be re-inspected again. Agencies have to define the criteria to determine the allowable variation between the reference value and the data measured.

The most commonly used quality management techniques are indicated in Figure 6. It is shown that quality acceptance can be implemented before, during, and after data collection. For instance, the testing of control and verification sites can be applied before or during data acquisition. At control sites, data can be evaluated in terms of accuracy and repeatability, based on a well-defined manual or any other standard procedure. Repeatability represents the capacity of the equipment or the ability of the raters to get the same values on repeated measurements under the same conditions (Transportation Research Circular E-C037, 2002).

At verification sites, data can be evaluated in terms of repeatability and reproducibility. When data repeatability is evaluated, the procedure is called oversampling. When the personnel or the equipment used are different, but the verification site is the same, data reproducibility is analyzed and the procedure is referred to as cross testing. Reproducibility is the capacity of different equipment or the ability of different raters to accurately replicate pavement condition values on the same section (Transportation Research Circular E-C037, 2002).

Before data collection, another quality acceptance activity is the review of qualifications or certifications used to validate equipment, raters, and acquisition procedures. These standards need to be carefully set by the transportation agency because they serve to define the acceptance criteria. Other related techniques are sampling and re-rating of sections during data gathering, software database checks, quality acceptance reviews using Geographic Information Systems (GIS), and data comparison with existing time-series data.

MTC's quality acceptance plan includes a certification program to accredit and train raters and technicians. The program also includes data checks comparing the contractor's collected data against pavement deterioration models and curves from the Pavement Management System's database (Tan & Cheng, 2014).

2.4.3 Independent Assurance

Independent assurance refers to an independent assessment or audit of the quality of the data. A third party is needed to re-inspect or reevaluate a sample of the data and perform a comparison between the results obtained during and after the collection process. Generally, the techniques considered for this independent validation are similar to the ones described for the quality acceptance process. Consistency reviews, sampling and re-evaluation of data, completeness checks and time history comparisons are some examples of independent assurance techniques (see Figure 6).

2.5 Quality Management Techniques

Figure 6 displays the main quality management techniques grouped based on two conditions: the quality management process to which they correspond and its application during the data collection process. The most relevant techniques are described in the following section, focusing on its importance within the entire quality management process.

Personnel Training and Certification

The equipment operators and the raters responsible for conducting the visual inspections need to be trained before the data collection process is performed. One alternative to ensure acceptable knowledge and skills is requiring an official certification to verify that the personnel have the required training level. Equipment can also be certified, meaning that it has successfully passed formal verification testing and complies with a specific standard. Another option is to only use qualified personnel with extensive experience during pavement condition inspections. Less experienced raters can receive considerable training to improve their capabilities.

Personnel training and certification is an important quality management technique necessary to reduce rater's subjectivity and improve the consistency, reliability, and accuracy of data gathered from visual inspections. Another positive effect of implementing training and certification methods is the appropriate operation of equipment.

Equipment/Method Calibration

Equipment and method calibration techniques can be conducted before and during the data acquisition process. The objective is to verify that the collection methodology is being applied correctly and check if the equipment is working according to required specifications.

Calibration is a process to systematically validate equipment or a methodology by comparing the data collected with a reference value. If pavement distress data is collected, calibration is performed by evaluating control sites where the pavement condition has been measured and monitored by a group of experts (Chang-Albitres, Smith & Pendleton, 2007). Based on the condition of the control site, equipment can be calibrated and raters can be trained. The requirements for personnel criteria and equipment calibration are defined based on statistical confidence intervals (McQueen & Timm, 2005).

Data Verification

Data verification techniques consist of periodic testing of control or verification sites, oversampling, cross testing, and re-inspections of pavement sections, which can be known or blind to the data collection personnel. The objective of verifying pavement condition data during its acquisition is to prevent gathering a large amount of poor data.

If errors are detected, they can be corrected as soon as possible, avoiding the generation of more unsatisfactory data. Corrections may include equipment calibration (if automated or semi-automated methodologies are utilized) or standardization of rater's criteria (if manual inspections are performed). Additional rater training can also be recommended (Morian, Stoffels & Frith, 2002).

Testing of control sites are used to determine data accuracy as well as equipment and raters' repeatability. Accuracy is determined by comparing the data collected with a defined reference value. Repeatability is determined based on the standard deviation of repeated measurements taken by the raters or equipment operators, under the same conditions over a short period of time (Transportation Research Circular E-C037, 2002).

Testing of verification sites are used to determine repeatability and reproducibility, which is the equipment's capacity or rater's ability to accurately reproduce measurements. Reproducibility is determined based on the standard deviation of measurements taken with different equipment or using different methodologies (Transportation Research Circular E-C037, 2002).

Oversampling is performed at verification sites where repeatability is evaluated. The data collection personnel samples the same pavement section repeated times. Measures from the retest are compared (McQueen & Timm, 2005). Random errors can be detected, but systematic errors cannot.

Cross testing is performed at verification sites where reproducibility is evaluated. Different personnel or equipment is used to measure the same pavement section. Both random and systematic errors can be detected. Other data verification method consists of reanalyzing or re-

inspecting a sample of sections measured by an independent evaluator. This is usually performed as part of the independent assurance process.

Data Checks

Data checks are performed during data collection for quality control, quality acceptance, and independent assurance. When data have been included into the pavement management database, data checks are applied. Data checks can comprise format verification, missing data identification, logic checks, and data patterns identification (e.g. consecutive zero, null, repeated, or out of range values). These checks are important to prevent collecting large quantities of deficient data.

The purpose is to identify systematic errors, indicate if a pavement condition is underestimated or overestimated, search data that is out of expected ranges, and check for general data inconsistencies. These inconsistencies can include data that is missing, incorrectly identified, and inadequately formatted.

Data checks can include statistical analyses on the differences between mean values of a specific parameter, for the quality control or quality acceptance, and the mean values of the same parameter obtained from the data collected. Paired T-Tests are generally used for means comparison. For each sample, differences between measurements from the verification site and the data collected from the inspections are calculated to determine if the data from the survey is under or overestimating pavement condition.

Each individual observation from the sample is evaluated by utilizing defined criteria, which determine if the observation passes or fails minimum acceptable quality requirements (Selezneva, Mladenovic, Speir, Amenta & Kennedy, 2008). This data check is usually applied to project level analysis.

Time-history comparisons

Time-history techniques include the comparison of pavement data gathered from the field with existing time-series data previously collected. The purpose is to identify unusual behaviors in the condition that might be a sign of data collection errors. The measurements need to be consistent with the historical data, to ensure year-to-year reliability.

Time-history comparison is only possible if reliable data from the past has been documented. Documentation of the quality control and quality acceptance processes during data collection is needed to create and conserve historical information. Reporting quality management procedures and outcomes enables the ability to refer back to previous data and compare it with new collected data.

2.6 Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management

Pavement condition data collected from site inspections may contain certain levels of uncertainty and low quality. To compensate this scenario, a variety of solution alternatives can be used to improve data quality, diminish uncertainty, and, based on the improved data, design models to describe pavement deterioration in time. Depending on the amount of data available and its characteristics, the type and complexity of the solution alternatives will vary.

This section describes general features about different statistical techniques, mathematical models, data analysis techniques, quality approaches, and softwares applicable to data quality enhancement. Their definition, purpose, types, formulation, and applicability are also discussed.

2.6.1 Statistical Techniques

Cohen's Kappa Statistic

Cohen's Kappa is a statistical tool used to evaluate either interrater or intrarater reliability. Ideally, interrater reliability is evinced when two or more competent raters (experienced and/or certified), using the same manual or equipment on the same pavement sections, concur on identical

results. Intrarater reliability is evinced when a single rater, using the same manual or equipment on the same pavement sections, is able to obtain similar results each time the data is collected.

Raters may sometimes agree or disagree on the results obtained after data collection. When the raters are not completely sure about some specific aspect of the data they are collecting, the measures they make are based on guesses due to uncertainty (McHugh, 2012). Agreement between results might occur in this particular situation. Kappa statistic estimates the level of consistency between different raters, excluding the possibility that they could agree by chance.

There is some variability among raters, even though the manuals, equipment, and pavement sections are the same. Kappa calculation relies on the comparison between the agreement of rater's results due to a real representation of the pavement condition and the agreement due exclusively to chance. Table 1 shows the interpretation of the possible Kappa values (Viera & Garrett, 2005).

Table 1: Interpretation of Kappa Values.
Source: Viera and Garret (2005).

Level of Agreement	**Poor**	**Slight**		**Fair**		**Moderate**		**Substantial**		**Almost Perfect**		**Perfect**
Kappa value	≤ 0.00	0.01	0.20	0.21	0.40	0.41	0.60	0.61	0.80	0.81	0.99	1.00

Sui Tan and DingXin Cheng (2017) applied Kappa statistics during an independent assurance process. A company was hired to conduct a pavement condition visual inspection on the field to audit the results obtained by another agency who had previously performed the inspection. The manual used by both agencies was the Distress Identification Manual developed by MTC. The Kappa value was equal to 0.75, which represents a substantial agreement between both company's results.

Percent agreement

Percent Agreement is a statistical tool used to measure the percent of data that are consistent or similar between two or more raters, evaluating the same pavement section using the same methodology and equipment (interrater reliability) or a single rater evaluating several times the same pavement section using the same methodology and equipment (intrarater reliability).

This statistic is calculated as the number of agreement scores, interpreted as similar results divided by the total number of scores. Unlike Kappa statistics, the Percent Agreement does not consider the possibility that raters guessed on similar results. The consistency among raters is overestimated (McHugh, 2012).

Missing Data Techniques

Quality control should include the identification of missing data in a pavement condition dataset after the corresponding collection process has been performed. Incomplete data represents a problem in PMS applications because it has a negative impact on the forecast of pavement performance, the selection of maintenance and rehabilitation treatments, and the estimation of funding allocation (Al-Zou'Bi et al., 2015).

According to Tsikriktsis (2005), there are two alternatives to address missing data problems. One option is to eliminate the incomplete data cases, and the other is to complete the missing values. There are two approaches to fill in incomplete data with estimated values: utilizing model-free or model-based replacement techniques.

Model-free replacement techniques use one or more known values of the same amount to estimate new values and complete the missing ones. Depending on the available data used to estimate the new values, three substitution techniques can be defined: case substitution techniques, if the values are part of the same dataset; subgroup substitution techniques, if the values belong to different datasets; and total substitution techniques, if all the available data is used. Table 2 shows the model-free techniques that are reliable in completing missing data. Only the moving average technique can be used for prediction.

Table 2: Statistical Model-Free Replacement Techniques.
Source: Al-Zou'Bi et al. (2015).

Substitution Technique	Statistical Model-Free Replacement Technique	Description
Case Substitution Techniques	Mean of Nearby Points	Missing value is replaced by the mean of surrounding known values
	Median of Nearby Points	Missing value is replaced by the median of surrounding known values
	Moving Average	Missing value is replaced by the moving average of the known values
Subgroup Substitution Techniques	Mean	Missing value is replaced by the mean substitution of the subgroup of the known values
	Median	Missing value is replaced by the median substitution of the subgroup of the known values
	Maximum	Missing value is replaced by the maximum substitution of the subgroup of the known values
	Minimum	Missing value is replaced by the minimum substitution of the subgroup of the known values

Model-based replacement techniques use a defined statistical model and model's parameters to estimate the new values based on known data. The parameters can be calculated considering the entire dataset or a part of it, from a specific time period (DeSarbo, 1986). Table 3 shows the model-based techniques that are reliable in completing missing data. All of the techniques can be used for prediction.

Al-Zou'Bi et al. (2015) conducted a case study to evaluate the efficiency of missing data statistical techniques in predicting pavement performance in terms of distress scores. The moving average technique obtained the highest accuracy compared to the other methods, but the authors recommended checking different statistical techniques depending on the characteristics of the data itself. The study concluded that statistical techniques have to be used to complete missing data in order to predict pavement performance more accurately.

Table 3: Statistical Model-Based Replacement Techniques.
Source: Al-Zou'Bi et al. (2015).

Statistical Model-Based Replacement Technique	Description
Linear Interpolation	Missing value is replaced by the linear interpolation of surrounding known values.
Linear Regression	Missing value is replaced by the linear regression of the known values.
Cubic Regression	Missing value is replaced by the cubic regression of the known values.
Cubic Spline	Missing value is replaced by the cubic spline interpolation of the known values. Different values of the parameters are used to describe different time periods.

2.6.2 Mathematical Models

Artificial Neural Network (ANN)

Artificial Neural Network are artificial intelligence techniques that studies the processes that occur in the human's brain when information is being processed. ANN has the purpose of creating models based on mathematical relationships, capable of replicating brain-related processes (Smith, 1993). Human brains think and learn through perception, reasoning, and interpretation. The brain is composed of interconnected neurons, organized as a network, that receive signals from other neurons. After reaching a certain level of excitation, a neuron sends an output signal to other neurons, which receives the signal as an input. This process is modeled by an ANN, which describes the relationship between neurons using algorithms (Freeman & Skapura, 1991). ANNs are nonlinear regression models, which are appropriate for processing numerical information and recovering data from historical numerical information (Sundin & Braban-Ledoux, 2001).

The dataset in an ANN can be composed of pavement data gathered from the field and its quality is important to develop accurate models. There are several applications of ANN in pavement engineering. ANNs can estimate the current pavement condition, predict the future deterioration, and provide decision-maker engineers with maintenance and rehabilitation actions, so they can select the optimal one.

Attoh-Okine (1994) applied ANNs for the estimation of the progression of International Roughness Index (IRI) in flexible pavements in terms of structural deformations, surface distresses, and environmental and non-traffic-related factors. Banan and Hjelmstad (1996) developed an ANNs model to estimate the present serviceability index (PSI) and compared the results with the PSI calculated from the American Association of State Highway Officials (AASHO).

Eldin and Senouci (1995) created an ANN-based pavement condition-rating model to determine the condition rate of flexible pavements. The output of the model must be equal to the output provided by the condition-rating scheme established by the Oregon Department of Transportation (ODOT). The input for the ANN model were the types and severity levels of patching, bleeding, rutting, alligator, transverse, and block cracks distresses. ANNs were used for the detection and quantification of surface pavement cracks based on pavement images. Alligator, longitudinal, transverse, and block cracking were identified from the images utilizing ANNs (Kaseko & Ritchie, 1993).

Many authors have applied ANNs in the prediction of future pavement condition. Roberts and Attoh-Okine (1996) predicted IRI for composite, full-depth, and partial-design asphalt concrete pavements based on an ANN quadratic function. The inputs for the prediction were the equivalent axle loads, IRI, rutting, fatigue cracking, transverse cracking, and block cracking. La Torre, Domenichini, and Darter (1998) predicted the IRI of flexible pavements four years forward using ANNs. The inputs were the thickness of pavement layers and its modulus of elasticity, climatic data, the average annual equivalent axle loads, the age of the pavement, and the IRI value in the current year of analysis. Another application of ANNs in predicting pavement condition is the study performed by Huang and Moore (1997). They used ANNs models for estimating the probability of occurrence of a specific roughness distress level in the future. Abdallah, Melchor, Ferregut, and Nazarian (2000) applied ANN models to predict the remaining life of flexible pavements based on layer thickness and surface deflections measured with the Falling Weight Deflectometer (FWD).

ANNs have also been successfully applied for data quality control when data recorded by hand is digitalized. Data checks of the information transferred from paper to a digital database is performed by ANNs which are capable to solve missing data problems and time series' outliers (Benvenuto & Marani, 2001). Dai, Yoshigoe, and Parsley (2018) used ANN algorithms to improve traditional time-consuming data quality control methods, which have limited performance and low accuracy. ANNs and statistical quality control models were integrated for improving data quality.

Fuzzy Logic

Fuzzy logic is an artificial intelligence technique, in which simple logic is extended beyond true and false values. Partial or continuous truths are defined. Fuzzy Logic depend on the Fuzzy Set Theory, where a continuum of probabilities, which may have values ranging from 0 to 1, represents the degrees of a membership in fuzzy sets (Sundin & Braban-Ledoux, 2001).

Grivas and Shen (1995) used Fuzzy Set Theory to manage uncertain information with the purpose of deciding the most convenient maintenance and rehabilitation strategies. The knowledge of an expert decision maker on the relationship between deficient road conditions and its corresponding treatments is represented as knowledge graphs. Fuzzy Theory was used to establish the relationships in the knowledge graphs for condition analysis and treatment identification.

Fwa and Shanmugam (1998) used Fuzzy Logic techniques for pavement condition rating and maintenance needs assessment in a road network. Subjectivity and uncertainty were taken into account to develop the Fuzzy Logic-based pavement distress condition rating procedure. Another application of Fuzzy Logic was presented by Wee and Kim (2006). Fuzzy Logic was used along with Expert Systems to automate a Pavement Management System and develop reliable and consistent strategies for maintenance, rehabilitation, and reparation of pavement structures (Wee & Kim, 2006).

Fuzzy Logic was also used in Transportation Asset Management to improve the probabilistic approach of Life-Cycle Cost Analysis. Fuzzy Logic was included into the risk analysis process of a Life-Cycle Cost Analysis model to determine the timing and strategies for

pavement maintenance, rehabilitation, and reconstruction, based on performance curves and Fuzzy Logic triggering models. Fuzzy Logic systems showed good results when making inferences from uncertain, ambiguous, and subjective data (Chen & Flintsch, 2007).

Matía, Aguilar-Crespo, Jiménez, Sanz, and Domínguez (1995) applied Fuzzy Logic to data quality by solving data validation problems. Fuzzy Logic can represent human expert knowledge and also manage linguistic terms, uncertainty, and imprecision. Janta-Polczyhki and Roventa (1999) studied three dimensions of data quality based on Fuzzy Logic representation: quality of data conceptual component, quality of the information to be stored in the database, and quality of data representation.

Bayesian Methods

Bayesian Methods are models based on a mathematical approach to manage uncertainty. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that are related. Equation 1 shows the mathematical formulation of Bayesian theorem (Heller, 2007):

$$P(\theta|x) = \frac{P(x|\theta)\,P(\theta)}{P(x)} \qquad \text{Eq. 1}$$

Where:

P (θ): probability of θ, which is referred to as the prior. It represents the prior probability of θ before observing any information about x.

$P(x|\theta)$: probability of x conditioned on θ. It is referred to as the likelihood.

$P(\theta|x)$: probability of θ conditioned on x, or posterior probability of θ after observing x.

P (x): probability of x, independently from θ.

Letting P (x, θ) be the joint probability of x and θ, it is possible to marginalize out θ to obtain the following equation for P(x): $P(x) = \int P(x,\theta)d\theta = \int P(x|\theta)\,P(\theta)\,d\theta$. Assuming that a random sample $x = (x_1, \ldots, x_n)$, with $P(x,\theta)$ as the distribution function, which describes the random variables $x_1, \ldots, x_n$. Therefore, the likelihood function is given by $P(x|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$,

which represents the probability of observing x_i under different values of the θ parameter. The Bayes Theorem takes into account the information already collected, the prior knowledge for the parameters, represented by one or more prior distributions, then considers the observed data and makes an inference (Pateras, 2013).

If θ is continuous, equation 2 represents the distribution for P(θ|x).

$$P(\theta|x) = \frac{P(x|\theta)\ P(\theta)}{\int P(x|\theta)P(\theta)d\theta} \qquad \text{Eq. 2}$$

Where:

$\int P(x|\theta)P(\theta)d\theta$: marginal likelihood of the data, which is equal to P(x).

P (θ): prior probability of parameter θ

P(x|θ): likelihood of the data given θ

If θ is discrete, equation 2 changes to equation 3:

$$P(\theta|x) = \frac{P(x|\theta)\ P(\theta)}{\sum P(x|\theta)P(\theta)d\theta} \qquad \text{Eq. 3}$$

Where:

$\sum P(x|\theta)P(\theta)d\theta$: marginal likelihood of the data, which is equal to P (x).

According to Bayes Theorem, the posterior distribution of the parameter given the data can be expressed as $P(\theta|x) \propto P(\theta)P(x|\theta)$. Given the previous equations, it is possible to conclude that the posterior distribution considers information from both the prior knowledge – represented by the prior distribution P(x) – and the data being observed – represented by the likelihood P(x|θ) (Pateras, 2013).

Park, Smith, Freeman, and Spiegelman (2008) implemented a Bayesian approach for improved pavement performance prediction to plan appropriate road repairs. The authors presented a methodology for forecasting future longitudinal cracking on pavements based on small data with high variability. Theoretical pavement distresses were modeled based on prior

engineering knowledge. Bayesian formulation was able to obtain sensible predictions and provide reasonable uncertainty statements for predictions.

A Bayesian model was applied for predicting the IRI of pavements that have been rehabilitated with thin hot mix asphalt overlays. The probabilistic models had two components: the ANNs that forecast the IRI not considering any treatment, and the Bayesian regression models that forecast the decrement in IRI due to rehabilitation. The results indicated a good fit of the models with a low percentage of outliers (Liu & Gharaibeh, 2013).

The quality of the data is a relevant component for the development of Bayesian Networks. Data accuracy has an important impact on the efficiency and results of a Bayesian Network algorithm. Therefore, Sessions and Valtorta (2006) have created Bayesian algorithms that incorporate data quality assessment and decrease the effect of inaccurate data. Similarly, to overcome the challenges of quality assessment, such as uncertainty, threshold value definition, and metric combination; Caro, Calero, Sahraoui, Malak, and Piattini (2007) have develop a Bayesian Network model using a Data Quality Management model.

Bayesian Network model was considered by Taware and Kolhe (2014) as an "end-to-end system for form design, entry and data quality assurance". Thus, Bayesian Network model has been used to improve data quality at every step of the data entry process. At the three stages (before, during, and after data entry), the model is capable of identifying possibly erroneous inputs and inaccurate data.

2.6.3 Data Analysis Techniques

Big Data Analysis Techniques

Data collected on-site related to pavement condition is important because it represents the basis for making cost-effective infrastructure management decisions regarding maintenance, rehabilitation, and reconstruction of transportation assets. Pavement condition data is recurrently gathered by transportation agencies, and important volumes of data is being stored. The concept

of Big Data is applicable when relevant management information is obtained from the pavement condition data collected (Kobayashi & Kaito, 2017).

The diagram presented in Figure 7 shows the types of data based on two criteria: the volume of data and the quality of data. Data volume is represented in the horizontal axis, and it can vary from small to large volumes. Data quality is represented in the vertical axis, and it can range from low (or poor) to high quality. Quality and volume of data are used in the decision-making process to obtain pavement management final decision (Kobayashi & Kaito, 2017).

In the past, statistical techniques were focused on the incomplete small data region, which represents a small volume of poor quality data. Based on that type of data, substantial information had to be obtained to make appropriate management decisions. Technological development can improve the incomplete small data and move from this region towards others: complete small data region or incomplete Big Data region. Big Data region corresponds to the case where extensive amounts of data are available to make a decision, but data quality needs to be improved (Kobayashi & Kaito, 2017).

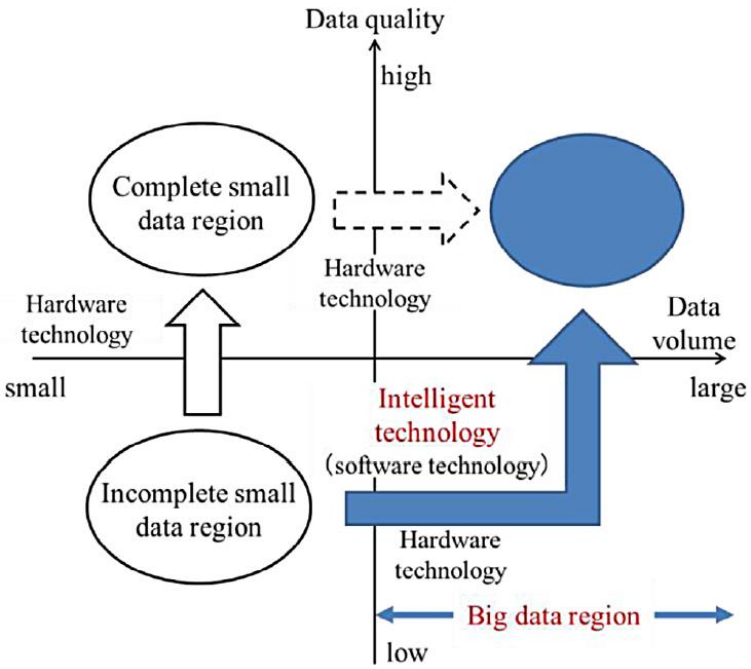


Figure 7: Big Data Concept Applied to Pavement Management.
Source: Kobayashi and Kaito (2017).

Statistical methods designed to solve problems for the incomplete small data region are not applicable in the Big Data area. Hardware technology cannot achieve high quality information from large amounts of poor quality data, but intelligent technology can analyze great amounts of accumulated data (Big Data) and extract valuable information for the pavement management decision-making process (Kobayashi & Kaito, 2017).

The analysis and use of Big Data must rely on good quality data to generate value from this massive amount of information. Unfortunately, there is a lack of quality standards and quality assessment methods for Big Data. Poor data quality might bring about low data efficiency utilization and decision-making mistakes. Researchers have focused on how to overcome this challenge. For instance, Cai and Zhu (2015) proposed a data quality assessment framework and process for Big Data analysis based on five dimensions: availability, usability, reliability, relevance, and presentation quality. Merino, Caballero, Rivas, Serrano, and Piattini (2015) propose a quality model for Big Data, based on three data quality characteristics: contextual adequacy, operational adequacy, and temporal adequacy.

Data Mining

Data Mining is a data analysis technique that discovers patterns from accumulated historical data with the purpose of detecting and taking advantage of success patterns, evading failure patterns, and enhancing business processes. Generally, Data Mining techniques are used in business applications, but can also be applied in Pavement Management Systems. One requirement for the applicability of Data Mining techniques is the availability of historical data to be analyzed (Nassar, 2007). “Data Mining can provide a great tool for discovering the wealth of information contained in this data” (Cabena, 1997). Transportation agencies need to assure those databases have good quality and are accessible, in case they are considering using Data Mining techniques effectively in their decision-making process.

Data Mining techniques do not replace long-established statistical methods. On the other hand, these techniques are a continuation of traditional statistical techniques, which are not able to

reveal important information within the database. Data Mining techniques complement the statistical analysis by extracting knowledge from data that can be useful for managerial decisions (Nassar, 2007).

Data Mining techniques can be grouped into four main categories based on their functionality: classification, clustering, numeric prediction, and association-learning techniques. Their principal differences are their mechanism of deriving information, based on algorithms or other methods, and the way results are presented, in terms of rules or knowledge.

Classification techniques are forecasting models that establish patterns or categorical groups of information from an existing database. These techniques indicate the class where each value of the database corresponds. Data Mining classification methods are centered on the identification of the characteristics for each group or class. However, when the identification of classes or groups are not possible, the records of the database are grouped using the clustering techniques that associate the items that fall naturally together. The numeric prediction techniques are similar to the classification techniques, but the result to be predicted is a numerical value instead of a category or discrete class. Finally, association-learning techniques have the objective of discovering significant patterns in the data by identifying association rules. Each rule has a probability of occurrence and a number of cases in which it is found. Association rules can forecast any attribute, and are not limited to the prediction of a group or class as in the classification rules. Association rules can also predict more than one value of an attribute at a time.

Data Mining techniques have been implemented principally in business applications, such as fraud discovery, market segmentation, retail promotions evaluation, customer profiling, and credit risk evaluation, among others. Nevertheless, Data Mining can also be applied in Transportation Asset Management. Attoh-Okine (1997), used Data Mining techniques to analyze a PMS database created with information based on objective and subjective methods. The author successfully applied those techniques to derive decision rules for pavement maintenance and rehabilitation decision-making process.

"As quality data is important for Data Mining, reversely Data Mining is necessary to measure the quality of data" (Anam & Shahriar, 2008). While considerable amount of information is being collected and stored, it is necessary to gather good quality data for Data Mining and also to use Data Mining for quality measurement. Researches have shown interest in the relationship between Data Mining and data quality. Athanasiadis, Rizzoli , and Beard (2010) investigated how to incorporate Data Mining techniques into the quality assurance decision-making process. Hipp, Güntzer, and Grimmer (2001) applied Data Mining techniques to identify, quantify, describe and correct data quality deficiencies in high-volume databases.

2.6.4 Quality Approaches

Six Sigma

Six Sigma is a concept conceived by the Motorola Corporation in 1986 that represents a high quality level the company was trying to accomplish for its production process. "The focus of Six Sigma is reducing variability in key product quality characteristics to the level at which failure or defects are extremely unlikely" (Montgomery, 2013). This philosophy emphasizes reducing the variability in product quality features to a certain level where failure or defects are very improbable to occur (Montgomery, 2012). Extremely high quality objectives are defined based on the Six Sigma approach. Data collection and fine analysis of the results are performed with the purpose of reducing defects in services and products. If defects in a process can be measured, it is possible to determine how to systematically eliminate them and achieve perfection (Summers, 2006).

Six Sigma level of quality is related to approximately 3.4 defective parts per million, which correspond to six standard deviations of the process (Reid & Sanders, 2012). This is equivalent to a success rate of 99.9997% in the fabrication of products that meet specifications (FHWA, 2010).

Figure 8 shows the number of defects produced, in parts per million, when three and six standard deviations are considered. Lower Specification Limits (LSL) and Upper Specification Limits (USL) are indicated in the figure (Reid & Sanders, 2012).

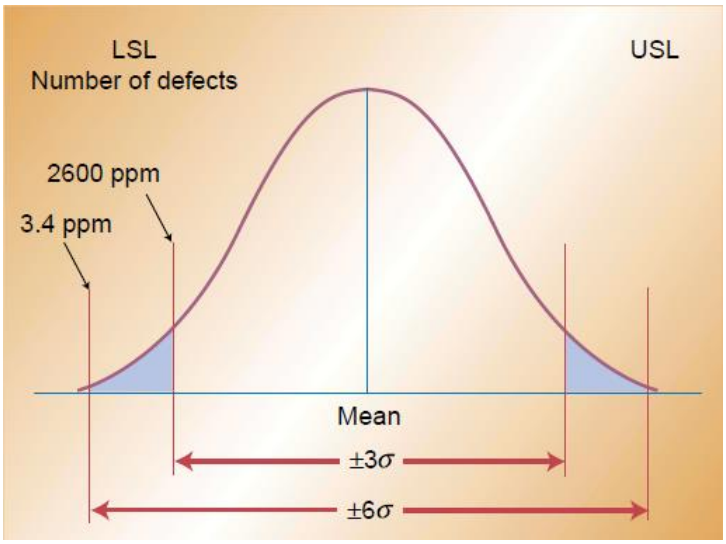


Figure 8: Parts per Million Defective for ±3σ and ±6σ Quality.
Source: Montgomery, 2013.

The implementation of Six Sigma involves the application of statistical quality control tools to detect and eliminate causes of quality issues. These technical tools include analysis of variance, correlation, process capability analysis, control charts, ANOVA analysis, histograms, root cause analysis, design of experiments, cost-benefit analysis, cause and effect diagrams, acceptance sampling, etc. These tools have to be completely integrated in the organizational system of the company (Reid & Sanders, 2012).

Another important aspect of the implementation of Six Sigma is the commitment of all employees to understand the process and continually apply the technical tools needed to eliminate quality problems. Different levels of knowledge are defined to motivate staff members to improve its Six Sigma skills and achieve the highest level (Reid & Sanders, 2012).

The procedure used by Six Sigma for solving problems is called DMAIC and consists of five steps: define opportunities, measure performance, analyze opportunity, improve performance, and control performance. DMAIC process can also be applied to project management and quality improvement (Montgomery, 2013).

Six Sigma quality approach was adopted by Motorola in all aspects of the organization including marketing, product design, manufacturing, finance and accounting, among others.

Nowadays, Six Sigma quality standards are being adopted by many manufacturing industries concern in quality improvement (Reid & Sanders, 2012).

FHWA (2010) recommend the Six Sigma approach to Transportation Asset Management for root-cause analysis of pavements with poor performance. Pavements with an extraordinarily good performance could be analyzed, as well as pavements with an exceptional bad performance. The core causes of those performances regarding treatment practices, materials selection, treatment timing, construction practices, and maintenance history, among others could be determined. Once the causes had been identified, the practices that led to good pavement performance could be standardize and promoted, while the practices that led to bad pavement performance, could be reduced and avoided.

Total Quality Management

Total Quality Management (TQM) is one of the key concepts that has affected the development of quality control standards. TQM is a philosophy applied by organizations that aims to provide customers with services and products that satisfy its needs. TQM synchronizes all organizational processes (e.g. engineering, design, and production) in order to align their objectives in achieving customer quality needs (Martin, 1993).

Total Quality Management can be defined as an integrated effort to improve quality performance in all organizational levels. TQM is a philosophy of continuous improvement, which focuses on the identification of the root sources of quality problems and its correction. The concept of quality is considered in every aspect of the organization, including people: customers, employees, and suppliers (Reid & Sanders, 2012).

The seven concepts that TQM philosophy rely on, are: a) customer emphasis, b) constant improvement, c) employee empowerment, d) quality tools application, e) product design, f) process management, and g) supplier quality management (Reid & Sanders, 2012).

a) Quality is represented as the accomplishment or surpass of customer expectations; consequently, the customer needs have to be well known by the organization through focus groups, surveys, interviews, etc.
b) A company can always improve its level of quality through learning and problem solving. TQM states that performance has to be constantly evaluated in order to make measurements and improve performance.
c) Employees are empowered to identify quality problems and correct them; thus, quality tools are used for quality assessment.
d) These technical tools include cause and effect diagrams, flowcharts, checklists, control charts, scatter plots, Pareto charts, and histograms, among others.
e) Quality function development tool is used in TQM to translate customer expectations in technical requirements for the design of the product.
f) TQM considers that a quality product derives from a quality process; therefore, quality must be incorporated in the process.
g) Suppliers of the materials to fabricate a product have to meet quality standards. Supplier quality is also taken into account in TQM.

Poister and Harris (1996), studied the impact of TQM-related techniques on a highway maintenance program. The correlation between TQM indicators and variations on the highway's performance measures were analyzed. TQM exhibited a positive correlation with the attitudes of the employees, the quality of the highway maintenance activities, and the condition of the highway. Furthermore, TQM presented a negative correlation with complaints, sick-leave usages, and injuries.

The Federal Highway Administration FHWA (2012b) concludes that the adoption of the TQM framework as part of a transportation agency's processes is important to show responsibility towards the users, maximize the resources, and assure its viability in a long-term period.

2.6.5 Softwares

Expert System (ES)

Expert systems based on knowledge have been, historically, the first artificial intelligence technique applied in pavement management (Sundin & Braban-Ledoux, 2001). ESs are decision support technologies capable of incorporating the expertise, understanding, and knowledge of expert decision-maker engineers in the identification of maintenance and rehabilitation techniques. In that sense, ESs constitute a design and analysis tool for transportation agencies in the creation and implementation of maintenance and rehabilitation strategies.

The most experienced decision makers in an agency are valuable elements throughout the entire decision-making process. These experts are hard to find and sometimes difficult to keep as part of the organization. The commonly used technology that can substitute human expertise are called Expert Systems. An ES is a software capable of replicating skills and reasoning processes of a human expert while making a decision to solve specific problems. No software nor ESs can replace expert decision makers in an organization. ESs aim to make their knowledge and experience more accessible to other members of the agency. ESs are highly efficient in problems where the knowledge of an expert decision maker is required. When the human reasoning process is not simple but complex to be defined and executed analytically, ESs can be the appropriate tool to use.

An ES contains two mechanisms: a knowledge based mechanism and an inference mechanism. The knowledge based mechanism includes facts and rules for drawing conclusions and knowledge used by experts to solve a problem. The inference mechanism arrives at conclusions based on the knowledge mechanism. The program can also describe the reasoning process utilized to infer a conclusion (Sundin & Braban-Ledoux, 2001).

Many authors have implemented ESs in multiple pavement management applications. Lee and Galdiero (1989) determined the most appropriate maintenance and rehabilitation strategies based on ESs, depending on current pavement conditions. Kotb and Moore (1996) used ESs with the same purpose, but incorporated cost estimations to calculate a required budget for maintenance

and rehabilitation activities. The inputs for the software are volume of traffic, types and severity of distresses, PCI values of pavement sections, deterioration rates, levels of past maintenance, skid resistance, applicable maintenance and rehabilitation activities, unit cost, and projected service life for each activity. Kuncheria and Veeragavana (1996) applied ESs for the estimation of the causes of deterioration and the treatment technique needed. In addition, the system determines the appropriate type of material for rehabilitation and overlay thickness, based on overlays' life cycle cost.

According to Allen and Kathawala (1992), Expert System technologies are applicable to quality management in multiple areas including Transportation Asset Management. The identification of pavement irregularities and the detection of fatigue problems in bridges are some examples of the applicability of ESs. Expert Systems can also be used in Total Quality Management areas, such as "statistical process control, quality costing, goods receiving, corrective action procedures, supplier development, quality function deployment and field failure analysis", among others (Crossfield & Dale, 1991). Paladini (2000) proposed a Decision Supporting Expert System for quality control to assist with the decisions associated with activities for inspection development.

Genetic Algorithm (GA)

Genetic Algorithms are artificial intelligence techniques that can be defined as intelligent heuristic search programs. The solution of a problem is represented as a chromosome that contains 0s and 1s. These are the values of a vector of decision variables, which describes the possible alternatives to solve the problem. Based on an arbitrary population of solutions, the GAs combine parts of chromosomes together to create new solutions with eventual mutations. A fitness function is used to test if the new solutions are feasible. A selection process is then performed to identify the best solutions from the current and previous generations. The best solution is determined after a great quantity of iterations (Sundin & Braban-Ledoux, 2001).

Kwasi and Attoh-Okine (1999) applied generic algorithms in the estimation of roughness evolution in flexible pavements, based on a roughness index, equivalent axles, age of the pavement, rut depth, thickness of the asphalt layer, and the overall structural number of the pavement system. Fwa, Chan, and Tan (1996) developed a Genetic Algorithm model for programming pavement maintenance and rehabilitation activities at the network level, considering an analysis period of 20 years.

Fwa, Chan, and Hoque (1998a) applied Genetic Algorithms to pavement maintenance programming at the network level. The problem of the constraints in programming of pavement management activities was solved by the powerful search capability of the GAs. More applications of Genetic Algorithms to Transportation Asset Management were presented by Fwa et al. (1998b). The authors developed a computer model based on GAs for setting maintenance warning levels, defining trade-offs between maintenance and rehabilitation, planning budget, and estimating impacts of different investment strategies.

Yang, Remenyte-Prescott, and Andrews (2015) made a similar application of GAs. They developed optimal maintenance and rehabilitation strategies for highway networks using a multi-objective and multi-constrained Genetic Algorithm. The GA's final result was the optimal pavement maintenance and rehabilitation strategy. This optimal alternative is the one that minimizes maintenance or rehabilitation costs and maximizes the pavement condition during the planning period.

Genetic Algorithms were also used for improving data quality. Das and Saha (2009) applied GAs to identify, quantify, explain and correct data quality deficiencies in databases that contained important amounts of information. The GA is used to search association rules among all the data, based on accuracy, comprehensibility, and completeness. Vizhi and Bhuvaneswari (2012) also used association rules to measure data quality. A GA was developed to generate high quality association rules based on four metrics: confidence, completeness, interestingness, and comprehensibility.

A summary of the statistical techniques, mathematical models, data analysis techniques, quality approaches, and softwares applicable to quality management is displayed in Table 4. The description of these solution alternatives are presented including its applicability to pavement management and quality management.

Table 4: Summary of Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management.

Type	Solution Alternatives	Description	Pavement Management Applicability	Quality Management Applicability
Statistical Techniques	Cohen's Kappa	Kappa statistic estimates the level of consistency between different raters, excluding the possibility that they could agree by chance.	Verification of the reasonableness of pavement condition survey data for Pavement Management Systems.	Measurement of interrater or intrarater reliability during for quality control.
	Percent Agreement	Statistical tool used to measure the percent of data that are consistent or similar between two or more raters.	Verification of the reasonableness of pavement condition survey data for Pavement Management Systems.	Measurement of interrater or intrarater reliability during for quality control.
	Missing Data Techniques	Techniques that identify and complete missing values in a dataset based on statistical replacement techniques.	Missing Data techniques can complete missing values in pavement condition datasets, which represents a problem in Pavement Management Systems.	Statistical model-free and model-bases replacement techniques are used to create new values for the missing data and improve the quality.
Mathematical Models	Artificial Neural Network	Artificial Neural Network are models based on mathematical relationships, capable of replicating brain-related processes, such as thinking and learning through perception, reasoning, and interpretation.	Artificial Neural Network can estimate the current pavement condition, predict the future deterioration, and provide decision-maker engineers with maintenance and rehabilitation actions, so they can select the optimal alternative.	Artificial Neural Network can be applied to improve data quality while solving missing data problems and time series' outliers.

Table 4: (Continued). Summary of Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management.

Type	Solution Alternatives	Description	Pavement Management Applicability	Quality Management Applicability
Mathematical Models	Fuzzy Logic	Fuzzy logic is a model based on partial or continuous truths rather than binary logic, which only considers false (0) and true (1) values. Fuzzy logic considers truth values between 0 and 1.	Fuzzy Logic can be used to manage subjective and uncertain information with the purpose of deciding the most convenient maintenance and rehabilitation strategies.	Since Fuzzy Logic can represent human expert knowledge and also manage linguistic terms, uncertainty, and imprecision, problems regarding data validation and data quality can be solved.
	Bayesian Methods	Mathematical models designed for uncertainty management. The probability of an event is described based on prior knowledge of conditions that are related.	Bayesian Methods can be used to improve pavement performance prediction to plan appropriate road repairs and predict pavement condition.	Bayesian algorithms can incorporate data quality assessment to decrease the effect of inaccurate data.
Data Analysis Techniques	Big Data Analysis Techniques	Analysis techniques that obtain valuable information from great volumes of data that are continuously collected.	Big Data techniques can analyze great amounts of accumulated pavement condition data and extract valuable information for the pavement management decision-making process.	Data quality assessment for Big Data analysis is needed to improve the quality of the dataset in order to generate better quality information from this massive amount of information.
	Data Mining	Data analysis technique that identifies success patterns from accumulated historical data to enhance business processes.	Data Mining techniques can be applied to Pavement Management Systems' databases to derive decision rules for pavement maintenance and rehabilitation decision-making process.	Data Mining can be used for quality measurement and can be incorporated into the quality assurance decision-making process.

Table 4: (Continued). Summary of Statistical Techniques, Mathematical Models, Data Analysis Techniques, Quality Approaches, and Softwares Applicable to Quality Management.

Type	Solution Alternatives	Description	Pavement Management Applicability	Quality Management Applicability
Quality Approaches	Six Sigma	Philosophy that consists of the reduction of variability in key product quality characteristics to the level at which failure or defects are extremely unlikely.	Analysis of the root causes of pavements that had very good and very bad performances in order to standardize good practices regarding treatment application, materials selection, maintenance and rehabilitation, among others.	Six Sigma involve the application of statistical quality control tools to detect and eliminate causes of quality issues.
	Total Quality Management	Philosophy that consists in an integrated effort to improve quality performance in all levels of an organization.	Total Quality Management techniques can be applied on highway maintenance programs to improve performance in maintenance activities, highway condition, and working environment.	Quality is achieved based on a continue performance improvement using quality tools that can be applied to an entire process.
Softwares	Expert System	Software capable of replicating skills, expertise, understanding, knowledge, and reasoning processes of a human expert while making a decision to solve specific problems.	Expert Systems can be used determined the most appropriate maintenance and rehabilitation strategies, causes of deterioration, and treatment techniques required based on current pavement conditions.	Expert Systems can be applied in Total Quality Management, statistical process control, quality costing, corrective action procedures, failure analysis, and quality control to assist decision-making processes, among others.
	Genetic Algorithm	Intelligent heuristic search programs that combine parts of the solution of a problem, represented as a chromosome, to create new solutions with eventual mutations.	Genetic Algorithm models can be used for programming optimal pavement maintenance and rehabilitation strategies.	Genetic Algorithm can be applied to identify, quantify, explain and correct data quality deficiencies in databases that contained important amount of information.

2.7 Pavement Performance Prediction Models in Pavement Management Systems

"A key factor for the success of a Pavement Management System is that it contains accurate and reliable pavement performance models" (Serigos, 2015). These models can predict future functional and structural behavior of pavements based on its current condition. Pavement Management Systems can only be effectively implemented when estimations of future pavement conditions are in concordance with real observed behavior. Present pavement performance can be determined with periodical measurements of its ride quality, complemented with historical traffic data. This information serves as inputs for the performance models to estimate future pavement conditions. Performance models can be utilized in a life-cycle cost analysis and maintenance and rehabilitation timing, including the type of technique to be applied (Sundin & Braban-Ledoux, 2001).

Based on the management level in which the decision will take place, the complexity of the pavement performance model is going to be different. For instance, network level models are less complex than project level models. At the network level, the predictions represent the general and common behavior of a group of pavements. The forecasts of network level models related to pavement conditions are necessary to establish multi-year treatment strategies based on the future estimated treatment needs. On the contrary, project level models are more detailed than network level models. The forecasts of project level models are more accurate due to their applicability regarding the definition of corrective actions for a specific pavement section. Consequently, decision makers usually use network level models for selecting maintenance and rehabilitation treatments and then, at the project level, the level of detail of the model is refined and its complexity is increased (Gallegos, 2012).

Pavement performance models must idealistically represent the deterioration process of a pavement structure, taking into account all factors that contribute to decrease pavement performance. Some of these factors are moisture, radiation, temperature, freeze-thaw cycles, maintenance and rehabilitation treatment and strategies, traffic loads, materials, construction

methods, structural design, etc. (Haas, 2003). However, the deterioration of pavement structures is a complicated process and, thus, hard to model.

Performance Models have many applications in Transportation Asset Management. Models were applied in Pavement Management Plans for identifying adequate treatment levels and for predicting pavement condition enhancements after the assigned treatments have been applied. The treatment levels identification are based on decision trees, while the prediction of pavement conditions are based on performance models. The data needed consists of location, pavement type and characteristics, pavement distress scores, ride scores, geometry of the pavement section, and traffic. Quality of pavement condition data is important for the estimation of needs and for selecting appropriate treatment levels that will result in adequate funding allocation decisions for Pavement Management Plans (Chang-Albitres et al., 2013).

Pavement performance prediction models can also be applied for timely application of maintenance and rehabilitation treatments and future needs estimation, based on data related to distresses and pavement condition indexes. Pavement performance can be expressed in terms of indicators, such as density of individual distress types (obtained from visual inspections), distress scores, and condition scores, among others. Gharaibeh et al. (2012) proposed performance models with adjusted coefficients based on the comparison of measured and predicted performance in terms of distress scores. The quality of rater's pavement condition data collected from the visual inspections has an impact on the calibration process because new values for the model's coefficients are calculated to minimize the difference between predicted and observed performance.

Currently, there are several pavement prediction models available. These models were developed based on different methodologies, input data, and considering different factors that influence pavement performance. Therefore, if the same inputs were provided, the models would generate different predictions. Pavement performance models have to be designed based on appropriate and fundamental engineering principles to be reliable and acceptable. Furthermore, the models have to be easily calibrated or adjusted depending on historical data and information

related to materials, climatic effects, construction, maintenance and rehabilitation activities, among others (Al-Zou'Bi et al., 2015). It is imperative that the data used for modeling pavement performance has good quality and truly reflects real pavement condition.

Pavement management agencies use different types of pavement performance prediction models depending on their management needs and the availability of resources (Gallegos, Chang-Albitres & Nazarian, 2013). These models can be classified in two major groups: deterministic and probabilistic models.

Deterministic Models

A deterministic model predicts specific pavement condition measures (e.g. level of distresses) throughout the analysis period. According to Lytton (1987), deterministic models can be classified as primary response, structural performance, functional performance, and damage models.

When a pavement is subjected to traffic loads as well as climatic effects, and the purpose is to forecast the primary responses of the structure, the more suitable deterministic model to use is the primary response model. Deflections, stresses, strains, moisture, and temperature are examples of primary responses of a pavement structure. Structural performance models forecast pavement structural behavior in terms of distresses (e.g. rutting, alligator cracking, punchout, faulting) or pavement condition, such as the PCI. Functional performance models forecast pavement's serviceability (e.g. pavement surface friction; present serviceability index). Lastly, the forecast of the normalized distress or loss of serviceability index of a pavement is predicted based on damage models. These models can result from functional or structural models by dividing them by the permissible serviceability index or distress values, respectively (Lytton, 1987).

Probabilistic Models

Probabilistic models predict pavement condition distribution during the analysis period. These models can be classified into more subcategories as survival curves and transition process

models (Lytton, 1987). Survival models calculate the percentage of sections of a pavement network that continue in service at the end of the analysis period or after the passes, of a standard load, a specific number of times. Transition process probabilistic models can be classified into two additional subcategories: Markov and Semi-Markov models.

The basis for the majority of probabilistic models are the Markov models, which define the probability that a pavement system (group of pavements with similar features, such as age, surface type or similar traffic loads) will change from one condition to another during the analysis period. The new condition of the pavement system depends on the previous one, but is independent on how the preceding condition was obtained. A limitation of Markov models is the unrealistic time independency of the transition process, while changing from one condition to another (Saba, 2007).

When a pavement structure is functioning and subjected to traffic loads and climatic factors, the processes of pavement performance reduction or pavement deterioration exhibit stochastic characteristics. Some factors that influence these processes, such as traffic loads and environmental conditions, are hard to predict due to its high variability in time. Therefore, pavement deterioration or performance is constantly changing with time. This changeable situation creates randomness and uncertainty, which can derive from the data collection process while performing measurements or inspections, from the incapacity in the quantification of factors that influence the deterioration process, and from models that should represent the true deterioration process of pavement materials.

The stochastic characteristics of the loss of pavement performance and the pavement deterioration can be addressed through the development of probabilistic models. As mentioned before, the majority of probabilistic models rely on Markov process modeling where pavement condition X_{i+1} at time "i+1" depends on the pavement condition X_i at some previous time "i" but does not depend on how the condition X_i was obtained (Al-Zou'Bi, 2013). A limitation for the majority of probabilistic models is the stationary assumption they rely on, which denote that pavement deterioration rate is not dependent on time. To improve this limitation, some

probabilistic models are based on time dependent Markov models (Zhen, 2005). These types of probabilistic models are the Semi-Markov models, which are more realistic because they consider that the state of a pavement depends on time, so variations in weather and in traffic will have an impact on the transition process from one condition to another.

Probabilistic models involve the definition of a Transition Probability Matrix (TPM), that is a square [s x s] matrix where "s" is the number of probable states in a system, such as the number of possible conditions in a pavement structure. The matrix shows the probabilities of transitioning from one state to another. For instance, the probability for a pavement in good condition to change into a bad condition over three years without applying any maintenance nor rehabilitation. The data needed to establish a TPM could be obtained from historical data or opinions of experts. The process of defining a TPM while developing a probabilistic model represents an important challenge due to the difficulties while collecting the data regarding quality, time demand, and expenses. The majority of matrices of current probabilistic models are designed based on a great amount of observed long-term pavement performance data. The tools used for data analysis are regression and the Markov model (Al-Zou'Bi, 2013).

Karan (1979), applied probabilistic models for determining pavement deterioration functions and model pavement maintenance of the Waterloo (Ontario) regional pavement system. Time-independent Markov process modeling with a constant TPM, were used to model pavement deterioration related to the age of the structure, throughout the entire analysis period. The data used to define the matrix was obtained from opinions of experts through interviews and questionnaires. The development of the TPM implied the demand of significant time and expenditures due to the subjectivity of the data and the way it was collected and processed.

Ramirez (2015) applied probabilistic models to analyze different pavement deterioration and performance scenarios to improve the decision-making process regarding the development of maintenance and rehabilitation strategies including treatment selection and budget estimation. A stochastic approach was considered to develop performance curves and performance-based

scenarios to determine treatment and budget needs, predict pavement performance, and estimate pavement deterioration rates.

Another example of a probabilistic model is the Highway Investment Planning System (HIPS). This system uses Markov chains and mathematical optimization methods to model the development of pavement condition. The purpose of the system is trying to find the best condition distribution in terms of rutting and roughness, considering user costs. HIPS is commonly used in Finland and Norway for network level pavement management (Al-Zou'Bi, 2013).

According to Gallegos (2012), another classification considers that deterministic and probabilistic models can be described as empirical, mechanistic or mechanistic-empirical. The criterion used for this classification was the type of data used for the development of the models.

Empirical Models

Empirical models are equations formulated according to experience, experiments, and observation. Therefore, these models cannot express theoretical mechanisms of pavement response. In an empirical model, measured or estimated variables (deflection, traffic loads, etc.) are related to pavement deterioration measures (loss of serviceability) and pavement age. These variables are typically related through regression analysis. Empirical models are generally regression equations that predict pavement performance under certain conditions. If conditions are different, models are no longer valid. The scope of the data used for developing the model defines its validity.

An example of empirical models are the ones developed in the AASHTO Guide for Design of Pavement Structures (AASHTO, 1993). Seven miles of asphalt pavement and concrete pavement, forming six loops and a tangent, were constructed to perform traffic tests for 2 years and determine the consequences of traffic loads and climate over the pavement structure. This test was called the American Association of State Highway Officials (AASHO) Road Test. Based on the test results, empirical relationships for pavement structural designs were established according to expected loadings during the life of the structure.

The equation for flexible pavement design relates future estimated traffic with reliability and standard deviation factors, subgrade resilient properties, loss of serviceability in terms of PSI, and structural pavement parameters. Concrete pavement's equation relates future estimated traffic with reliability and standard deviation factors, subgrade resilient properties, concrete characteristics, slabs' load transfer coefficient, drainage coefficient, loss of serviceability in terms of PSI, and thickness of the slab.

Another example of empirical models, are the ones incorporated in the World Bank's Highway Design and Maintenance Standards Model (HDM-4). HDM-4 is a tool for pavement management used in infrastructure planning of investments at the strategic and project levels. The software includes deterioration models for different distresses, traffic congestion models, cold climate effects, road safety, and environmental effects. For instance, the HDM-4 model described for total incremental change in roughness is expressed in terms of incremental changes due to structural deformation, cracking, rutting, potholing, and environmental effects. The model for plastic deformation is another example of an empirical formulation and is part of HDM-4 models as well. For this case, the incremental increase in plastic deformation is expressed in terms of a construction defects indicator, heavy vehicle speed, and total thickness of bituminous surfacing. All models in HDM-4 include a calibration factor in the equations.

The majority of Departments of Transportation in the United States have designed empirical performance models for different performance indicators. Some of these indicators are roughness, initiation of cracks, plastic deformation, longitudinal and transverse profile, surface and structural cracking, deflections, surface distresses, and skid resistance. The independent variable of the models can generally be load repetitions or age.

Mechanistic Models

Mechanistic models are equations that represent a pavement response, such as stress, strain, or deflection, based on theoretical knowledge. Mechanistic models are deterministic-based equations that directly predict pavement serviceability (Queiroz, 1983) or predict pavement

distresses that can be simplified and associated with serviceability. These mechanistic models can be applied to regional or local PMSs, but it is important to emphasize that deterministic models cannot be applied to all pavement management scenarios due to three main reasons. First, one reason is related to the uncertainty of pavement's behavior when traffic load and environmental conditions vary in time. Second, it is hard to quantify the factors that influence pavement deterioration. Finally, the last reason is the poor quality of pavement condition data due to errors in pavement condition measurements or bias during a subjective evaluation of pavement condition (Al-Zou'Bi, 2013).

Mechanistic-Empirical Models

The combination of mechanistic models and empirical data results in an equation considered as a Mechanistic-Empirical model. The form of the model as well as its variables, can be determined based on theoretical knowledge, but the coefficients are defined based on regression analysis constituted from observed or measured data (Rauhut & Gendel, 1987).

Mechanistic-Empirical models can predict pavement performance and deterioration (e.g. reduction of serviceability throughout the design period) based on regression analysis. The calculated response variables in a mechanistic-empirical model characterize the mechanistic behavior of the pavement structure. For instance, some variables could be the tensile stresses between the asphalt concrete layer and the granular base layer, the compression strains at the bottom of the last granular layer and on top of the subgrade, and the expected accumulated axle loads applied to the structure during the entire analysis period.

The performance of a pavement structure is generally represented based on fatigue cracking, rut depth, and other particular distress. The stresses and strains due to traffic loading are determined applying linear elastic theory for multilayer systems and finite element methodology. Other variables considered in the response calculation are the properties of the materials, such as the elastic modulus of the pavement' layers, and the environmental factors, such as temperature and moisture effects.

An example of mechanistic-empirical models are the ones developed by the National Cooperative Highway Research Program NCHRP Project 1-37A: Guide for Mechanistic-Empirical Design of New and Rehabilitated Pavement Structures (Hallin, McGhee & Schwartz, 2004). Distresses such as rutting, bottom-up cracking, top-down cracking, thermal cracking, and pavement' superficial properties such as smoothness, expressed in terms of IRI, are some examples of pavement condition metrics used in the model. The MEPDG proposes using either linear elastic multilayer theory or finite element approach to estimate pavement responses.

Each of the previously described models have to be carefully selected to adequately represent pavement performance. The selection process must include the advantages and disadvantages of the models, data availability, and the complexity of the problem to solve. For instance, applying empirical models are only convenient if pavements have similar characteristics, such as material types or climatic conditions. Mechanistic and mechanistic-empirical models can extrapolate outside of the data from which the models were calibrated (Lytton, 1987).

Regarding the information needed for the models design, the complexity of the data varies depending on the type of model to develop. For example, empirical models require data gathered from PMSs. On the other hand, mechanistic-empirical models are based on more complex data, which cannot be found in PMS's databases. The information needed for probabilistic models are historical data or subjective data, such as opinions of expert engineers. These types of data may represent a challenge for some transportation agencies due to the data collection and quality control processes. Finally, the amount and quality of data required for developing the models represents a limitation. The selection of the appropriate model to represent pavement condition behavior depends on the good quality of the data, which has to be reliable and complete to develop a model that truly reflects pavement performance (Lytton, 1987).

Summary of Chapter 2

Pavement management is a decision-making process which aims to provide, evaluate, and maintain pavements infrastructure in a serviceable condition throughout their entire life cycle. Pavement condition data is a critical component in pavement management due to the costs associated with data collection and the impact of data on decisions regarding funding allocation for the implementation of cost-effective maintenance and rehabilitation strategies. As a result, the quality of pavement condition data has to be complete, accurate and reliable. To achieve these quality features, a quality management approach has to be defined in the context of pavement management considering the processes related to quality control, quality acceptance, and independent assurance. These three quality management processes comprises the use of quality management techniques, such as personnel training, personnel certification, equipment calibration, data verification, and time-history comparisons, among others. Other alternatives available for quality assessment of condition data in pavement management include statistical techniques (e.g. Cohen's Kappa, Percent Agreement, Missing Data Techniques), mathematical models (e.g. Artificial Neural Networks, Fuzzy Logic, Bayesian Methods), data analysis techniques (Big Data, Data Mining), quality approaches (Six Sigma, Total Quality Management), and softwares (e.g. Expert Systems, Genetic Algorithms). Finally, prediction models are used to forecast pavement performance based on quality condition data. In conclusion, there is not a systematic methodology for quality control throughout the entire pavement management process, which is needed to obtain reliable information to make better-informed investment decisions that will lead to the implementation of the most cost-effective maintenance and rehabilitation strategies.

Chapter 3: Framework to Incorporate Quality Control in Pavement Management Systems

3.1 Framework for Quality Control in Pavement Management Systems

Good quality data is crucial for effective pavement management decisions. Quality control is one of the three components of quality management (quality acceptance and independent assurance are the other two) in which the data and its collection process are evaluated and adjusted. The purpose is to obtain data with acceptable levels of quality previously defined by the management agency (e.g. percentage of out of range data, closeness to ground truth values, percentage of missing data, etc.). As a result, quality control must be incorporated into the pavement management process to integrate quality management procedures and expand its application.

Figure 9 shows the generic asset management system components (FHWA, 1999). All of the elements and their relationships are shown graphically. This vertical organization can be incorporated in any asset management process, with specific differences depending on the type of asset to manage. This decision framework is composed of structured decision steps supported by information related to goals, policies, and budgets.

Figure 10 displays an asset management approach to resource allocation and project delivery, similar to the general asset management framework presented in Figure 9. The framework shown in Figure 10 is more focused on the decision-making process of a transportation agency concerning its investments in terms of system preservation, transportation system management and operation, and capacity expansion (NCHRP, 2006).

Both frameworks complement each other at different stages of their sequences. The program implementation process indicated in the framework regarding asset management components (Figure 9); involve the management functions of programming, construction program delivery, maintenance, and operations, which are included in the framework related to resource

allocation and delivery (Figure 10). The definition of performance measures and targets considered in the framework of Figure 10, involve the development of asset inventories, condition assessment, and performance modelling, which are mentioned in the framework of Figure 9.

The decision-making processes displayed in Figure 9 and Figure 10 do not explicitly include quality improvement of pavement data. Good quality information is important to support main components of pavement management, such as asset inventory, condition assessment, performance modeling, and budget allocations, among others. Based on the general Transportation Asset Management elements flowchart, a proposed framework to incorporate quality control in PMS is presented in Figure 11.

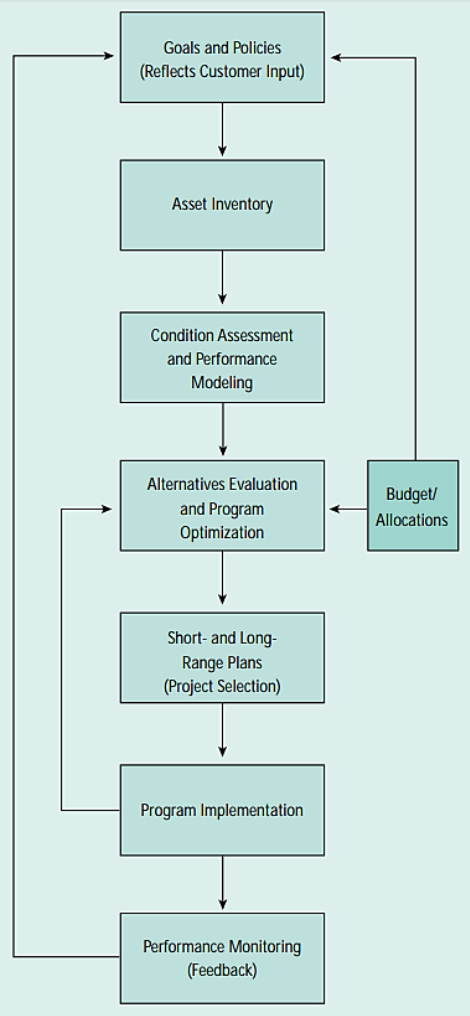


Figure 9: Generic Asset Management System Components. Source: FHWA, 1999.

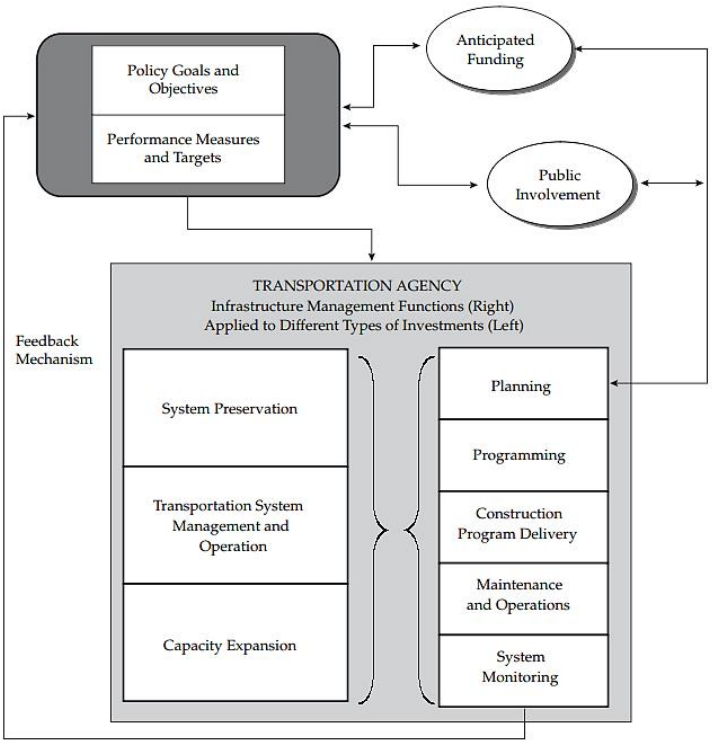


Figure 10: An Asset Management Approach to Resource Allocation and Project Delivery. Source: NCHRP, 2006.

The framework for quality control in Pavement Management Systems has ten major elements or stages. Stage number four is a new process proposed for data quality control in the context of pavement management, which includes a series of statistical tools organized in a sequential manner to identify poor quality data when comparing pavement condition data (data from the condition assessment stage) with reference values. The sequence of application of these statistical quality control tools is described in section 3.5.

I. Policy goals and objectives
II. Pavement inventory
III. Condition assessment
IV. Statistical quality control tools
V. Performance modeling
VI. Determination of needed work and funds
VII. Identification of candidate projects
VIII. Determination of impacts of funding alternatives
IX. Budget allocation
X. Feedback

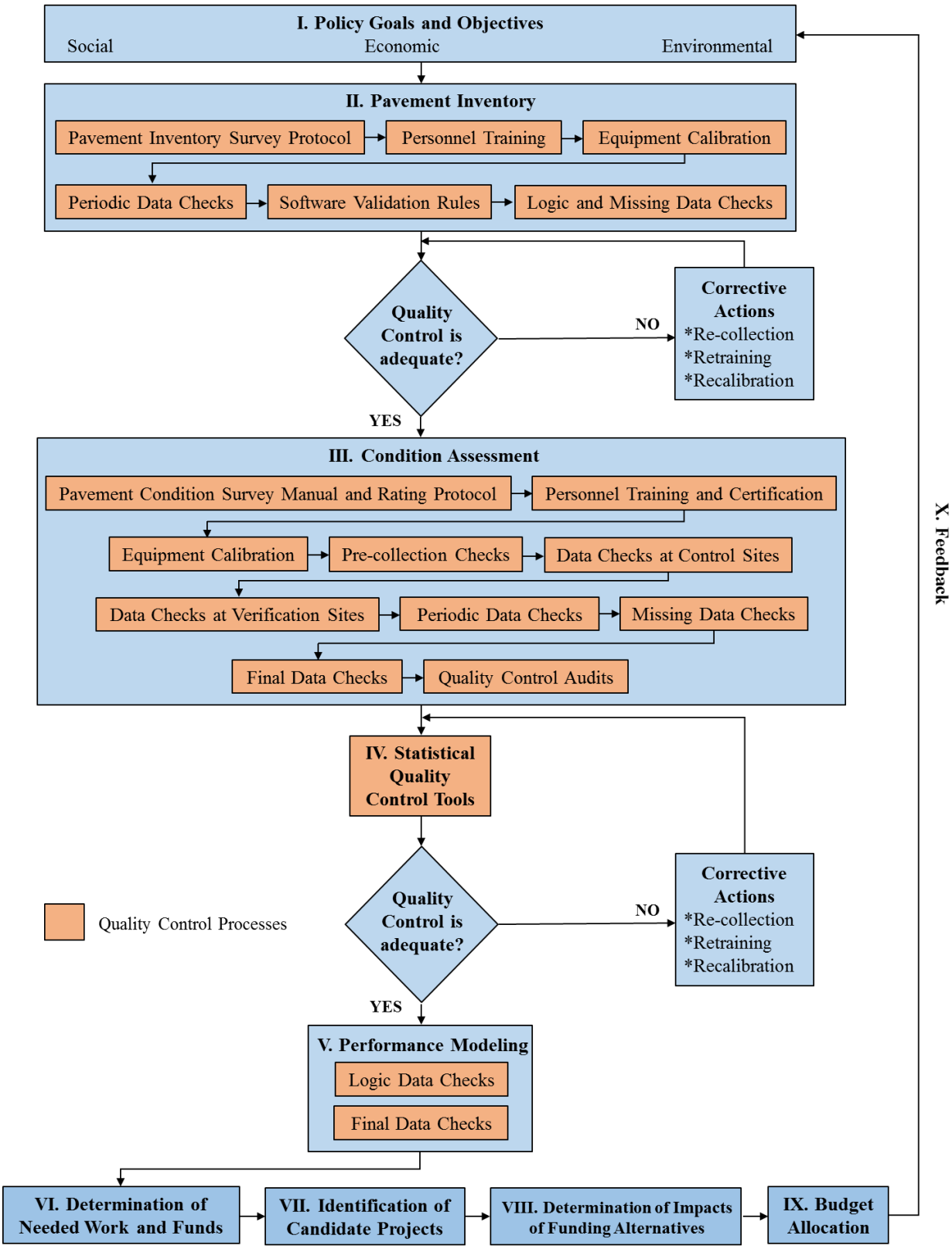


Figure 11: Framework for Quality Control in Pavement Management Systems.

A description for each stage is presented in the following sections. Furthermore, the relationships between the elements involved are explained in detail. Quality control is incorporated into the pavement management decision-making process, in two stages primarily: at the collection of data for pavement inventory and for condition assessment. The proposed methodology will focus on these two stages of the framework including one additional stage corresponding to the models to forecast pavement performance.

3.2 Stage I. Policy Goals and Objectives

The establishment of policy goals and objectives are the first stage of the framework. All the decisions on infrastructure management reflect the policy goals and objectives that define asset condition, performance levels, and quality of services to achieve user's needs (NCHRP, 2006). Management agencies have to establish policy goals and objectives while considering account sustainable principles, regarding social, economic, and environmental impact. Goals are the desired outcomes, broadly defined, not directly measurable, that an organization is willing to accomplish in a long term (FHWA, 2007). ISO 55000 (2014) defines objectives as "the results to be achieved". An objective can be defined as "the translation of a policy goal into a more specific measure of attainment" (FHWA, 2007). Objectives have to be consistent with the asset management policy in order to achieve a specific measurable result in a mid to short term. The goals and objectives have to be aligned to the organization's mission, intentions and directions. Clearly defined policy goals and objectives are needed for the assessment of physical facilities, for short and long term planning, and for budget allocation. An example of a policy goal for a transportation agency regarding data quality management, might be to improve the quality management plan to enhance the pavement data collection process. One objective associated to the previous goal can be the periodically upgrade of agency's manual used for pavement condition rating.

3.3 Stage II. Pavement Inventory

According to FHWA (2007), an inventory is "a compilation of the infrastructure assets of an agency and their relevant characteristics". These relevant features are referred to location and physical attributes of the transportation asset such as physical cross section, quantity or count, location, size, functional classification, miles of paved roads, materials, history, traffic usage, load data, district responsibility, etc. A complete inventory of highway infrastructure is an important component of any transportation management system (Flintsch & Bryant, 2006).

The entire inventory data collected resides in a central database that must be updated on a regular basis in order to manage current information regarding the assets under the jurisdiction of the transportation agency. Budget constraints limit the inventory extension regarding the level of detail of the information and the amount of assets being covered. The agency has to define breadth and depth of the inventory based on the availability of resources (ODOT, 2011).

Quality control is important at this stage to improve reliability and completeness of inventory data. In the next steps, the quality control methodology for pavement inventory is described:

Step 2.1. Develop a pavement inventory survey protocol

A quality management procedure is the development and updating of an agency's inventory survey protocol. This document describes the procedure for appropriately gathering inventory data taking into account agency's criteria. Examples of the guidelines contemplated in the inventory manuals are the element identification of the cross section of a typical road, the usage of equipment to measure linear dimensions (measuring tape, laser distance estimator, etc.), and the location of a drainage structure using geographic positioning system (GPS). The protocol can include forms to fill manually (sheet of paper) or digitally (by using smartphones, tablets, or similar devices), depending on the availability of resources. The field crew responsible for inventory data collection has to study and be familiar with the inventory protocol established by the transportation agency.

Step 2.2. Train the personnel

Training the personnel in charge of data collection for asset inventory is an important task to achieve good quality information. The training has to be based on the pavement inventory survey protocol of the agency. Data gathering for inventory does not demand the use of sophisticated equipment or expertise as pavement condition data. However, training is needed to assure that the collection crew is capable of capturing relevant information of the assets that truly reflect in site characteristics. A certification could be included for the personnel that successfully pass the training.

Step 2.3. Calibrate equipment

Prior to initializing pavement data collection for the asset inventory, the equipment has to be calibrated. Equipment for inventory might include odometers, GPSs, video cameras, photographic cameras, and portable computers, among others. The calibration of the equipment is performed according to the manufacturer's manual. All the data collection crew must learn how to operate this basic equipment and how to calibrate it appropriately.

Step 2.4. Perform periodic data checks

Once data is being collected for asset inventory, the collection crew can make periodic checks to review the collected data and identify possible inconsistencies or errors. If photos are taken, field crew should check the images periodically throughout the collection process and identify any bad quality or poor resolution. Video checks can also be performed to assure good visibility of the recordings. Collection crew must be capable of making simple checks and calibrations on the equipment used for inventory data gathering. For example, GPS devises have to be set at adequate geographic coordinate system or a minimum number of satellites connected. A GPS calibration on site can be done by comparing the coordinates and altitude from the device with a known Benchmark point.

Step 2.5. Define software validation rules

Software validation rules can be defined if the personnel are collecting inventory data digitally in portable computers. If distances are going to be recorded, the software must not accept zeros nor negative values. Drop-down menus can be set for the selection of road surface' material types used in a pavement network, such as asphalt or concrete.

Step 2.6. Perform logic and missing data checks

Data checks at the end of the collection process are recommended before leaving the area of interest just in case some errors are detected and corrective actions need to be performed. It is important to assure an appropriate format for the inventory data collected. Logic data checks are also needed, for example, limiting the range of values corresponding to the width of a lane in a pavement structure. Missing data checks can be applied to identify and complete missing information or null values.

If quality control is acceptable, based on the agency criteria (e.g. 10% maximum of missing data after data collection), the next step corresponds to asset condition assessment' data collection. If quality control is not acceptable, equipment used to gather the inventory information has to be re-calibrated and the collection crew has to re-collect part of pavements while staying near to the inspected area. If needed, personnel can be retrained until they comply with the agency's pavement inventory survey protocol.

3.4 Stage III. Condition Assessment

Condition assessment is a measure of the physical state of a transportation asset affected by aging and deterioration over time. Historical maintenance and rehabilitation of the asset also affects its condition by improving its performance when those treatments are applied (FHWA, 2007). The condition is typically expressed based on performance measures, such as the PCI. Other performance measures are IRI, Present Serviceability Rating (PSR), and percentage of individual distresses, among others.

Periodic data collection for pavement condition assessment is required to monitor pavement performance and identify maintenance and rehabilitation needs. The periodicity of condition surveys depends on the availability of resources, such as personnel, equipment, time, and funds. Budget constraints are the strongest reason for an agency to plan the number of interventions for data collection. For instance, Maryland State Highway Administration performs a condition assessment of its 16,000 lane-miles of highway network yearly, rating ride quality as poor, mediocre, fair, good, and very good (FHWA, 2007). For different transportation assets with different service life, the periodicity of data acquisition for condition assessment varies. For example, Thompson, Ford, Arman, Labi, Sinha, and Shirole (2012) recommends yearly inspections for Portland cement concrete pavement, and traffic sign inspections every one to two years.

The information available for the condition of an asset is updated regularly for routine or corrective maintenance identification, and for the selection of adequate treatment types and its appropriate timing. Quality control is required to improve the consistency, reliability, and accuracy of data gathered from condition assessment surveys. In the next steps, the quality control methodology for condition assessment is described:

Step 3.1. Define pavement condition survey manual and rating protocol

Transportation agencies must clearly define the pavement condition survey manuals and rating protocols for the collection crew and distress raters. Data collection personnel must know how to operate and adjust computer hardware, softwares, and automated data gathering systems. In case of the distress raters, they need to have a very good understanding about the agency's particular distress rating protocol, which can include different distresses or criteria for its evaluation.

For instance, the MTC's "Distress Identification Manual for Flexible Pavements" (MTC, 2016a) considers eight distress types including alligator cracking, block cracking, distortions, longitudinal and transverse cracking, patching, rutting and depressions, weathering, and raveling.

Almost all the distresses have three severity levels (low, medium, and high), except for raveling that has two severity levels (medium and high).

Another example is the "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys" (ASTM, 2018), that considers 20 distress types for asphalt pavements. These distresses are alligator cracking, bleeding, block cracking, bumps and sags, corrugation, depression, edge cracking, joint reflection cracking, lane/shoulder drop off, longitudinal and transversal cracking, patching, polished aggregate, potholes, railroad crossing, rutting, shoving, slippage cracking, swell, weathering, and raveling. Almost all the distresses have three severity levels (low, medium, and high), except for raveling that has two severity levels (medium and high) and polished aggregate that do not have any severity level.

Step 3.2. Train and certify the personnel

The training of the collection crew and distress raters is crucial for an effective data gathering. The collection crew is responsible for data acquisition regarding the condition of a transportation asset, e.g. when data collection crew collects deflections and radius of curvature of a pavement structure using the Falling Weight Deflectometer (FWD). Distress raters are in charge of collecting features of different distresses (manually or automated) that can be visually detected on the pavement surface, e.g. when graders collect the type, extension, and severity of a distress. Training programs for manual or automated data gathering rely on the agency's condition survey manuals for collection crew and distress raters. Agency's manuals for data condition acquisition include the operation of specialized equipment, softwares, and the knowledge of rating criteria, among other specifications. Certification programs are reliable alternatives to ascertain that the collection crew and raters have the knowledge and technical capabilities to perform the data collection process adequately in accordance with standards. For instance, the use of the FWD requires a training and certification program where operators have to demonstrate their proficiency (Irwin, Orr, and Atkins, 2011). Another example is the inertial profiler certification program

conducted by the Minnesota Department of Transportation (MnDOT) where operators and equipment are certified before collecting the data (MnDOT, 2017).

Step 3.3. Calibrate equipment

Equipment calibration for data collection has to be performed prior to start the pavement condition survey. Calibration comprises physical apparatus, methodologies and rater's criteria to ensure proper equipment utilization and proper application of rating manuals. Special sites with known length, pavement type, and condition values can be used to perform the calibration process. For specialized equipment (e.g. inertial profilers, light and heavy weight deflectometers), certified technicians must be handling and supervising the calibration process directly.

Step 3.4. Perform pre-collection checks

The final verification of equipment calibration and acceptance of data acquisition procedures are necessary before the collection starts. These pre-collection checks take place in control sites with known condition values. The collection crew must demonstrate that all its equipment is properly calibrated and that the correct procedures based on predefined manuals are being applied. Once the pre-collection checks are successfully passed, the collection crew will be qualified to begin the data collection process. Pre-collection checks at control sites are also used for distress ratings to check if raters are properly applying rating protocols.

Step 3.5. Perform data checks at control sites

Once data is being collected for condition assessment, control sites are periodically used to evaluate collection crew's procedure and validate equipment calibration. Control sites are road segments whose condition has been carefully measured by the transportation agency. This condition data is considered as a ground truth and is compared against the data obtained by the crew being evaluated. The collection crew has to return regularly to the control sites and check equipment calibration (e.g. accelerometers or sensors have to be recalibrated on regular basis) and the methodology for data acquisition.

Step 3.6. Perform data checks at verification sites

Verification sites are road segments also used for collection crew's evaluation, as well as distress raters' evaluation. Unlike control sites, verification sites have not been measured by the transportation agency. Once the collection crew has proven that its equipment is calibrated, another road segment (called verification site) is measured and the results are considered as reference values for future data checks. Multiple calibration checks are performed on verification sites to determine repeatability and accuracy of collection crew's equipment. Distress raters can also be evaluated periodically (e.g. weekly) on verification sites to verify their proficiency while collecting pavement condition data. Inter-rater reliability and intra-rater reliability are evaluated as well. In both cases, similar results are expected based on the agency's criteria. If any of the raters do not pass, additional training would be needed.

Step 3.7. Perform periodic data checks

Periodic data checks have to be performed during the collection process, in order to identify errors or inconsistencies. The collection crew has to detect if measurements (e.g. IRI, rut depth, cross slope, GPS, faulting) are outside of the expected range of admissible values. The majority of data collection vehicles have incorporated monitoring systems to identify these types of problems. If a video is being recorded during the collection process, personnel has to check for adequate clarity, focus, lighting, and visibility of the video. For distress raters that collect data using a rating software installed on a portable computer, tablet or other similar devices, values for specific data elements can be previously set to avoid the insertion of erroneous data. For instance, the units for each distress can be defined (e.g. square meters, linear meters), as well as the severity levels available for them (e.g. for raveling, ASTM D6433 considers two severity levels, and for longitudinal and transverse cracking the same standard considers three severity levels). In addition, validation rules can also be defined into the software, for example, constrain the insertion of a distress area greater than the total area of the pavement being inspected.

Step 3.8. Perform missing data checks

Missing data checks can be applied to identify incomplete information regarding specific data elements (that can be completed) or entire road segments (that have to be re-collected). Software routines checks during data acquisition can be performed to identify these type of problems or other inconsistencies in the data.

Step 3.9. Perform final data checks

Final data checks have to be implemented at the end of the day, to verify proper format, incomplete video recordings, and to identify multiple errors, such as values of zeros, null values, repeated values, out of range values, and negative values, among others. For manual distress data collection, final checks can detect inconsistencies such as asphalt cracking on Portland cement concrete pavements, or vice versa. Final checks also include missing data analysis to solve for incomplete values.

Step 3.10. Perform quality control audits

Sample audits are another technique that can be implemented for quality control once the data is collected. Random samples of road segments are selected by the agency and its condition is measured. These results are compared with the values obtained by the collection crew or distress raters. If the ratings do not meet agency's quality standards, corrective actions have to be implemented.

If quality control is acceptable based on the agency criteria (e.g. assuming a maximum of 10% of the raters' PCI values greater than +/- 20 PCI points of the ground truth's PCI values before data collections starts), the next step corresponds to performance modeling. If quality control is not acceptable, corrective actions have to be performed. Before data collection, if personnel does not pass the certification exam, additional training will be necessary until they were able to pass the test. If equipment is not calibrated, adjustments are needed to recalibrate it. If data checks on verification or control sites are not successful, the crew can re-collect any segment of the pavement

before retiring from the site. Similarly, distress graders can re-collect any road segment needed. Recalibration of defective equipment or additional training for the collection crew or distress raters are also included as corrective procedures during and after data collection.

3.5 Stage IV. Statistical Quality Control Tools

This stage is a new process proposed for data quality control in the context of pavement management. Table 5 shows a list of statistical tools proposed for quality control of pavement condition data. The definition, applicability, advantages, and disadvantages are described for each statistical tool. Quality control is based on the comparison of ground truth measurements versus pavement condition data collected from pavement sections of training programs, certification programs, pre-collection sites, control sites, verification sites, highway networks being assessed or sample audits. Statistical quality control tools can be applied to data gathered manually by raters or automatically with specialized equipment.

The process for the application of statistical tools proposed for quality control quantification of pavement condition data is described in this section. Statistical quality control tools are used to determine quality by comparing ground truth values with pavement condition data gathered manually or automatically. These statistical tools may be applicable to training programs, certification programs, pre-collection sites, control sites, verification sites, and sample audits, among other quality control processes.

The two data quality approaches proposed which were explained in the literature review, are incorporated in the methodology. Quality measured in terms of the closeness of pavement condition data and ground truth's values are considered, as well as quality measured in terms of the closeness between the mean and the standard deviation of the datasets. The proposed statistical quality control tools identify cases when the collected data do not reflect real pavement condition in the field. Figure 12 shows a flowchart of the proposed statistical quality control tools for condition data collected manually (by raters) or automatically (by equipment).

Table 5: Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
F-Test	F-Test compares the variances of two populations and determines if they are equal or not. A confidence level of 95% is defined, which means that the confidence interval includes the true value in 95 out of 100 studies performed (Bender & Lange, 2007). With a defined level of significance, p-values allow a decision about the rejection or maintenance of a previously formulated null hypothesis (Du Prel, Hommel, Röhrig, & Blettner, 2009). If the p-value is lower than 0.05, the null hypothesis of equal variances can be rejected, that is, the variances are significantly different. If the p-value is greater than 0.05, the null hypothesis of equal variances cannot be rejected, that is, the variances are not significantly different. If the calculated F value is equal to or greater than the F critical value, the null hypothesis of equal variances can be rejected, that is, the variances are significantly different. If the calculated F value is lower than the F critical value, the null hypothesis of equal variance cannot be rejected, that is, the variances are not significantly different.	The test can be applied to normally distributed populations.	The test can determine whether the variances are significantly different or not.	Normality tests are required (e.g. Shapiro-Wilk Test, QQ plot).

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Levene's Test	Levene's Test compares the variances of two populations and determines if they are equal or not. If the p-value is lower than 0.05, the null hypothesis of equal variances can be rejected, that is, the variances are significantly different. If the p-value is greater than 0.05, the null hypothesis of equal variances cannot be rejected, that is, the variances are not significantly different.	The test can be applied to non-normally distributed populations.	The test can determine whether the variances are significantly different or not.	Normality tests are required (e.g. Shapiro-Wilk Test, QQ plot).
T-Test	T-Test compares the means of two populations and determines if they are equal or not. If the p-value is lower than 0.05, the null hypothesis of equal means can be rejected, that is, the means are significantly different. If the p-value is greater than 0.05, the null hypothesis of equal means cannot be rejected, that is, the means are not significantly different.	The test can be applied to normally distributed populations.	The test can determine whether the means are significantly different or not.	Normality tests are required (e.g. Shapiro-Wilk Test, QQ plot).

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
T-Test (Continued)	If the calculated "t" value is equal to or greater than the "t" critical value, the null hypothesis of equal means can be rejected, that is, the means are significantly different. If the calculated "t" value is lower than the "t" critical value, the null hypothesis of equal means cannot be rejected, that is, the means are not significantly different.	The test can be applied to normally distributed populations.	The test can determine whether the means are significantly different or not.	Normality tests are required (e.g. Shapiro-Wilk Test, QQ plot).
Mann-Whitney Test	Mann-Whitney Test compares the means of two populations and determines if they are equal or not. If the p-value is lower than 0.05, the null hypothesis of equal means can be rejected, that is, the means are significantly different. If the p-value is greater than 0.05, the null hypothesis of equal means cannot be rejected, that is, the means are not significantly different.	The test can be applied to non-normally distributed populations.	The test can determine whether the means are significantly different or not.	Normality tests are required (e.g. Shapiro-Wilk Test, QQ plot).
Percent Agreement	Percent Agreement measures the absolute agreement between raters/measurements. Percent Agreement is calculated by dividing the number of agreements by the total number of ratings.	Two or more raters/measurements can be evaluated.	The percentage of exact agreement or the percentage of specific agreement between raters/measurements is calculated.	The agreement due to chance among raters is not considered.

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Percent Agreement (Continued)	Percent Agreement can range from 0% to 100%. The higher the percentage, the higher the agreement.	The ratings can be numerical (e.g. IRI values from 0 to 400 inches/mile) or categorical variables (e.g. condition categories such as very poor, poor, good, and very good).	An agreement per rater/measurement and an overall agreement can be calculated.	Therefore, Percent Agreement's results may be overestimated. No standard error exists.
Cohen's Kappa	Cohen's Kappa measures the agreement between two raters, disregarding the agreement due to chance. Cohen's Kappa is calculated by dividing the number of agreements (not considering the agreement due to chance, which is subtracted) by the total number of ratings not including the chance agreements. Cohen's Kappa can range from -100% to 100%. Degrees of agreement are: <0%: poor agreement (agreement due to chance) 0-20%: slight agreement 21-40%: fair agreement 41-60%: moderate agreement 61-80%: substantial agreement 81%-99%: almost perfect agreement 100%: perfect agreement	Two raters can be evaluated.	The agreement by chance is detected and eliminated from the total agreements.	Cohen's Kappa cannot be applied to more than two raters.

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Cohen's Kappa (Continued)	A Standard Error (SE) of Cohen's Kappa can be calculated as follow (Watson & Petrie, 2010): $$SE = \sqrt{\frac{p_0(1-p_0)}{n(1-p_E)^2}}$$ Where: p_0: observed agreement p_E: chance agreement n: total number of subjects (e.g. road sections) to be rated	The ratings are categorical variables.	The agreement by chance is detected and eliminated from the total agreements.	There are no degrees of disagreement. All disagreements have the same weight.
Weighted Cohen's Kappa	Weighted Cohen's Kappa measures the agreement between two raters, disregarding the agreement due to chance. Weights are applied depending on the degree of disagreement. The higher the disagreement, the higher the weight. Weighted Cohen's Kappa can range from 0% to 100%, same as the Cohen's Kappa. The same degrees of agreement are considered. A Standard Error can also be determined, similarly to the equation used for Cohen's Kappa, but including the disagreements' weights.	Two raters can be evaluated. The ratings are categorical variables.	The agreement by chance is detected and eliminated from the total agreements. The disagreements have different weights depending on its closeness to the categories where the raters agreed.	Weighted Cohen's Kappa cannot be applied to more than two raters.

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Fleiss' Kappa	Fleiss' Kappa measures the overall agreement between more than two raters, disregarding the agreement due to chance. Fleiss' Kappa can range from 0% to 100%, same as the Cohen's Kappa coefficient. The same degrees of agreement are considered. A Standard Error can also be determined, similarly to the equation used for Cohen's Kappa, but considering more than two raters.	More than two raters can be evaluated. The ratings are categorical variables.	The agreement by chance is detected and eliminated from the total agreements. An agreement per category is calculated, as well as an overall agreement, considering all categories.	There are no degrees of disagreement. All disagreements have the same weight.
Intraclass Correlation (ICC)	Intraclass Correlation measures the overall agreement between two or more raters/measurements. Intraclass Correlation compares the different ratings' variability of one specific subject with the total variability including the whole ratings and subjects. Intraclass Correlation can range from 0% to 100%. The degrees of agreement are: 0%: no agreement 1-39%: poor agreement 40%-59%: fair agreement 60%-74%: good agreement 75%-99%: strong agreement 100%: perfect agreement	Two or more raters/measurements can be evaluated. The ratings are quantitative variables.	Intraclass Correlation represents the variation in ratings due to raters/equipment performance. The difference of 100 minus ICC represents the variation in ratings due to rater/equipment disagreement. An agreement for two raters/measurements is calculated, as well as an overall agreement, including all raters/measurements.	No standard error exists.

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Kendall's Coefficient of Concordance (W)	Kendall's Coefficient of Concordance or Kendall's W measures the agreement between raters based on a rank order (Shweta, Himanshu, & Ram, 2015). Kendall's Coefficient of Concordance is calculated using the following equation: $$W = \frac{12R}{m^2(k^3 - k)}$$ Where: m: number of raters k: subjects to be ranked $$R = \sum_{i=1}^{k} (R_i - \bar{R})^2$$ Where: $R_i = \sum_{j=1}^{m} r_{ij}$, for each subject "i" r_{ij}: rating of subject "i" assigned by rater "j" $\bar{R}$: mean of the R_i Kendall's Coefficient of Concordance can range from 0% to 100%. The higher the percentage, the higher the agreement.	Two or more raters can be evaluated. The ratings are ordinal variables.	An agreement per rater and an overall agreement can be calculated.	In case of ties in the ratings, the average rank is used. If a number of ties are obtained, the calculation of W has to be redefined. No standard error exists.

Table 5: (Continued). Statistical Quality Control Tools.

Statistical Tool	Definition	Applicability	Advantages	Disadvantages
Bland-Altman Diagram	Bland-Altman Diagram is a plot that evaluates the agreement between raters/measurements. Bland-Altman Diagram is constructed by plotting the ratings' difference on the x-axis versus the average of the ratings on the y-axis. A limit of agreement of 95% is assumed. This means that the confidence interval covers the ratings' difference in 95 out of 100 of future measurements pairs. The mean of the ratings' differences and the 95% limits of agreement are represented as horizontal lines. The lower limit is equal to the mean of the ratings' differences minus 1.96 times the standard deviation of the differences. The upper limit is equal to the mean of the ratings' differences plus 1.96 times the standard deviation of the differences.	Two raters/measurements can be evaluated. The differences of the ratings have to be normally distributed to define the limits of agreement. A non-parametric approach can be used in case of non-normally distributed values.	The mean of the ratings' differences represent the estimated bias. Graphically it is possible to identify outliers, which are points located outside of the limits of the agreement, which represent the highest systematic errors. A trend of the points (above or below the mean difference line) is an indicator of proportional bias.	Normality tests are required for the ratings' differences (e.g. Shapiro-Wilk Test, QQ plot). ANOVA analysis (between the means and the differences) is needed to determine the existence of a proportional bias. If there is a statistical significant result, there is a proportional bias. If there is not a statistical significant result, there is not a proportional bias.

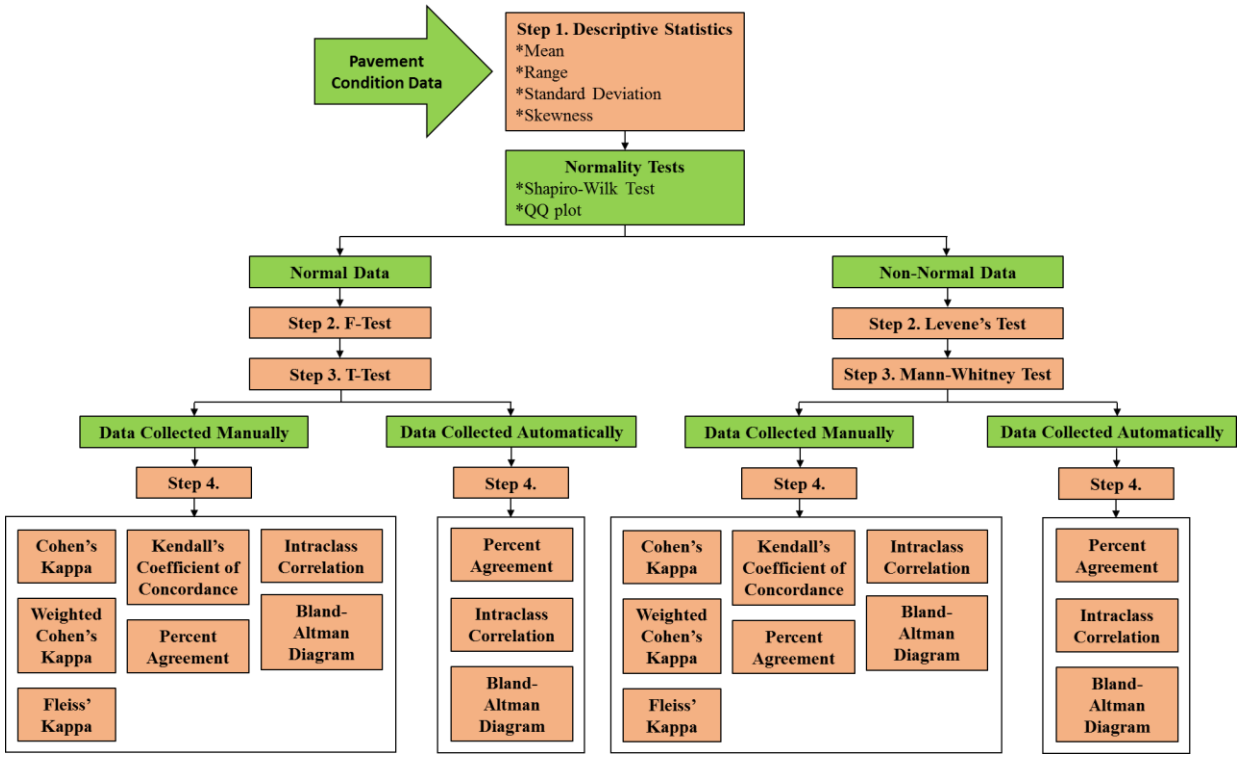


Figure 12: Flowchart of the Statistical Quality Control Tools.

Descriptive Statistics

Descriptive statistics can be used to determine some broad characteristics associated with quality (e.g. central tendency, variability) from a set of data. These quality statistics include the mean, range, the standard deviation, and a data distribution measurement (Montgomery, 2012).

The mean or average measures the central tendency of a dataset. The range can provide an idea of the amount of variation in the data. Standard deviation measures the amount of dispersion of the data around the mean. The lower the range and the standard deviation, the closest the clustering of values around the mean.

Finally, the shape of the distribution of the data can also reflect quality features. If the number of values below and above the mean are the same, the distribution is symmetric. On the other hand, a skewed distribution has a disproportionate number of values below the mean or above the mean. Skewness is a value that can be used as a measure of the asymmetry of a distribution.

The mean, range, standard deviation, and the skewness can be determined from the data that is being evaluated. These values provide an idea of the general characteristics of the data in terms of central tendency, variability, and distribution.

F-Test

The F-Test compares if the variance of the pavement condition data that is being evaluated is significantly different from the variance of the ground truth's values. It is expected that if the pavement condition data has good quality the values should be similar to the ones defined by the ground truth. In that case, the variances of the two datasets should not be significantly different. On the contrary, significantly different variances might indicate poor data quality.

F-Tests can only be applied to normally distributed datasets. Normality tests are required in the first place in order to determine if the data has a normal distribution. The analytical test proposed is the Shapiro-Wilk Test due to its high popularity for normality assumption diagnostics and because of its excellent power properties (Das & Imon, 2016). As a complement of the analytical test, the graphical test proposed is the QQ plot (quantile-quantile plot).

Levene's Test

Similarly to the F-Test, the Levene's test compares the variances of the pavement condition data and the ground truth's data with the purpose of determining if their variances are significantly different or not. The only difference between the tests is that the applicability of the Levene's Test is limited to non-normally distributed datasets.

Normality tests are required in the first place in order to determine if the data has a non-normal distribution. The tests proposed are the Shapiro-Wilk Test (analytical approach) and the QQ plot (graphical approach).

T-Test

The T-Test compares if the mean of the pavement condition data that is being evaluated is significantly different from the mean of the ground truth's values. It is expected that, if the pavement condition data has good quality, the values should be similar to the ones defined by the ground truth. In that case, the means of the two datasets should not be significantly different. On the contrary, significantly different means might indicate poor data quality.

T-Tests can only be applied to normally distributed datasets. The Shapiro-Wilk Test and the QQ plot are required to determine if the data has a normal distribution.

Mann-Whitney Test

Similarly to the T-Test, the Mann-Whitney Test compares the means of the pavement condition data and the ground truth's data with the purpose of determining if their means are significantly different or not. The only difference between the tests is that the applicability of the Mann-Whitney Test is limited to non-normally distributed datasets.

Normality tests are required in the first place in order to determine if the data has a normal distribution or not. The Shapiro-Wilk Test and the QQ plot are required to determine if the data has a non-normal distribution.

Percent Agreement

The absolute agreement between the ground truth's values and the pavement condition data can be quantified with the Percent Agreement. This coefficient is equal to the exact number of agreements divided by the total number of dataset's values, expressed as a percentage. Data quality is measured in terms of that percentage, which can range from 0% to 100%. Percent Agreement can be calculated for numerical or categorical data gathered either manually or automatically.

Cohen's Kappa, Weighted Cohen's Kappa, and Fleiss' Kappa

Cohen's Kappa measures the agreement between a rater and the ground truth disregarding the agreement due to chance. Data quality is measured in terms of degrees of agreement, which can range from 0% (slight agreement) to 100% (perfect agreement). Cohen's Kappa can be calculated for categorical data gathered manually by a rater.

Weighted Cohen's Kappa measures the agreement between a rater and the ground truth disregarding the agreement due to chance, but taking into consideration specific weights assigned for the degrees of disagreements. Data quality is measured in terms of degrees of agreement, which can range from 0% (slight agreement) to 100% (perfect agreement). Weighted Cohen's Kappa can be calculated for categorical data gathered manually by a rater.

Fleiss' Kappa measures the overall agreement between two or more raters and the ground truth disregarding the agreement due to chance. Data quality is measured in terms of degrees of agreement, which can range from 0% (slight agreement) to 100% (perfect agreement). Fleiss' Kappa can be calculated for categorical data gathered manually by a rater.

Intraclass Correlation

Intraclass Correlation measures the overall agreement between the ground truth and one or more than one rater, by comparing the different ratings' variability of one specific subject with the total variability, including the whole ratings and subjects. Equipment agreement can also be calculated for data collected automatically. Data quality is measured in terms of degrees of agreement, which can range from 0% (no agreement) to 100% (perfect agreement). Intraclass Correlation can be calculated for numerical data gathered either manually or automatically.

Kendall's Coefficient of Concordance

Kendall's Coefficient of Concordance measures the agreement between the ground truth and one or more than one rater based on a rank order. Data quality is measured in terms of a percentage, which can range from 0% to 100%. The higher the percentage, the higher the agreement. The ratings have to be ordinal variables.

Bland-Altman Diagram

The Bland-Altman Diagram is a graphical representation of the agreement between a rater and the ground truth. It can also be applicable to compare measurements made by two equipment. The diagram provides three indicators than can be related to the quality of the data.

One indicator is the estimated bias, which is the mean of the ratings' differences, and indicates if the mean values of the rater are below or above the mean values of the ground truth. This is only applicable if the differences of the ratings are normally distributed. For non-normally distributed differences, the estimated bias is equal to the median of the differences. The Shapiro-Wilk Test and the QQ plot are required to determine if the ratings' differences have a normal distribution.

Another indicator is the number of points located below the lower limit or above the upper limit. These outliers are points where the difference between the ratings of the rater and the ground truth surpass the limits of agreement. If the differences of the ratings have a normal distribution, the limits of agreement are equal to the mean of the ratings' differences +/– 1.96 times the standard deviation of the differences. If the differences of the ratings are not normally distributed, the lower limit and upper limit are the 2.5th percentile and 97.5th percentile, respectively. The Shapiro-Wilk Test and the QQ plot are required to determine if the ratings' differences have a normal distribution.

The third indicator is the existence of proportional bias. An ANOVA analysis has to be performed between the means and the differences of the rater's measurements and the ground truth's ratings. If there is a statistical significant result, there is a proportional bias, and vice versa.

Performance modeling is the next step whether quality control is adequate. If quality control is not acceptable, corrective actions have to be taken into account until the agency criteria for quality control is accomplished. Corrective actions include data re-collection, retraining of personnel, and equipment recalibration.

The first step of the statistical quality control tools flowchart is recommended to be applied to any pavement condition data to determine its general characteristics, such as central tendency, variability, and distribution. In the next steps, a minimum of two datasets are being compared, that is, ground truth values versus one or more pavement condition datasets collected from the field. Tests for variances comparisons (step 2) and means comparisons (step 3) are recommended to evaluate if the variance and mean of ground truth's dataset are significantly different or not from the variance and mean of another dataset. Similar values are expected for acceptable quality datasets. Step number four includes seven or three tools for data collected manually or automatically, respectively. These tools do not follow a strict order. Percent Agreement is used to determine the exact agreement between datasets. Agreement based on the different ratings' variability is calculated with the Intraclass Correlation coefficient. Agreement based on a rank order is calculated with the Kendall's Coefficient of Concordance. If agreement due to chance is disregarded from the analysis, the Cohen's Kappa coefficient can be used. Based on the previous tool, if weights are assigned for the degrees of disagreement, Weighted Cohen's Kappa coefficient can be calculated. If an overall agreement is needed neglecting agreement due to chance, the Fleiss' Kappa can be applied. The higher the agreement, the best the quality of the data. Finally, a graphical representation of the agreement between datasets can be obtained from the Bland-Altman diagram, which includes an estimated bias, outliers according to the ratings' difference, and the existence of proportional bias.

3.6 Stage V. Performance Modeling

Performance models are selected in this step to predict the condition of transportation assets expressed in terms of performance measures, such as PCI, IRI, bridge's deck area condition, fatality rate, and gas emissions, among others. Models can predict asset performance and deterioration rate under different maintenance scenarios, for instance, varying the time of treatment's application during the analysis period.

The types of models used to forecast transportation performance depends on the needs of the management agency (e.g. if network or project level analysis is required), the nature of the model (e.g. empirical, mechanistic, mechanistic-empirical, etc.), and the availability of resources to collect the input data for the model. The deterioration rate of the asset, the adequacy of the data to the model, the significance of the variables considered in the model, and the precision and accuracy of the model's estimates are factors that influence the selection of the performance models (Chang et al., 2017).

Once pavement condition data is collected from the field, it is used as an input for the models to predict assets performance. Quality control can be applied by analyzing the models' output. In the next steps, the quality control methodology for performance modeling is described:

Step 5.1. Perform logic data checks

Logic data checks have to be performed when out of range values are identified at some point of the asset's analysis period (e.g. PCI values above the maximum admissible value of 100, IRI values below the minimum admissible value of 0 in/mi).

Step 5.2. Perform final data checks

Logic data checks have to be performed when negative values, null values, or inconsistencies are identified (e.g. a significant increase in the deterioration rate that cannot be explained by traffic load nor climate conditions).

The majority of Pavement Management Systems include performance models to forecast deterioration considering preventive maintenance, routine maintenance, and rehabilitation, among other treatments. These predictions are compared to the results obtained with the same model but not considering any treatments. The objective is to quantify the effect of maintenance, reconstruction, and minor or mayor rehabilitation based on the performance measure considered in the model (Chang et al., 2017).

Maintenance improves asset condition and reduces deterioration rate at the time it is applied. The effect of maintenance depend on the treatment types, on the asset types and characteristics, and on the condition of the asset before maintenance is applied. In order to include the consequences of maintenance treatments on the model, good quality data regarding the condition of the asset have to be collected before and after the treatments are applied to measure the change in condition (Chang at al., 2017).

Models working with poor data quality, that do not represent the real condition of the asset, might generate wrong predictions. This may lead to inadequate decisions about maintenance treatments and budget allocation. Data quality control is important to overcome these negative consequences.

3.7 Stage VI. Determination of Needed Work and Funds

According to the projected pavement condition along the analysis period, transportation agencies are interested in the determination of the work needed for the asset to provide a minimum level of service. The funds and resources necessary to complete the work are an important aspect to consider because the available budget is limited.

Decision criteria has to be defined based on the evolution of pavement sections' condition throughout the analysis period. A periodic comparison (e.g. annually, semiannually) of the condition of a pavement section is important to establish decision criteria. These criteria or trigger

values (AASHTO, 2001a) contribute to the identification of pavement sections that require maintenance or rehabilitation.

For instance, if the condition of a pavement section is monitored using a performance measure that varies from 0 to 100 (such as the PCI), trigger values can be defined at specific PCI points in order to define condition ranges in which maintenance, rehabilitation or reconstruction are required. The breaking PCI points between maintenance and rehabilitation and between rehabilitation and reconstruction are trigger values. If the PCI of a pavement section reaches the maintenance, rehabilitation or reconstruction range, then the corresponding treatment is applied.

Once the treatment needs have been identified, the budget is quantified. The condition of the pavement improves after a treatment is applied. The summation of all the individual treatments applied at specific years represent the total budget needed for the entire analysis period. Nevertheless, funds are limited and it is unfeasible to apply all the treatments.

3.8 Stage VII. Identification of Candidate Projects

All the pavement sections that have been previously identified to receive treatment (maintenance or rehabilitation) have to be prioritized, in order to allocate the funds available. Commonly, transportation agencies do not count with unlimited funds. Thus, prioritization is important to provide the maximum benefit.

The prioritization process considers that maintenance and rehabilitation costs differ depending on the pavement's condition. The worse the condition, the higher the cost to improve it. Another consideration in prioritizing projects is the cost-effectiveness of treatments over time, which depends on the type of pavement, level of traffic loads, and importance of the road. Different ranking approaches are used in PMS to prioritizing projects; some examples are (Chang, 2007):

- Ranking based on damage measures. The higher the quality of the damage, the higher the priority.
- Ranking based on performance function. The better the performance function (e.g. serviceability, roughness), the higher the priority.

- Ranking based on life cycle cost. The lowest the life cycle treatment cost, the higher the priority.

The optimal funding allocation can be performed with optimization tools, such as integer programming, linear programing, dynamic programming, and Markov decision analysis, among others (AASHTO, 2001a). The purpose of these techniques is to identify the best set of pavement sections to treat, the type of treatments to apply, and the time of the intervention throughout the analysis period. This selection will bring about the maximum benefits in terms of economic values (e.g. user costs), minimum pavement condition, minimum treatment costs, etc.

The prioritization or optimization techniques used by the transportation agency will define the type of data needed to identify candidate projects. The higher the complexity of the technique, the higher the accuracy, cost, and level of detail of the data (AASHTO, 2001a).

3.9 Stage VIII. Determination of Impacts of Funding Alternatives

According to FHWA (2007), alternatives are "available choices or courses of action that can be considered at each stage of resource allocation". Alternatives are developed to sustain transportation assets during their entire life cycle. For example, different alternatives can consider the application of maintenance treatments at different times over the life of the asset, based on the availability of funds. These alternatives might generate distinct impacts in terms of remaining life and condition of the asset, fund needs, and other consequence considered by the transportation agency.

The impact of different strategies and funding levels has to be estimated taking into account the entire analysis period. The initial conditions of the asset can be projected and variations in condition can be detected depending on the funding strategies considered. Good quality data is important to define the initial conditions, which are then projected and analyzed based on the strategies available. For instance, the percentage change of pavement sections in good, fair, and poor condition due to different funding scenarios is a way to visualize the consequences of the alternatives. The initial condition of pavement sections have to reflect real in site conditions to

expect that projected conditions might also be realistic and possible to occur. Not only can the condition of the pavement be analyzed but similar projections can be performed with backlog costs, future economic needs, and user costs with the purpose of justifying the funds needed for specific alternatives.

3.10 Stage IX. Budget Allocation

Transportation agencies justify budget allocation by communicating the impacts of funding alternatives to the corresponding authorities. It is important to show the effects that different alternatives might generate during the analysis period. The percent change of pavement sections needing maintenance or rehabilitation, the changes in remaining life, and the changes in user costs are examples of the information that funding authorities are provided to allocate the funds. Budget allocation involve the assignation of resources such as money, equipment, and personnel, to the various areas of investment, taking into account the most optimal distribution.

The availability of resources is one of the most critical factors that is important for the success of implementing a pavement management system. The implementation of a PMS requires a formal financing structure to allocate budget and support the development of the programs (Akofio-Sowah & Amekudzi-Kennedy, 2016).

An engineering economic analysis is required before budget is allocated to the candidate projects. Some of the tools used in engineering economics include benefit/cost analysis, lifecycle cost analysis, prioritization, optimization, and risk analysis. The budget is allocated to the alternative that will accomplish performance objectives providing the highest benefit or the lowest cost at a long-term analysis (FHWA, 1999).

3.11 Stage X. Feedback

Feedback is important to evaluate the implementation and performance of a Pavement Management System. At this stage, projections are compared to real observed values. Based on these comparisons, prediction techniques are modified and enhanced in order to improve its

reliability. A continuous updating process is required for all the decision-making processes, prediction algorithms, and costs (AASHTO, 2001a).

The performance of the Pavement Management System is monitored during the feedback process. Transportation assets' condition and data related to assets' performance is studied and monitored throughout the analysis period. Performance monitoring also includes the identification of problems, the evaluation of improvements regarding investments, and the evolution of performance targets. "Performance monitoring provides a feedback mechanism for resource allocation and utilization decisions" (FHWA, 2007).

The entire framework that incorporates quality control into the TAM decision-making process is reevaluated periodically. The agency evaluates if the performance objectives are being accomplished through monitoring the condition of the assets. This procedure includes the collection of condition data, processing, and analysis. Hence, good quality data is required to ensure a reliable monitoring process.

Monitoring of performance measures must be a recurrent process to guarantee continuous feedback. Transportation agencies obtain important information from performance measures to detect problems and propose solutions to address these challenges. This feedback is needed to quantify the impact of past decisions applied during the life of the asset and consider them on future decisions.

Summary of Chapter 3

A framework to incorporate quality control in Pavement Management Systems is proposed. The framework contains ten stages including policy goals and objectives, pavement inventory, condition assessment, statistical quality control tools, performance modeling, determination of needed work and funds, identification of candidate projects, determination of impacts of funding alternatives, budget allocation, and feedback. Quality control is considered at pavement inventory, condition assessment, and performance modeling stages. For each stage, quality control methodologies are defined, such as the development of protocols, personnel training, equipment

calibration, and data checks, among others. The stage corresponding to the statistical quality control tools is a new process proposed for data quality control in the context of pavement management, which includes a series of statistical tools organized in a sequential manner to identify poor quality data when comparing pavement condition data with reference values. Descriptive statistics is the first recommended tool to determine general characteristics of the data, such as central tendency, variability, and distribution. The second and third tools are related to variances comparisons (F-Test for normally distributed data or Levene's Test for non-normally distributed data) and means comparisons (T-Test for normally distributed data or Mann-Whitney Test for non-normally distributed data). Not significantly different variances and means are expected for acceptable quality datasets. The next tools evaluate agreement among datasets under specific criteria. The higher the agreement, the better the quality of the data. These tools are Percent Agreement, Intraclass Correlation, Kendall's Coefficient of Concordance, Cohen's Kappa, Weighted Cohen's Kappa, and Fleiss' Kappa. All tools are applicable to manually collected data and only two tools (Percent Agreement and Intraclass Correlation) are applicable to automatically collected data. Bland-Altman diagram (applicable for data collected manually or automatically) is the final tool that graphically represents agreement in terms of an estimated bias, outliers according to the ratings' difference, and the existence of proportional bias.

Chapter 4: Case Studies Analysis

4.1 Raters Comparison Case Study

Certification exams for raters that will perform a pavement condition data collection process is a good alternative for quality control before data acquisition starts. An example of a certification exam is the MTC Rater Certification Exam designed by the Metropolitan Transportation Commission (MTC) of California. As mentioned on the official webpage of the agency, "The Rater Certification Exam is designed to improve the quality of pavement management data collected in the field by pavement raters" (StreetSaver Academy, 2018).

The exam includes two examinations: a theoretical exam and a practical survey exam. Both examinations evaluate the knowledge of the raters regarding the identification and quantification (in extension and severity) of pavement distresses in asphalt pavements and concrete pavements. The specifications are defined in a distress protocol created by the MTC in which eight distresses for asphalt pavements (MTC, 2016a) and seven distresses for concrete pavements (MTC, 2016b) are described and characterized. The procedures for an adequate manual collection process is also explained for each distress, indicating how to identify and measure them, and how to assign severity levels. A minimum score is set for the online written exam to get certified

The practical survey examination consists of 24 pavement sections, which contain different distresses at different severity levels. The raters must evaluate those sites manually (by walking), identifying the distresses, assigning a severity level for each distress, and measuring its extension. Based on the ratings, the Pavement Condition Index (PCI) is calculated. PCI values obtained from the raters' grading are compared to the PCI values obtained by agency's expert inspectors that are considered as ground truth values.

After comparing the ground truth's PCI values versus the raters' PCI values, it is possible to determine if the passing criteria is accomplished. In order to pass the practical survey exam, a rater must meet two criteria (StreetSaver Academy, 2018):

a) At least 50% of the PCI values for the inspected sections must be within +/- 8 PCI points of the reference or ground truth's PCI values.

b) No more than 12% of the PCI values for the inspected sections can be greater than +/- 18 PCI points of the reference or ground truth's PCI values.

Raters who want to be certified have to pass a field survey exam and an online written exam.

Eighteen raters took the MTC Rater Certification Exam to prove its proficiency for pavement condition data collection. The raters performed a visual inspection of the 24 pavement sections: twenty sections of asphalt pavement (A-1 to A-23) and the remaining sections of concrete pavement (P-1 to P-7).

The PCI values were calculated based on the data gathered during the surveys: type, extension, and severity level of the distresses. The software used to calculate the PCIs was Streetsaver. StreetSaver® is a pavement management software developed by the MTC. The program has approximately more than 400 users in the United States, most of them local transportation agencies. Particularly in the San Francisco Bay Area, MTC is in charge of the management of 43,000 lane miles of streets and roads for the 109 cities and counties. MTC performs pavement condition monitoring and maintenance needs assessments using StreetSaver® software (Tan & Cheng, 2014).

The pavement sections, the PCI values obtained by the raters, and the PCI values of the ground truth are shown in Table 6.

Table 6: PCI Values of the Ground Truth (GT) and Raters.

Section ID	GT	Raters																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A-1	13	4	18	30	28	22	24	16	25	24	26	23	20	19	17	8	16	10	21
A-2	17	13	32	34	35	38	33	33	33	27	25	43	37	37	24	24	29	9	23
A-3	64	62	60	68	82	73	54	54	64	69	71	84	67	69	72	71	59	63	73
A-4	85	79	66	93	76	95	69	92	87	95	88	95	87	88	95	87	81	64	80
A-5	32	15	33	49	56	53	35	38	50	51	58	52	37	49	40	37	52	20	37
A-6	23	13	27	28	28	26	23	32	30	28	29	12	31	28	5	13	31	4	13
A-7	13	8	28	35	34	37	31	38	29	24	38	35	34	26	25	13	30	28	16
A-8	59	36	36	53	41	62	38	60	49	51	47	63	57	34	61	46	45	26	34
A-9	33	25	17	54	51	56	40	43	43	46	47	56	53	31	52	39	38	25	31
A-10	49	21	48	50	59	53	45	65	51	50	60	54	67	49	52	55	60	38	48
A-11	26	6	39	36	45	48	25	15	28	19	37	44	28	29	30	29	27	4	9

Table 6: (Continued) PCI Values of the Ground Truth (GT) and Raters.

Section ID	GT	Raters																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A-12	30	20	33	34	35	37	32	49	35	34	41	36	35	34	26	27	36	3	16
A-14	63	60	45	60	43	64	49	62	54	60	61	47	67	41	63	55	39	38	37
A-15	21	19	37	51	49	51	44	58	44	47	57	39	42	31	32	21	32	39	30
A-17	48	51	45	55	53	55	43	22	49	51	49	55	24	47	53	53	37	22	42
A-18	49	36	28	44	77	68	41	34	52	58	55	66	32	26	65	44	47	29	55
A-19	50	40	43	48	69	58	38	40	57	61	59	59	45	43	59	49	48	34	56
A-21	94	98	91	96	97	97	85	80	96	100	94	97	83	95	97	97	94	92	95
A-22	93	90	93	100	100	100	87	83	98	100	96	100	78	97	100	99	96	97	97
A-23	96	94	91	97	98	98	85	78	94	99	88	98	84	94	97	97	94	89	96
P-1	59	49	33	61	82	61	52	48	47	42	61	59	48	59	59	47	45	43	43
P-5	74	81	69	83	80	75	74	70	83	59	79	77	67	62	77	79	70	51	55
P-6	48	50	41	55	68	61	35	32	56	46	53	61	34	45	56	58	48	38	31
P-7	77	68	61	85	85	75	72	68	65	81	74	75	65	59	76	69	58	52	66

The PCI values can range from 100 to 0, according to the distress protocols defined by the MTC for asphalt pavements (MTC, 2016a) and concrete pavements (MTC, 2016b). Four pavement condition categories are defined depending on the PCI values: very good (71-100), good (51-70), poor (26-50), and very poor (0-25).

Statistical quality control tools were used to evaluate the quality of the data collected by the raters while comparing them with the ground truth values. Since the distress data was collected manually, steps 1 through 4 of the proposed flowchart for manually collected data (Figure 12) can be applied depending if the raters' data has a normal or a non-normal distribution.

Through the majority of the cases, the statistical tools have been applied to the ground truth values and the values obtained by each of the raters. That is, ground truth's ratings versus values of rater 1, ground truth's ratings versus values of rater 2, and so on. In other cases, statistical techniques have been applied to all pavement condition data of the raters as a whole to obtain a general vision of the level of agreement of their results. On the rest of the cases, statistical techniques have been applied individually to the ratings of the ground truth and the ratings of all the raters.

The tools applied are alternatives with statistical basis that quantify the agreement between raters or equipment and quantify the quality of pavement condition data when comparing the results with ground truth values. Those agencies that would like to use these tools, would need to set criteria to identify poor data quality based on the tools' results and dismiss the rater or

equipment used during the collection of that poor quality data. Criteria selection might depend on each tools' results, pavement condition data that has been collected (manually or automatically) and each statistical tools' purpose of application.

As a complement, MTC's passing criteria has been taken as a reference to compare some of the results of the statistical quality control tools with the results of the MTC. The MTC's passing criteria results are indicated in Table 7.

Table 7: MTC Passing Criteria Results for Raters.

Raters	Criterion 1		Criterion 2	
	PCI within +/- 8 points	**PCI within +/- 8 points ≥ 50%**	**PCI values within +/- 18 points**	**PCI values within +/- 18 points ≥ 88%**
Rater 1	54%	Pass	87%	Fail
Rater 2	54%	Pass	83%	Fail
Rater 3	67%	Pass	87%	Fail
Rater 4	33%	Fail	62%	Fail
Rater 5	54%	Pass	71%	Fail
Rater 6	50%	Pass	92%	Pass
Rater 7	25%	Fail	83%	Fail
Rater 8	54%	Pass	96%	Pass
Rater 9	54%	Pass	92%	Pass
Rater 10	58%	Pass	87%	Fail
Rater 11	42%	Fail	79%	Fail
Rater 12	46%	Fail	79%	Fail
Rater 13	63%	Pass	83%	Fail
Rater 14	71%	Pass	96%	Pass
Rater 15	83%	Pass	100%	Pass
Rater 16	58%	Pass	87%	Fail
Rater 17	29%	Fail	58%	Fail
Rater 18	54%	Pass	87%	Fail

Step 1: Descriptive Statistics

Descriptive Statistics are determined for the PCI values of all the raters, including the ground truth. The summary contains 26 statistics commonly calculated to characterize any dataset. The statistics considered were the mean, variance, standard deviation, skewness, kurtosis, median, median absolute deviation, mode, minimum, maximum, range, count, sum, first quartile, third

quartile, interquartile range, and the percentiles for 1%, 2.5%, 5%, 10%, 20%, 80%, 90%, 95%, 97.5%, and 99%. The statistics are presented in Table 8. The complete statistical analyses are found in Appendix A. The software used to perform the statistical analyses was StatTools version 7.5.

Table 8: Descriptive Statistics of PCI values.

Raters	Mean	Range	Standard Deviation	Skewness
GT	50.67	83.00	26.33	0.2573
Rater 1	43.25	94.00	30.01	0.4304
Rater 2	46.42	76.00	22.08	0.9780
Rater 3	58.29	72.00	22.67	0.6198
Rater 4	61.29	72.00	23.01	0.1772
Rater 5	60.96	78.00	21.70	0.2705
Rater 6	48.08	64.00	20.01	0.8162
Rater 7	50.42	77.00	21.47	0.1754
Rater 8	54.96	73.00	22.18	0.7083
Rater 9	55.08	81.00	24.83	0.5792
Rater 10	58.04	71.00	20.72	0.2726
Rater 11	59.58	88.00	23.72	0.1381
Rater 12	50.92	67.00	20.57	0.3268
Rater 13	49.67	78.00	23.67	0.9130
Rater 14	55.54	95.00	26.90	0.0260
Rater 15	50.71	91.00	27.41	0.3175
Rater 16	50.50	80.00	22.36	0.8518
Rater 17	38.25	94.00	27.14	0.8010
Rater 18	46.00	88.00	26.79	0.6472

The mean, range, standard deviation, and skewness are general features that characterized each rater's data. Figure 13 shows a scatterplot of the mean, range, and standard deviation values of the pavement condition data collected by all the raters. These points can be compared to the mean, range, and standard deviation values obtained by the ground truth.

The central tendency of the raters' values are located below and above the ground truth's mean. This behavior is also noticeable for the range and the standard deviation. The closeness of any of these parameters to the ones related to the ground truth does not represent a level of agreement.

For instance, Rater 7 does not meet MTC's passing criteria and the mean is very close to the ground truth's value. The range of Rater 6 is very far from the ground truth's range, but Rater 6 does meet both criteria. The standard deviation of Rater 17 is very close to the ground truth's value, but Rater 17 does not meet the criteria.

The three previous statistics, as well as the skewness are general statistical features to characterize the data. The skewness for all the raters including the ground truth are positive, indicating that the right tail of the distribution is longer. Rater's data has different values for skewness that means that the shape of the distribution is different, however, in all the cases the shape is similar to the one indicated in Figure 14.

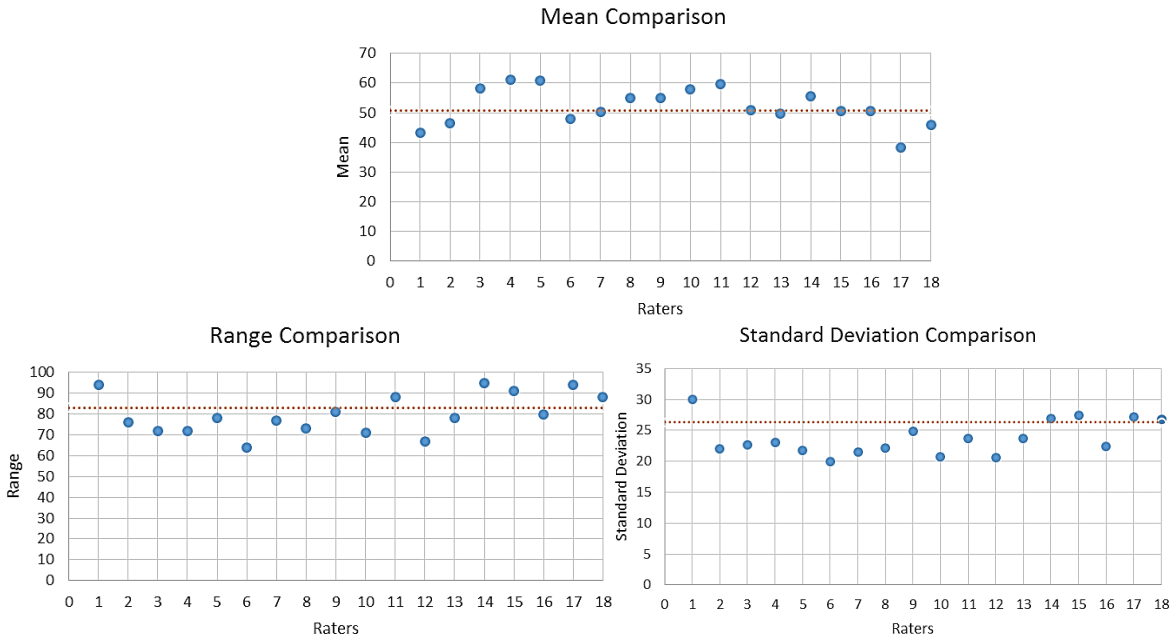


Figure 13: Mean, Range, and Standard Deviation Comparison between the Ground Truth (orange horizontal line) and the Raters (blue points).

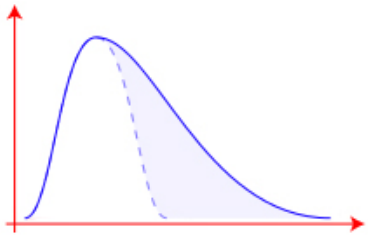


Figure 14: Distribution with a positive skewness.

Step 2: F-Test / Levene's Test

The F-Test and Levene's Test are applied to two sets of data: PCI values of the ground truth with PCI values of each rater. These tests are important to determine if the variance of one rater's dataset is significantly different from the variance of the ground truth values. The application of these tests depend on the distribution of the data. If the data is normally distributed, the F-Test is more suitable; if not, the Levene's test has to be used.

Normality tests are required. The Shapiro-Wilk test and the QQ plot are applied to all the PCI values per rater to evaluate its distribution. The Real Statistics Resource Pack developed by Charles Zaiontz was used to perform the calculations. Real Statistics Resource Pack is an Excel add-in which extends Excel's standard statistics capabilities by providing an advanced worksheet functions and statistical data analysis tools (Zaiontz, 2018).

The Shapiro-Wilk test results are shown in Table 9 and Figure 15 shows the QQ plots of two raters' data. The points on the diagram of Rater 7 are located more closely to the line y=x. PCI values of Rater 6 are not normally distributed, while PCI values of Rater 7 are normally distributed. The rest of the plots are found in Appendix B.

The results of the F-Test and Levene's Test indicate if the variances are significantly different or not (See Table 10). The F-Test and the Levene's Test were performed using the Real Statistics Resource Pack. The complete analyses for the F-Test and the Levene's Test are found in Appendix C and D, respectively.

Table 9: Shapiro-Wilk Test Results.

Raters	W-stat	P-value	Alpha	Normal
GT	0.942	0.183	0.05	Yes
Rater 1	0.926	0.077	0.05	Yes
Rater 2	0.886	0.011	0.05	No
Rater 3	0.903	0.025	0.05	No
Rater 4	0.940	0.159	0.05	Yes
Rater 5	0.953	0.310	0.05	Yes
Rater 6	0.887	0.011	0.05	No
Rater 7	0.971	0.701	0.05	Yes
Rater 8	0.907	0.030	0.05	No
Rater 9	0.916	0.047	0.05	No
Rater 10	0.958	0.400	0.05	Yes
Rater 11	0.964	0.530	0.05	Yes
Rater 12	0.929	0.093	0.05	Yes
Rater 13	0.882	0.009	0.05	No
Rater 14	0.960	0.429	0.05	Yes
Rater 15	0.952	0.306	0.05	Yes
Rater 16	0.908	0.032	0.05	No
Rater 17	0.919	0.055	0.05	Yes
Rater 18	0.925	0.077	0.05	Yes

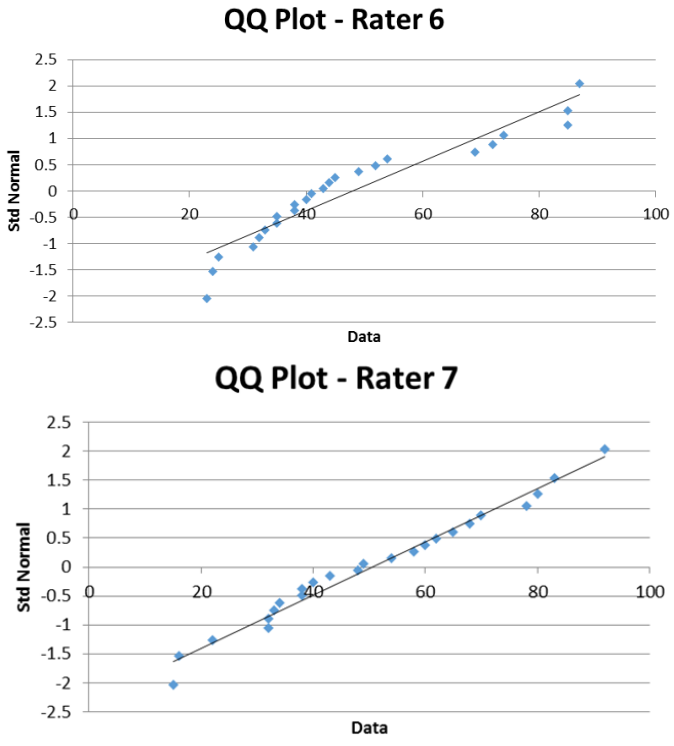


Figure 15: QQ plots for Rater 6 and Rater 7.

Table 10: Variances Comparison Based on the F-Test and the Levene's Test.

Raters	F-Test	Levene's Test
Rater 1	Variances are not significantly different	-
Rater 2	-	Variances are not significantly different
Rater 3	-	Variances are not significantly different
Rater 4	Variances are not significantly different	-
Rater 5	Variances are not significantly different	-
Rater 6	-	Variances are not significantly different
Rater 7	Variances are not significantly different	-
Rater 8	-	Variances are not significantly different
Rater 9	-	Variances are not significantly different
Rater 10	Variances are not significantly different	-
Rater 11	Variances are not significantly different	-
Rater 12	Variances are not significantly different	-
Rater 13	-	Variances are not significantly different
Rater 14	Variances are not significantly different	-
Rater 15	Variances are not significantly different	-
Rater 16	-	Variances are not significantly different
Rater 17	Variances are not significantly different	-
Rater 18	Variances are not significantly different	-

The results show that the variances of all the raters' data were not significantly different from the variance of the ground truth's data. This means that the variability of the raters' PCI values is close to the variability of the ground truth's values. It is possible to assume that the raters who took the exam have a certain level of knowledge that make their data's variances similar to the variance of the ground truth's data. All the raters' data could be considered as good quality data. On the contrary, if the variances were significantly different, it is possible to sustain that the rater's data quality is bad.

Step 3: T-Test / Mann-Whitney Test

The T-Test and the Mann-Whitney Test are applied to the PCI values of the ground truth and the PCI values of each rater. These tests are important to determine if the mean of one rater's dataset is significantly different from the mean of the ground truth values. The application of these tests depends on the distribution of the data. If the data is normally distributed, the T-Test is more suitable; if not, the Mann-Whitney test has to be used.

Table 11 indicates the results of the T-Test and the Mann-Whitney Test: if the means are significantly different or not for normally distributed and not normally distributed data, respectively. The T-Test and the Mann-Whitney Test were performed using the Real Statistics Resource Pack. The complete analyses for the T-Test and the Man-Whitney Test are found in Appendix E and F, respectively.

Table 11: Means Comparison Based on the T-Test and the Man-Whitney Test.

Raters	**T-Test**	**Man-Whitney Test**
Rater 1	Means are not significantly different	-
Rater 2	-	Means are not significantly different
Rater 3	-	Means are not significantly different
Rater 4	Means are not significantly different	-
Rater 5	Means are not significantly different	-
Rater 6	-	Means are not significantly different
Rater 7	Means are not significantly different	-
Rater 8	-	Means are not significantly different
Rater 9	-	Means are not significantly different
Rater 10	Means are not significantly different	-
Rater 11	Means are not significantly different	-
Rater 12	Means are not significantly different	-
Rater 13	-	Means are not significantly different
Rater 14	Means are not significantly different	-
Rater 15	Means are not significantly different	-
Rater 16	-	Means are not significantly different
Rater 17	Means are not significantly different	-
Rater 18	Means are not significantly different	-

The results show that the means of all the raters' data were not significantly different from the mean of the ground truth's data. The central tendency of the raters' PCI values is close to the central tendency of the ground truth's values. It is possible to assume that the raters who took the exam have a certain level of knowledge that make their data's means similar to the mean of the ground truth's data. All the raters' data could be considered as good quality data. On the other hand, if the means were significantly different, it is possible to sustain that the rater's data quality is bad.

Step 4: Percent Agreement

The exact number of agreements between the ground truth and each of the raters is determined with the Percent Agreement coefficient. Since an exact coincidence of the numerical PCI values is improbable to accomplish, PCI categories (very good, good, poor, and very poor) are used instead, in order to calculate the absolute agreement.

Percent Agreement was calculated by counting the number of coincident categories among the ground truth and the rater, and dividing that number by 24. The results per rater of the percentage of coincident categories are displayed in Table 12.

Table 12: Percent Agreement Values when comparing each Rater with the Ground Truth.

Raters	Percent Agreement
Rater 1	62.50%
Rater 2	54.17%
Rater 3	66.67%
Rater 4	33.33%
Rater 5	50.00%
Rater 6	70.83%
Rater 7	58.33%
Rater 8	54.17%
Rater 9	58.33%
Rater 10	54.17%
Rater 11	50.00%
Rater 12	58.33%
Rater 13	66.67%
Rater 14	66.67%
Rater 15	70.83%
Rater 16	54.17%
Rater 17	45.83%
Rater 18	54.17%

The agency should set the appropriate Percent Agreement value to accept or dismiss a rater due to its poor data quality. An option to define an appropriate value could be comparing the Percent Agreement percentages with an existing passing criteria. If a Percent Agreement value equal to or greater than 47% is considered acceptable, there would be 83% similarity with the results obtained according to MTC's criterion 1 (Raters 7, 11, and 12 that failed based on the MTC's criterion were accepted). A similar situation occurs for criterion 2 when a Percent

Agreement value equal to or greater than 68% is considered acceptable, obtaining a 83% similarity with MTC's results (Raters 8, 9, and 14 that passed based on the MTC's criterion were dismissed). A recommendation would be selecting a Percent Agreement value with the highest percentage of similarity and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative: accepting all raters with Percent Agreement values equal to or greater than 68%.

Step 4: Cohen's Kappa

Cohen's Kappa coefficient was calculated by comparing the results of the ground truth with each of the raters' values. The level of agreement per rater is determined using PCI categorical data and disregarding the agreement due to chance. Table 13 shows the Kappa coefficients and the interpretation of its corresponding level of agreement. Appendix G shows the complete results of the Real Statistic Resource Pack software used for the analyses.

Table 13: Cohen's Kappa Coefficients: Comparison of Raters versus the Ground Truth.

Raters	Cohen's Kappa	Agreement
Rater 1	50.34%	Moderate
Rater 2	33.50%	Fair
Rater 3	54.07%	Moderate
Rater 4	7.91%	Slight
Rater 5	34.55%	Fair
Rater 6	58.21%	Moderate
Rater 7	42.72%	Moderate
Rater 8	36.23%	Fair
Rater 9	43.79%	Moderate
Rater 10	37.88%	Fair
Rater 11	33.64%	Fair
Rater 12	42.17%	Moderate
Rater 13	51.52%	Moderate
Rater 14	56.56%	Moderate
Rater 15	60.28%	Moderate
Rater 16	34.16%	Fair
Rater 17	25.89%	Fair
Rater 18	36.99%	Fair

Similarly to the Percent Agreement, the agency should set the appropriate Cohen's Kappa value to accept or dismiss a rater due to its poor data quality. An option to define an appropriate value could be comparing the Cohen's Kappa percentages with an existing passing criteria. If a Cohen's Kappa value equal to or greater than 27% is considered acceptable (corresponding to a minimum fair agreement), there would be 83% similarity with the results obtained according to MTC's criterion 1 (Raters 7, 11, and 12 that failed based on the MTC's criterion were accepted). A similar situation occurs for criterion 2 when a Cohen's Kappa value equal to or greater than 56% (corresponding to a minimum moderate agreement) is considered acceptable, obtaining a 89% similarity with MTC's results (Raters 8, and 9 that passed based on the MTC's criterion were dismissed). A recommendation would be selecting a Cohen's Kappa value with the highest percentage of similarity and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative: accepting all raters with Cohen's Kappa values equal to or greater than 56%.

Step 4: Weighted Cohen's Kappa

Weighted Cohen's Kappa coefficient was calculated by comparing the results of the ground truth with each of the raters' values. The level of agreement per rater is determined using PCI categorical data and disregarding the agreement due to chance. The weights assigned for the PCI categories are indicated in Table 14. The diagonal represents perfect agreement and the weights are equal to 0. The weights increase depending on the distance from the diagonal.

Table 14: Weights Used for Weighted Cohen's Kappa Calculation.

PCI Categories	Very Good	Good	Poor	Very Poor
Very Good	0	1	2	3
Good	1	0	1	2
Poor	2	1	0	1
Very Poor	3	2	1	0

Table 15 shows the Weighted Kappa coefficients and the interpretation of its corresponding level of agreement. Appendix H shows the complete results of the Real Statistic Resource Pack software used for the analyses.

Table 15: Weighted Cohen's Kappa Coefficients: Raters versus Ground Truth Comparison.

Raters	Weighted Cohen's Kappa	Agreement
Rater 1	70.81%	Substantial
Rater 2	55.70%	Moderate
Rater 3	65.82%	Substantial
Rater 4	39.29%	Fair
Rater 5	34.55%	Fair
Rater 6	73.67%	Substantial
Rater 7	58.62%	Moderate
Rater 8	57.83%	Moderate
Rater 9	63.53%	Substantial
Rater 10	56.50%	Moderate
Rater 11	57.65%	Moderate
Rater 12	61.54%	Substantial
Rater 13	67.89%	Substantial
Rater 14	72.88%	Substantial
Rater 15	75.65%	Substantial
Rater 16	56.15%	Moderate
Rater 17	54.12%	Moderate
Rater 18	61.40%	Substantial

Similarly to the Cohen's Kappa, the agency should set the appropriate Weighted Cohen's Kappa value to accept or dismiss a rater due to its poor data quality. An option to define an appropriate value could be comparing the Weighted Cohen's Kappa percentages with an existing passing criteria. If a Weighted Cohen's Kappa value equal to or greater than 56% is considered acceptable (corresponding to a minimum moderate agreement), there would be 78% similarity with the results obtained according to MTC's criterion 1 (Rater 5 that passed based on the MTC's criterion was dismissed and Raters 7, 11, and 12 that failed based on the MTC's criterion were accepted). A similar situation occurs for criterion 2 when a Weighted Cohen's Kappa value equal to or greater than 69% (corresponding to a minimum substantial agreement) is considered acceptable, obtaining an 83% similarity with MTC's results (Rater 1 that failed based on the

MTC's criterion was accepted and Raters 8, and 9 that passed based on the MTC's criterion were dismissed). A recommendation would be selecting a Weighted Cohen's Kappa value with the highest percentage of similarity and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative: accepting all raters with Cohen's Kappa values equal to or greater than 69%.

Step 4: Fleiss' Kappa

Fleiss' Kappa coefficient is calculated taking into account the results of all the raters together. This coefficient represents the overall agreement of all the group of raters between each other, disregarding the agreement due to chance and considering the PCI categorical data. Real Statistic Resource Pack software was used for the analyses. Table 16 shows the overall agreement and the agreements per category.

Table 16: Fleiss' Kappa Results for All the Raters.

PCI Categories	Total	Very Good	Good	Poor	Very Poor
Kappa	38.52%	75.11%	21.54%	29.25%	30.80%
Agreement	Fair	Substantial	Fair	Fair	Fair
Std. Error	0.010024	0.016502	0.016502	0.016502	0.016502
Z-stat	38.43054	45.51142	13.05123	17.72662	18.66379
P-value	0	0	0	0	0
Lower	0.365563	0.718706	0.183033	0.260189	0.275654
Upper	0.404854	0.783395	0.247722	0.324877	0.340343

The Fleiss' Kappa was determined for all the PCI values of the raters. The total agreement of the entire group of raters was fair with a coefficient of 38.52%. This level of agreement can be used to compare the group of raters with others that have taken the same certification exam. This general level of agreement among raters can also be calculated for each condition category. The highest level of agreement among raters, which corresponds to a substantial agreement, was found for the very good category. For the remaining three categories (good, poor, and very poor) the agreement among raters correspond to a lower level that was a fair agreement. The raters tend to have a higher agreement to those pavement sections in very good condition, due probably to the

low quantity of the distresses, to the type of distresses found (distresses with no important structural damage impact) or maybe because of the low severity level of the distresses found. For those pavement sections in good, poor, and very poor condition, more distresses, extension, and severity levels are encountered. These factors make the rating more challenging and additional differences among raters is expected.

Step 4: Interclass Correlation

Interclass Correlation coefficient is calculated for each rater when its results are compared to the ground truth's rating. In addition, an overall agreement is calculated taking into account the values of all the raters together. Real Statistic Resource Pack software was used for the analyses. Table 17 shows the overall agreement and the agreements per rater.

Table 17: Interclass Correlation Coefficients for each Rater and Overall Agreement.

Raters	**Interclass Correlation**
Rater 1	92.45%
Rater 2	86.19%
Rater 3	89.14%
Rater 4	79.38%
Rater 5	85.42%
Rater 6	88.52%
Rater 7	80.12%
Rater 8	91.66%
Rater 9	91.59%
Rater 10	86.19%
Rater 11	86.05%
Rater 12	84.61%
Rater 13	89.09%
Rater 14	94.83%
Rater 15	97.10%
Rater 16	90.40%
Rater 17	79.94%
Rater 18	90.18%
Overall Agreement	83.15%

Similarly to the Weighted Cohen's Kappa, the agency should set the appropriate Interclass Correlation value to accept or dismiss a rater due to its poor data quality. An option to define an

appropriate value could be comparing the Interclass Correlation values with an existing passing criteria. If an Interclass Correlation value equal to or greater than 82% is considered acceptable (corresponding to a minimum strong agreement), there would be 89% similarity with the results obtained according to MTC's criterion 1 (Raters 11, and 12 that failed based on the MTC's criterion were accepted). A similar situation occurs for criterion 2 when an Interclass Correlation value equal to or greater than 92% (corresponding to a minimum strong agreement) is considered acceptable, obtaining an 89% similarity with MTC's results (Rater 1 that failed based on the MTC's criterion was accepted and Raters 6 that passed based on the MTC's criterion were dismissed). A recommendation would be selecting an Interclass Correlation value with the highest percentage of similarity and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative: accepting all raters with Interclass Correlation values equal to or greater than 92%.

An overall Interclass Correlation coefficient was also calculated taking into account all the raters. The agreement obtained was 83.15% corresponding to a strong agreement among raters. The Intraclass Correlation coefficient is calculated by comparing the different ratings' variability of one rater with the total variability including the whole raters. Since the variance of the raters are not significantly different, high values of the coefficient are expected.

Step 4: Kendall's Coefficient of Concordance

Kendall's Coefficient of Concordance is calculated for each rater when its results are compared to the ground truth's rating. In addition, an overall agreement is calculated taking into account the values of all the raters together. Pavement sections were ranked based on the numerical PCI values assigned by the raters. In case of ties, the average of the rankings were calculated. Real Statistic Resource Pack software was used for the analyses. Table 18 shows the overall agreement and the agreements per rater.

Similarly to the Interclass Correlation, the agency should set the appropriate Kendall's W value to accept or dismiss a rater due to its poor data quality. An option to define an appropriate value could be comparing the Kendall's W values with an existing passing criteria. If a Kendall's W value equal to or greater than 94% is considered acceptable, there would be 83% similarity with the results obtained according to MTC's criterion 1 (Raters 2 and 13 that passed based on MTC's criterion were dismissed and Rater 11 that failed based on the MTC's criterion was accepted). A similar situation occurs for criterion 2 when a Kendall's W value equal to or greater than 97% is considered acceptable, obtaining a 72% similarity with MTC's results (Raters 1 and 5 that failed based on MTC's criterion were accepted and Raters 6, 8, and 9 that have passed based on MTC's criterion were dismissed). A recommendation would be selecting a Kendall's W value with the highest percentage of similarity and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative: accepting all raters with Kendall's W values equal to or greater than 94%. An overall Kendall's W coefficient was also calculated taking into account all the raters. The agreement obtained was 86.25%.

Table 18: Kendall's Coefficient of Concordance for Each Rater and Overall Agreement.

Raters	**Kendall's Coefficient of Concordance**
Rater 1	97.11%
Rater 2	91.30%
Rater 3	94.80%
Rater 4	92.36%
Rater 5	97.57%
Rater 6	94.53%
Rater 7	89.90%
Rater 8	95.90%
Rater 9	94.98%
Rater 10	95.38%
Rater 11	95.65%
Rater 12	91.33%
Rater 13	91.93%
Rater 14	98.08%
Rater 15	96.99%
Rater 16	94.35%
Rater 17	91.34%
Rater 18	95.01%
Overall Agreement	86.25%

Step 4: Bland-Altman Diagram

Bland-Altman diagrams were plotted comparing the ground truth's results with the PCI values of each rater. First, a normality test was performed to the differences of the PCI values between the raters and the ground truth, in order to determine if those differences were normally distributed. The estimated bias and the limits of agreement of the diagram depends on the distribution of the differences of the ratings. The normality tests performed were the Shapiro-Wilk Test and the QQ plot, and the results are found in Appendix I.

Figure 16 shows Bland-Altman diagrams for two raters. The mean of the differences between PCI values of Rater 3 and the ground truth's PCI values is equal to 7.625 (estimated bias). Rater 3 tendency is to overestimate the real PCI values. In the diagram of rater 3, one outlier can be identified.

The mean of the differences between PCI values of Rater 17 and the ground truth's PCI values is equal to –12.417 (estimated bias). Rater 17's tendency is to underestimate the real PCI values. In the diagram of Rater 17, two outliers can be identified. The complete diagrams for all the raters are found in Appendix J.

An ANOVA analysis was applied to the means and differences of the PCI values between the ground truth and the raters. The raters that result in proportional bias are Rater 1, Rater 3, Rater 5, Rater 6, Rater 8, Rater 10, and Rater 12. Appendix K shows the results of the ANOVA analysis. All the statistical analyses of this step were performed using the Real Statistic Resource Pack software.

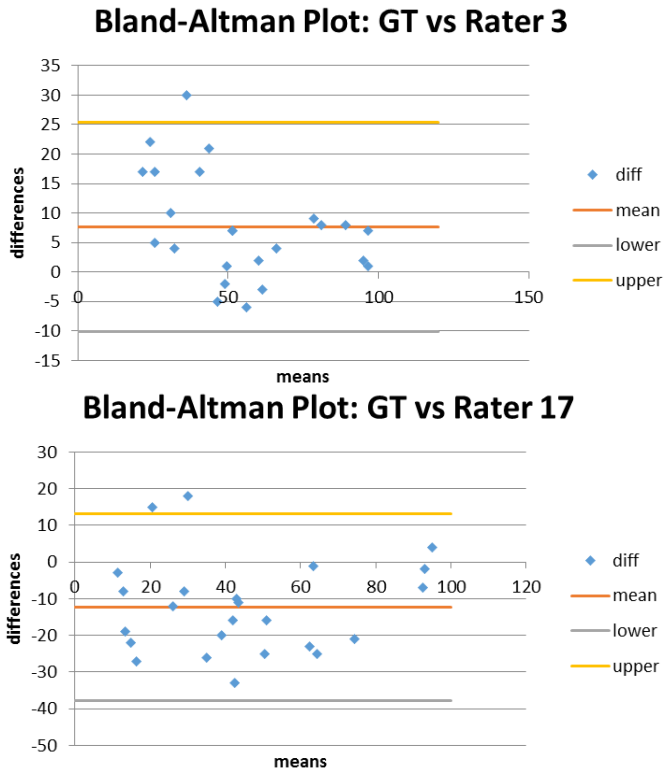


Figure 16: Bland-Altman Diagrams for Rater 3 and Rater 17.

Table 19 shows the estimated bias, the number of outliers, and the existence of proportional bias based on the Bland-Altman diagrams displayed on Appendix J, when comparing ground truth's PCI values against each raters' PCI values.

According to the estimated bias sign (positive or negative), it is possible to identify if a rater has the tendency to underestimate or overestimate the condition of pavement sections. A positive value corresponds to an overestimated tendency and vice versa. 56% of the raters tend to overestimate the condition of pavement sections, and the remaining 44% of the raters tend to underestimate the condition.

Table 19: Estimated Bias, Outliers, and Proportional Bias from the Bland-Altman Plot.

Raters	Estimated Bias	Number of Outliers	Proportional Bias
Rater 1	-7.417	1	yes
Rater 2	-4.250	0	no
Rater 3	7.625	1	yes
Rater 4	13.000	1	no
Rater 5	8.000	2	yes
Rater 6	-2.583	1	yes
Rater 7	-0.250	1	no
Rater 8	4.292	1	yes
Rater 9	4.417	3	no
Rater 10	7.375	1	yes
Rater 11	8.917	1	no
Rater 12	0.250	0	yes
Rater 13	-1.000	1	no
Rater 14	4.875	1	no
Rater 15	0.042	1	no
Rater 16	-0.167	1	no
Rater 17	-12.417	2	no
Rater 18	-2.000	1	no

The outliers provide information regarding the pavement sections in which the rater's PCI value had a considerable difference with the ground truth's result. The difference between the ratings is higher than the limits of agreements defined for each rater's dataset. Those sections can be identified per rater and further analysis can be performed to determine the errors in the quantification of pavement condition (type, extension, and severity level of distresses).

The existence of proportional bias can be detected if the points on the diagram follow a certain tendency, which is visually noticeable. Statistically, an ANOVA analysis can be performed to the mean and the differences of the PCI values, which corresponds to the x-axis and the y-axis of the Bland-Altman plot, respectively.

It is recommended to use the information provided in the Bland-Altman diagram as a complement of the other statistical quality control tools. Once the raters have been accepted or dismissed based on previous results, the Bland-Altman plot would be beneficial to identify raters'

tendency to underestimate or overestimate the condition of a pavement section, detect pavement sections with outliers, and determine the existence of proportional bias.

In Table 20, a summary of the recommended values to accept or dismiss a rater due to its poor data quality for Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Interclass Correlation, and Kendall's Coefficient of Concordance, are displayed. The data of raters with results equal to or greater than the acceptance value, can be considered as good quality data. The acceptance values were determined based on the highest percentage of similarity with the existing MTC's criteria and the lower number of raters accepted after being dismissed by the MTC. This would be the most conservative alternative.

Furthermore, Table 20 shows the raters that had the lowest and highest results for each statistical tool, which represent the raters with the worst and best quality data, respectively. The worst and best raters based on the Bland-Altman diagram results are also included. Percent Agreement, Cohen's Kappa, Interclass Correlation, and Bland-Altman Diagram agreed in considering Rater 4 as que worst rater. Weighted Cohen's Kappa considered Rater 5 as the worst one and Kendall's W considered Rater 7 as the worst rater.

Table 20: Recommended Acceptance Values for Statistical Quality Control Tools.

Statistical Tools	**Acceptance Value**	**Percent Similarity with MTC's results**	**Worst Rater**	**Best Rater**
Percent Agreement	68%	83%	Rater 4	Rater 6 Rater 15
Cohen's Kappa	56%	89%	Rater 4	Rater 15
Weighted Cohen's Kappa	69%	83%	Rater 5	Rater 15
Interclass Correlation	92%	89%	Rater 4	Rater 15
Kendall's Coefficient of Concordance	94%	83%	Rater 7	Rater 14
Bland-Altman Diagram	-	-	Rater 4	Rater 15

Almost all the statistical tools agreed in considering Rater 15 as the best rater. Percent Agreement added Rater 6 as the best one, tied with Rater 15, and Kendall's Coefficient of

Concordance considered Rater 14 as the best rater. As a reference, all the best raters indicated in Table 20 passed MTC's passing criteria, and all the worst raters failed MTC's passing criteria.

In Table 21 a summary of the results of the statistical quality control tools regarding the F-Test, Levene's Test, T-Test, Mann-Whitney Test, Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Interclass Correlation, Kendall's Coefficient of Concordance, and Bland-Altman Diagram output (estimated bias, number of outliers, and proportional bias) are displayed. Overall results were also calculated with three statistical tools. Fleiss' Kappa, Interclass Correlation, and Kendall's Coefficient of Concordance showed overall percentages of agreement of 38.52%, 83.15%, and 86.25%, respectively. MTC's passing criteria can be considered as a reference. Since five out of 18 raters passed both criteria (27.78%), it is recommended to only consider the general result obtained with Fleiss' Kappa.

Table 21: Summary of the Quality Control Tools Results.

Rater	F-Test or Levene's Test	T-Test or Mann-Whitney Test	Percent Agreement	Cohen's Kappa	Weighted Cohen's Kappa	Interclass Correlation	Kendall's Coefficient of Concordance	Bland-Altman Diagram		
	Variances are significantly different?	Means are significantly different?						Estimated Bias	Number of Outliers	Proportional Bias
Rater 1	No	No	62.50%	50.34%	70.81%	92.45%	97.11%	-7.417	1	yes
Rater 2	No	No	54.17%	33.50%	55.70%	86.19%	91.30%	-4.250	0	no
Rater 3	No	No	66.67%	54.07%	65.82%	89.14%	94.80%	7.625	1	yes
Rater 4	No	No	33.33%	7.91%	39.29%	79.38%	92.36%	13.000	1	no
Rater 5	No	No	50.00%	34.55%	34.55%	85.42%	97.57%	8.000	2	yes
Rater 6	No	No	70.83%	58.21%	73.67%	88.52%	94.53%	-2.583	1	yes
Rater 7	No	No	58.33%	42.72%	58.62%	80.12%	89.90%	-0.250	1	no
Rater 8	No	No	54.17%	36.23%	57.83%	91.66%	95.90%	4.292	1	yes
Rater 9	No	No	58.33%	43.79%	63.53%	91.59%	94.98%	4.417	3	no
Rater 10	No	No	54.17%	37.88%	56.50%	86.19%	95.38%	7.375	1	yes
Rater 11	No	No	50.00%	33.64%	57.65%	86.05%	95.65%	8.917	1	no
Rater 12	No	No	58.33%	42.17%	61.54%	84.61%	91.33%	0.250	0	yes
Rater 13	No	No	66.67%	51.52%	67.89%	89.09%	91.93%	-1.000	1	no
Rater 14	No	No	66.67%	56.56%	72.88%	94.83%	98.08%	4.875	1	no
Rater 15	No	No	70.83%	60.28%	75.65%	97.10%	96.99%	0.042	1	no
Rater 16	No	No	54.17%	34.16%	56.15%	90.40%	94.35%	-0.167	1	no
Rater 17	No	No	45.83%	25.89%	54.12%	79.94%	91.34%	-12.417	2	no
Rater 18	No	No	54.17%	36.99%	61.40%	90.18%	95.01%	-2.000	1	no

4.2 Rating Protocol Update Case Study

The first distress identification manuals developed by the Metropolitan Transportation Commission (MTC & ERES Consultants, 1986), were derived from the manual titled "Pavement Maintenance Management for Roads and Parking Lots" developed by the United States Army Corps of Engineers (Shahin & Kohn, 1981). Eighteen years later, this document was adapted and published as the "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys, ASTM D6433-99" (ASTM, 1999).

The ASTM standard of 1999 describes a methodology to quantify pavement condition after performing a visual inspection and registering type, severity level and extent of 19 distresses for both flexible and rigid pavements (ASTM, 1999). In 2009, the standard was modified by separating a distress called "weathering and raveling" into two different distresses, resulting a total of 20 distresses for flexible pavements (ASTM, 2009). Consequently, the MTC distress identification manual for flexible pavements was also updated considering weathering and raveling as two separated distresses (MTC, 2016).

The distresses considered in MTC's manual for flexible pavement condition data collection are: alligator cracking, block cracking, distortions, longitudinal and transverse cracking, patching, rutting and depressions, weathering, and raveling. The PCI values obtained based on these eight distresses are not necessarily the same as the PCI values obtained with seven distresses (when weathering and raveling were considered as one single distress).

To determine the impact of this update, three transportation agencies have collected pavement distress data from three different networks using the two standards. One standard was the MTC's manual in which seven distresses were considered (weathering and raveling together) and the other was the MTC's manual in which eight distresses were considered (weathering separated from raveling). PCI values were calculated based on the data collected using StreetSaver® software. PCI_7 refers to a PCI value obtained based on data collected with seven

distress types and PCI_8 refers to a PCI value obtained based on data collected with eight distress types.

The PCI values of each agency were analyzed using the proposed statistical quality control flowchart presented in Figure 12. The results obtained should reflect the impact of the rating protocol update over the PCI values calculated based on the collected data. Furthermore, final data checks were performed to identify duplicated values, out of range values, null values, and other inconsistencies.

Step 1: Descriptive Statistics

A total of 6,370 PCI values were calculated. Figure 17 displays a scatterplot that shows the relationship between PCI_7 and PCI_8 for each pavement section. Each point represents a PCI value of a particular pavement section calculated considering seven and eight distresses. The regression line of all the data is shown in Figure 17 and its equation and R^2 value are indicated as well. There is a positive correlation between both PCI values. Some dispersion is observed.

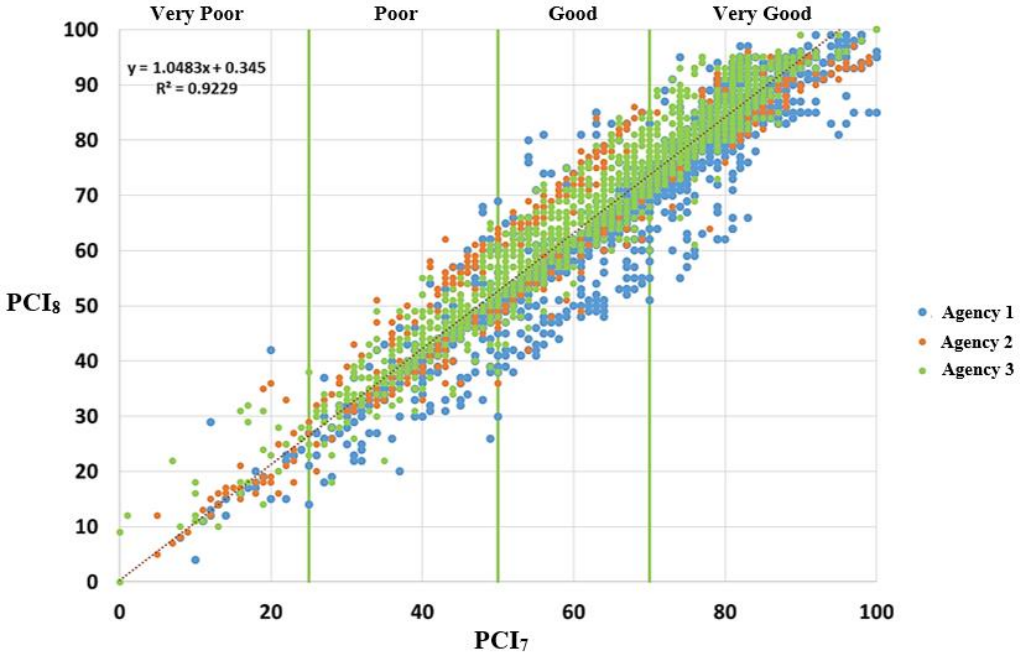


Figure 17: PCI_7 and PCI_8 Values of Pavement Sections Including all the Data.

Descriptive Statistics parameters of PCI_7 and PCI_8 values are displayed in Table 22. The mean and median of PCI values when eight distresses are considered, are higher. Comparing the standard deviations, more dispersion in the data is presented with PCI_8 values.

Table 22: Descriptive Statistics of PCI_7 and PCI_8 Values Including all the Data.

Parameter	PCI_7	PCI_8
Mean	70.28	74.02
Variance	282.51	336.41
Standard Deviation	16.81	18.34
Median	76	79
Minimum	0	4
Maximum	100	100

Pavement sections that had differences between PCI_8 and PCI_7 above 10 points and below -10 points, called outliers, were removed. 795 pavement sections had differences above 10 points and 91 pavement sections had differences below -10 points. A total of 886 outliers (13.9%) were dismissed. The scatterplot of the remaining 5,484 points is displayed in Figure 18 with a R^2 of 90.96%, which indicates a positive correlation between both PCI_7 and PCI_8 values.

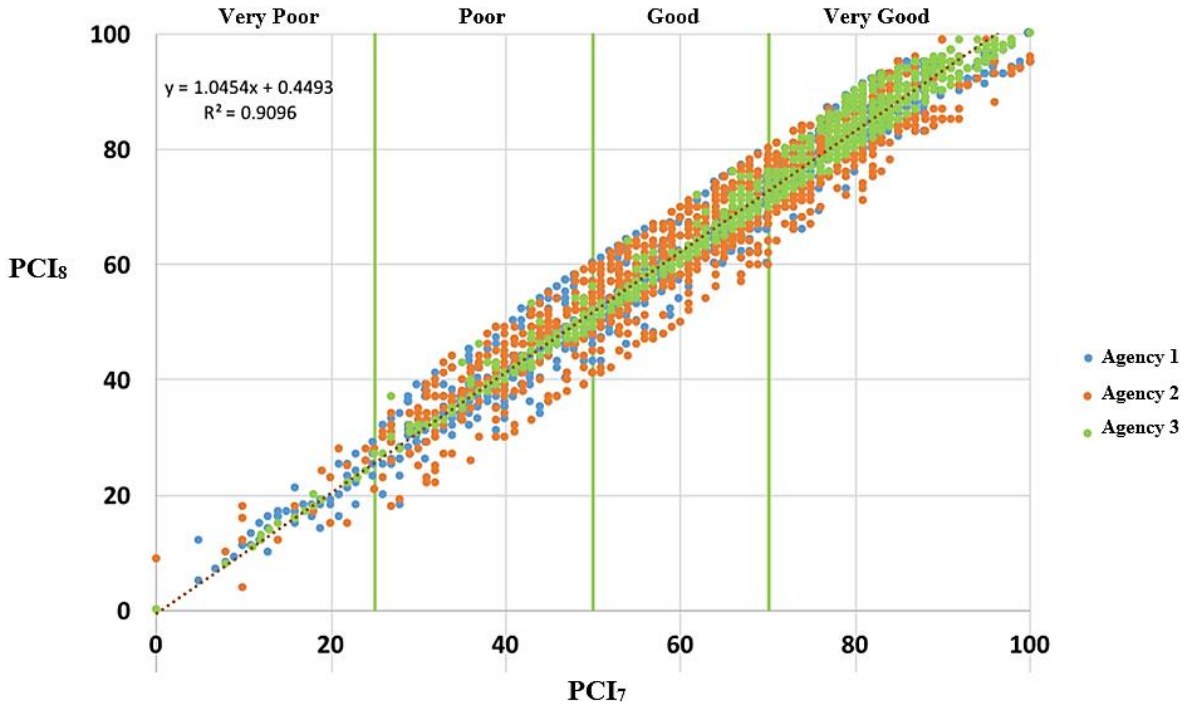


Figure 18: PCI_7 and PCI_8 Values of Pavement Sections Disregarding Outliers.

Descriptive statistics are shown in Table 23. The mean and median of PCI_8 are higher, and also the variance and the standard deviation indicating more dispersion in the data.

Table 23: Descriptive Statistics of PCI_7 and PCI_8 Disregarding Outliers.

Parameter	**PCI_7**	**PCI_8**
Mean	70.14	72.85
Variance	294	332.14
Standard Deviation	17.15	18.22
Median	75	78
Minimum	0	4
Maximum	100	100

The information is organized by agency to determine if the general tendency of the data is maintained for each contractor. Descriptive statistics of PCI_7 values and PCI_8 values are displayed in Table 24. For the three agencies, the mean and median of PCI_8 values are higher than PCI_7 values. The variance and standard deviation of PCI_8 values are also higher, indicating more dispersion in the data. The general tendency is maintained in each agency.

Table 24: Descriptive Statistics of PCI_7 and PCI_8 for Each Agency (outliers are not considered).

Parameters	**Agency 1**		**Agency 2**		**Agency 3**	
	PCI_7	**PCI_8**	**PCI_7**	**PCI_8**	**PCI_7**	**PCI_8**
Mean	70.93	73.72	71.65	72.90	67.90	71.80
Variance	267.24	299.31	311.60	370.03	301.77	334.81
Standard Deviation	16.35	17.30	17.65	19.24	17.37	18.30
Median	76	79	77	79	72	77
Minimum	5	5	8	4	0	9
Maximum	99	98	100	99	100	100

Step 2: Levene's Test

Normality tests were applied to the PCI_7 and PCI_8 values of the three agencies. In Appendix L, the Shapiro-Wilk Test results and the QQ plots for each agency's data are presented. The results evidence that none of the PCI values are normally distributed. Therefore, a Levene's Test is applied to determine if the variances of PCI_8 are significantly different or not from the variances of PCI_7. The resulting p-values were 0.01282, 0.00008, and 0.00869 for agencies 1, 2, and 3, respectively.

All the p-values are lower than 0.05, indicating that the variances of PCI_8 are significantly different from the variances of PCI_7.

Step 3: Mann-Whitney Test

Since none of the PCI values are normally distributed, the Mann-Whitney test was applied to determine if the means of PCI_8 are significantly different or not from the variances of PCI_7. The complete statistical analyses are found in Appendix M. The results indicate that, for all the agencies, the means of PCI_8 are significantly different from the means of PCI_7.

Step 4: Percent Agreement

The Percent Agreement was calculated comparing the PCI_8 and PCI_7 values of each agency, taking into account the exact numerical agreement for each pavement section. In addition, the number of pavement sections with PCI_8 values greater than PCI_7 values, and the number of pavement sections with PCI_8 values lower than PCI_7 values, were determined. Table 25 shows the number of pavement sections with the previously stated relations between PCI_8 and PCI7. The Percent Agreement for agencies 1, 2, and 3 were 14.34%, 4.28%, and 6.30%, respectively. The overall Percent Agreement was 7.87%.

Table 25: PCI_8 versus PCI_7 Values for Individual Sections.

Agency	**Number of pavement sections**	**$PCI_8 > PCI_7$**	**$PCI_8 < PCI_7$**	**$PCI_8 = PCI_7$**
Agency 1	2,078	60.05%	25.61%	14.34%
Agency 2	1,597	87.01%	8.71%	4.28%
Agency 3	1,809	87.40%	6.30%	6.30%
All data	5,484	79.29%	12.84%	7.87%

As shown in Table 23, PCI_8 values tends to be higher than PCI_7 values for all the three datasets. For Agency 2 and Agency 3 the percentage of pavement sections in which PCI_8 values

are greater than PCI_7 values is almost the same value. Agency 1 has the highest number of pavement sections with PCI_8 values equal to PCI_7 values (highest Percent Agreement).

Step 4: Cohen's Kappa

The Cohen's Kappa coefficient was calculated for each agency when comparing PCI_8 categories with PCI_7 categories for each of the pavement sections. The Cohen's Kappa for agencies 1, 2, and 3 were 86.51%, 89.19%, and 79.27%, respectively. The results indicate that the agreement among raters is almost perfect, for Agency 1 and Agency 2, and substantial for Agency 3. The agreement due to chance for all the agencies was low.

Based on this results, additional analysis was made. Transition matrices are used to analyze the migration of pavement sections among condition categories for PCI_8 values. The transition matrices indicate the quantity and the percentage of pavement sections that move from one condition category to another. Figure 19 shows the transition condition matrices for each agency. The diagonals of the matrices indicate the percentage of pavement sections that remain in the same condition category. In all the agencies, the majority of pavement sections remain in the same condition category.

Agency 1		PCI_8			
		Very Good	Good	Poor	Very Poor
		(70-100]	(50-70]	(25-50]	[0-25]
PCI_7	Very Good	99.90%	0.10%	-	-
	Good	14.70%	84.00%	1.30%	-
	Poor	-	14.30%	84.40%	1.30%
	Very Poor	-	-	8.80%	91.20%

Agency 2		PCI_8			
		Very Good	Good	Poor	Very Poor
		(70-100]	(50-70]	(25-50]	[0-25]
PCI_7	Very Good	97.80%	2.20%	-	-
	Good	8.60%	83.30%	8.10%	-
	Poor	-	5.20%	90.10%	4.70%
	Very Poor	-	-	3.70%	96.30%

Agency 3		PCI_8			
		Very Good	Good	Poor	Very Poor
		(70-100]	(50-70]	(25-50]	[0-25]
PCI_7	Very Good	99.80%	0.20%	-	-
	Good	20.20%	77.70%	2.10%	-
	Poor	-	18.70%	80.30%	1.00%
	Very Poor	-	-	20.00%	80.00%

Figure 19: Transition Condition Matrices for each Agency.

Step 4: Weighted Cohen's Kappa

The Weighted Cohen's Kappa was calculated for each agency when comparing PCI_8 categories with PCI_7 categories for each of the pavement sections. The weights assigned for the PCI categories are the same as the ones used in the previous case study (Table 13). The Weighted Cohen's Kappa for agencies 1, 2, and 3 were 91.96%, 93.56%, and 87.57%, respectively. The results indicate that the agreement among raters is almost perfect for the three agencies. This is expected due to the high percentage of pavement sections that stay in the same condition category when comparing PCI_8 and PCI_7 values.

Step 4: Fleiss' Kappa

Fleiss' Kappa analysis was not performed because this tool is only applicable when more than two raters are being evaluated. In this case study, the PCI_8 values are compared to the PCI_7 values. There are two datasets per agency that are compared, therefore, Fleiss' Kappa was not applied.

Step 4: Intraclass Correlation

The Intraclass Correlation coefficient was calculated for each agency when comparing PCI_8 with PCI_7 values for each of the pavement sections. Intraclass Correlation coefficients for agencies 1, 2, and 3 were 97.19%, 97.76%, and 95.92%, respectively. The high level of agreement denote a good performance of the raters that took the measurements.

Step 4: Kendall's Coefficient of Concordance

Kendall's Coefficient of Concordance was not performed because of the great amount of pavement sections analyzed that have to be ranked. It was unfeasible to rank 2,078 pavement sections for Agency 1 for example.

Step 4: Bland-Altman Diagram

Bland-Altman diagrams were plotted comparing the means of PCI_8 and PCI_7 values with the differences between both PCI values (PCI_8 minus PCI_7). Figure 20 shows the Bland-Altman diagrams for the three agencies. The means of the differences between PCI_8 and PCI_7 (estimated bias) for agencies 1, 2, and 3 were 2.79, 1.24, and 3.90, respectively. The tendency of all the agencies is to obtain PCI_8 values greater than PCI_7 values. Several outliers can be identified from the diagrams. An ANOVA analysis was performed to the means and differences of the PCI values. In all the three agencies a proportional bias was identified. Appendix N shows the results of the ANOVA analysis.

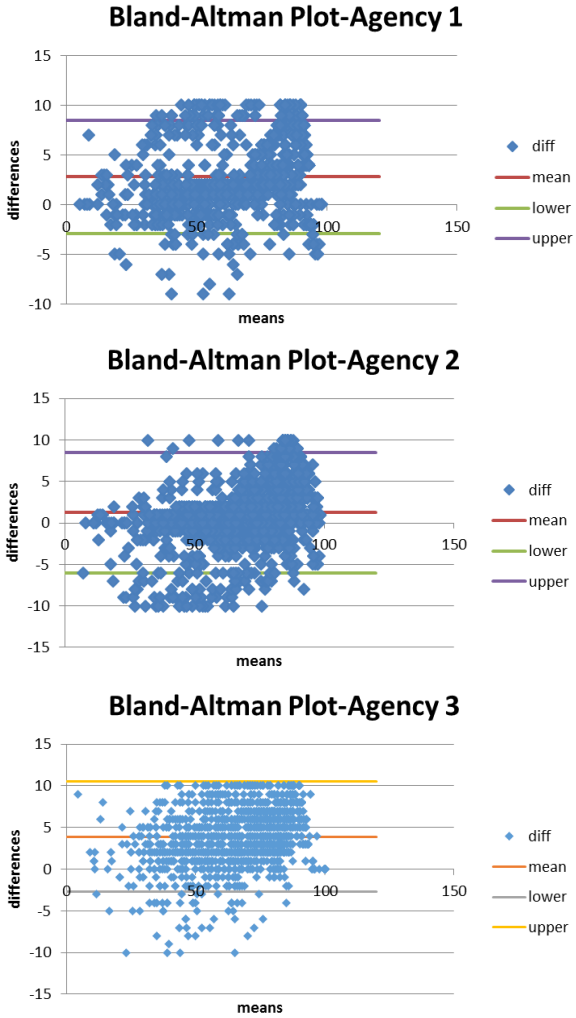


Figure 20: Bland-Altman Diagrams for each Agency.

The results obtained in each statistical quality control tools reflect the impact of considering weathering and raveling as two separated distresses instead of considering weathering and raveling as one single distress. The variances and the means of PCI_8 values were significantly different to PCI_7 values. For the same pavement section, different PCI values are expected if seven or eight distresses are considered.

PCI values calculated when weathering and raveling are considered as separate distresses tend to be higher. Although the majority of the individual pavement sections in the case study remain in the same condition category, the average PCI values are expected to increase by one to four PCI points approximately.

Finally, proportional bias was detected when comparing PCI_8 and PCI_7 in all the agency. This bias might be due to the modification in the rating protocol when eight distresses are considered instead of seven. All the agencies performed the rating for PCI_8 calculation based on eight distresses, different from the previous rating when seven distresses were considered.

Summary of Chapter 4

The proposed quality control statistical tools were applied in two case studies. In case study 1, "Raters Comparison Case Study", 18 raters collected distress data in 24 pavement sections as part of the MTC's certification exam. PCI values were calculated. Statistical quality control tools were used to evaluate the quality of the raters' PCI data while comparing them with ground truth PCI values. Data of each rater was characterized using Descriptive Statistics while calculating the mean, range, standard deviation, and skewness. The F-Test (for normally distributed data) and Levene's Test (for non-normally distributed data) were used to conclude that variances were not significantly different when comparing ground truth's values with raters' values. The T-Test (for normally distributed data) and Mann-Whitney Test (for non-normally distributed data) were used to conclude that means were not significantly different. So far, based on the last four tests, raters' data can be considered as good quality data. Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Interclass Correlation, and Kendall's Coefficient of Concordance were applied

comparing raters' values with ground truth' values. Different percentages were obtained for each rater indicating different levels of agreement with the ground truth. Acceptance values for each tool were recommended to accept or dismissed a rater based on the quality of its data and taking into account percentage similarity with MTC's results. Similarities among statistical tools' results are encountered while identifying best and worst raters. Fleiss' Kappa, Interclass Correlation and Kendall's Coefficient of Concordance were used to calculate an overall agreement of all raters. Fleiss' Kappa was recommended due to its closeness to the percentage of raters that had passed based on MTC's passing criteria. Bland-Altman Diagram was used as a complement to identify raters' tendency to underestimate or overestimate the condition of a pavement section, detect pavement sections with outliers, and determine the existence of proportional bias.

In case study 2, "Rating Protocol Update Case Study", three transportation agencies collected distress data from three different networks using two different rating standards. One standard considered seven distresses (weathering and raveling together), and the updated version considered eight distresses (weathering separated from raveling). Each agency inspected the same pavement sections two times, using both standards. Two datasets were analyzed for each agency. PCI values were calculated based on seven distresses (PCI_7) or eight distresses (PCI_8). Statistical quality control tools were used to evaluate the impact of the rating standard update. Descriptive Statistics (e.g. mean, variance, standard deviation, median, etc.) were calculated to characterize the data. More dispersion in the data is encountered when eight distresses were considered. Pavement sections with differences between PCI_8 and PCI_7 above 10 points and below -10 point were considered outliers and were disregarded. Since all the data was non-normally distributed, Levene's Test and Mann-Whitney Test were used. These tests indicated that the variances and means of PCI_8 were significantly different from the variances and means of PCI_7, respectively. This can be explained due to the difference in standards used for data collection. Based on the low Percent Agreement coefficients obtained, it was possible to conclude that PCI_8 values tends to be higher than PCI_7 values for all the three agencies. The Cohen's Kappa and Weighted Cohen's Kappa tools were used to evaluate the agreement among PCI_8 and PCI_7 values for each agency. It

was concluded that majority of pavement sections remain in the same condition category. Intraclass Correlation coefficients were calculated for each agency. The high level of agreement indicated a good performance of the raters that took the measurements. Finally, Bland-Altman Diagrams were plotted. All the agencies tended to obtain PCI_8 values greater than PCI_7 values due to positive estimated bias values. Several outliers were identified in the diagrams of all the agencies. Finally, proportional bias was detected in all the cases.

Chapter 5: Summary of Research Findings, Conclusions, and Recommendations

5.1 Summary of Research Findings

1. A framework to incorporate quality control in Pavement Management Systems was developed to identify poor quality data and take corrective actions to improve its quality. The framework is composed of ten stages including: policy goals and objectives, pavement inventory, condition assessment, statistical quality control tools, performance modeling, determination of needed work and funds, identification of candidate projects, determination of impact of funding alternatives, budget allocation, and feedback.
2. Quality control practices are incorporated in the framework at the following stages: inventory, condition assessment, statistical quality control tools and performance modeling. The stage corresponding to statistical quality control tools represent a new process proposed for data quality control in the context of pavement management. This stage includes a series of statistical tools organized in a sequential manner to identify poor quality data when comparing pavement condition values with ground truth values. After pavement inventory and condition assessment, two verification steps are considered. If quality control is adequate, the process continues. If quality control is inadequate, corrective actions have to be made, such as re-collecting the data, retraining the raters that have performed the data collection or recalibrating the equipment used in the gathering process.
3. A methodology to include quality control into pavement management practices was developed for the stages of pavement inventory, condition assessment, statistical quality control tools, and performance modeling.

   The quality control methodology for pavement inventory (stage II) includes the following steps:

- Step 2.1. Development of a pavement inventory survey protocol
- Step 2.2. Personnel training
- Step 2.3. Equipment calibration
- Step 2.4. Periodic data checks
- Step 2.5. Definition of software validation rules
- Step 2.6. Logic and missing data checks

The quality control methodology for condition assessment (stage III) includes the following steps:

- Step 3.1: Definition of a pavement condition survey manual and rating protocol
- Step 3.2. Personnel training and certification
- Step 3.3. Equipment calibration
- Step 3.4. Pre-collection checks
- Step 3.5. Data checks at control sites
- Step 3.6. Data checks at verification sites
- Step 3.7. Periodic data checks
- Step 3.8. Missing data checks
- Step 3.9. Final data checks
- Step 3.10. Quality control audits

Statistical quality control tools (stage IV) were proposed to evaluate data quality when comparing pavement condition data with known reference values. The tools are applicable to training programs, certification programs, pre-collection sites, control sites, verification sites, highway networks being assessed or sample audits. Statistical quality control tools can be applied to data gathered manually by raters or automatically with specialized equipment.

The steps consider for data collected manually are:

- Step 4.1. Descriptive Statistics
- Step 4.2. F-Test or Levene's Test

- Step 4.3. T-Test or Mann-Whitney Test
- Step 4.4. Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Fleiss' Kappa, Intraclass Correlation, Kendall's Coefficient of Concordance, Bland-Altman Diagram.

The steps consider for data collected automatically are:

- Step 4.1. Descriptive Statistics
- Step 4.2. F-Test or Levene's Test
- Step 4.3. T-Test or Mann-Whitney Test
- Step 4.4. Percent Agreement, Intraclass Correlation, and Bland-Altman Diagram.

The quality control methodology for performance modeling (stage V) includes the following steps:

- Step 5.1. Perform logic data checks
- Step 5.2. Perform final data checks

5.2 Conclusions from the Case Studies

1. Two case studies were developed to apply the quality control methodology using the proposed statistical quality control tools described in stage IV. The first case study, "Raters Comparison", was related to the personnel training and certification process (step III.2), from the condition assessment stage of the framework (stage III). Pavement condition data from 18 raters were compared to ground truth values. The second case study, "Rating Protocol Update" was related to the final data checks process (step III.9), from the condition assessment stage of the framework (stage III). Three transportation agencies collected pavement condition data from three different pavement networks using two distress rating protocols. The proposed statistical quality control methodology was successfully applied to both cases, showing a favorable contribution to the identification of data quality problems.

2. It is concluded from the first case study that statistical quality control tools can be used to accept or dismiss raters based on the quality of their data. Quality is measured in terms of the agreement between raters' data and the reference values or ground truth values.
    - Descriptive statistics provide a general idea of the characteristics of the data regarding central tendency, variability, and distribution shape.
    - The tests for equality of variances and means can identify poor quality data if the variance and the mean of a rater's data are significantly different from the variance and mean of ground truth's data. These tests are indicators of data quality issues. If the variance and mean were not significantly different, the rater's data could be considered as good quality data, based on variances and means comparisons. However, further statistical analysis is recommended to identify data quality issues based on the level of agreement between raters and the ground truth.
    - Fleiss' Kappa, Interclass Correlation, and Kendall's' Coefficient of Concordance provide an overall agreement coefficient of all the raters evaluated and compared to the ground truth. Fleiss' Kappa overall agreement is recommended for data quality evaluation. The other two tools tend to provide higher values of agreement that might not correspond to the data under evaluation.
    - Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Interclass Correlation, and Kendall's Coefficient of Concordance are statistical tools that evaluate the level of agreement between raters and the ground truth. Acceptance values for each tool could be determined to accept or dismissed a rater based on the quality of its data. It is recommended to consider previous passing criteria as a reference to define the acceptance values.
    - Bland-Altman diagram represents a graphical tool to identify raters that tends to overestimate or underestimate the ratings, detect pavement sections with outliers, and determine the existence of proportional bias. Once the raters have been

accepted or dismissed, it is recommended to use the information provided by the Bland-Altman diagram as a complement.

3. It is concluded from the second case study that the statistical quality control tools can capture the differences in the condition assessment results when using an updated version of the PCI distress protocol (weathering and raveling definitions and deduct PCI curves).
    - A positive correlation was evidenced between PCI values calculated with eight distresses (PCI_8) and PCI values calculated with seven distresses (PCI_7). More dispersion in the data was also observed when using eight distresses.
    - The tests for equality of variances and means found that these two datasets were significantly different, at 95% confidence level, when comparing PCI_8 with PCI_7.
    - The maximum Percent Agreement obtained was 14.34% for Agency 1. The majority of pavement sections had different PCI values when PCI_8 was compared to PCI_7. The PCI_8 values tends to be higher than PCI_7 values for all the three datasets.
    - The Cohen's Kappa, Weighted Cohen's Kappa, and Intraclass Correlation coefficients resulted in high values for the three agencies. This tendency may be due to a large number of pavement sections classified in the same condition category with PCI_7 and PCI_8.
    - From the Bland-Altman diagrams per agency it is observed that all the raters reported PCI_8 values greater than PCI_7 values. In addition, proportional bias was detected for each of the agency's datasets. This bias may be due to the updated protocol used in the collection of the data.
4. The statistical quality control tools recommended for quality evaluation of data collected manually are Percent Agreement, Cohen's Kappa, Weighted Cohen's Kappa, Interclass Correlation, and Kendall's Coefficient of Concordance because of the similar results found in the analysis. These tools estimate the agreement for each individual rater compared to a ground truth. For an overall agreement of all raters, Fleiss' Kappa is recommended.

5. The statistical quality control tools recommended for quality evaluation of data collected automatically are Percent Agreement and Intraclass Correlation. For an overall agreement, Interclass Correlation is recommended.
6. Bland-Altman Diagram is recommended as a graphical tool for quality control since it includes estimated bias, outliers, and existence of proportional bias.

Quality control practices are traditionally applied to data collection processes regarding pavement inventory and condition assessment. However, quality control is not limited to those stages and it can also be used at other stages. Quality control must be implemented for logic and data validation checks of pavement performance and budget estimates.

5.3 Recommendations for Future Research

1. A methodology that includes quality control into asset management practices is needed, not only focusing on pavements. Other infrastructure (e.g. bridges, tunnels, airports, channels, guardrails, etc.) can be considered for data quality management to detect poor quality data that will influence managerial decisions.
2. A quality index could be calculated based on the results of the statistical quality control tools proposed in the methodology. This index might integrate all the statistical tools' output to express data quality as one single value. Different criteria could be set, depending on the transportation agencies, to accept or dismiss a rater or an equipment measurement.
3. The proposed quality control methodology could be applied to more case studies, for instance, cases in which automated equipment had been used to collect the data. Descriptive statistics, test for equality of variances and means, Percent Agreement, Intraclass Correlation, and Bland-Altman Diagram could be applied to evaluate the quality of the data gathered by automated equipment, such as profilometers.
4. The statistical quality control tools considered in the framework can be replaced by mathematical models, data analysis techniques, quality approaches, and softwares applicable to quality management. Further research is needed to adequate artificial neural

networks, fuzzy logic, Bayesian methods, big data analysis techniques, data mining, six sigma, total quality management, expert systems, genetic algorithms, etc. to the quality control practices required for pavement condition data.

References

Abdallah, I., Melchor, O., Ferregut, C., and Nazarian, S. (2000). "Artificial Neural Network Models for Assessing Remaining Life of Flexible Pavements". Texas Department of Transportation.

Akofio-Sowah, M. and Amekudzi-Kennedy, A. (2016). "Identifying Factors to Improve Transportation Asset Management Program Sustainment". Transportation Research Record: Journal of the Transportation Research Board, No. 2593, Washington, D.C., pp. 1–7.

Allen, B. and Kathawala, Y. (1992). "Expert Systems Applications in Quality Management: An Integrative Approach". ISA Transactions®. Artificial Intelligence.

Al-Zou'Bi, M. M. (2003). "A Systematic Approach to Manage Missing Data in Pavement Management Systems". Doctoral Thesis: University of Texas at El Paso.

Al-Zou'Bi, M. M., Chang, C. M., Nazarian, S., and Kreinovich, V. (2015). "Systematic Statistical Approach to Populate Missing Performance Data in Pavement Management Systems". Journal of Infrastructure Systems, 21(4), 04015002.

American Association of State Highway and Transportation Officials (AASHTO). (1990). "Guidelines on Pavement Management". Washington, D.C., 45 pp.

American Association of State Highway and Transportation Officials (AASHTO). (1993). "Guide for Design of Pavement Structures". Washington, D.C., 624 pp.

American Association of State Highway and Transportation Officials (AASHTO). (2001a). "Pavement Management Guide. Executive Summary Report". Washington, D.C., 254 pp.

American Association of State Highway and Transportation Officials (AASHTO). (2001b). "Pavement Management Guide". First Edition. Washington, D.C., 254 pp.

American Association of State Highway and Transportation Officials (AASHTO). (2012). "Pavement Management Guide". Second Edition. Washington, D.C., 161 pp.

American Association of State Highway and Transportation Officials (AASHTO). (2011). "Transportation Asset Management Guide: A Focus on Implementation". Washington, D.C., 48pp.

American Society for Testing and Materials (ASTM). (1999). "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys". ASTM D6433-99.

American Society for Testing and Materials (ASTM). (2009). "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys". ASTM D6433-09.

American Society for Testing and Materials. (ASTM). (2016). "Standard Guide for Classification of Automated Pavement Condition Survey Equipment". ASTM E1656/E1656M.

American Society for Testing and Materials (ASTM). (2018). "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys". ASTM D6433-18.

Athanasiadis, I.N., Rizzoli, A. E., and Beard, D.W. (2010). "Data Mining Methods for Quality Assurance in an Environmental Monitoring Network". Artificial Neural Networks. ICANN 2010. Lecture Notes in Computer Science, vol. 6354. Springer, Berlin, Heidelberg.

Attoh-Okine, N. O. (1994). “Predicting roughness progression in flexible pavements using artificial neural networks”. Third International Conference on Managing Pavements, San Antonio, TX, 1994.

Attoh-Okine, N. O. (1997). “Rough set application to data-mining principles in pavement management database”. Journal of Computing in Civil Engineering. American Society of Civil Engineers (ASCE). 11 (4) 231-237.

Banan, M. R., and Hjelmstad, K. D. (1996). “Neural networks and AASHO road test”. Journal of Transportation Engineering, 122 (5), 358–66.

Bender, R. and Lange, S. (2007). “What is a confidence interval?” Deutsche Medizinische Wochenschrift. 132 Suppl. 1:e17-8.

Bennett, C.R., Chammoro, A., Chen, C., Solminihac, H., and Flintsch, G. W. (2005). “Data Collection Technologies for Road Management”. The World Bank, Washington, D.C., 124 pp.

Benvenuto, F. and Marani, A. (2001). “Neural Networks for Environmental Problems: Data Quality Control and Air Pollution Nowcasting”. Global Nest: the Int. J. Vol 2, No 3, pp 281-292.

Cabena, P. (1997). “Discovering data mining: From concept to implementation”. Prentice Hall, NJ.

Cai, L. and Zhu, Y. (2015). “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era”. Data Science Journal, 14: 2, pp. 1-10.

Caro, A., Calero, C., Sahraoui, H., Malak, G., and Piattini, M. (2007). “A Bayesian Network to Structure a Data Quality Model for Web Portals”. International Journal of Information Quality 1(3):272-294.

Chang A., C. M. (2007). “Development of a Multi-Objective Strategic Management Approach to Improve Decisions for Pavement Management Practices in Local Agencies”. Doctoral Thesis: Texas A&M University.

Chang A., C. M. (2016). “Infrastructure Management”. [Class handout]. The University of Texas at El Paso. El Paso, Texas. College of Engineering. CE 6301.

Chang-Albitres, C. M., Smith, R. E., and Pendleton, O. J. (2007). “Comparison of Automated Pavement Distress Data Collection Procedures for Local Agencies in San Francisco Bay Area, California,” Transportation Research Record: Journal of the Transportation Research Board, No. 1990, Washington, D.C., pp. 119–126.

Chang-Albitres, C. M., Saenz, D., Nazarian, S., Abdallah, I. N., Wimsatt, A., Freeman, T., and Fernando, E. G. (2013). “Improvements in Pavement Ride, Distress and Condition Based on Different Treatment Types”. Technical Report TX 0-6673-R1. Texas Department of Transportation. 90 pp.

Chang, C. M., Nazarian, S., Vavrova, M., Yapp, M. T., Pierce, L. M., Robert, W., and Smith, R. E. (2017). “Consequences of Delayed Maintenance of Highway Assets”. National Cooperative Highway Research Program (NCHRP). Research Report 859.

Chen, C. and Flintsch, G, W. (2007). “Fuzzy Logic Pavement Maintenance and Rehabilitation Triggering Approach for Probabilistic Life-Cycle Cost Analysis”. Transportation Research Record: Journal of the Transportation Research Board, No. 1990, Washington, D.C., 2007, pp. 80–91.

Crossfield, R. T. and Dale, B. G. (1991). “The use of expert systems in total quality management: An exploratory study”. Quality and Reliability Engineering International, v.7, no.1, pp. 19-26.

Dai, W., Yoshigoe, K. and Parsley, W. (2018). “Improving Data Quality through Deep Learning and Statistical Models”. Advances in Intelligent Systems and Computing.

Das, S. and Saha, B. (2009). “Data Quality Mining using Genetic Algorithm”. International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (2), pp. 105-112.

Das, K. R. and Imon, A. H. M. R. (2016). “A Brief Review of Tests for Normality”. American Journal of Theoretical and Applied Statistics. Vol. 5, No. 1, 2016, pp. 5-12.

DeSarbo, W. S., and Rao, V. R. (1986). "A constrained unfolding model for product positioning analysis." (Marketing Science). 5, 1-19.

Du Prel, J., Hommel, G., Röhrig, B., and Blettner, M. (2009). “Confidence Interval or P-Value?” Deutsches Ärzteblatt International. Part 4 of a Series on Evaluation of Scientific Publications; 106(19): 335–339.

Ecola, L. and Wachs, M. (2012). “Exploring the Relationship between Travel Demand and Economic Growth”. Federal Highway Administration (FHWA).

Eldin, N. N. and Senouci, A. B. (1995). “A pavement condition rating model using backpropagation neural networks”. Microcomputers in Civil Engineering, 10 (6), 433–41.

Federal Highway Administration (FHWA). (1999). “Asset Management Primer”. U.S. Department of Transportation. Office of Asset Management.

Federal Highway Administration (FHWA). (2007). “Asset Management Overview”. FHWA-IF-08-008. U.S. Department of Transportation. Office of Asset Management.

Federal Highway Administration (FHWA). (2007). “Relationships between Asset Management and Travel Demand: Findings and Recommendations from Four State DOT Site Visits”. U.S. Department of Transportation. Office of Asset Management.

Federal Highway Administration (FHWA). (2010). “Beyond the Short Term. Transportation Asset Management for Long-Term Sustainability, Accountability and Performance”. U.S. Department of Transportation. Office of Asset Management.

Federal Highway Administration (FHWA). (2012a). “Moving Ahead for Progress in the 21st Century Act
(MAP-21): A Summary of Highway Provisions”. Office of Policy and Governmental Affairs.

Federal Highway Administration (FHWA). (2012b). “Advancing a Transportation Asset Management Approach”. U.S. Department of Transportation. Office of Asset Management.

Federal Highway Administration (FHWA). (2014). “Distress Identification Manual for the Long-Term Pavement Performance Program”. FHWA-HRT-13-092. U.S. Department of Transportation.

Federal Highway Administration (FHWA). (2016). “Fixing America’s Surface Transportation Act (FAST Act): A Summary of Highway Provisions”. Office of Policy and Governmental Affairs.

Flintsch, G. W. and Bryant, J. W. (2006). “Asset Management Data Collection for Supporting Decision Processes”. U.S. Department of Transportation. Federal Highway Administration, Washington, D.C., 97 pp.

Flintsch, G.W. and Bryant, J. W. (2008). “Asset Management Data Collection for Supporting Decision Processes”. U.S. Department of Transportation. Federal Highway Administration, Washington, D.C., 95 pp.

Flintsch, G. and McGhee, K. K. (2009). “NCHRP Synthesis 401: Quality Management of Pavement Condition Data Collection”. Transportation Research Board, Washington, DC.

Freeman, J.A. and Skapura, M. D. (1991), “Neural Networks: Algorithms, Applications, and Programming Techniques”, Addison-Wesley Publishers Company, Massachusetts.

Fwa, T. F. and Shanmugam, R. (1998). “Fuzzy Logic Technique for Pavement Condition Rating and Maintenance-Needs Assessment”. 4th International Conference on Managing Pavements.

Fwa, T. F., Chan W. T., and Hoque, K. Z. (1998a). “Analysis of Pavement Management Activities Programming by Genetic Algorithms”. Transportation Research Record 1643, Paper No. 98-0019.

Fwa, T. F., Chan W. T., and Hoque, K. Z. (1998b). “Network Level Programming for Pavement Management Using Genetic Algorithms”. 4th International Conference on Managing Pavements.

Fwa, T. F., Chan, W. T. and Tan, C. Y. (1996). “Genetic algorithm programming of road maintenance and rehabilitation”. Journal of Transportation Engineering, 122 (3), 246–53.

Gallegos, A. (2012). “Calibration of Concrete Pavement Performance Models”. Master in Science Thesis: The University of Texas at El Paso.

Gallegos, A., Chang-Albitres, C. M., and Nazarian, S. (2013). “Hybrid Technique for Calibrating Network-Level Performance Models of Continuously Reinforced Concrete Pavements”. Journal of Transportation Engineering Vol. 139, Issue 12.

Gharaibeh, N., Freeman, T., Saliminejad, S., Chang-Albitres, C., Weissman, J., Weissman, A., Wimsatt, A., and Gurganus, C. (2012). “Evaluation and Development of Pavement Scores, Performance Models and Needs Estimates for the TxDOT Pavement Management Information System–Final Report”. Technical Report TX 0-6386-3. Texas Department of Transportation. 304 pp.

Gregory, D. C., Shahin, M. Y., Burkhalter, J. A. (2003). "Automated Data Collection for Pavement Condition Index Survey". Transportation Research Board (TRB).

Grivas, D. A. and Shen, Y. C. (1995). “Use of fuzzy relations to manage decisions in preserving civil infrastructure”. Transportation Research Record, 10–26.

Haas, R., Hudson, W. R., and Zaniewski, J. P. (1994). “Modern Pavement Management”. Krieger Publishing Company, Malabar, Fla.

Haas, R. (2003). “Good technical foundations are essential for successful pavement management”. Guimaraes, Portugal: Haas, Ralph, 2003. Good technical foundations are essential for proceedings of MAIREPAV’ 03.

Hallin, J., McGhee, K., and Schwartz, C.W. (2004). “NCHRP Project 1-37A: Guide for Mechanistic-Empirical Design of New and Rehabilitated Pavement Structures”. Transportation Research Board, Washington, DC.

Heller, K. A. (2007). “Efficient Bayesian Methods for Clustering”. Doctoral Thesis: University of London.

Hipp, J., Güntzer, U., and Grimmer, U. (2001). “Data Quality Mining – Making a Virtue of Necessity”. Data Mining and Knowledge Discovery, Santa Barbara, CA.

Huang, Y. and Moore, R. K. (1997). “Roughness level probability prediction using artificial neural networks”. Transportation Research Record, 89–97.

Huang, Y. H. (2004). “Pavement analysis and design”. Second Edition. Pearson Prentice Hall.

International Organization for Standardization (ISO). (2005). “Quality Management Systems—Fundamentals and Vocabulary”. ISO 9000. Geneva, Switzerland.

International Organization for Standardization (ISO). (2008). “International Vocabulary of Metrology”. International Organization for Standardization, Geneva, Switzerland.

Irwin, L. H., Orr, D. P., and Atkins, D. (2011). “Falling Weight Deflectometer Calibration Center and Operational Improvements: Redevelopment of the Calibration Protocol and Equipment”. Federal Highway Administration (FHWA). Final Report, 268 pp.

Janta-Polczyhki, M. and Roventa, E. (1999). “Fuzzy Measures for Data Quality”. 18th International Conference of the North American Fuzzy Information Processing Society - NAFIPS (Cat. No.99TH8397).

Karan, M. A. (1979). “A System for Priority Programming of investments for Road Network Improvements”. Waterloo: University of Waterloo, Waterloo, Canada.

Kaseko, M. S., Lo, Z. P., and Ritchie, S. G. (1994). “Comparison of traditional and neural classifiers for pavement-crack detection”. ASCE Journal of Transportation Engineering, 120 (4), 552–69.

Kobayashi, K. and Kiyoyuki, K. (2017). “Big data-based deterioration prediction models and infrastructure management: towards assetmetrics”. Structure and Infrastructure Engineering. 13:1, 84-93, DOI: 10.1080/15732479.2016.1198407.

Kotb, A. S. and Moore, C. J. (1996). “A decision support expert system for maintenance and rehabilitation needs”. World Transport Research, Proceedings of the 7th World Conference, Vol. 4: Transport Management, 1996, pp. 397–410.

Kuncheria, P. I. and Veeragavana, A. (1996). “PADMA: Expert system for flexible pavement deterioration and maintenance”. Indian Highways, 24 (9), 27–47.

Kwasi, A. A. and Attoh-Okine, N. O. (1999). “Effects of cross probabilities in evolutionary algorithm trained to predict pavement performance”. 78th Transportation Research Board Annual Meeting, Washington.

La Torre, F., Domenichini, L., and Darter, M. I. (1998). “Roughness prediction model based on the artificial neural network approach”. Fourth International Conference on Managing Pavements, pp. 599–612.

Lee, H. and Galdiero, V. (1989). “Enhancing the existing pavement management systems by using expert system techniques”. First International Conference on Applications of Advanced Technologies in Transportation Engineering, San Diego, CA, February 1989, pp. 76–81.

Pierce, L. M., McGovern, G., and Zimmerman, K. A. (2013). “Practical Guide for Quality Management of Pavement Condition Data Collection”. Federal Highway Administration (FHWA).

Liu, L. and Gharaibeh, N. (2013). “Bayesian Model for Predicting the Performance of Pavements Treated with Thin HMA Overlays”. Transportation Research Board 2014 Annual Meeting.

Lytton, R. L. (1987). “Concepts of pavement performance prediction and modeling”. Proc., 2nd North American Pavement Management Conf., Vol. 2, Canada, 2.4–2.19.

Martin, L. (1993). “Total Quality Management in the Public Sector”. National Productivity Review, Vol. 10, pp. 195–213.

Matía, F., Aguilar-Crespo, J. A., Jiménez, A., Sanz, R., and Domínguez, J. M. (1995). “Fuzzy Logic and Data Quality in Real-time Systems”. Integrated Computer-Aided Engineering - Special issue: real-time intelligent control systems archive. Volume 2 Issue 3, 1995. Pages 229 – 239.

McGhee, K. (2004). “Automated Pavement Distress Collection Techniques”. NCHRP Synthesis of Highway Practice 334. Transportation Research Board, Washington, DC.

McHugh, M. L. (2012). “Interrater reliability: the kappa statistic”. Biochemia Medica, 276-282.

McQueen, J.M. and Timm, D. H. (2005). “Statistical Analysis of Automated Versus Manual Pavement Condition Surveys”. Transportation Research Record: Journal of the Transportation Research Board, No. 2004, Transportation Research Board of the National Academies, Washington, D.C., pp. 55–62.

Merino, J., Caballero, I., Rivas, B., Serrano, M., and Piattini, M. (2015). “A Data Quality in Use model for Big Data”. Future Generation Computer Systems 63 (2016) 123–130.

Metropolitan Transportation Commission (MTC). (2016a). “Pavement Condition Index Distress Identification Manual for Flexible Pavements”. Fourth Edition. Metropolitan Transportation Commission, Oakland, CA.

Metropolitan Transportation Commission (MTC). (2016b). “Pavement Condition Index Distress Identification Manual for Rigid Pavements”. Third Edition. Metropolitan Transportation Commission, Oakland, CA.

Metropolitan Transportation Commission (MTC) and ERES Consultants, INC. (1986). “Pavement Condition Index Distress Identification Manual for Asphalt and Surface Treatments Pavements”. Oakland, CA.

Minnesota Department of Transportation (MnDOT). (2017). “MnDOT Inertial Profiler Certification Program”. Office of Materials and Road Research, Maplewood, MN, USA.

Montgomery, D. C. (2013). “Introduction to Statistical Quality Control”. Seventh Edition. John Wiley and Sons, Inc. ISBN: 978-1-118-14681-1.

Morian, D., Stoffels, S, and Frith, D. J. (2002). “Quality Management of Pavement Performance Data”. Pavement Evaluation Conference, Roanoke, Virginia, USA.

Muench, S. T., Mahoney, J. P., Pierce, L. M. (2003). “Pavement Management”. Washington Department of Transportation Pavement Guide. Washington D.C.

Nassar, K. (2007). “Application of Data-Mining to State Transportation Agencies’ Projects Databases”. Journal of Information Technology in Construction. ITcon Vol 12. Pg. 139.

National Cooperative Highway Research Program (NCHRP). (2006). “Performance Measures and Targets for Transportation Asset Management”. Transportation Research Board. Report 551.

Neural Ware Inc. (1993), “Neural Computing”, Technical Publications Group, Pittsburgh, PA.

New Hampshire Department of Transportation (NHDOT). (2014). “Transportation Asset Management Implementation Plan”. Concord, NH.

Office of Energy Efficiency and Renewable Energy (EERE). (2018). “Gross Domestic Product Continues to Outpace Vehicle Miles Traveled”. FOTW #1023. URL: https://www.energy.gov/eere/vehicles/articles/fotw-1023-april-2-2018-gross-domestic-product-continues-outpace-vehicle-miles. Date accessed April 09, 2018.

Oregon Department of Transportation (ODOT). (2011). “Asset Management Strategic Plan”. Asset Management Integration, Oregon, USA.

Paladini, E. P. (2000). “An expert system approach to quality control”. Expert Systems with Applications 18 (2000) 133–151.

Park, E. S., Smith, R. E., Freeman, T. J, and Spiegelman, C. H. (2008). “A Bayesian approach for improved pavement performance prediction”. Journal of Applied Statistics. Vol. 35, No. 11, November 2008, 1219–1238.

Pateras, K. (2013). "Bayesian Inference and Variable Selection in Normal and Binomial Regression Models with Applications in Medical Research". Master in Science thesis: National and Kapodistrian University of Athens.

Poister, T. H. and Harris, R. H. (1996). “Service Delivery Impacts of TQM. A Preliminary Investigation”. Public Productivity & Management Review, Vol. 20, No. 1 (Sep., 1996), pp. 84-100.

Ponniah, J., Sharma, B. N., and Kazmierowski, T. J. (2001). “A Critical Review of an Existing Pavement Condition Rating System”. Fifth International Conference on Managing Pavements, Transportation Research Board, Seattle, Wash., 15 pp.

Queiroz, C. (1983). “A Mechanistic Analysis of Asphalt Pavement Performance in Brazil,” Asphalt Paving. Volume 52.

Rada, G.R., Simpson A. L., and Hunt, J. E. (2004). "Collecting and Interpreting Long-Term Pavement Performance Photographic Distress Data: Quality Control–Quality Assurance Processes". Transportation Research Record: Journal of the Transportation Research Board No. 1889, Washington, D.C., pp. 97–105.

Ramirez, R. (2015). "A Stochastic Approach for Pavement Condition Projections and Budget Needs for the MTC Pavement Management System". Doctoral Thesis: University of Texas at El Paso.

Rauhut, J. B., and Gendel, D. S. (1987). "Proposed Development of Pavement Performance Prediction Models from SHRP/LTPP Data." 2nd North American Pavement Management Conference.

Reid, R. D. and Sanders, N. R. (2012). "Operations Management: An Integrated Approach". 5th Edition. John Wiley & Sons Inc.

Roberts, C. A. and Attoh-Okine, N. O. (1996). "A comparative analyses of two artificial neural networks using pavement performance prediction". Conference Proceedings, Neural Network Applications in Highway and Vehicle Engineering, pp. 59–73

Saba, R.G. (2007). "Performance Prediction Models for Flexible Pavements." Nordic Road and Transport Research No. 1, Norway.

Saliminejad, S. and Gharaibeh, N. (2013). "Impact of Error in Pavement Condition Data on Output of

Network-Level Pavement Management Systems". TRB Paper No. 13-4466. The 2013 Transportation Research Board Meeting. Washington D.C.

Selezneva, O. I., Mladenovic, G., Speir, R., Amenta, J., and Kennedy, J. (2008). "National Park Service Road Inventory Program—Quality Assurance Sampling Considerations for Automated Collection and Processing of Distress Data". Transportation Research Record: Journal of the Transportation Research Board, No. 1889, Washington, D.C., pp. 106–115.

Serigos, P. A. (2015). "Bayesian Hierarchical Modelling of Pavement Performance". Master of Science thesis: The University of Texas at Austin.

Sessions, V. and Valtorta, M. (2006). "The Effects of Data Quality on Machine Learning Algorithms". Proceedings of the 11th International Conference on Information Quality, MIT, Cambridge, MA, USA, November 10-12, 2006.

Shahin, M. Y., and Kohn, S. D. (1981). "Pavement Maintenance Management for Roads and Parking Lots". United States Army Corps of Engineers. Technical Report M-294.

Shahriar, S. and Anam, S. (2008). "Quality Data for Data Mining and Data Mining for Quality Data: A Constraint Based Approach in XML". Second International Conference on Future Generation Communication and Networking Symposia.

Shekharan, R., Frith, D., Chowdhury, T., Larson, C., and Morian, D. (2006). "The Effects of a Comprehensive QA/QC Plan on Pavement Management". Transportation Research Board, Washington, DC.

Shekharan, R., Frith, D., Chowdhury, T., Larson, C., and Morian, D. (2007). “Effects of Comprehensive Quality Assurance/Quality Control Plan on Pavement Management”. Transportation Research Record: Journal of the Transportation Research Board, No. 1990, Transportation Research Board of the National Academies, Washington, D.C., pp. 65–71.

Shweta, B., Himanshu, K. C., and Ram, B. (2015). “Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods”. Journal of the Indian Academy of Applied Psychology. 9pp.

Simpson, A., Rada, G., Bryce, J., Serigos, P., Visintine, B., and Groeger, J. (2018). “Interstate Highway Pavement Sampling Data Quality Management Plan”. Federal Highway Administration (FHWA). U.S. Department of Transportation.

Smith, M., (1993), “Neural Networks for Statistical Modeling”, Van Nostrand Reinhold, 115 Fifth Ave., New York, NY, 10003.

StreetSaver Academy. (2018). “MTC Rater Certification Exam”. Retrieved from https://www.streetsaver.com/academy/academy-rater-certification

Summers, D. (2006). “Quality”. 4th ed., Prentice Hall, Upper Saddle River, N.J.

Sundin, S. and Braban-Ledoux, C. (2001) “Artificial Intelligence−Based Decision Support Technologies in Pavement Management”. Computer-Aided Civil and Infrastructure Engineering. Pp 143-157.

Sundquist, E. and McCahill, C. (2015). “For the first time in a decade, U.S. per capita highway travel ticks up”. State Smart Transportation initiative.

Tan, S., and Cheng, D. (2014). “Improving Data Quality for Pavement Management System”. 9th International Conference on Managing Pavement Assets. Washington, D.C.

Tan, S., and Cheng, D. (2017). “Pavement Condition Data Quality Verification Methods for Pavement Management System”. World Conference on Pavement and Asset Management WCPAM 2017. Milan, Italy. June 12/16, 2017.

Taware, S. N. and Kolhe, V. (2014). “A System to Improve Data Quality in healthcare using Naïve Bayes Classifier”. International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5206-5209.

The Data Warehousing Institute. (2002). <http://www.twdi.org>.

Thompson, P. D., Ford, K. M., Arman, M. H. R., Labi, S., Sinha, K. C., and Shirole, A. M. (2012). “Estimating Life Expectancies of Highway Assets”. NCHRP Report 713. Volume 1: Guidebook. Transportation Research Board of the National Academies, Washington, D.C.

Transportation Research Board (TRB). (2002). “Transportation Research Circular E-C037: Glossary of Highway Quality Assurance Terms”. National Research Council, Washington, D.C., 30 pp.

Tsikriktsis, N. (2005). "A review of techniques for treating missing data in OM survey research." (Journal of Operations Management) 24 (2005) 53–62.

Uddin, W., Hudson, W. R., and Haas, R. (2013). “Public infrastructure asset management”. New York: McGraw-Hill Education.

Viera, A. and Garrett, J. (2005). “Understanding Interobserver Agreement: The Kappa Statistic”. Family Medicine. Vol. 37, No. 5, pp. 360-363.

Vizhi, J. M. and Bhuvaneswari, T. (2012). “Data Quality Measurement with Threshold Using Genetic Algorithm”. International Journal of Engineering Research and Applications (IJERA). Vol. 2, Issue4, July-August 2012, pp.1197-1203

Watson, P.F. and Petrie, A. (2010). “Method agreement analysis: A review of correct methodology”. Theriogenology 73, 1167–1179, 13 pp.

Wee, S. and Kim, N. (2006). “Angular Fuzzy Logic Application for Pavement Maintenance and Rehabilitation Strategy in Ohio”. KSCE Journal of Civil Engineering. Vol 10, No. 2, pp. 81-89.

Yang, C., Remenyte-Prescott, R., and Andrews, J. D. (2015). “Pavement Maintenance Scheduling using Genetic Algorithms”. International Journal of Performability Engineering Vol. 11, No. 2, March, 2015, pp. 135-152.

Zaiontz C. (2018). “Real Statistics Using Excel”. www.real-statistics.com

Appendix A

Descriptive Statistics of PCI Values from Ground Truth and Raters

Case Study 1

StatTools Report

StatTools Report

Analysis: One Variable Summary
Performed By: Rodriguez Velasquez, Edgar D
Date: -
Updating: Live

	GT	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
One Variable Summary	Data Set #1	Data Set #1	Data Set #1	Data Set #1	Data Set #1	Data Set #1	Data Set #1
Mean	50.67	43.25	46.42	58.29	61.29	60.96	48.08
Variance	693.19	900.72	487.47	513.78	529.26	470.91	400.25
Std. Dev.	26.33	30.01	22.08	22.67	23.01	21.70	20.01
Skewness	0.2573	0.4304	0.9780	0.6198	0.1772	0.2705	0.8162
Kurtosis	1.9898	1.9173	3.1374	2.2226	1.7752	2.6509	2.4857
Median	49.00	36.00	39.00	53.00	56.00	58.00	41.00
Mean Abs. Dev.	21.44	25.52	17.31	18.20	19.98	16.46	16.19
Mode	13.00	13.00	33.00	34.00	28.00	37.00	35.00
Minimum	13.00	4.00	17.00	28.00	28.00	22.00	23.00
Maximum	96.00	98.00	93.00	100.00	100.00	100.00	87.00
Range	83.00	94.00	76.00	72.00	72.00	78.00	64.00
Count	24	24	24	24	24	24	24
Sum	1216.00	1038.00	1114.00	1399.00	1471.00	1463.00	1154.00
1st Quartile	26.00	15.00	32.00	36.00	41.00	48.00	33.00
3rd Quartile	64.00	62.00	60.00	68.00	80.00	73.00	54.00
Interquartile Range	38.00	47.00	28.00	32.00	39.00	25.00	21.00
1.00%	13.00	4.00	17.00	28.00	28.00	22.00	23.00
2.50%	13.00	4.00	17.00	28.00	28.00	22.00	23.00
5.00%	13.00	6.00	18.00	30.00	28.00	26.00	24.00
10.00%	17.00	8.00	27.00	34.00	34.00	37.00	25.00
20.00%	23.00	13.00	28.00	35.00	35.00	38.00	32.00
80.00%	77.00	79.00	66.00	85.00	82.00	75.00	72.00
90.00%	93.00	90.00	91.00	96.00	97.00	97.00	85.00
95.00%	94.00	94.00	91.00	97.00	98.00	98.00	85.00
97.50%	96.00	98.00	93.00	100.00	100.00	100.00	87.00
99.00%	96.00	98.00	93.00	100.00	100.00	100.00	87.00

Figure A.1. Descriptive Statistics of Ground Truth and Raters 1 to 6.

StatTools Report

Analysis: One Variable Summary
Performed By: Rodriguez Velasquez, Edgar D
Date: -
Updating: Live

One Variable Summary	Rater 7 Data Set #1	Rater 8 Data Set #1	Rater 9 Data Set #1	Rater 10 Data Set #1	Rater 11 Data Set #1	Rater 12 Data Set #1
Mean	50.42	54.96	55.08	58.04	59.58	50.92
Variance	461.12	491.78	616.69	429.17	562.69	423.04
Std. Dev.	21.47	22.18	24.83	20.72	23.72	20.57
Skewness	0.1754	0.7083	0.5792	0.2726	0.1381	0.3268
Kurtosis	2.1847	2.5747	2.5134	2.3489	2.5649	1.8081
Median	48.00	50.00	51.00	57.00	56.00	45.00
Mean Abs. Dev.	17.95	17.11	19.26	16.05	18.35	17.91
Mode	32.00	49.00	51.00	47.00	59.00	67.00
Minimum	15.00	25.00	19.00	25.00	12.00	20.00
Maximum	92.00	98.00	100.00	96.00	100.00	87.00
Range	77.00	73.00	81.00	71.00	88.00	67.00
Count	24	24	24	24	24	24
Sum	1210.00	1319.00	1322.00	1393.00	1430.00	1222.00
1st Quartile	33.00	35.00	34.00	41.00	43.00	34.00
3rd Quartile	65.00	64.00	61.00	71.00	75.00	67.00
Interquartile Range	32.00	29.00	27.00	30.00	32.00	33.00
1.00%	15.00	25.00	19.00	25.00	12.00	20.00
2.50%	15.00	25.00	19.00	25.00	12.00	20.00
5.00%	16.00	28.00	24.00	26.00	23.00	24.00
10.00%	22.00	29.00	24.00	29.00	35.00	28.00
20.00%	32.00	33.00	28.00	38.00	39.00	32.00
80.00%	70.00	83.00	81.00	79.00	84.00	67.00
90.00%	80.00	94.00	99.00	88.00	97.00	83.00
95.00%	83.00	96.00	100.00	94.00	98.00	84.00
97.50%	92.00	98.00	100.00	96.00	100.00	87.00
99.00%	92.00	98.00	100.00	96.00	100.00	87.00

Figure A.2. Descriptive Statistics of Raters 7 to 12.

StatTools Report

Analysis: One Variable Summary
Performed By: Rodriguez Velasquez, Edgar D
Date: -
Updating: Live

One Variable Summary	Rater 13 Data Set #1	Rater 14 Data Set #1	Rater 15 Data Set #1	Rater 16 Data Set #1	Rater 17 Data Set #1	Rater 18 Data Set #1
Mean	49.67	55.54	50.71	50.50	38.25	46.00
Variance	560.23	723.48	751.43	500.00	736.54	717.48
Std. Dev.	23.67	26.90	27.41	22.36	27.14	26.79
Skewness	0.9130	0.0260	0.3175	0.8518	0.8010	0.6472
Kurtosis	2.7592	2.2465	2.2186	2.9458	3.0714	2.4588
Median	43.00	56.00	47.00	45.00	34.00	37.00
Mean Abs. Dev.	18.81	21.25	21.85	17.46	20.48	21.75
Mode	26.00	52.00	13.00	45.00	38.00	16.00
Minimum	19.00	5.00	8.00	16.00	3.00	9.00
Maximum	97.00	100.00	99.00	96.00	97.00	97.00
Range	78.00	95.00	91.00	80.00	94.00	88.00
Count	24	24	24	24	24	24
Sum	1192.00	1333.00	1217.00	1212.00	918.00	1104.00
1st Quartile	31.00	30.00	27.00	32.00	20.00	23.00
3rd Quartile	59.00	72.00	69.00	59.00	51.00	56.00
Interquartile Range	28.00	42.00	42.00	27.00	31.00	33.00
1.00%	19.00	5.00	8.00	16.00	3.00	9.00
2.50%	19.00	5.00	8.00	16.00	3.00	9.00
5.00%	26.00	17.00	13.00	27.00	4.00	13.00
10.00%	26.00	24.00	13.00	29.00	4.00	16.00
20.00%	29.00	26.00	24.00	31.00	10.00	21.00
80.00%	69.00	77.00	79.00	70.00	63.00	73.00
90.00%	94.00	97.00	97.00	94.00	89.00	95.00
95.00%	95.00	97.00	97.00	94.00	92.00	96.00
97.50%	97.00	100.00	99.00	96.00	97.00	97.00
99.00%	97.00	100.00	99.00	96.00	97.00	97.00

Figure A.3. Descriptive Statistics of Raters 13 to 18.

Appendix B

Normality Test of PCI Values from Ground Truth and Raters

QQ Plot

Case Study 1

Real Statistics Report

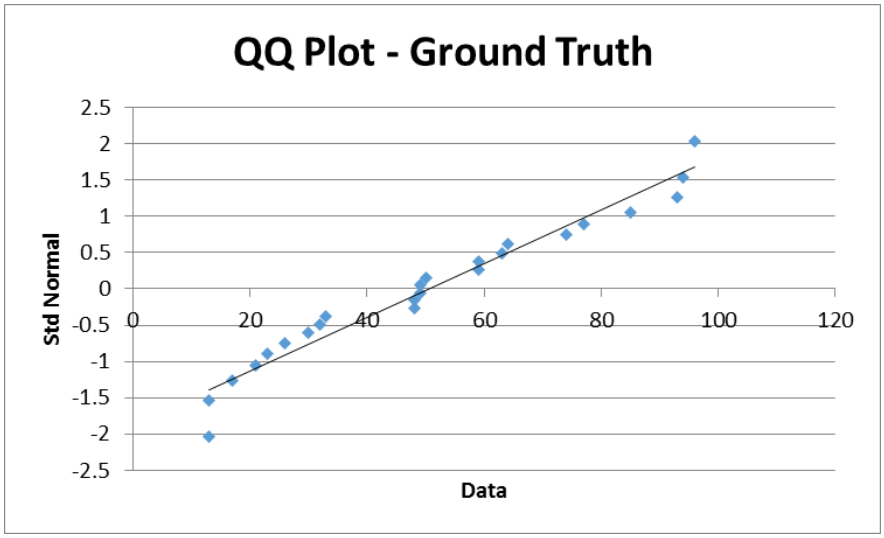


Figure B.1. QQ Plot for Ground Truth.

QQ Plot - Ground Truth

Count	24	48
Mean	51	
Std Dev	26.32847	

Interval	Data	Std Norm	Std Data
1	13	-2.036834132	-1.430643888
3	13	-1.534120544	-1.430643888
5	17	-1.258161561	-1.278717103
7	21	-1.054472452	-1.126790319
9	23	-0.887146559	-1.050826926
11	26	-0.741594044	-0.936881838
13	30	-0.61029461	-0.784955053
15	32	-0.488776411	-0.708991661
17	33	-0.37409541	-0.671009965
19	48	-0.264146977	-0.101284523
21	48	-0.157310685	-0.101284523
23	49	-0.05224518	-0.063302827
25	49	0.05224518	-0.063302827
27	50	0.157310685	-0.025321131
29	59	0.264146977	0.316514134
31	59	0.37409541	0.316514134
33	63	0.488776411	0.468440919
35	64	0.61029461	0.506422615
37	74	0.741594044	0.886239576
39	77	0.887146559	1.000184665
41	85	1.054472452	1.304038234
43	93	1.258161561	1.607891803
45	94	1.534120544	1.645873499
47	96	2.036834132	1.721836891

Figure B.2. QQ Plot Results for Ground Truth.

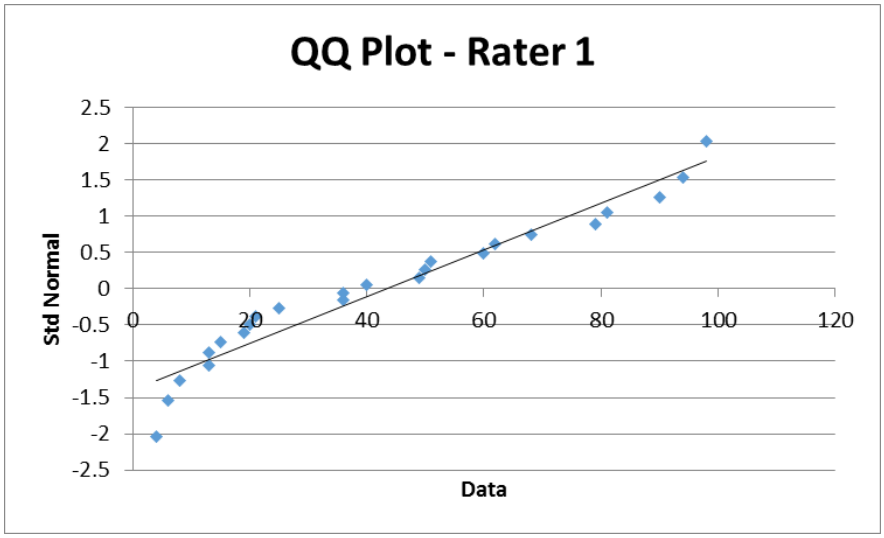


Figure B.3. QQ Plot for Rater 1.

QQ Plot - Rater 1

Count	24	48
Mean	43	
Std Dev	30.01195	

Interval	Data	Std Norm	Std Data
1	4	-2.036834132	-1.307812208
3	6	-1.534120544	-1.241172095
5	8	-1.258161561	-1.174531983
7	13	-1.054472452	-1.007931701
9	13	-0.887146559	-1.007931701
11	15	-0.741594044	-0.941291589
13	19	-0.61029461	-0.808011364
15	20	-0.488776411	-0.774691308
17	21	-0.37409541	-0.741371251
19	25	-0.264146977	-0.608091026
21	36	-0.157310685	-0.241570408
23	36	-0.05224518	-0.241570408
25	40	0.05224518	-0.108290183
27	49	0.157310685	0.191590323
29	50	0.264146977	0.22491038
31	51	0.37409541	0.258230436
33	60	0.488776411	0.558110942
35	62	0.61029461	0.624751055
37	68	0.741594044	0.824671392
39	79	0.887146559	1.191192011
41	81	1.054472452	1.257832123
43	90	1.258161561	1.55771263
45	94	1.534120544	1.690992854
47	98	2.036834132	1.824273079

Figure B.4. QQ Plot Results for Rater 1.

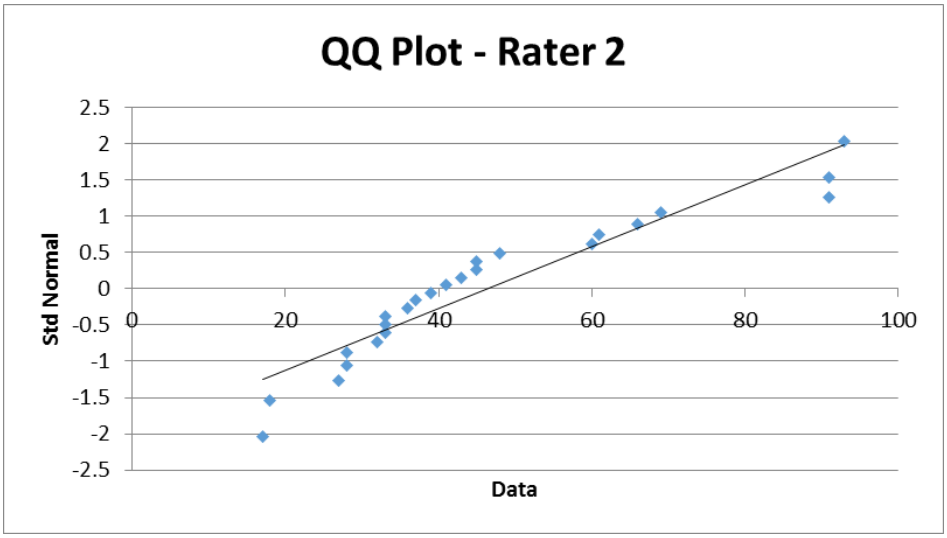


Figure B.5. QQ Plot for Rater 2.

QQ Plot - Rater 2

Count	24	48
Mean	46	
Std Dev	22.07874576	

Interval	Data	Std Norm	Std Data
1	17	-2.036834132	-1.332352253
3	18	-1.534120544	-1.287059825
5	27	-1.258161561	-0.879427975
7	28	-1.054472452	-0.834135547
9	28	-0.887146559	-0.834135547
11	32	-0.741594044	-0.652965835
13	33	-0.61029461	-0.607673407
15	33	-0.488776411	-0.607673407
17	33	-0.37409541	-0.607673407
19	36	-0.264146977	-0.471796124
21	37	-0.157310685	-0.426503696
23	39	-0.05224518	-0.33591884
25	41	0.05224518	-0.245333984
27	43	0.157310685	-0.154749129
29	45	0.264146977	-0.064164273
31	45	0.37409541	-0.064164273
33	48	0.488776411	0.071713011
35	60	0.61029461	0.615222145
37	61	0.741594044	0.660514573
39	66	0.887146559	0.886976712
41	69	1.054472452	1.022853996
43	91	1.258161561	2.019287409
45	91	1.534120544	2.019287409
47	93	2.036834132	2.109872265

Figure B.6. QQ Plot Results for Rater 2.

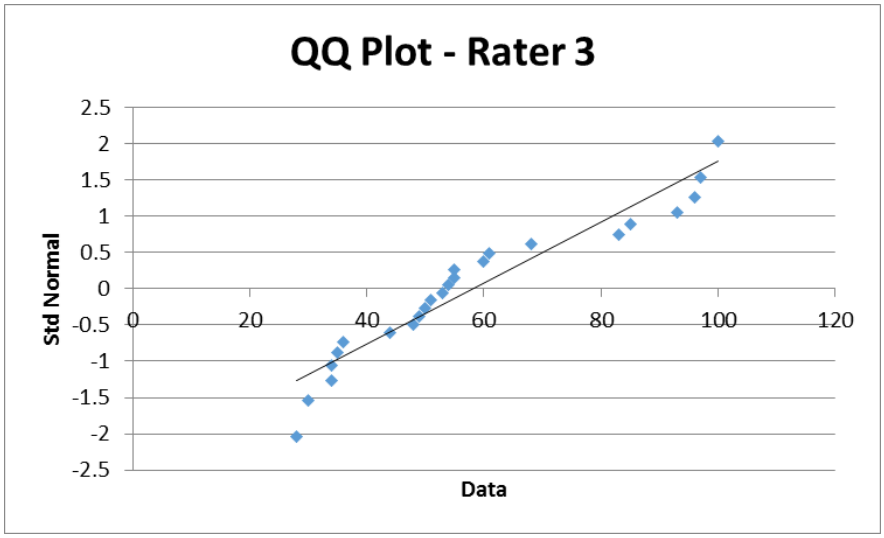


Figure B.7. QQ Plot for Rater 3.

QQ Plot - Rater 3

Count	24	48
Mean	58	
Std Dev	22.66673327	

Interval	Data	Std Norm	Std Data
1	28	-2.036834	-1.336393132
3	30	-1.534121	-1.248158097
5	34	-1.258162	-1.071688027
7	34	-1.054472	-1.071688027
9	35	-0.887147	-1.02757051
11	36	-0.741594	-0.983452993
13	44	-0.610295	-0.630512853
15	48	-0.488776	-0.454042784
17	49	-0.374095	-0.409925266
19	50	-0.264147	-0.365807749
21	51	-0.157311	-0.321690231
23	53	-0.052245	-0.233455196
25	54	0.052245	-0.189337679
27	55	0.157311	-0.145220162
29	55	0.264147	-0.145220162
31	60	0.374095	0.075367426
33	61	0.488776	0.119484943
35	68	0.610295	0.428307565
37	83	0.741594	1.090070326
39	85	0.887147	1.178305361
41	93	1.054472	1.531245501
43	96	1.258162	1.663598053
45	97	1.534121	1.70771557
47	100	2.036834	1.840068123

Figure B.8. QQ Plot Results for Rater 3.

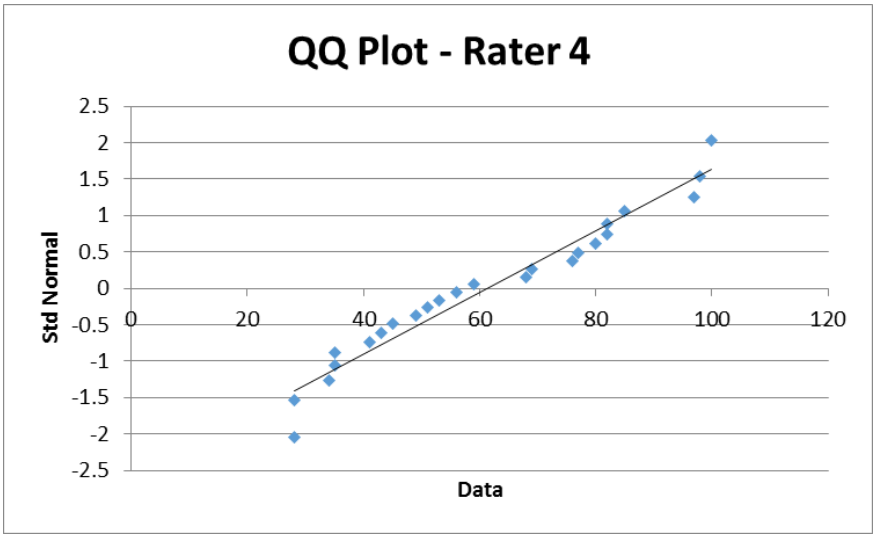


Figure B.9. QQ Plot for Rater 4.

QQ Plot - Rater 4

Count	24	48
Mean	61	
Std Dev	23.00563	

Interval	Data	Std Norm	Std Data
1	28	-2.036834132	-1.447109478
3	28	-1.534120544	-1.447109478
5	34	-1.258161561	-1.186303764
7	35	-1.054472452	-1.142836146
9	35	-0.887146559	-1.142836146
11	41	-0.741594044	-0.882030433
13	43	-0.61029461	-0.795095195
15	45	-0.488776411	-0.708159957
17	49	-0.37409541	-0.534289482
19	51	-0.264146977	-0.447354244
21	53	-0.157310685	-0.360419006
23	56	-0.05224518	-0.23001615
25	59	0.05224518	-0.099613293
27	68	0.157310685	0.291595276
29	69	0.264146977	0.335062895
31	76	0.37409541	0.639336227
33	77	0.488776411	0.682803846
35	80	0.61029461	0.813206703
37	82	0.741594044	0.90014194
39	82	0.887146559	0.90014194
41	85	1.054472452	1.030544797
43	97	1.258161561	1.552156223
45	98	1.534120544	1.595623842
47	100	2.036834132	1.68255908

Figure B.10. QQ Plot Results for Rater 4.

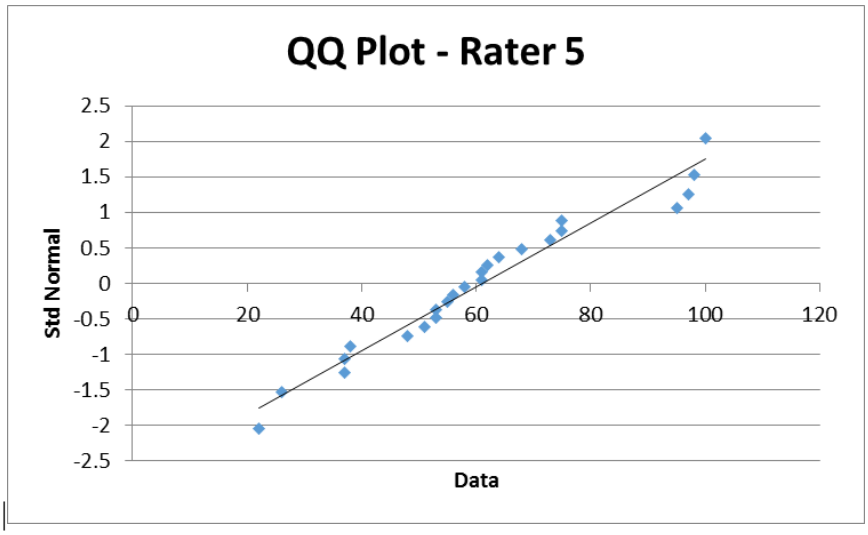


Figure B.11. QQ Plot for Rater 5.

QQ Plot - Rater 5

Count	24	48
Mean	61	
Std Dev	21.7004892	

Interval	Data	Std Norm	Std Data
1	22	-2.036834	-1.7952744
3	26	-1.534121	-1.6109468
5	37	-1.258162	-1.1040458
7	37	-1.054472	-1.1040458
9	38	-0.887147	-1.0579639
11	48	-0.741594	-0.5971448
13	51	-0.610295	-0.458899
15	53	-0.488776	-0.3667352
17	53	-0.374095	-0.3667352
19	55	-0.264147	-0.2745714
21	56	-0.157311	-0.2284895
23	58	-0.052245	-0.1363257
25	61	0.0522452	0.00192008
27	61	0.1573107	0.00192008
29	62	0.264147	0.04800199
31	64	0.3740954	0.14016581
33	68	0.4887764	0.32449345
35	73	0.6102946	0.554903
37	75	0.741594	0.64706683
39	75	0.8871466	0.64706683
41	95	1.0544725	1.56870503
43	97	1.2581616	1.66086886
45	98	1.5341205	1.70695077
47	100	2.0368341	1.79911459

Figure B.12. QQ Plot Results for Rater 5.

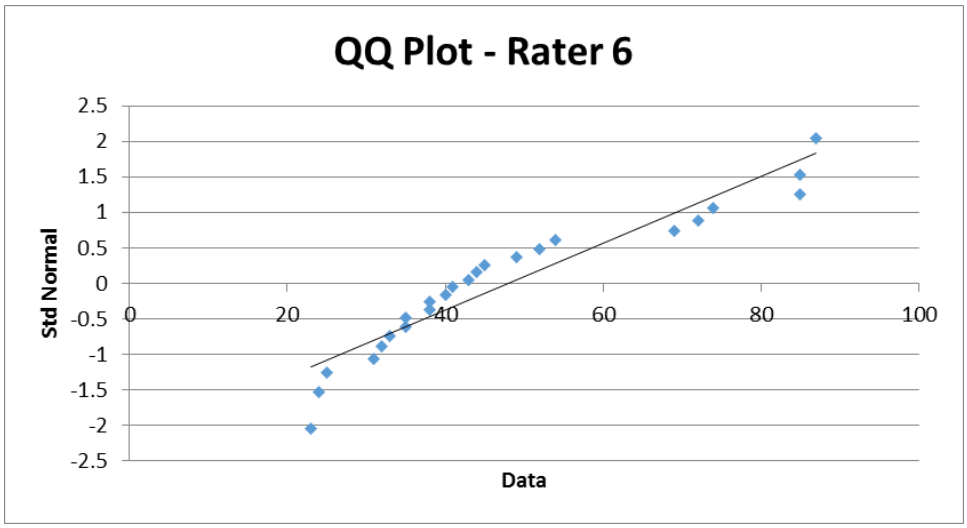


Figure B.13. QQ Plot for Rater 6.

QQ Plot - Rater 6

Count	24	48
Mean	48	
Std Dev	20.00633957	

Interval	Data	Std Norm	Std Data
1	23	-2.036834132	-1.253769248
3	24	-1.534120544	-1.203785092
5	25	-1.258161561	-1.153800936
7	31	-1.054472452	-0.853896
9	32	-0.887146559	-0.803911844
11	33	-0.741594044	-0.753927688
13	35	-0.61029461	-0.653959375
15	35	-0.488776411	-0.653959375
17	38	-0.37409541	-0.504006907
19	38	-0.264146977	-0.504006907
21	40	-0.157310685	-0.404038595
23	41	-0.05224518	-0.354054439
25	43	0.05224518	-0.254086127
27	44	0.157310685	-0.204101971
29	45	0.264146977	-0.154117815
31	49	0.37409541	0.04581881
33	52	0.488776411	0.195771278
35	54	0.61029461	0.29573959
37	69	0.741594044	1.045501931
39	72	0.887146559	1.1954544
41	74	1.054472452	1.295422712
43	85	1.258161561	1.845248429
45	85	1.534120544	1.845248429
47	87	2.036834132	1.945216741

Figure B.14. QQ Plot Results for Rater 6.

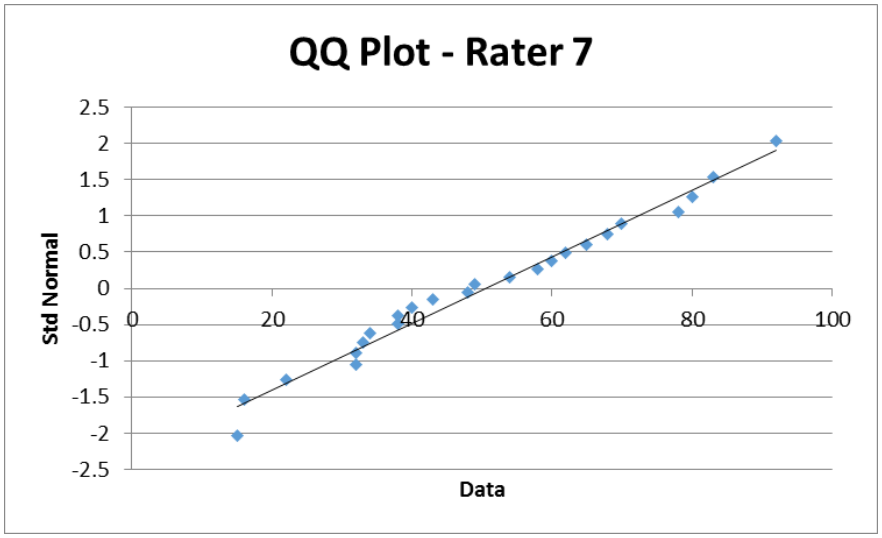


Figure B.15. QQ Plot for Rater 7.

QQ Plot - Rater 7

Count	24	48
Mean	50	
Std Dev	21.47378	

Interval	Data	Std Norm	Std Data
1	15	-2.03683413	-1.649298268
3	16	-1.53412054	-1.602729847
5	22	-1.25816156	-1.323319316
7	32	-1.05447245	-0.8576351
9	32	-0.88714656	-0.8576351
11	33	-0.74159404	-0.811066678
13	34	-0.61029461	-0.764498256
15	38	-0.48877641	-0.578224569
17	38	-0.37409541	-0.578224569
19	40	-0.26414698	-0.485087726
21	43	-0.15731068	-0.345382461
23	48	-0.05224518	-0.112540352
25	49	0.05224518	-0.065971931
27	54	0.15731068	0.166870178
29	58	0.26414698	0.353143865
31	60	0.37409541	0.446280708
33	62	0.48877641	0.539417551
35	65	0.61029461	0.679122816
37	68	0.74159404	0.818828081
39	70	0.88714656	0.911964925
41	78	1.05447245	1.284512298
43	80	1.25816156	1.377649142
45	83	1.53412054	1.517354407
47	92	2.03683413	1.936470202

Figure B.16. QQ Plot Results for Rater 7.

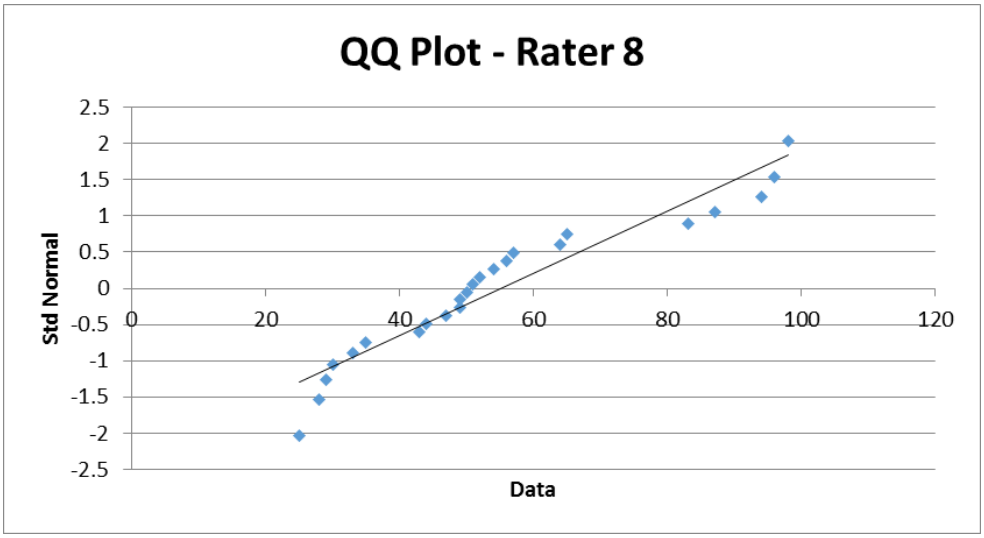


Figure B.17. QQ Plot for Rater 8.

QQ Plot - Rater 8

Count	24	48
Mean	55	
Std Dev	22.17613125	

Interval	Data	Std Norm	Std Data
1	25	-2.036834132	-1.350926949
3	28	-1.534120544	-1.215646365
5	29	-1.258161561	-1.170552836
7	30	-1.054472452	-1.125459308
9	33	-0.887146559	-0.990178724
11	35	-0.741594044	-0.899991667
13	43	-0.61029461	-0.539243442
15	44	-0.488776411	-0.494149913
17	47	-0.37409541	-0.358869329
19	49	-0.264146977	-0.268682272
21	49	-0.157310685	-0.268682272
23	50	-0.05224518	-0.223588744
25	51	0.05224518	-0.178495216
27	52	0.157310685	-0.133401688
29	54	0.264146977	-0.043214631
31	56	0.37409541	0.046972425
33	57	0.488776411	0.092065953
35	64	0.61029461	0.407720651
37	65	0.741594044	0.452814179
39	83	0.887146559	1.264497687
41	87	1.054472452	1.4448718
43	94	1.258161561	1.760526497
45	96	1.534120544	1.850713554
47	98	2.036834132	1.94090061

Figure B.18. QQ Plot Results for Rater 8.

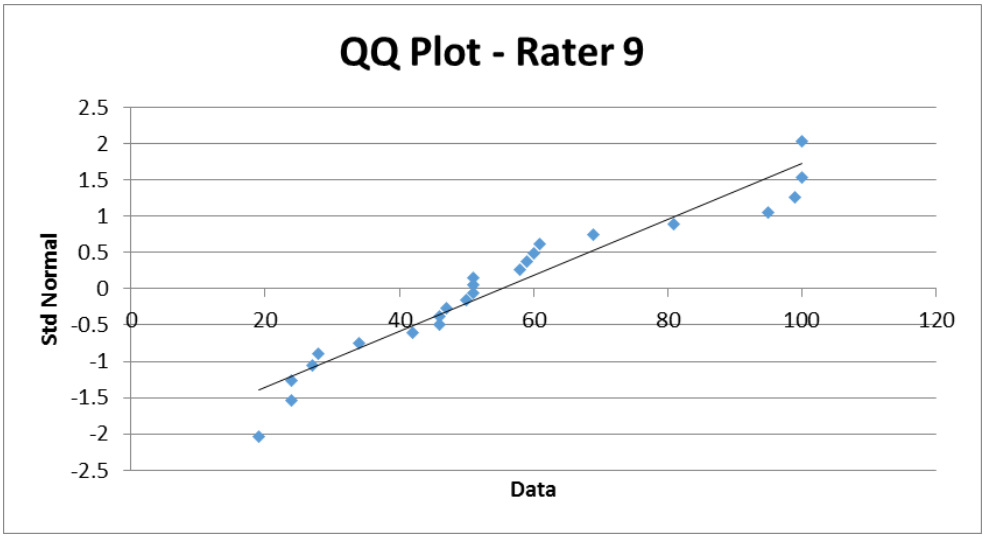


Figure B.19. QQ Plot for Rater 9.

QQ Plot - Rater 9

Count	24	48
Mean	55	
Std Dev	24.83321175	

Interval	Data	Std Norm	Std Data
1	19	-2.036834132	-1.453027248
3	24	-1.534120544	-1.251683981
5	24	-1.258161561	-1.251683981
7	27	-1.054472452	-1.13087802
9	28	-0.887146559	-1.090609366
11	34	-0.741594044	-0.848997445
13	42	-0.61029461	-0.526848217
15	46	-0.488776411	-0.365773603
17	46	-0.37409541	-0.365773603
19	47	-0.264146977	-0.325504949
21	50	-0.157310685	-0.204698989
23	51	-0.05224518	-0.164430335
25	51	0.05224518	-0.164430335
27	51	0.157310685	-0.164430335
29	58	0.264146977	0.117450239
31	59	0.37409541	0.157718893
33	60	0.488776411	0.197987547
35	61	0.61029461	0.2382562
37	69	0.741594044	0.560405428
39	81	0.887146559	1.043629271
41	95	1.054472452	1.60739042
43	99	1.258161561	1.768465034
45	100	1.534120544	1.808733688
47	100	2.036834132	1.808733688

Figure B.20. QQ Plot Results for Rater 9.

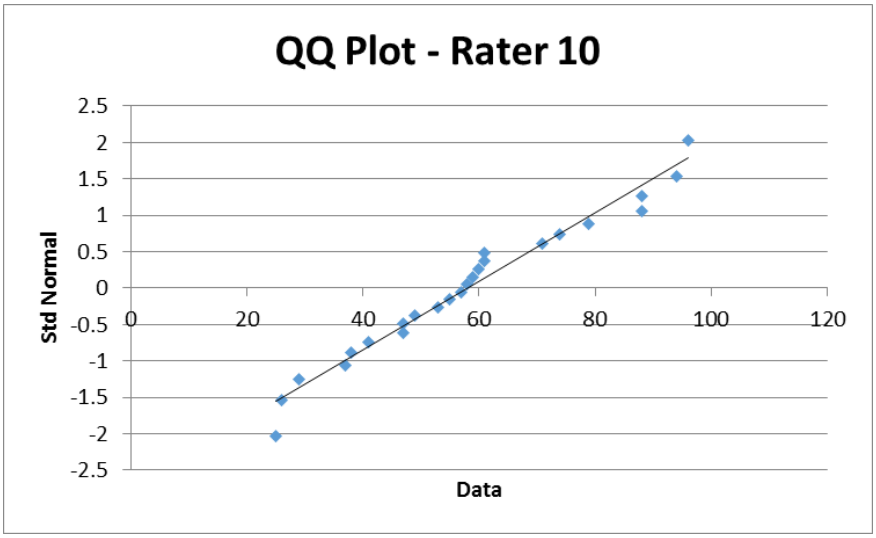


Figure B.21. QQ Plot for Rater 10.

QQ Plot - Rater 10

Count	24	48
Mean	58	
Std Dev	20.7164693	

Interval	Data	Std Norm	Std Data
1	25	-2.036834132	-1.59494681
3	26	-1.534120544	-1.546676036
5	29	-1.258161561	-1.401863716
7	37	-1.054472452	-1.015697527
9	38	-0.887146559	-0.967426753
11	41	-0.741594044	-0.822614433
13	47	-0.61029461	-0.532989791
15	47	-0.488776411	-0.532989791
17	49	-0.37409541	-0.436448244
19	53	-0.264146977	-0.24336515
21	55	-0.157310685	-0.146823603
23	57	-0.05224518	-0.050282056
25	58	0.05224518	-0.002011282
27	59	0.157310685	0.046259491
29	60	0.264146977	0.094530265
31	61	0.37409541	0.142801038
33	61	0.488776411	0.142801038
35	71	0.61029461	0.625508774
37	74	0.741594044	0.770321095
39	79	0.887146559	1.011674963
41	88	1.054472452	1.446111925
43	88	1.258161561	1.446111925
45	94	1.534120544	1.735736566
47	96	2.036834132	1.832278113

Figure B.22. QQ Plot Results for Rater 10.

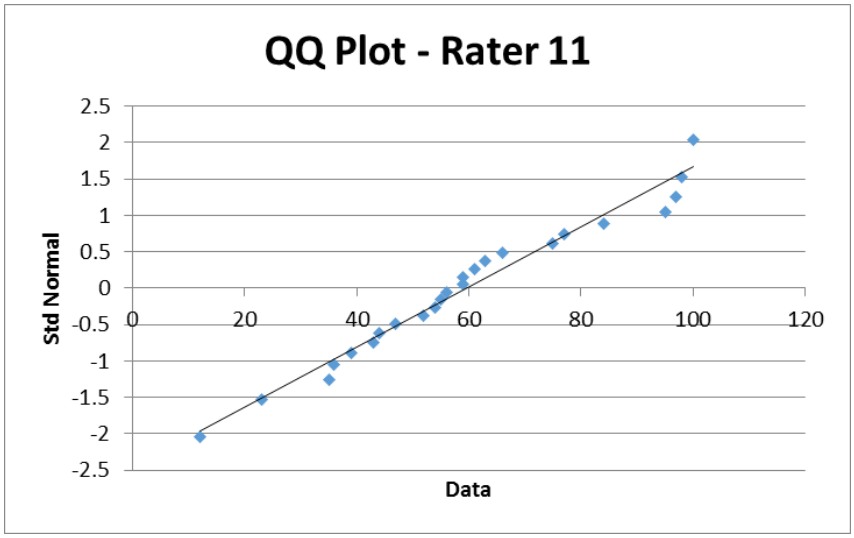


Figure B.23. QQ Plot for Rater 11.

QQ Plot - Rater 11

Count	24	48
Mean	60	
Std Dev	23.72105406	

Interval	Data	Std Norm	Std Data
1	12	-2.036834132	-2.005953581
3	23	-1.534120544	-1.542230511
5	35	-1.258161561	-1.036350799
7	36	-1.054472452	-0.994194156
9	39	-0.887146559	-0.867724228
11	43	-0.741594044	-0.699097658
13	44	-0.61029461	-0.656941015
15	47	-0.488776411	-0.530471087
17	52	-0.37409541	-0.319687874
19	54	-0.264146977	-0.235374588
21	55	-0.157310685	-0.193217946
23	56	-0.05224518	-0.151061303
25	59	0.05224518	-0.024591375
27	59	0.157310685	-0.024591375
29	61	0.264146977	0.05972191
31	63	0.37409541	0.144035196
33	66	0.488776411	0.270505124
35	75	0.61029461	0.649914908
37	77	0.741594044	0.734228193
39	84	0.887146559	1.029324692
41	95	1.054472452	1.493047761
43	97	1.258161561	1.577361047
45	98	1.534120544	1.61951769
47	100	2.036834132	1.703830975

Figure B.24. QQ Plot Results for Rater 11.

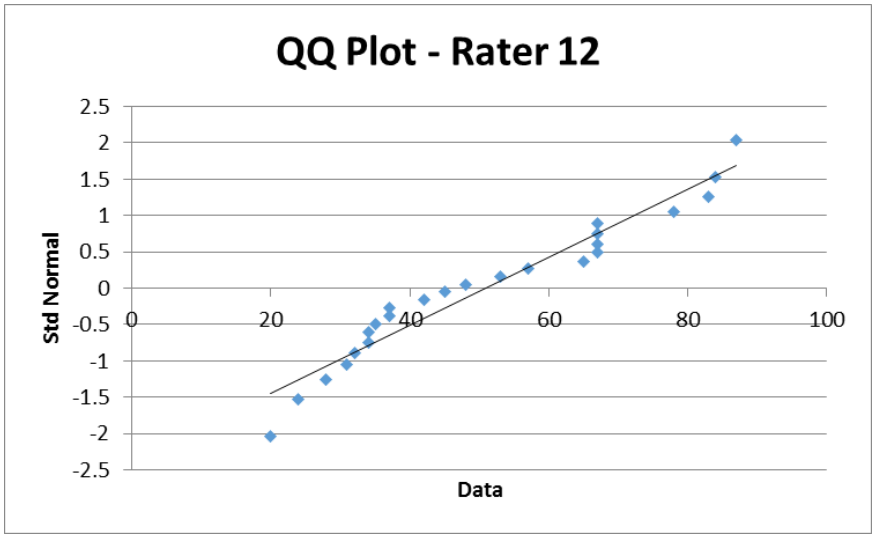


Figure B.25. QQ Plot for Rater 12.

QQ Plot - Rater 12

Count	24	48
Mean	51	
Std Dev	20.56784461	

Interval	Data	Std Norm	Std Data
1	20	-2.036834132	-1.503155399
3	24	-1.534120544	-1.308677072
5	28	-1.258161561	-1.114198746
7	31	-1.054472452	-0.968340001
9	32	-0.887146559	-0.919720419
11	34	-0.741594044	-0.822481256
13	34	-0.61029461	-0.822481256
15	35	-0.488776411	-0.773861674
17	37	-0.37409541	-0.676622511
19	37	-0.264146977	-0.676622511
21	42	-0.157310685	-0.433524603
23	45	-0.05224518	-0.287665858
25	48	0.05224518	-0.141807113
27	53	0.157310685	0.101290795
29	57	0.264146977	0.295769122
31	65	0.37409541	0.684725775
33	67	0.488776411	0.781964938
35	67	0.61029461	0.781964938
37	67	0.741594044	0.781964938
39	67	0.887146559	0.781964938
41	78	1.054472452	1.316780336
43	83	1.258161561	1.559878244
45	84	1.534120544	1.608497826
47	87	2.036834132	1.754356571

Figure B.26. QQ Plot Results for Rater 12.

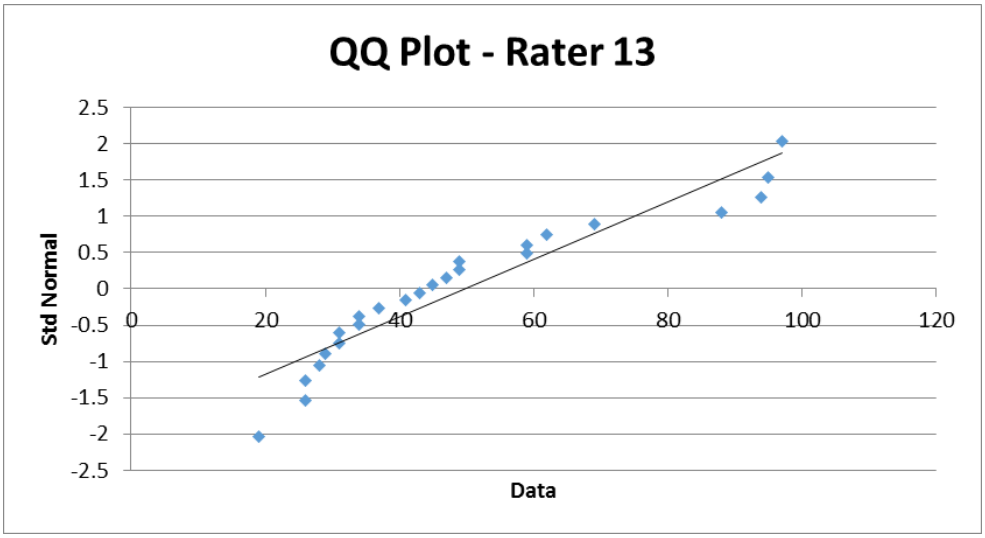


Figure B.27. QQ Plot for Rater 13.

QQ Plot - Rater 13

Count	24	48
Mean	50	
Std Dev	23.66921807	

Interval	Data	Std Norm	Std Data
1	19	-2.036834132	-1.295634971
3	26	-1.534120544	-0.999892206
5	26	-1.258161561	-0.999892206
7	28	-1.054472452	-0.915394273
9	29	-0.887146559	-0.873145307
11	31	-0.741594044	-0.788647374
13	31	-0.61029461	-0.788647374
15	34	-0.488776411	-0.661900474
17	34	-0.37409541	-0.661900474
19	37	-0.264146977	-0.535153575
21	41	-0.157310685	-0.366157709
23	43	-0.05224518	-0.281659776
25	45	0.05224518	-0.197161843
27	47	0.157310685	-0.112663911
29	49	0.264146977	-0.028165978
31	49	0.37409541	-0.028165978
33	59	0.488776411	0.394323687
35	59	0.61029461	0.394323687
37	62	0.741594044	0.521070586
39	69	0.887146559	0.816813351
41	88	1.054472452	1.619543714
43	94	1.258161561	1.873037512
45	95	1.534120544	1.915286479
47	97	2.036834132	1.999784412

Figure B.28. QQ Plot Results for Rater 13.

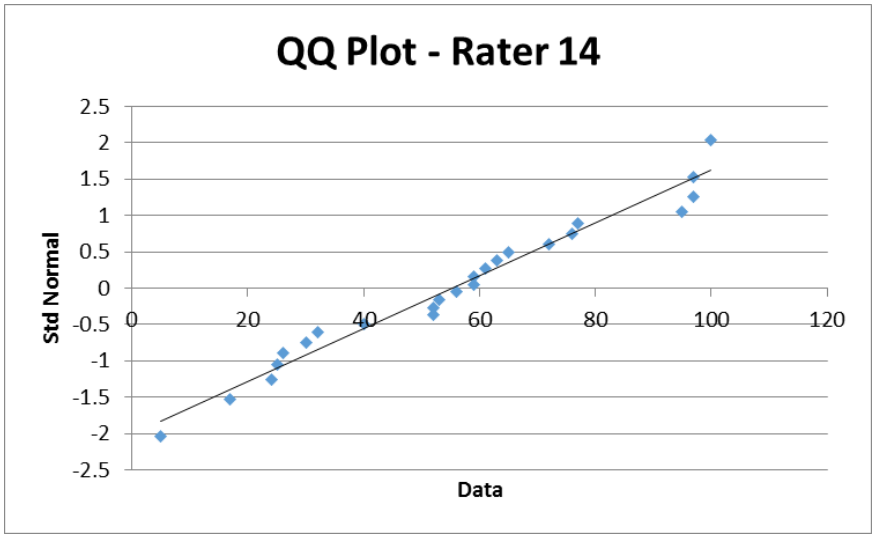


Figure B.29. QQ Plot for Rater 14.

QQ Plot - Rater 14

Count	24	48
Mean	56	
Std Dev	26.89751753	

Interval	Data	Std Norm	Std Data
1	5	-2.036834132	-1.879045775
3	17	-1.534120544	-1.432907949
5	24	-1.258161561	-1.172660883
7	25	-1.054472452	-1.135482731
9	26	-0.887146559	-1.098304579
11	30	-0.741594044	-0.94959197
13	32	-0.61029461	-0.875235666
15	40	-0.488776411	-0.577810448
17	52	-0.37409541	-0.131672622
19	52	-0.264146977	-0.131672622
21	53	-0.157310685	-0.09449447
23	56	-0.05224518	0.017039986
25	59	0.05224518	0.128574443
27	59	0.157310685	0.128574443
29	61	0.264146977	0.202930747
31	63	0.37409541	0.277287052
33	65	0.488776411	0.351643356
35	72	0.61029461	0.611890421
37	76	0.741594044	0.76060303
39	77	0.887146559	0.797781182
41	95	1.054472452	1.466987921
43	97	1.258161561	1.541344226
45	97	1.534120544	1.541344226
47	100	2.036834132	1.652878682

Figure B.30. QQ Plot Results for Rater 14.

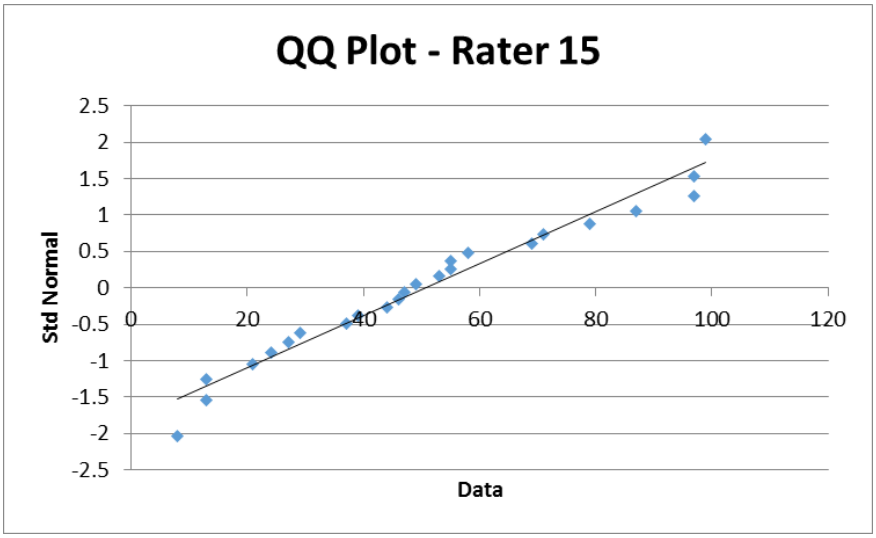


Figure B.31. QQ Plot for Rater 15.

QQ Plot -Rater 15

Count	24	48
Mean	51	
Std Dev	27.41227774	

Interval	Data	Std Norm	Std Data
1	8	-2.036834132	-1.558000168
3	13	-1.534120544	-1.375600149
5	13	-1.258161561	-1.375600149
7	21	-1.054472452	-1.083760117
9	24	-0.887146559	-0.974320105
11	27	-0.741594044	-0.864880093
13	29	-0.61029461	-0.791920085
15	37	-0.488776411	-0.500080054
17	39	-0.37409541	-0.427120046
19	44	-0.264146977	-0.244720026
21	46	-0.157310685	-0.171760019
23	47	-0.05224518	-0.135280015
25	49	0.05224518	-0.062320007
27	53	0.157310685	0.083600009
29	55	0.264146977	0.156560017
31	55	0.37409541	0.156560017
33	58	0.488776411	0.266000029
35	69	0.61029461	0.667280072
37	71	0.741594044	0.74024008
39	79	0.887146559	1.032080111
41	87	1.054472452	1.323920143
43	97	1.258161561	1.688720182
45	97	1.534120544	1.688720182
47	99	2.036834132	1.76168019

Figure B.32. QQ Plot Results for Rater 15.

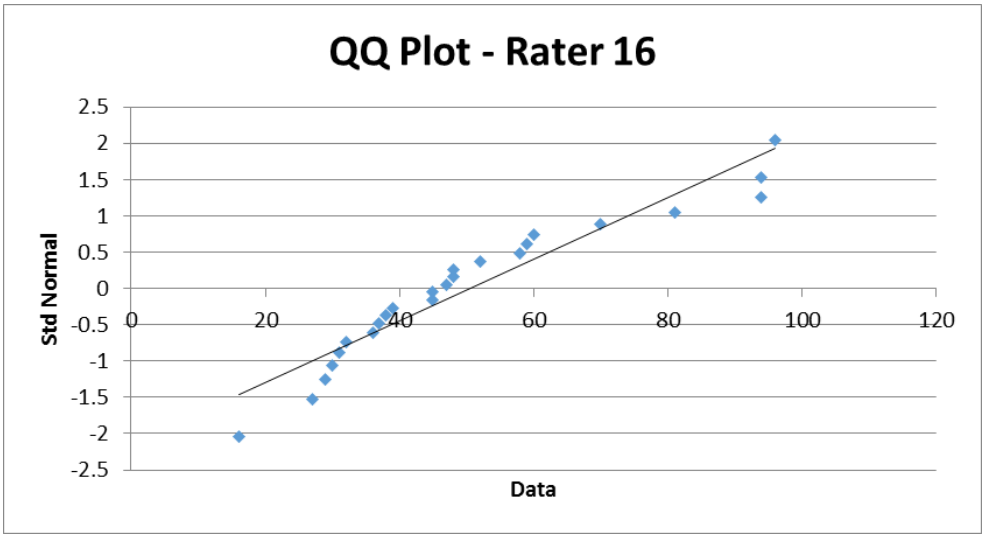


Figure B.33. QQ Plot for Rater 16.

QQ Plot - Rater 16

Count	24	48
Mean	51	
Std Dev	22.36067977	

Interval	Data	Std Norm	Std Data
1	16	-2.036834132	-1.542886904
3	27	-1.534120544	-1.050951949
5	29	-1.258161561	-0.96150923
7	30	-1.054472452	-0.916787871
9	31	-0.887146559	-0.872066511
11	32	-0.741594044	-0.827345152
13	36	-0.61029461	-0.648459713
15	37	-0.488776411	-0.603738354
17	38	-0.37409541	-0.559016994
19	39	-0.264146977	-0.514295635
21	45	-0.157310685	-0.245967478
23	45	-0.05224518	-0.245967478
25	47	0.05224518	-0.156524758
27	48	0.157310685	-0.111803399
29	48	0.264146977	-0.111803399
31	52	0.37409541	0.067082039
33	58	0.488776411	0.335410197
35	59	0.61029461	0.380131556
37	60	0.741594044	0.424852916
39	70	0.887146559	0.872066511
41	81	1.054472452	1.364001466
43	94	1.258161561	1.94537914
45	94	1.534120544	1.94537914
47	96	2.036834132	2.03482186

Figure B.34. QQ Plot Results for Rater 16.

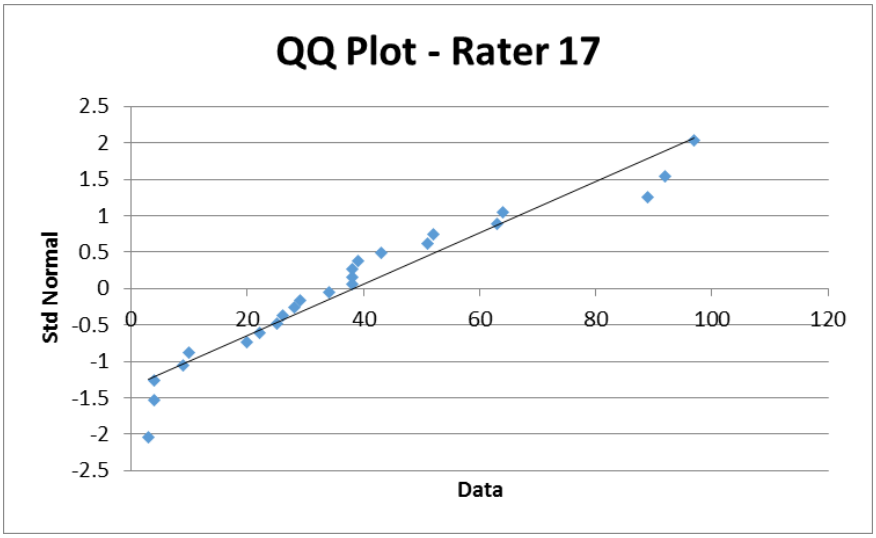


Figure B.35. QQ Plot for Rater 17.

QQ Plot -Rater 17

Count	24	48
Mean	38	
Std Dev	27.13933452	

Interval	Data	Std Norm	Std Data
1	3	-2.036834132	-1.298852777
3	4	-1.534120544	-1.262005889
5	4	-1.258161561	-1.262005889
7	9	-1.054472452	-1.077771453
9	10	-0.887146559	-1.040924566
11	20	-0.741594044	-0.672455693
13	22	-0.61029461	-0.598761918
15	25	-0.488776411	-0.488221256
17	26	-0.37409541	-0.451374369
19	28	-0.264146977	-0.377680595
21	29	-0.157310685	-0.340833707
23	34	-0.05224518	-0.156599271
25	38	0.05224518	-0.009211722
27	38	0.157310685	-0.009211722
29	38	0.264146977	-0.009211722
31	39	0.37409541	0.027635165
33	43	0.488776411	0.175022715
35	51	0.61029461	0.469797813
37	52	0.741594044	0.5066447
39	63	0.887146559	0.91196046
41	64	1.054472452	0.948807347
43	89	1.258161561	1.86997953
45	92	1.534120544	1.980520191
47	97	2.036834132	2.164754628

Figure B.36. QQ Plot Results for Rater 17.

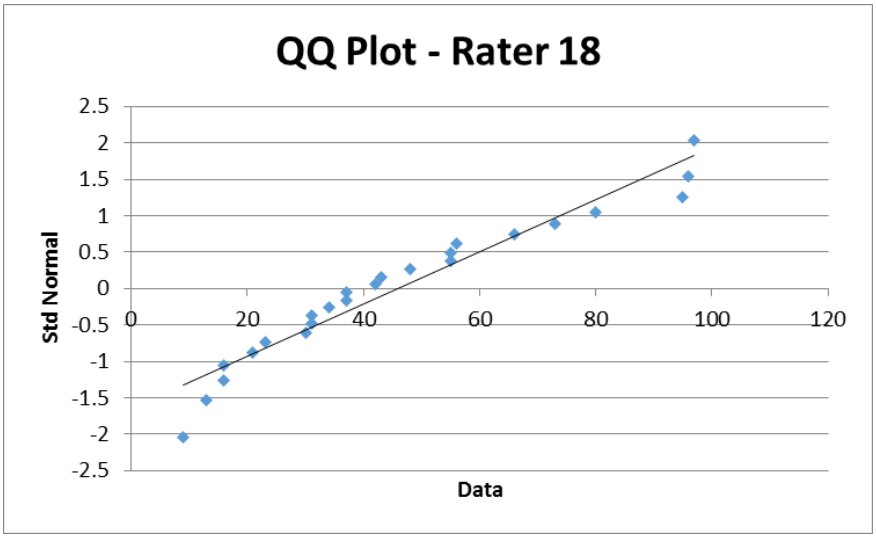


Figure B.37. QQ Plot for Rater 18.

QQ Plot - Rater 18

Count	24	48
Mean	46	
Std Dev	26.78578468	

Interval	Data	Std Norm	Std Data
1	9	-2.036834132	-1.381329703
3	13	-1.534120544	-1.231996762
5	16	-1.258161561	-1.119997057
7	16	-1.054472452	-1.119997057
9	21	-0.887146559	-0.933330881
11	23	-0.741594044	-0.85866441
13	30	-0.61029461	-0.597331764
15	31	-0.488776411	-0.559998528
17	31	-0.37409541	-0.559998528
19	34	-0.264146977	-0.447998823
21	37	-0.157310685	-0.335999117
23	37	-0.05224518	-0.335999117
25	42	0.05224518	-0.149332941
27	43	0.157310685	-0.111999706
29	48	0.264146977	0.07466647
31	55	0.37409541	0.335999117
33	55	0.488776411	0.335999117
35	56	0.61029461	0.373332352
37	66	0.741594044	0.746664704
39	73	0.887146559	1.007997351
41	80	1.054472452	1.269329998
43	95	1.258161561	1.829328526
45	96	1.534120544	1.866661761
47	97	2.036834132	1.903994996

Figure B.38. QQ Plot Results for Rater 18.

Appendix C

F-Test for Variances Comparison

Normally Distributed Data

Case Study 1

Excel Report

	GT	*Rater 1*
Mean	50.66666667	43.25
Variance	693.1884058	900.71739
Observations	24	24
df	23	23
F	0.769595894	
p-value	0.267520872	
F-crit	0.496419613	

	GT	*Rater 4*
Mean	50.66666667	61.291667
Variance	693.1884058	529.25906
Observations	24	24
df	23	23
F	1.309733665	
p-value	0.261389941	
F-crit	2.014424842	

	GT	*Rater 5*
Mean	50.66666667	60.958333
Variance	693.1884058	470.91123
Observations	24	24
df	23	23
F	1.472015019	
p-value	0.180281975	
F-crit	2.014424842	

	GT	*Rater 7*
Mean	50.66666667	50.416667
Variance	693.1884058	461.12319
Observations	24	24
df	23	23
F	1.503260784	
p-value	0.167614546	
F-crit	2.014424842	

Figure C.1. F-Test Results for Raters 1, 4, 5, and 7.

	GT	*Rater 10*
Mean	50.66666667	58.04166667
Variance	693.1884058	429.1721014
Observations	24	24
df	23	23
F	1.615175831	
p-value	0.12883322	
F-crit	2.014424842	

	GT	*Rater 11*
Mean	50.66666667	59.58333
Variance	693.1884058	562.6884
Observations	24	24
df	23	23
F	1.231922319	
p-value	0.310547294	
F-crit	2.014424842	

	GT	*Rater 12*
Mean	50.66666667	50.91666667
Variance	693.1884058	423.0362319
Observations	24	24
df	23	23
F	1.638602922	
p-value	0.121889733	
F-crit	2.014424842	

	GT	*Rater 14*
Mean	50.66666667	55.54166667
Variance	693.1884058	723.4764493
Observations	24	24
df	23	23
F	0.958135412	
p-value	0.459601567	
F-crit	0.496419613	

Figure C.2. F-Test Results for Raters 10, 11, 12, and 14.

	GT	*Rater 15*
Mean	50.66666667	50.70833333
Variance	693.1884058	751.432971
Observations	24	24
df	23	23
F	0.922488675	
p-value	0.424126562	
F-crit	0.496419613	

	GT	*Rater 17*
Mean	50.66666667	38.25
Variance	693.1884058	736.5434783
Observations	24	24
df	23	23
F	0.941137117	
p-value	0.442793698	
F-crit	0.496419613	

	GT	*Rater 18*
Mean	50.66666667	46
Variance	693.1884058	717.4782609
Observations	24	24
df	23	23
F	0.966145518	
p-value	0.467445961	
F-crit	0.496419613	

Figure C.3. F-Test Results for Raters 15, 17, and 18.

Appendix D

Levene's-Test for Variances Comparison

Not Normally Distributed Data

Case Study 1

Real Statistics Report

Rater 2

type	p-value
means	0.30886949
medians	0.275802786
trimmed	0.293778518

Rater 3

type	p-value
means	0.41970537
medians	0.36574293
trimmed	0.41435554

Rater 6

type	p-value
means	0.1693056
medians	0.15446114
trimmed	0.16661798

Rater 8

type	p-value
means	0.29409291
medians	0.27671833
trimmed	0.28916965

Rater 9

type	p-value
means	0.61411521
medians	0.56137751
trimmed	0.61489626

Rater 13

type	p-value
means	0.5235721
medians	0.48128228
trimmed	0.50242569

Rater 16

type	p-value
means	0.33113925
medians	0.30585533
trimmed	0.32768154

Figure D.1. Levene's-Test for Raters 2, 3, 6, 8, 9, 13, and 16

Appendix E

T-Test for Means Comparison

Normally Distributed Data

Case Study 1

Excel Report

	GT	*Rater 1*
Mean	50.6666667	43.25
Variance	693.188406	900.717391
Observations	24	24
Hypothesized Mean Difference	0	
df	45	
t Stat	0.91008731	
P(T<=t) one-tail	0.1838128	
t Critical one-tail	1.67942739	
P(T<=t) two-tail	0.3676256	
t Critical two-tail	2.01410339	

	GT	*Rater 4*
Mean	50.6666667	61.2916667
Variance	693.188406	529.259058
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-1.48874206	
P(T<=t) one-tail	0.07168981	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.14337963	
t Critical two-tail	2.0128956	

	GT	*Rater 5*
Mean	50.6666667	60.9583333
Variance	693.188406	470.911232
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-1.47773396	
P(T<=t) one-tail	0.07314685	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.14629369	
t Critical two-tail	2.0128956	

Figure E.1. T-Test for Raters 1, 4, and 5.

	GT	*Rater 7*
Mean	50.6666667	50.4166667
Variance	693.188406	461.123188
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	0.03604824	
P(T<=t) one-tail	0.48569994	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.97139987	
t Critical two-tail	2.0128956	

	GT	*Rater 10*
Mean	50.6666667	58.0416667
Variance	693.188406	429.172101
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-1.07845358	
P(T<=t) one-tail	0.14322712	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.28645424	
t Critical two-tail	2.0128956	

	GT	*Rater 11*
Mean	50.6666667	59.5833333
Variance	693.188406	562.688406
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-1.23263539	
P(T<=t) one-tail	0.11198734	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.22397469	
t Critical two-tail	2.0128956	

Figure E.2. T-Test for Raters 7, 10, and 11.

	GT	*Rater 12*
Mean	50.6666667	50.9166667
Variance	693.188406	423.036232
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-0.03665809	
P(T<=t) one-tail	0.48545812	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.97091625	
t Critical two-tail	2.0128956	

	GT	*Rater 14*
Mean	50.6666667	55.5416667
Variance	693.188406	723.476449
Observations	24	24
Hypothesized Mean Difference	0	
df	46	
t Stat	-0.63452199	
P(T<=t) one-tail	0.26444176	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.52888352	
t Critical two-tail	2.0128956	

	GR	*Rater 15*
Mean	50.6666667	49.9130435
Variance	693.188406	769.719368
Observations	24	23
Hypothesized Mean Difference	0	
df	45	
t Stat	0.09544206	
P(T<=t) one-tail	0.46219373	
t Critical one-tail	1.67942739	
P(T<=t) two-tail	0.92438747	
t Critical two-tail	2.01410339	

Figure E.3. T-Test for Raters 12, 14, and 15.

	GT	*Rater 17*
Mean	50.66666667	38.25
Variance	693.1884058	736.543478
Observation	24	24
Hypothesize	0	
df	46	
t Stat	1.608730672	
P(T<=t) one-	0.057258944	
t Critical one	1.678660414	
P(T<=t) two-	0.114517887	
t Critical two	2.012895599	

	GT	*Rater 18*
Mean	50.6666667	45.1304348
Variance	693.188406	731.118577
Observations	24	23
Hypothesized Mean Difference	0	
df	45	
t Stat	0.71076343	
P(T<=t) one-tail	0.24044933	
t Critical one-tail	1.67942739	
P(T<=t) two-tail	0.48089867	
t Critical two-tail	2.01410339	

Figure E.4. T-Test for Raters 17 and 18.

Appendix F

Man-Whitney Test for Means Comparison

Not Normally Distributed Data

Case Study 1

Real Statistics Report

	GT	Rater 2
count	24	24
median	49	40
rank sum	615	561
U	261	315

	one tail	two tail
alpha	0.05	
U	261	
mean	288	
std dev	48.49742261	
z-score	0.54642079	
effect r	0.078869048	
U-crit	207.7288385	192.446798
p-value	0.292388361	0.58477672
sig (norm)	no	no

	GT	Rater 3
count	24	24
median	49	54
rank sum	529.5	646.5
U	346.5	229.5

	one tail	two tail
alpha	0.05	
U	229.5	
mean	288	
std dev	48.4974226	
z-score	1.19593984	
effect r	0.17261905	
U-crit	207.728839	192.446798
p-value	0.11586002	0.23172004
sig (norm)	no	no

Figure F.1. Man-Whitney Test for Raters 2 and 3.

	GT	Rater 6
count	24	24
median	49	42
rank sum	601	575
U	275	301

	one tail	two tail
alpha	0.05	
U	275	
mean	288	
std dev	48.4974226	
z-score	0.25774566	
effect r	0.03720238	
U-crit	207.728839	192.446798
p-value	0.3983016	0.79660321
sig (norm)	no	no

	GT	Rater 8
count	24	24
median	49	51
rank sum	554	622
U	322	254

	one tail	two tail
alpha	0.05	
U	254	
mean	288	
std dev	48.4974226	
z-score	0.69075836	
effect r	0.09970238	
U-crit	207.728839	192.446798
p-value	0.2448587	0.48971741
sig (norm)	no	no

Figure F.2. Man-Whitney Test for Raters 6 and 8.

	GT	Rater 9
count	24	24
median	49	51
rank sum	556.5	619.5
U	319.5	256.5

	one tail	two tail
alpha	0.05	
U	256.5	
mean	288	
std dev	48.4974226	
z-score	0.63920923	
effect r	0.0922619	
U-crit	207.728839	192.446798
p-value	0.26134342	0.52268683
sig (norm)	no	no

	GT	Rater 13
count	24	24
median	49	44
rank sum	594.5	581.5
U	281.5	294.5

	one tail	two tail
alpha	0.05	
U	281.5	
mean	288	
std dev	48.4974226	
z-score	0.12371791	
effect r	0.01785714	
U-crit	207.728839	192.446798
p-value	0.45076931	0.90153863
sig (norm)	no	no

Figure F.3. Man-Whitney Test for Raters 9 and 13.

	GT	Rater 16
count	24	24
median	49	46
rank sum	592.5	583.5
U	283.5	292.5

	one tail	two tail
alpha	0.05	
U	283.5	
mean	288	
std dev	48.4974226	
z-score	0.08247861	
effect r	0.01190476	
U-crit	207.728839	192.446798
p-value	0.46713306	0.93426613
sig (norm)	no	no

Figure F.4. Man-Whitney Test for Rater 16.

Appendix G

Cohen's Kappa

Case Study 1

Real Statistics Report

Cohen's Kappa		Cohen's Kappa		Cohen's Kappa	
Rater 1		Rater 2		Rater 3	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.50344828	kappa	0.33501259	kappa	0.54066986
std err	0.13200622	std err	0.14598351	std err	0.12769361
lower	0.24472084	lower	0.04889017	lower	0.29039498
upper	0.76217571	upper	0.62113502	upper	0.79094473

Cohen's Kappa		Cohen's Kappa		Cohen's Kappa	
Rater 4		Rater 5		Rater 6	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.07913669	kappa	0.34545455	kappa	0.58208955
std err	0.1342295	std err	0.14142406	std err	0.13198696
lower	-0.18394829	lower	0.06826848	lower	0.32339986
upper	0.34222167	upper	0.62264061	upper	0.84077925

Cohen's Kappa		Cohen's Kappa		Cohen's Kappa	
Rater 7		Rater 8		Rater 9	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.42720764	kappa	0.362318841	kappa	0.43793911
std err	0.13691811	std err	0.146541739	std err	0.13994993
lower	0.15885307	lower	0.07510231	lower	0.16364229
upper	0.69556221	upper	0.649535372	upper	0.71223593

Cohen's Kappa		Cohen's Kappa		Cohen's Kappa	
Rater 10		Rater 11		Rater 12	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.37882353	kappa	0.33640553	kappa	0.42168675
std err	0.13877254	std err	0.14347956	std err	0.14038318
lower	0.10683434	lower	0.05519076	lower	0.14654076
upper	0.65081272	upper	0.6176203	upper	0.69683273

Figure G.1. Cohen's Kappa Results for Raters 1 to 12.

Cohen's Kappa	
Rater 13	
Alpha	0.05
kappa	0.51515152
std err	0.13510244
lower	0.25035561
upper	0.77994742

Cohen's Kappa	
Rater 14	
Alpha	0.05
kappa	0.56561086
std err	0.12552052
lower	0.31959516
upper	0.81162656

Cohen's Kappa	
Rater 15	
Alpha	0.05
kappa	0.60283688
std err	0.12534763
lower	0.35716004
upper	0.84851372

Cohen's Kappa	
Rater 16	
Alpha	0.05
kappa	0.34164589
std err	0.1482888
lower	0.05100518
upper	0.63228659

Cohen's Kappa	
Rater 17	
Alpha	0.05
kappa	0.25890736
std err	0.14457144
lower	-0.02444746
upper	0.54226218

Cohen's Kappa	
Rater 18	
Alpha	0.05
kappa	0.3699284
std err	0.13694685
lower	0.10151751
upper	0.63833929

Figure G.2. Cohen's Kappa Results for Raters 13 to 18.

Appendix H

Weighted Cohen's Kappa

Case Study 1

Real Statistics Report

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 1		Rater 2		Rater 3	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.708108108	kappa	0.55704698	kappa	0.658227848
std err	0.086313849	std err	0.102927702	std err	0.097698841
lower	0.538936073	lower	0.35531239	lower	0.466741638
upper	0.877280144	upper	0.758781569	upper	0.849714058

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 4		Rater 5		Rater 6	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.392857143	kappa	0.345454545	kappa	0.736677116
std err	0.103983748	std err	0.141424061	std err	0.089498629
lower	0.189052742	lower	0.06826848	lower	0.561263026
upper	0.596661544	upper	0.622640611	upper	0.912091206

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 7		Rater 8		Rater 9	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.586206897	kappa	0.57827476	kappa	0.635258359
std err	0.108929983	std err	0.108391946	std err	0.100341518
lower	0.372708054	lower	0.365830451	lower	0.438592597
upper	0.799705739	upper	0.79071907	upper	0.83192412

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 10		Rater 11		Rater 12	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.564954683	kappa	0.576470588	kappa	0.615384615
std err	0.10701869	std err	0.101957799	std err	0.09881022
lower	0.355201905	lower	0.376636974	lower	0.421720144
upper	0.77470746	upper	0.776304202	upper	0.809049087

Figure H.1. Weighted Cohen's Kappa Results for Raters 1 to 12.

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 13		Rater 14		Rater 15	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.678929766	kappa	0.728813559	kappa	0.756521739
std err	0.093638035	std err	0.087307704	std err	0.087898152
lower	0.49540259	lower	0.557693604	lower	0.584244527
upper	0.862456942	upper	0.899933515	upper	0.928798951

Cohen's Weighted Kappa		Cohen's Weighted Kappa		Cohen's Weighted Kappa	
Rater 16		Rater 17		Rater 18	
Alpha	0.05	Alpha	0.05	Alpha	0.05
kappa	0.561461794	kappa	0.541176471	kappa	0.614035088
std err	0.110516654	std err	0.10142699	std err	0.101049418
lower	0.344853133	lower	0.342383223	lower	0.415981868
upper	0.778070455	upper	0.739969718	upper	0.812088308

Figure H.2. Cohen's Kappa Results for Raters 13 to 18.

Appendix I

Normality Test of the Differences between PCI Values from the Ground Truth and Raters

Shapiro-Wilk Test

QQ Plot

Case Study 1

Real Statistics Report

Shapiro-Wilk Test

	Rater 1
W-stat	0.957775
p-value	0.395372
alpha	0.05
normal	yes

QQ Plot - Rater 1

Count	24	48
Mean	-7	
Std Dev	8.5002	

Interval	Data	Std Norm	Std Data
1	-28	-2.03683	-2.4215
3	-23	-1.53412	-1.8333
5	-20	-1.25816	-1.4804
7	-17	-1.05447	-1.1274
9	-13	-0.88715	-0.6568
11	-10	-0.74159	-0.3039
13	-10	-0.61029	-0.3039
15	-10	-0.48878	-0.3039
17	-10	-0.3741	-0.3039
19	-9	-0.26415	-0.1863
21	-9	-0.15731	-0.1863
23	-8	-0.05225	-0.0686
25	-6	0.052245	0.16666
27	-5	0.157311	0.28431
29	-4	0.264147	0.40195
31	-3	0.374095	0.51959
33	-3	0.488776	0.51959
35	-2	0.610295	0.63724
37	-2	0.741594	0.63724
39	-2	0.887147	0.63724
41	2	1.054472	1.10782
43	3	1.258162	1.22546
45	4	1.534121	1.3431
47	7	2.036834	1.69604

Figure I.1. Shapiro-Wilk Test and QQ Plot Results for Rater 1.

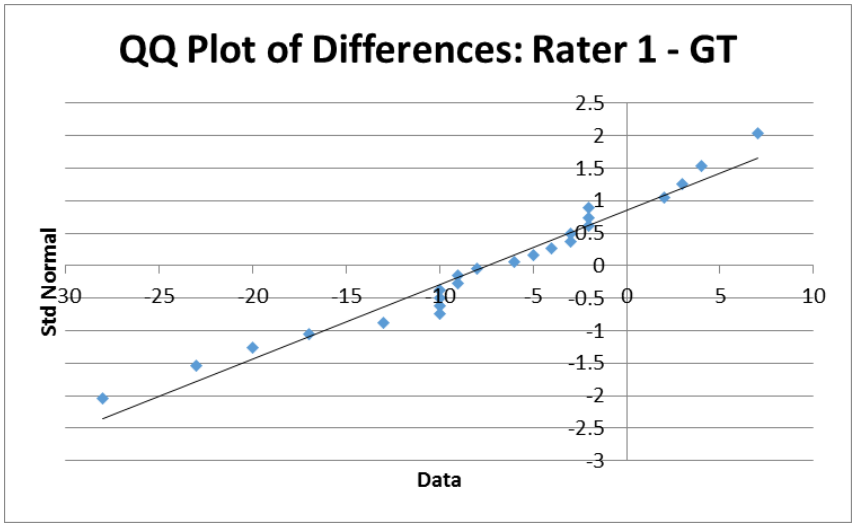


Figure I.2. QQ Plot for Rater 1.

Shapiro-Wilk Test
Rater 2

	Group 1
W-stat	0.952587
p-value	0.307962
alpha	0.05
normal	yes

QQ Plot - Rater 2

Count	24	48
Mean	-4	
Std Dev	12.368	

Interval	Data	Std Norm	Std Data
1	-26	-2.03683	-1.7585
3	-23	-1.53412	-1.516
5	-21	-1.25816	-1.3543
7	-19	-1.05447	-1.1926
9	-18	-0.88715	-1.1117
11	-16	-0.74159	-0.95
13	-16	-0.61029	-0.95
15	-7	-0.48878	-0.2223
17	-7	-0.3741	-0.2223
19	-5	-0.26415	-0.0606
21	-5	-0.15731	-0.0606
23	-4	-0.05225	0.02021
25	-3	0.052245	0.10106
27	-3	0.157311	0.10106
29	-1	0.264147	0.26277
31	0	0.374095	0.34362
33	1	0.488776	0.42447
35	3	0.610295	0.58617
37	4	0.741594	0.66702
39	5	0.887147	0.74787
41	13	1.054472	1.39468
43	15	1.258162	1.55638
45	15	1.534121	1.55638
47	16	2.036834	1.63723

Figure I.3. Shapiro-Wilk Test and QQ Plot Results for Rater 2.

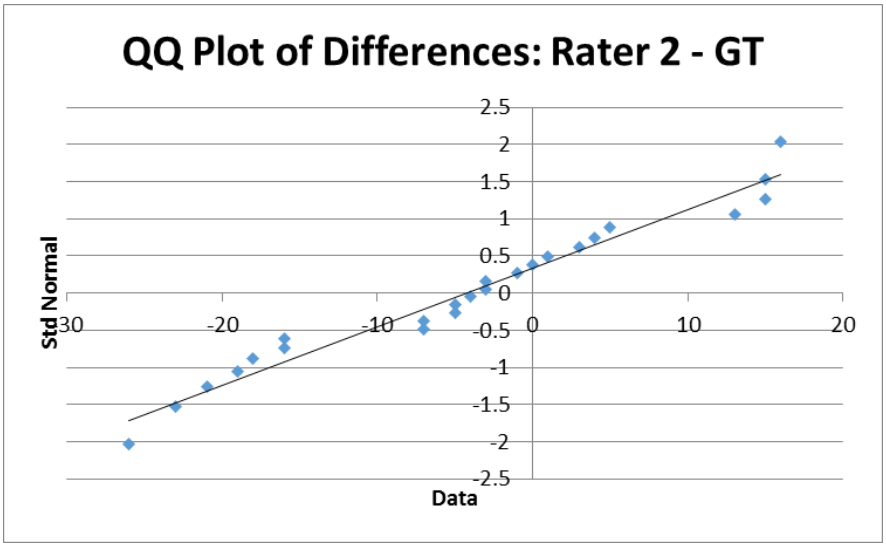


Figure I.4. QQ Plot for Rater 2.

Shapiro-Wilk Test
Rater 3

	Group 1
W-stat	0.946763
p-value	0.230295
alpha	0.05
normal	yes

QQ Plot - Rater 3

Count	24	48
Mean	8	
Std Dev	9.0737	

Interval	Data	Std Norm	Std Data
1	-6	-2.03683	-1.5016
3	-5	-1.53412	-1.3914
5	-3	-1.25816	-1.171
7	-2	-1.05447	-1.0608
9	1	-0.88715	-0.7301
11	1	-0.74159	-0.7301
13	2	-0.61029	-0.6199
15	2	-0.48878	-0.6199
17	4	-0.3741	-0.3995
19	4	-0.26415	-0.3995
21	5	-0.15731	-0.2893
23	7	-0.05225	-0.0689
25	7	0.052245	-0.0689
27	7	0.157311	-0.0689
29	8	0.264147	0.04133
31	8	0.374095	0.04133
33	9	0.488776	0.15154
35	10	0.610295	0.26175
37	17	0.741594	1.03321
39	17	0.887147	1.03321
41	17	1.054472	1.03321
43	21	1.258162	1.47404
45	22	1.534121	1.58425
47	30	2.036834	2.46593

Figure I.5. Shapiro-Wilk Test and QQ Plot Results for Rater 3.

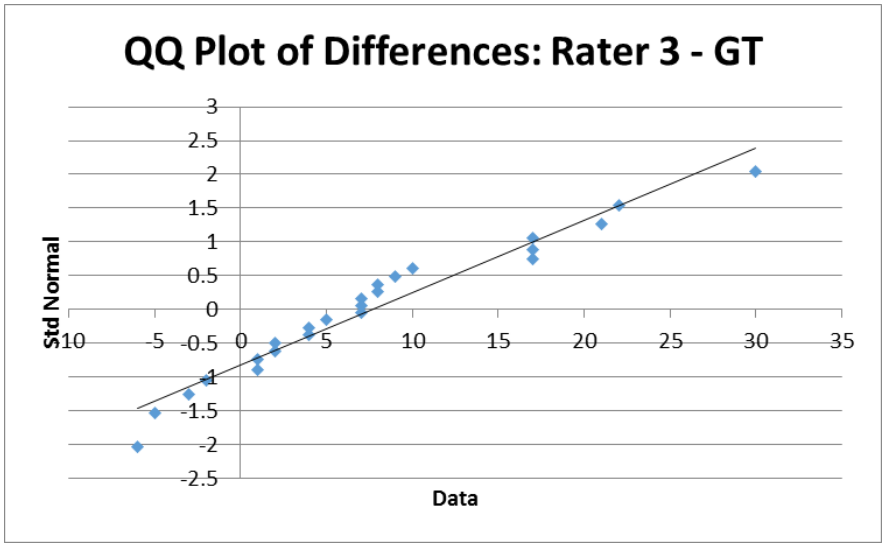


Figure I.6. QQ Plot for Rater 3.

QQ Plot - Rater 4

Count	24	48
Mean	11	
Std Dev	12.964	

Interval	Data	Std Norm	Std Data
1	-20	-2.03683	-2.3623
3	-18	-1.53412	-2.208
5	-9	-1.25816	-1.5138
7	2	-1.05447	-0.6653
9	3	-0.88715	-0.5882
11	5	-0.74159	-0.4339
13	5	-0.61029	-0.4339
15	5	-0.48878	-0.4339
17	6	-0.3741	-0.3568
19	7	-0.26415	-0.2796
21	8	-0.15731	-0.2025
23	10	-0.05225	-0.0482
25	15	0.052245	0.33747
27	18	0.157311	0.56887
29	18	0.264147	0.56887
31	18	0.374095	0.56887
33	19	0.488776	0.64601
35	19	0.610295	0.64601
37	20	0.741594	0.72314
39	21	0.887147	0.80028
41	23	1.054472	0.95455
43	24	1.258162	1.03169
45	28	1.534121	1.34023
47	28	2.036834	1.34023

Shapiro-Wilk Test
Rater 4

	Group 1
W-stat	0.91081
p-value	0.036728
alpha	0.05
normal	no

Figure I.7. Shapiro-Wilk Test and QQ Plot Results for Rater 4.

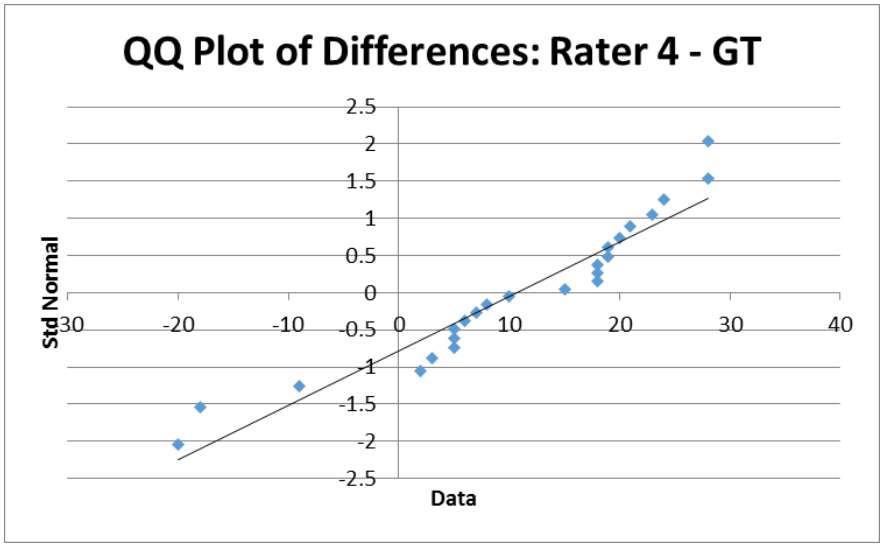


Figure I.8. QQ Plot for Rater 4.

Shapiro-Wilk Test Rater 5	
	Group 1
W-stat	0.898534
p-value	0.02004
alpha	0.05
normal	no

QQ Plot - Rater 5

Count	24	48
Mean	10	
Std Dev	9.0625	

Interval	Data	Std Norm	Std Data
1	-2	-2.03683	-1.3563
3	1	-1.53412	-1.0253
5	1	-1.25816	-1.0253
7	2	-1.05447	-0.9149
9	2	-0.88715	-0.9149
11	3	-0.74159	-0.8046
13	3	-0.61029	-0.8046
15	3	-0.48878	-0.8046
17	4	-0.3741	-0.6943
19	7	-0.26415	-0.3632
21	7	-0.15731	-0.3632
23	7	-0.05225	-0.3632
25	8	0.052245	-0.2529
27	9	0.157311	-0.1425
29	9	0.264147	-0.1425
31	10	0.374095	-0.0322
33	13	0.488776	0.29885
35	19	0.610295	0.96092
37	21	0.741594	1.18161
39	21	0.887147	1.18161
41	22	1.054472	1.29196
43	23	1.258162	1.4023
45	24	1.534121	1.51265
47	30	2.036834	2.17472

Figure I.9. Shapiro-Wilk Test and QQ Plot Results for Rater 5.

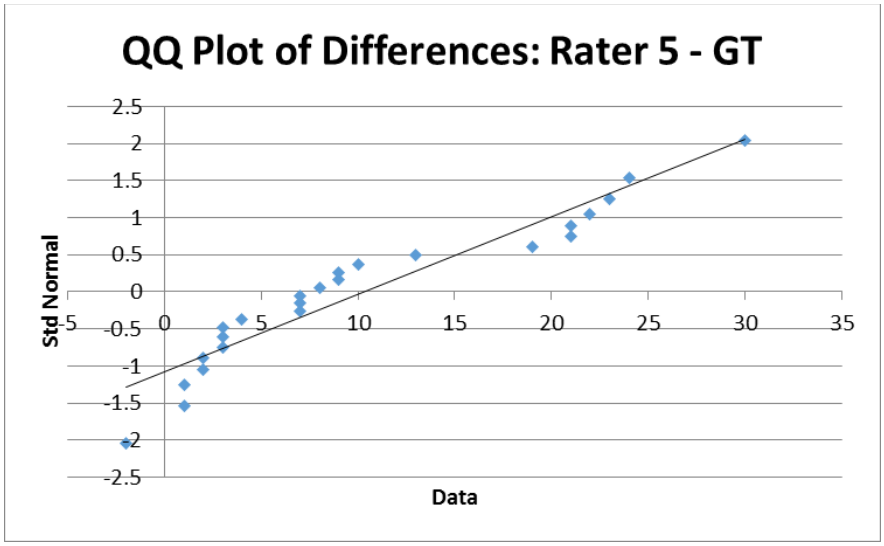


Figure I.10. QQ Plot for Rater 5.

Shapiro-Wilk Test
Rater 6

	Group 1
W-stat	0.949195
p-value	0.260263
alpha	0.05
normal	yes

QQ Plot - Rater 6

Count	24	48
Mean	-3	
Std Dev	11.143	

Interval	Data	Std Norm	Std Data
1	-21	-2.03683	-1.6528
3	-16	-1.53412	-1.204
5	-14	-1.25816	-1.0246
7	-13	-1.05447	-0.9348
9	-12	-0.88715	-0.8451
11	-11	-0.74159	-0.7553
13	-10	-0.61029	-0.6656
15	-9	-0.48878	-0.5758
17	-8	-0.3741	-0.4861
19	-7	-0.26415	-0.3964
21	-6	-0.15731	-0.3066
23	-5	-0.05225	-0.2169
25	-5	0.052245	-0.2169
27	-4	0.157311	-0.1271
29	-1	0.264147	0.14209
31	0	0.374095	0.23183
33	0	0.488776	0.23183
35	2	0.610295	0.41132
37	3	0.741594	0.50106
39	7	0.887147	0.86003
41	11	1.054472	1.219
43	16	1.258162	1.66771
45	18	1.534121	1.8472
47	23	2.036834	2.29591

Figure I.11. Shapiro-Wilk Test and QQ Plot Results for Rater 6.

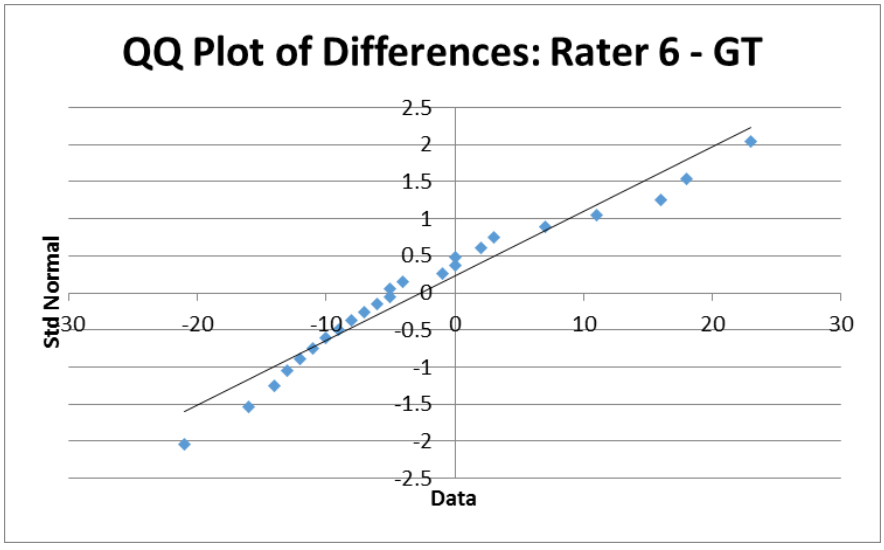


Figure I.12. QQ Plot for Rater 6.

Shapiro-Wilk Test Rater 7	
	Group 1
W-stat	0.956994
p-value	0.381058
alpha	0.05
normal	yes

QQ Plot - Rater 7

Count	24	48
Mean	0	
Std Dev	15.405	

Interval	Data	Std Norm	Std Data
1	-26	-2.03683	-1.6715
3	-18	-1.53412	-1.1522
5	-16	-1.25816	-1.0224
7	-15	-1.05447	-0.9575
9	-14	-0.88715	-0.8925
11	-11	-0.74159	-0.6978
13	-11	-0.61029	-0.6978
15	-10	-0.48878	-0.6329
17	-10	-0.3741	-0.6329
19	-10	-0.26415	-0.6329
21	-9	-0.15731	-0.568
23	-4	-0.05225	-0.2434
25	-1	0.052245	-0.0487
27	1	0.157311	0.08114
29	3	0.264147	0.21097
31	6	0.374095	0.4057
33	7	0.488776	0.47061
35	9	0.610295	0.60044
37	10	0.741594	0.66535
39	16	0.887147	1.05483
41	16	1.054472	1.05483
43	19	1.258162	1.24956
45	25	1.534121	1.63904
47	37	2.036834	2.41798

Figure I.13. Shapiro-Wilk Test and QQ Plot Results for Rater 7.

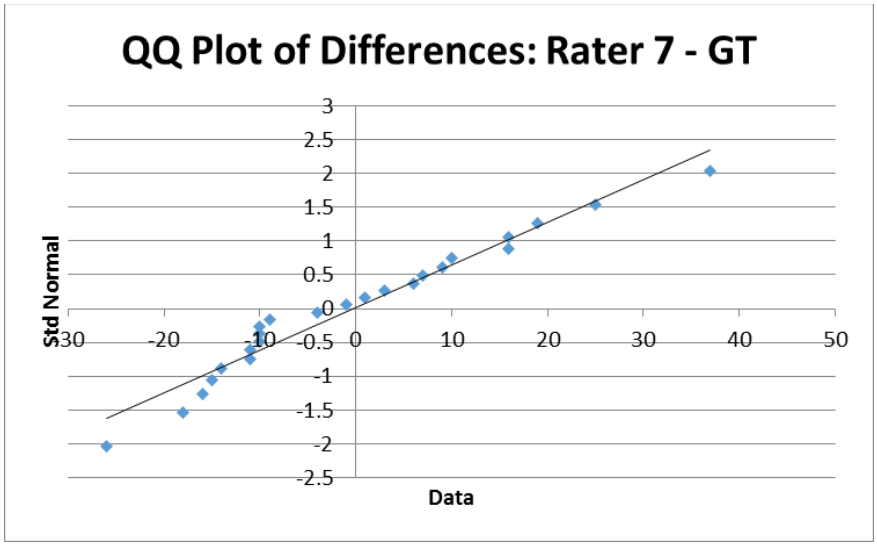


Figure I.14. QQ Plot for Rater 7.

Shapiro-Wilk Test
Rater 8

	Group 1
W-stat	0.963514
p-value	0.512975
alpha	0.05
normal	yes

QQ Plot - Rater 8

Count	24	48
Mean	4	
Std Dev	9.2289	

Interval	Data	Std Norm	Std Data
1	-12	-2.03683	-1.7653
3	-12	-1.53412	-1.7653
5	-10	-1.25816	-1.5486
7	-9	-1.05447	-1.4402
9	-2	-0.88715	-0.6817
11	0	-0.74159	-0.465
13	1	-0.61029	-0.3567
15	2	-0.48878	-0.2483
17	2	-0.3741	-0.2483
19	2	-0.26415	-0.2483
21	2	-0.15731	-0.2483
23	3	-0.05225	-0.14
25	5	0.052245	0.07675
27	5	0.157311	0.07675
29	7	0.264147	0.29346
31	7	0.374095	0.29346
33	8	0.488776	0.40182
35	9	0.610295	0.51017
37	10	0.741594	0.61853
39	12	0.887147	0.83524
41	16	1.054472	1.26866
43	16	1.258162	1.26866
45	18	1.534121	1.48537
47	23	2.036834	2.02715

Figure I.15. Shapiro-Wilk Test and QQ Plot Results for Rater 8.

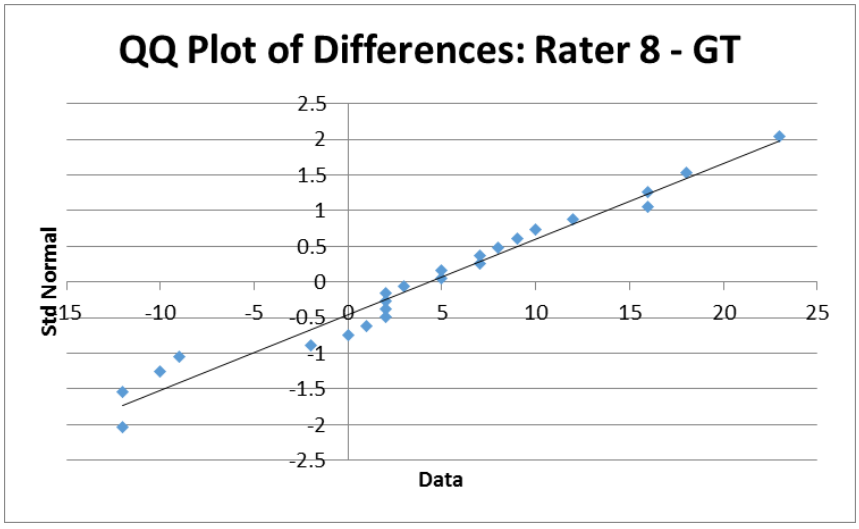


Figure I.16. QQ Plot for Rater 8.

QQ Plot - Rater 9

Count	24	48
Mean	4	
Std Dev	9.7976	

Interval	Data	Std Norm	Std Data
1	-17	-2.03683	-2.1859
3	-15	-1.53412	-1.9818
5	-8	-1.25816	-1.2673
7	-7	-1.05447	-1.1653
9	-3	-0.88715	-0.757
11	-2	-0.74159	-0.6549
13	1	-0.61029	-0.3487
15	3	-0.48878	-0.1446
17	3	-0.3741	-0.1446
19	4	-0.26415	-0.0425
21	4	-0.15731	-0.0425
23	5	-0.05225	0.05954
25	5	0.052245	0.05954
27	6	0.157311	0.1616
29	7	0.264147	0.26367
31	9	0.374095	0.4678
33	10	0.488776	0.56987
35	10	0.610295	0.56987
37	11	0.741594	0.67193
39	11	0.887147	0.67193
41	11	1.054472	0.67193
43	13	1.258162	0.87607
45	19	1.534121	1.48846
47	26	2.036834	2.20292

Shapiro-Wilk Test

Rater 9

	Group 1
W-stat	0.962354
p-value	0.487524
alpha	0.05
normal	yes

Figure I.17. Shapiro-Wilk Test and QQ Plot Results for Rater 9.

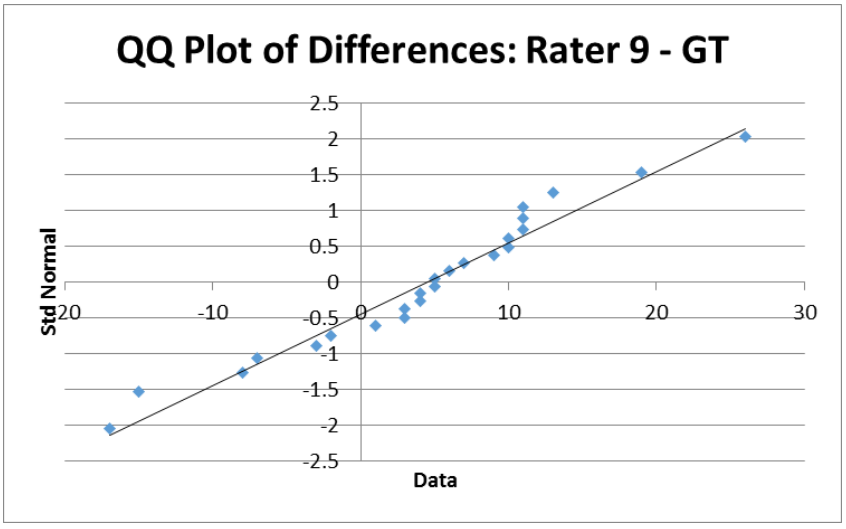


Figure I.18. QQ Plot for Rater 9.

Shapiro-Wilk Test Rater 10	
	Group 1
W-stat	0.936634
p-value	0.137197
alpha	0.05
normal	yes

QQ Plot - Rater 10

Count	24	48
Mean	7	
Std Dev	10.59	

Interval	Data	Std Norm	Std Data
1	-12	-2.03683	-1.8295
3	-8	-1.53412	-1.4518
5	-3	-1.25816	-0.9797
7	-2	-1.05447	-0.8852
9	0	-0.88715	-0.6964
11	1	-0.74159	-0.602
13	2	-0.61029	-0.5075
15	3	-0.48878	-0.4131
17	3	-0.3741	-0.4131
19	5	-0.26415	-0.2243
21	5	-0.15731	-0.2243
23	6	-0.05225	-0.1298
25	6	0.052245	-0.1298
27	7	0.157311	-0.0354
29	8	0.264147	0.05902
31	9	0.374095	0.15344
33	11	0.488776	0.34229
35	11	0.610295	0.34229
37	11	0.741594	0.34229
39	13	0.887147	0.53114
41	14	1.054472	0.62556
43	25	1.258162	1.66424
45	26	1.534121	1.75866
47	36	2.036834	2.70291

Figure I.19. Shapiro-Wilk Test and QQ Plot Results for Rater 10.

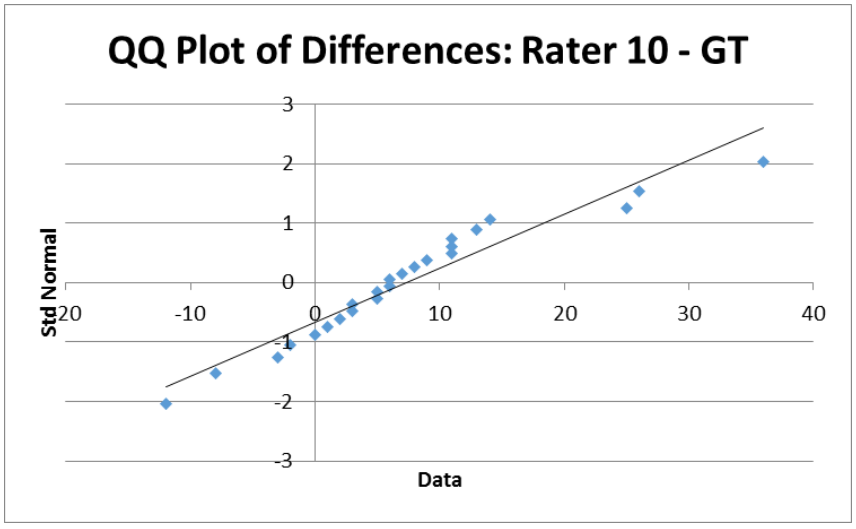


Figure I.20. QQ Plot for Rater 10.

Shapiro-Wilk Test Rater 11	
	Group 1
W-stat	0.961319
p-value	0.465502
alpha	0.05
normal	yes

QQ Plot - Rater 11

Count	24	48
Mean	9	
Std Dev	10.521	

Interval	Data	Std Norm	Std Data
1	-16	-2.03683	-2.3683
3	-11	-1.53412	-1.8931
5	-2	-1.25816	-1.0376
7	0	-1.05447	-0.8475
9	2	-0.88715	-0.6574
11	3	-0.74159	-0.5624
13	3	-0.61029	-0.5624
15	4	-0.48878	-0.4673
17	5	-0.3741	-0.3723
19	6	-0.26415	-0.2772
21	7	-0.15731	-0.1822
23	7	-0.05225	-0.1822
25	9	0.052245	0.00792
27	10	0.157311	0.10297
29	10	0.264147	0.10297
31	13	0.374095	0.38812
33	17	0.488776	0.76832
35	18	0.610295	0.86336
37	18	0.741594	0.86336
39	20	0.887147	1.05346
41	20	1.054472	1.05346
43	22	1.258162	1.24356
45	23	1.534121	1.33861
47	26	2.036834	1.62376

Figure I.21. Shapiro-Wilk Test and QQ Plot Results for Rater 11.

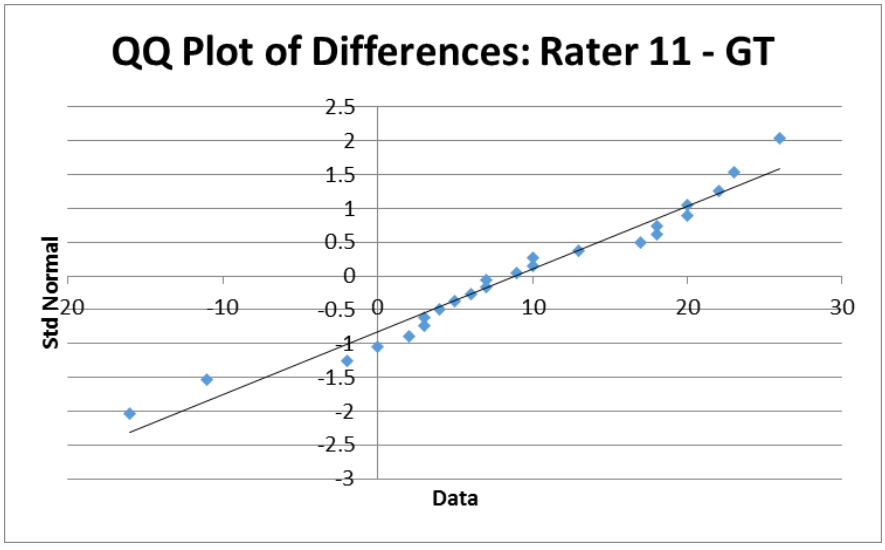


Figure I.22. QQ Plot for Rater 11.

Shapiro-Wilk Test Rater 12	
	Group 1
W-stat	0.941383
p-value	0.175115
alpha	0.05
normal	yes

QQ Plot - Rater 12

Count	24	48
Mean	0	
Std Dev	13.342	

Interval	Data	Std Norm	Std Data
1	-24	-2.03683	-1.8175
3	-17	-1.53412	-1.2929
5	-15	-1.25816	-1.143
7	-14	-1.05447	-1.068
9	-12	-0.88715	-0.9181
11	-12	-0.74159	-0.9181
13	-11	-0.61029	-0.8432
15	-11	-0.48878	-0.8432
17	-7	-0.3741	-0.5434
19	-5	-0.26415	-0.3935
21	-2	-0.15731	-0.1686
23	2	-0.05225	0.13116
25	2	0.052245	0.13116
27	3	0.157311	0.20611
29	4	0.264147	0.28106
31	5	0.374095	0.35601
33	5	0.488776	0.35601
35	7	0.610295	0.5059
37	8	0.741594	0.58085
39	18	0.887147	1.33034
41	20	1.054472	1.48023
43	20	1.258162	1.48023
45	21	1.534121	1.55518
47	21	2.036834	1.55518

Figure I.23. Shapiro-Wilk Test and QQ Plot Results for Rater 12.

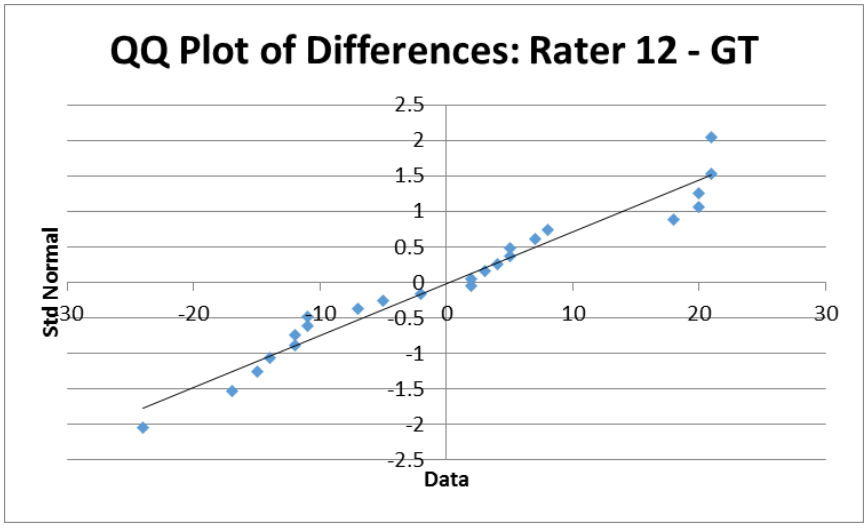


Figure I.24. QQ Plot for Rater 12.

Shapiro-Wilk Test Rater 13	
	Group 1
W-stat	0.933343
p-value	0.115806
alpha	0.05
normal	yes

QQ Plot - Rater 13

Count	24	48
Mean	-1	
Std Dev	11.876	

Interval	Data	Std Norm	Std Data
1	-25	-2.03683	-2.0209
3	-23	-1.53412	-1.8524
5	-22	-1.25816	-1.7682
7	-18	-1.05447	-1.4314
9	-12	-0.88715	-0.9262
11	-7	-0.74159	-0.5052
13	-3	-0.61029	-0.1684
15	-2	-0.48878	-0.0842
17	-2	-0.3741	-0.0842
19	-1	-0.26415	0
21	0	-0.15731	0.0842
23	0	-0.05225	0.0842
25	1	0.052245	0.1684
27	3	0.157311	0.33681
29	3	0.264147	0.33681
31	4	0.374095	0.42101
33	4	0.488776	0.42101
35	5	0.610295	0.50521
37	5	0.741594	0.50521
39	6	0.887147	0.58942
41	10	1.054472	0.92622
43	13	1.258162	1.17883
45	17	1.534121	1.51564
47	20	2.036834	1.76825

Figure I.25. Shapiro-Wilk Test and QQ Plot Results for Rater 13.

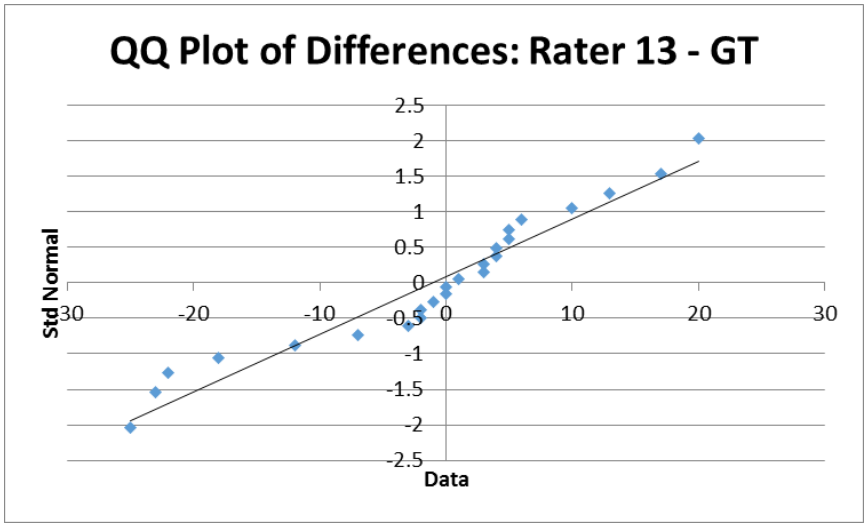


Figure I.26. QQ Plot for Rater 13.

Shapiro-Wilk Test Rater 14	
	Group 1
W-stat	0.923059
p-value	0.068272
alpha	0.05
normal	yes

QQ Plot - Rater 14

Count	24	48
Mean	5	
Std Dev	7.261	

Interval	Data	Std Norm	Std Data
1	-18	-2.03683	-3.1504
3	-4	-1.53412	-1.2223
5	-1	-1.25816	-0.8091
7	0	-1.05447	-0.6714
9	0	-0.88715	-0.6714
11	1	-0.74159	-0.5337
13	2	-0.61029	-0.3959
15	3	-0.48878	-0.2582
17	3	-0.3741	-0.2582
19	3	-0.26415	-0.2582
21	4	-0.15731	-0.1205
23	4	-0.05225	-0.1205
25	5	0.052245	0.01722
27	7	0.157311	0.29266
29	7	0.264147	0.29266
31	8	0.374095	0.43038
33	8	0.488776	0.43038
35	8	0.610295	0.43038
37	9	0.741594	0.5681
39	10	0.887147	0.70582
41	11	1.054472	0.84354
43	12	1.258162	0.98126
45	16	1.534121	1.53215
47	19	2.036834	1.94531

Figure I.27. Shapiro-Wilk Test and QQ Plot Results for Rater 14.

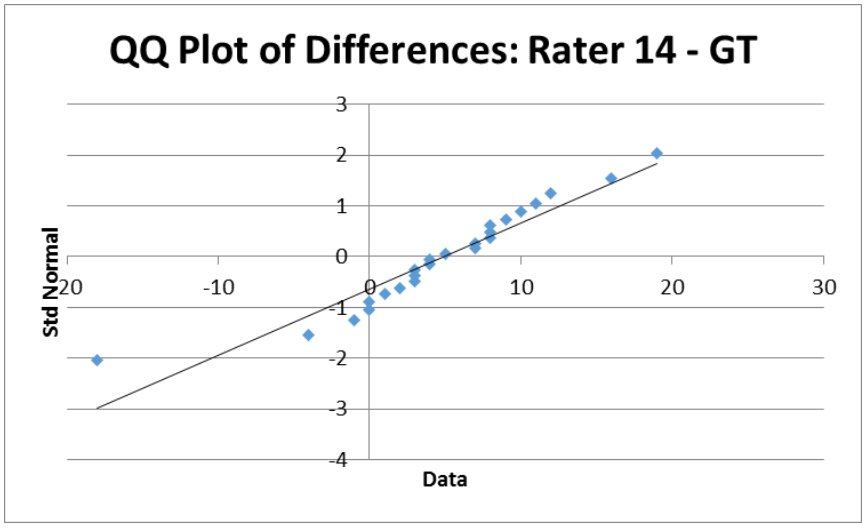


Figure I.28. QQ Plot for Rater 14.

Shapiro-Wilk Test Rater 15	
	Group 1
W-stat	0.927871
p-value	0.087375
alpha	0.05
normal	yes

QQ Plot - Rater 15

Count	24	48
Mean	0	
Std Dev	6.6101	

Interval	Data	Std Norm	Std Data
1	-13	-2.03683	-1.973
3	-12	-1.53412	-1.8217
5	-10	-1.25816	-1.5191
7	-8	-1.05447	-1.2166
9	-8	-0.88715	-1.2166
11	-5	-0.74159	-0.7627
13	-5	-0.61029	-0.7627
15	-3	-0.48878	-0.4602
17	-1	-0.3741	-0.1576
19	0	-0.26415	-0.0063
21	0	-0.15731	-0.0063
23	1	-0.05225	0.14498
25	2	0.052245	0.29626
27	3	0.157311	0.44755
29	3	0.264147	0.44755
31	5	0.374095	0.75011
33	5	0.488776	0.75011
35	5	0.610295	0.75011
37	6	0.741594	0.90139
39	6	0.887147	0.90139
41	6	1.054472	0.90139
43	7	1.258162	1.05268
45	7	1.534121	1.05268
47	10	2.036834	1.50653

Figure I.29. Shapiro-Wilk Test and QQ Plot Results for Rater 15.

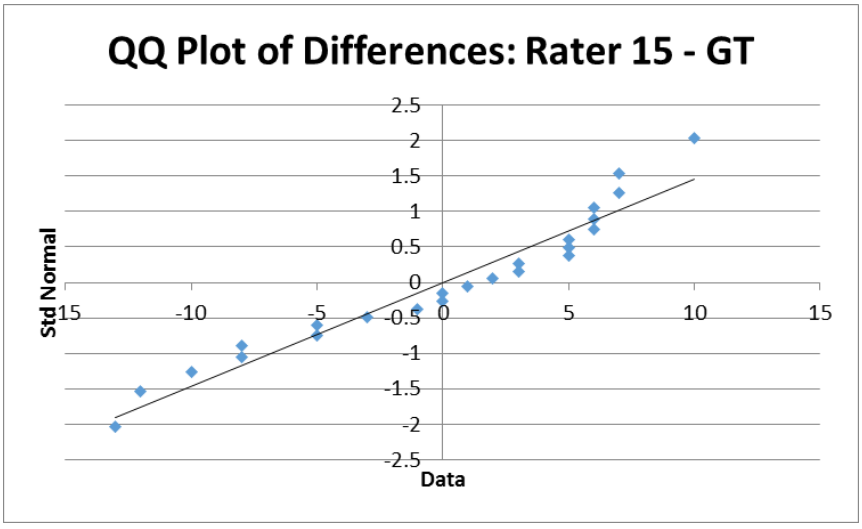


Figure I.30. QQ Plot for Rater 15.

Shapiro-Wilk Test
Rater 16

	Group 1
W-stat	0.979328
p-value	0.88316
alpha	0.05
normal	yes

QQ Plot - Rater 16

Count	24	48
Mean	0	
Std Dev	10.909	

Interval	Data	Std Norm	Std Data
1	-24	-2.03683	-2.1847
3	-19	-1.53412	-1.7263
5	-14	-1.25816	-1.268
7	-14	-1.05447	-1.268
9	-11	-0.88715	-0.993
11	-5	-0.74159	-0.443
13	-4	-0.61029	-0.3514
15	-4	-0.48878	-0.3514
17	-2	-0.3741	-0.1681
19	-2	-0.26415	-0.1681
21	-2	-0.15731	-0.1681
23	0	-0.05225	0.01528
25	0	0.052245	0.01528
27	1	0.157311	0.10694
29	3	0.264147	0.29027
31	3	0.374095	0.29027
33	5	0.488776	0.4736
35	6	0.610295	0.56526
37	8	0.741594	0.74859
39	11	0.887147	1.02358
41	11	1.054472	1.02358
43	12	1.258162	1.11525
45	17	1.534121	1.57357
47	20	2.036834	1.84856

Figure I.31. Shapiro-Wilk Test and QQ Plot Results for Rater 16.

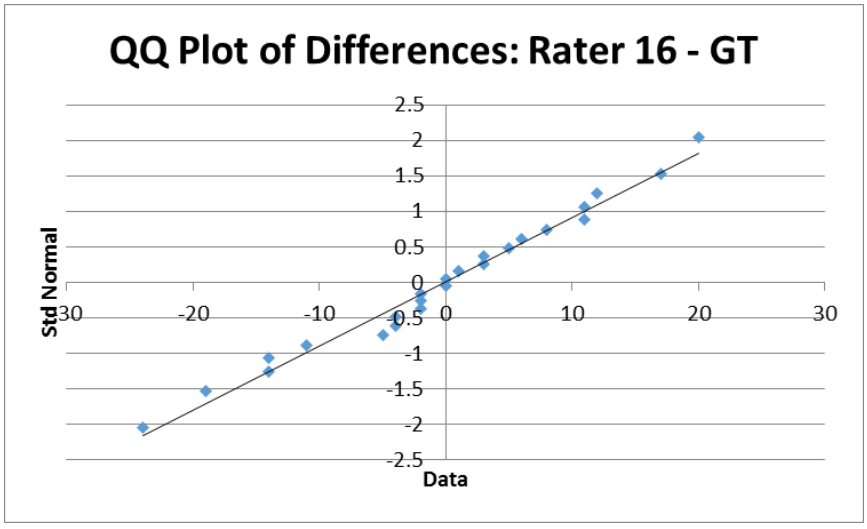


Figure I.32. QQ Plot for Rater 16.

Shapiro-Wilk Test
Rater 17

	Group 1
W-stat	0.947015
p-value	0.233247
alpha	0.05
normal	yes

QQ Plot - Rater 17

Count	24	48
Mean	-12	
Std Dev	13.008	

Interval	Data	Std Norm	Std Data
1	-33	-2.03683	-1.5823
3	-27	-1.53412	-1.1211
5	-26	-1.25816	-1.0442
7	-25	-1.05447	-0.9673
9	-25	-0.88715	-0.9673
11	-23	-0.74159	-0.8136
13	-22	-0.61029	-0.7367
15	-21	-0.48878	-0.6598
17	-20	-0.3741	-0.583
19	-19	-0.26415	-0.5061
21	-16	-0.15731	-0.2755
23	-16	-0.05225	-0.2755
25	-12	0.052245	0.03203
27	-11	0.157311	0.10891
29	-10	0.264147	0.18578
31	-8	0.374095	0.33953
33	-8	0.488776	0.33953
35	-7	0.610295	0.41641
37	-3	0.741594	0.72391
39	-2	0.887147	0.80078
41	-1	1.054472	0.87766
43	4	1.258162	1.26204
45	15	1.534121	2.10766
47	18	2.036834	2.33829

Figure I.33. Shapiro-Wilk Test and QQ Plot Results for Rater 17.

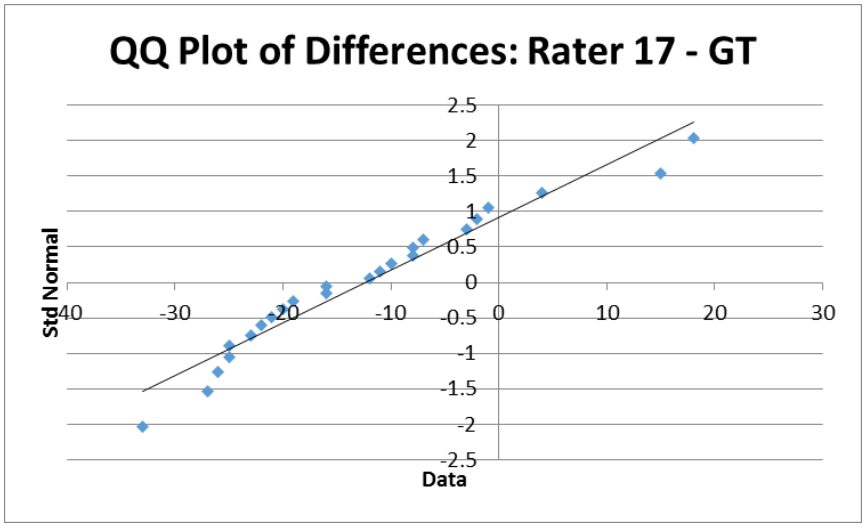


Figure I.34. QQ Plot for Rater 17.

Shapiro-Wilk Test
Rater 18

	Group 1
W-stat	0.915473
p-value	0.046429
alpha	0.05
normal	no

QQ Plot - Rater 18

Count	24	48
Mean	-5	
Std Dev	11.115	

Interval	Data	Std Norm	Std Data
1	-26	-2.03683	-1.9194
3	-25	-1.53412	-1.8294
5	-19	-1.25816	-1.2896
7	-17	-1.05447	-1.1096
9	-17	-0.88715	-1.1096
11	-16	-0.74159	-1.0197
13	-14	-0.61029	-0.8397
15	-11	-0.48878	-0.5698
17	-10	-0.3741	-0.4798
19	-6	-0.26415	-0.12
21	-5	-0.15731	-0.03
23	-2	-0.05225	0.23992
25	-1	0.052245	0.32989
27	0	0.157311	0.41986
29	1	0.264147	0.50984
31	3	0.374095	0.68978
33	4	0.488776	0.77975
35	5	0.610295	0.86972
37	6	0.741594	0.95969
39	6	0.887147	0.95969
41	6	1.054472	0.95969
43	8	1.258162	1.13963
45	9	1.534121	1.2296
47	9	2.036834	1.2296

Figure I.35. Shapiro-Wilk Test and QQ Plot Results for Rater 18.

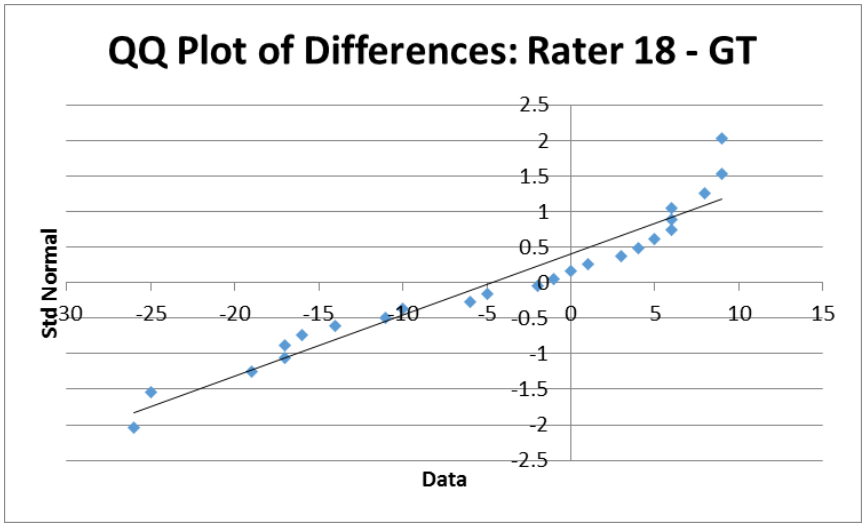


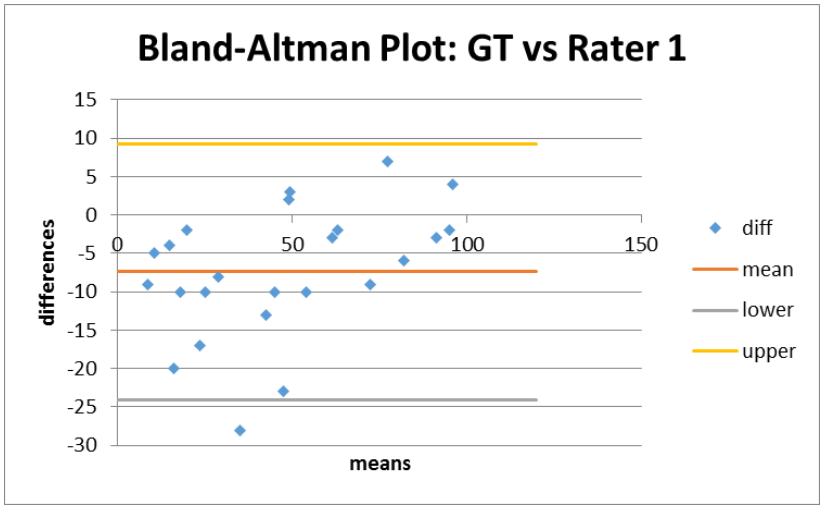
Figure I.36. QQ Plot for Rater 18.

Appendix J

Bland-Altman Diagrams

Case Study 1

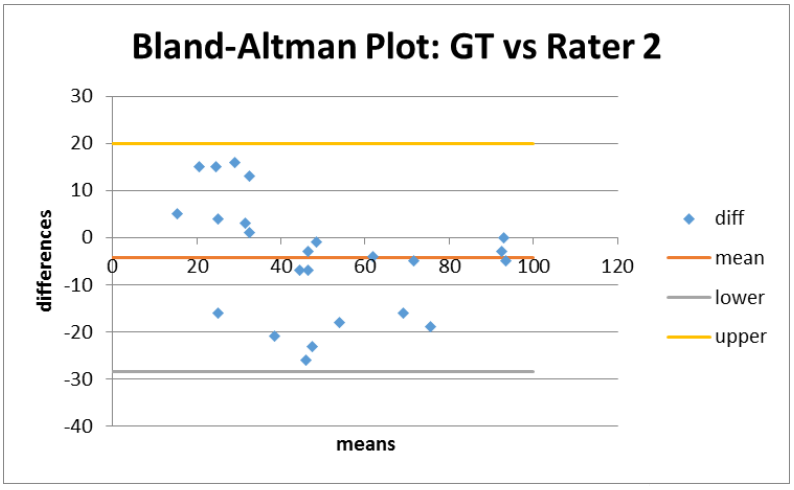
Real Statistics Report



	value	*s.e.*	*lower*	*upper*
mean diff	-7.4167	1.735	-11.00599	-3.827
lower limit	-24.077	3.007	-30.29806	-17.86
upper limit	9.2434	3.007	3.02216	15.465

stdev diff	8.5002
sample size	24
alpha	0.05

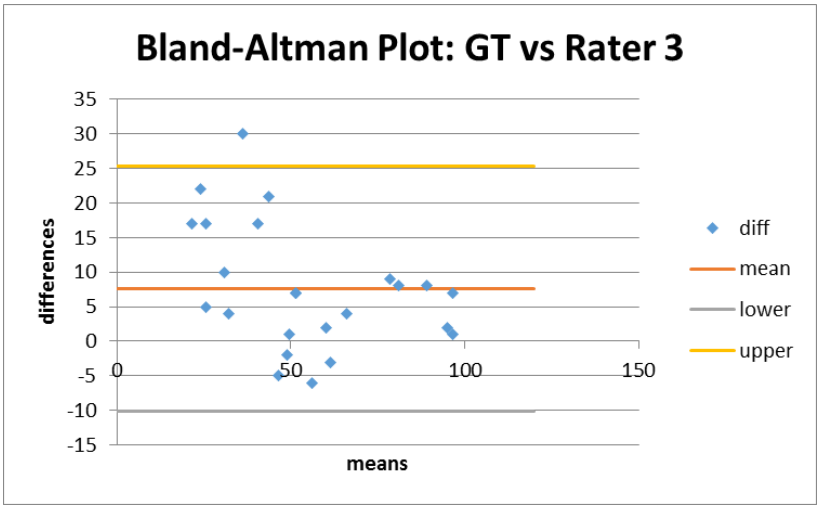
Figure J.1. Bland-Altman Diagram for Rater 1.



	value	*s.e.*	*lower*	*upper*
mean diff	-4.25	2.525	-9.472733	0.9727
lower limit	-28.492	4.376	-37.54412	-19.44
upper limit	19.992	4.376	10.93926	29.044

stdev diff	12.368
sample size	24
alpha	0.05

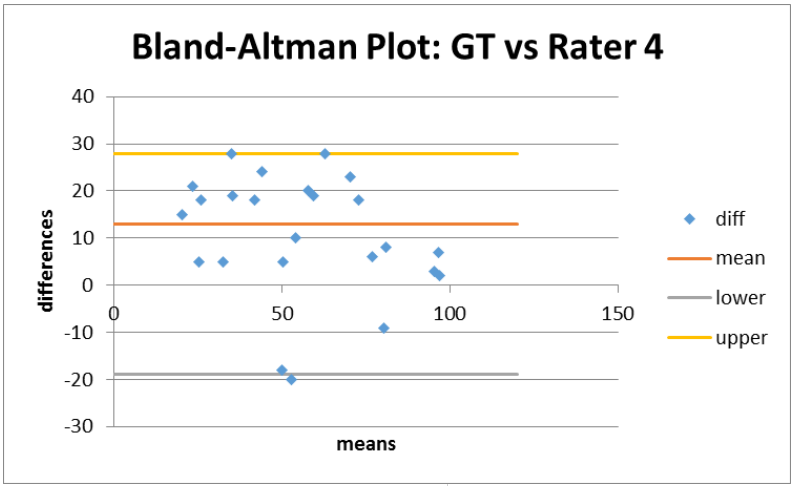
Figure J.2. Bland-Altman Diagram for Rater 2.



	value	*s.e.*	*lower*	*upper*
mean diff	7.625	1.852	3.793524	11.456
lower limit	-10.159	3.21	-16.80007	-3.518
upper limit	25.409	3.21	18.76807	32.05

stdev diff	9.0737
sample size	24
alpha	0.05

Figure J.3. Bland-Altman Diagram for Rater 3.

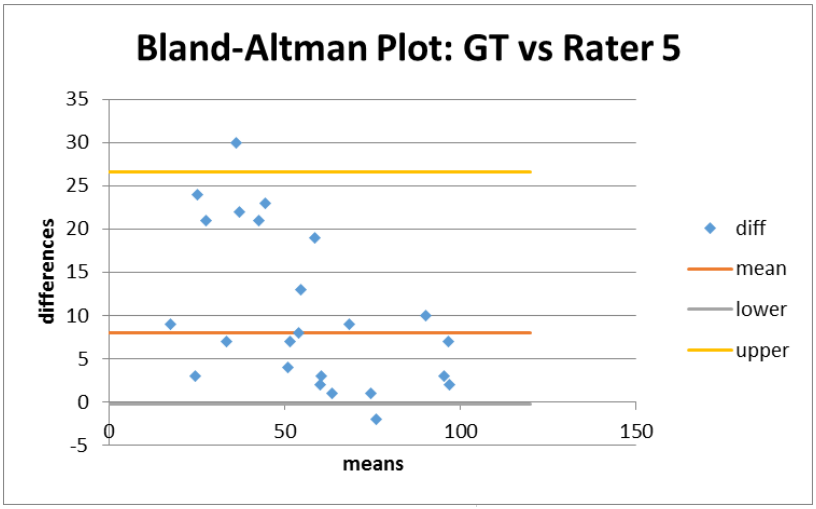


	value	*s.e.*	*lower*	*upper*
mean diff	10.625	2.646	5.150696	16.099
lower limit	-14.784	4.587	-24.27285	-5.296
upper limit	36.034	4.587	26.54591	45.523

stdev diff	12.964
sample size	24
alpha	0.05

13	Median
-18.85	2.5th percentile
28	97.5th percentile

Figure J.4. Bland-Altman Diagram for Rater 4.

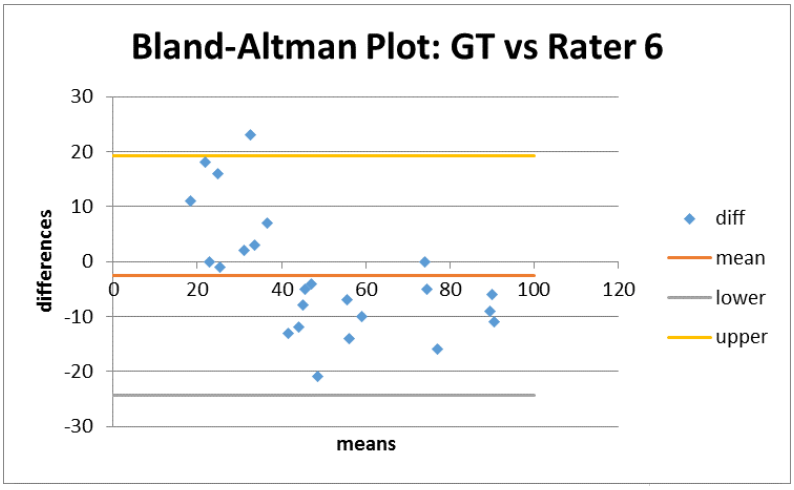


	value	*s.e.*	*lower*	*upper*
mean diff	10.292	1.85	6.464915	14.118
lower limit	-7.4705	3.206	-14.10329	-0.838
upper limit	28.054	3.206	21.421	34.687

stdev diff	9.0625
sample size	24
alpha	0.05

8	Median
-0.275	2.5th percentile
26.55	97.5th percentile

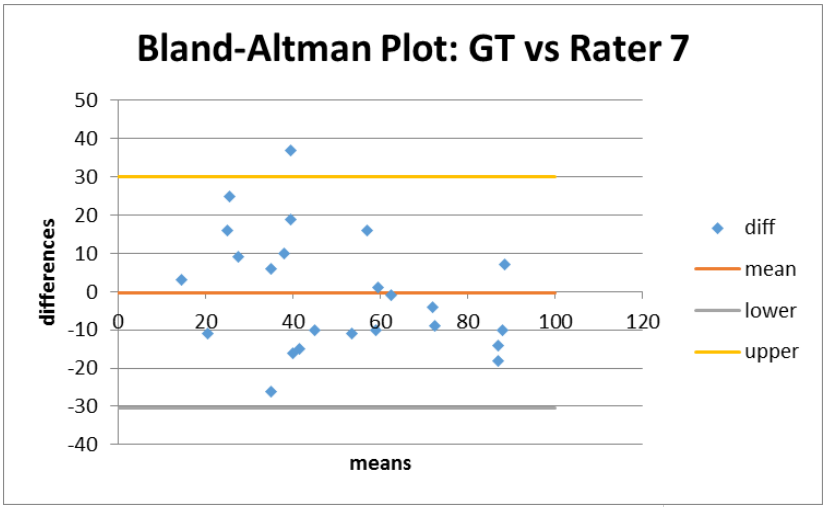
Figure J.5. Bland-Altman Diagram for Rater 5.



	value	*s.e.*	*lower*	*upper*
mean diff	-2.5833	2.275	-7.288614	2.1219
lower limit	-24.423	3.942	-32.57877	-16.27
upper limit	19.257	3.942	11.10102	27.412

stdev diff	11.143
sample size	24
alpha	0.05

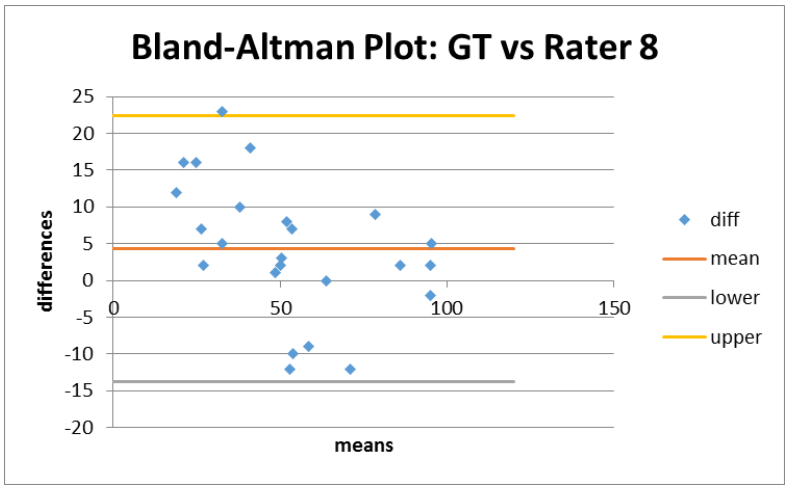
Figure J.6. Bland-Altman Diagram for Rater 6.



	value	s.e.	lower	upper
mean diff	-0.25	3.145	-6.755126	6.2551
lower limit	-30.444	5.45	-41.71918	-19.17
upper limit	29.944	5.45	18.66885	41.219

stdev diff	15.405
sample size	24
alpha	0.05

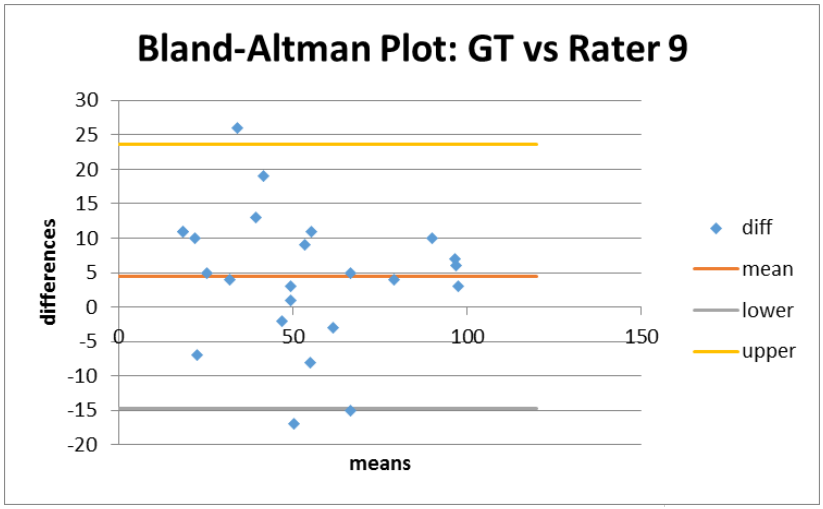
Figure J.7. Bland-Altman Diagram for Rater 7.



	value	s.e.	lower	upper
mean diff	4.2917	1.884	0.394655	8.1887
lower limit	-13.797	3.265	-20.55118	-7.042
upper limit	22.38	3.265	15.62534	29.135

stdev diff	9.2289
sample size	24
alpha	0.05

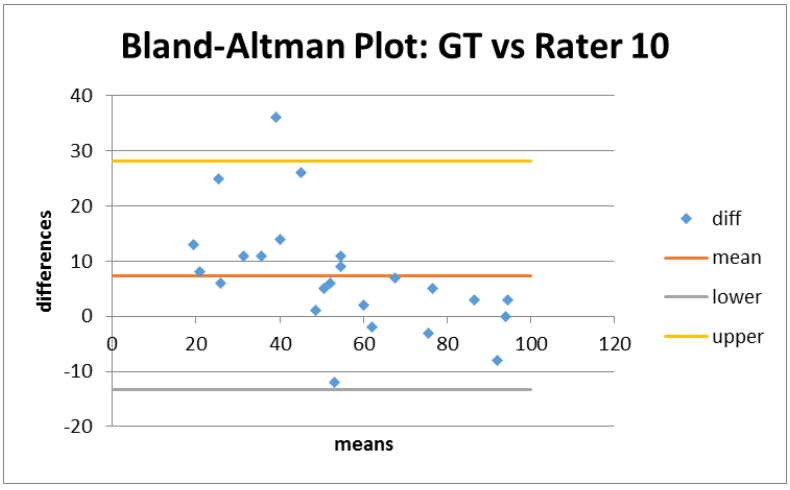
Figure J.8. Bland-Altman Diagram for Rater 8.



	value	*s.e.*	*lower*	*upper*
mean diff	4.4167	2	0.279508	8.5538
lower limit	-14.786	3.466	-21.95709	-7.615
upper limit	23.62	3.466	16.44876	30.79

stdev diff	9.7976
sample size	24
alpha	0.05

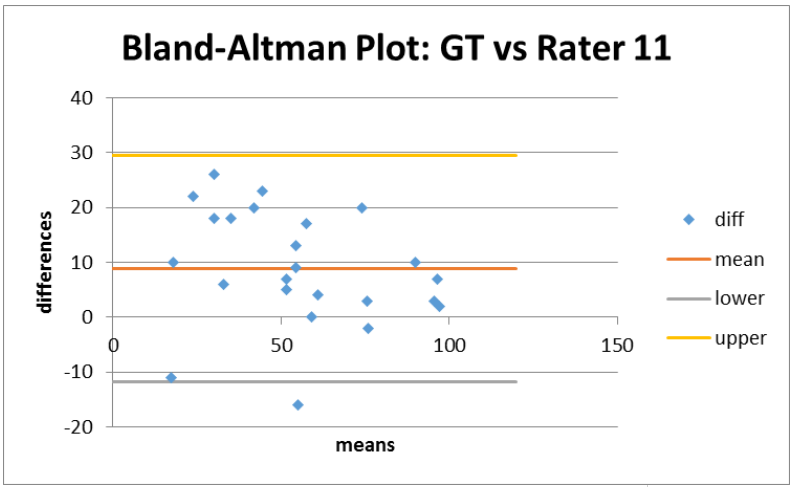
Figure J.9. Bland-Altman Diagram for Rater 9.



	value	*s.e.*	*lower*	*upper*
mean diff	7.375	2.162	2.903046	11.847
lower limit	-13.382	3.747	-21.13302	-5.631
upper limit	28.132	3.747	20.38078	35.883

stdev diff	10.59
sample size	24
alpha	0.05

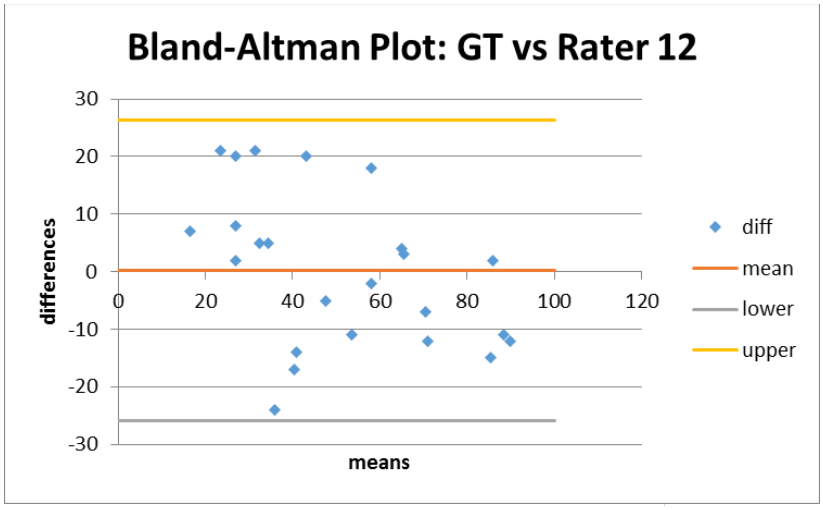
Figure J.10. Bland-Altman Diagram for Rater 10.



	value	*s.e.*	*lower*	*upper*
mean diff	8.9167	2.148	4.474099	13.359
lower limit	-11.704	3.722	-19.40402	-4.004
upper limit	29.537	3.722	21.83698	37.237

stdev diff	10.521
sample size	24
alpha	0.05

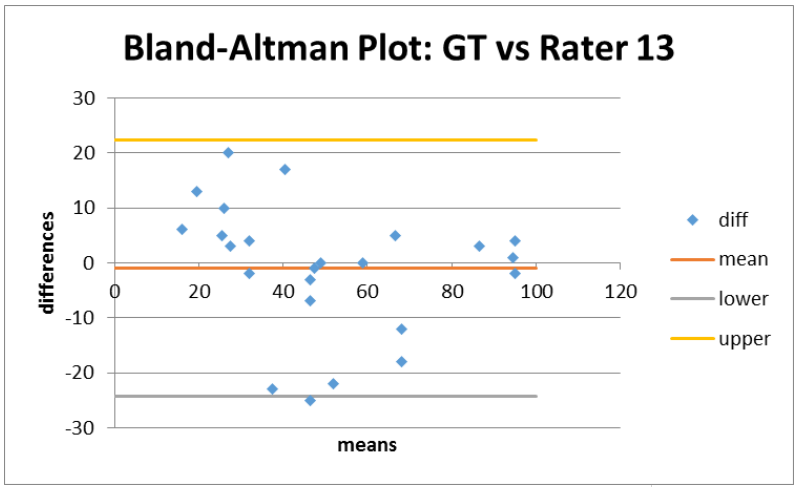
Figure J.11. Bland-Altman Diagram for Rater 11.



	value	*s.e.*	*lower*	*upper*
mean diff	0.25	2.724	-5.384035	5.884
lower limit	-25.901	4.721	-35.6661	-16.14
upper limit	26.401	4.721	16.63545	36.166

stdev diff	13.342
sample size	24
alpha	0.05

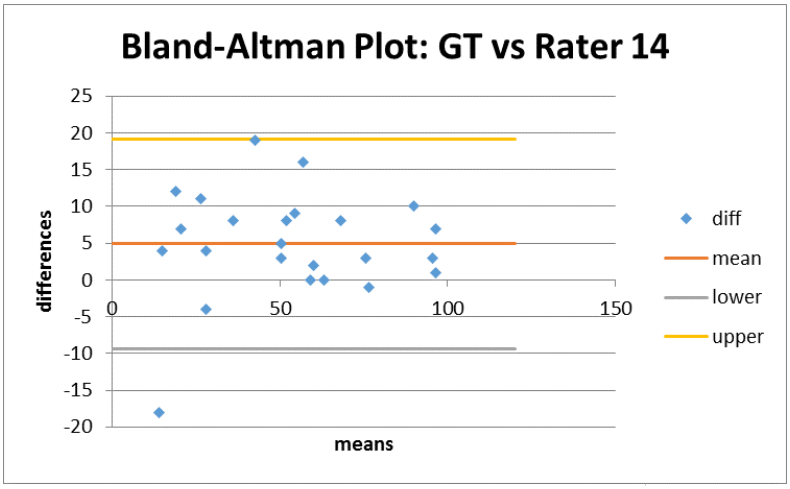
Figure J.12. Bland-Altman Diagram for Rater 12.



	value	*s.e.*	*lower*	*upper*
mean diff	-1	2.424	-6.014868	4.0149
lower limit	-24.277	4.202	-32.96901	-15.58
upper limit	22.277	4.202	13.58473	30.969

stdev diff	11.876
sample size	24
alpha	0.05

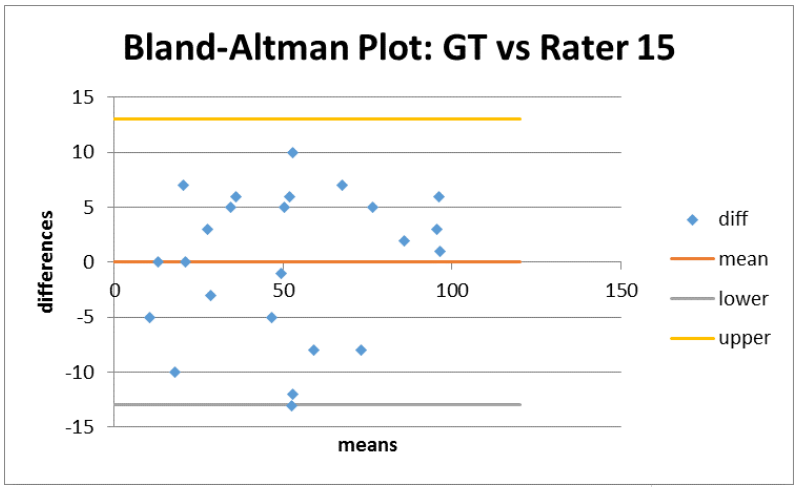
Figure J.13. Bland-Altman Diagram for Rater 13.



	value	*s.e.*	*lower*	*upper*
mean diff	4.875	1.482	1.808928	7.9411
lower limit	-9.3564	2.569	-14.67074	-4.042
upper limit	19.106	2.569	13.79205	24.421

stdev diff	7.261
sample size	24
alpha	0.05

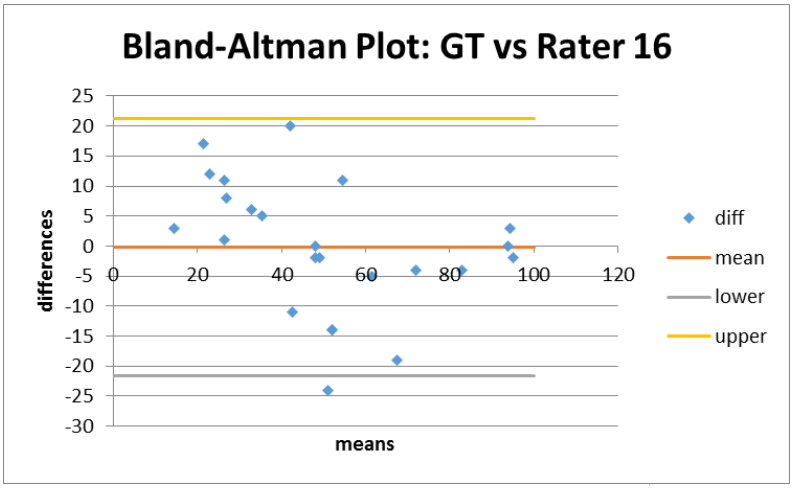
Figure J.14. Bland-Altman Diagram for Rater 14.



	value	*s.e.*	*lower*	*upper*
mean diff	0.0417	1.349	-2.749547	2.8329
lower limit	-12.914	2.339	-17.75189	-8.076
upper limit	12.997	2.339	8.159348	17.835

stdev diff	6.6101
sample size	24
alpha	0.05

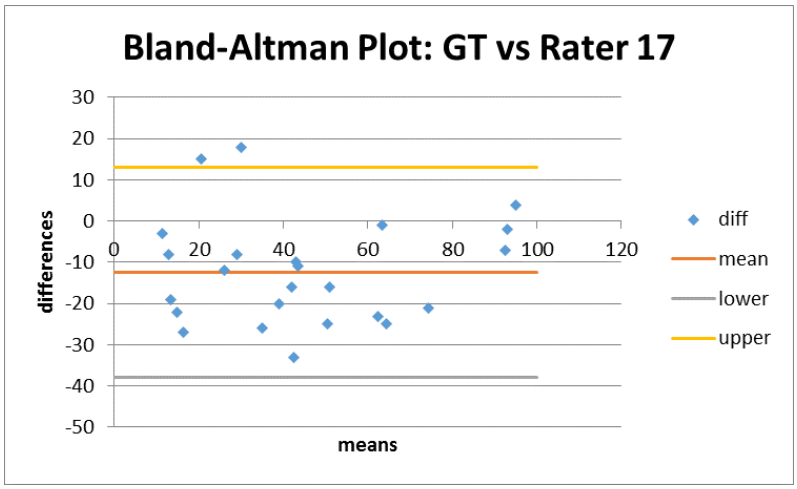
Figure J.15. Bland-Altman Diagram for Rater 15.



	value	*s.e.*	*lower*	*upper*
mean diff	-0.1667	2.227	-4.773292	4.44
lower limit	-21.549	3.86	-29.5332	-13.56
upper limit	21.215	3.86	13.23077	29.2

stdev diff	10.909
sample size	24
alpha	0.05

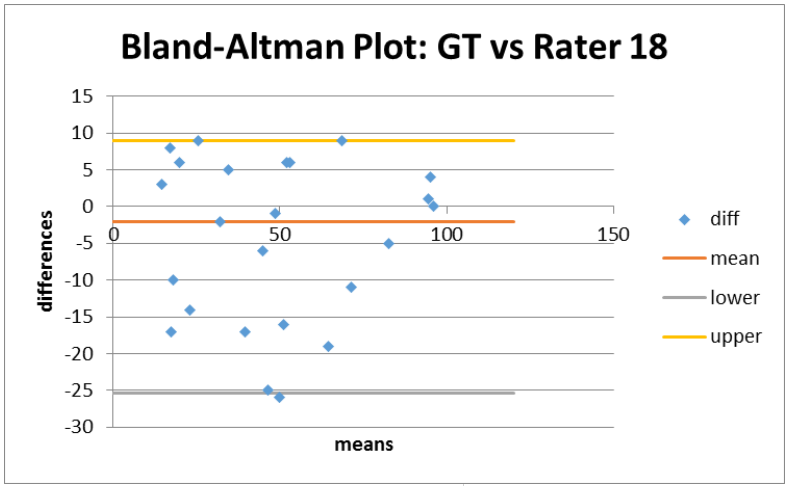
Figure J.16. Bland-Altman Diagram for Rater 16.



	value	*s.e.*	*lower*	*upper*
mean diff	-12.417	2.655	-17.9095	-6.924
lower limit	-37.912	4.602	-47.43262	-28.39
upper limit	13.079	4.602	3.55812	22.599

stdev diff	13.008
sample size	24
alpha	0.05

Figure J.17. Bland-Altman Diagram for Rater 17.



	value	*s.e.*	*lower*	*upper*
mean diff	-4.6667	2.269	-9.359987	0.0267
lower limit	-26.451	3.932	-34.58586	-18.32
upper limit	17.118	3.932	8.982906	25.253

stdev diff	11.115
sample size	24
alpha	0.05

-2	Median
-25.425	2.5th percentile
9	97.5th percentile

Figure J.18. Bland-Altman Diagram for Rater 18.

Appendix K

ANOVA Analysis

Case Study 1

Real Statistics Report

Rater 1

OVERALL FIT

Multiple R	0.437402267	AIC	100.61
R Square	0.191320743	AICc	101.81
Adjusted R Square	0.154562595	SBC	102.96
Standard Error	7.815747927		
Observations	24		

ANOVA Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	317.9431887	317.9431887	5.2049	0.032561341	yes
Residual	22	1343.890145	61.08591566			
Total	23	1661.833333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-13.67250615	3.172426946	-4.309793852	0.0003	-20.25171695	-7.093295343
Group 1	0.133221071	0.058394064	2.281414611	0.0326	0.012119194	0.254322949

Figure K.1. ANOVA Analysis for Rater 1.

Rater 2

OVERALL FIT

Multiple R	0.353934683	AIC	120.49
R Square	0.12526976	AICc	121.69
Adjusted R Square	0.085509294	SBC	122.85
Standard Error	11.82781458		
Observations	24		

ANOVA Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	440.7616497	440.7616497	3.1506	0.089737455	no
Residual	22	3077.73835	139.8971977			
Total	23	3518.5				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	4.793772621	5.638176812	0.850234532	0.4044	-6.899090421	16.48663566
Group 1	-0.186309479	0.10496332	-1.774996056	0.0897	-0.403990081	0.031371123

Figure K.2. ANOVA Analysis for Rater 2.

Rater 3

OVERALL FIT			
Multiple R	0.409477574	AIC	104.43
R Square	0.167671884	AICc	105.63
Adjusted R Square	0.129838788	SBC	106.79
Standard Error	8.464141822		
Observations	24		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	317.507671	317.507671	4.4319	0.046917894	yes
Residual	22	1576.117329	71.64169677			
Total	23	1893.625				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	16.00889789	4.341092717	3.687757653	0.0013	7.006022618	25.01177316
Group 1	-0.153891816	0.073100668	-2.105203966	0.0469	-0.305493322	-0.00229031

Figure K.3. ANOVA Analysis for Rater 3.

Rater 4

OVERALL FIT			
Multiple R	0.265000084	AIC	124.22
R Square	0.070225045	AICc	125.42
Adjusted R Square	0.027962547	SBC	126.57
Standard Error	12.78166533		
Observations	24		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	271.463688	271.463688	1.6616	0.210771871	no
Residual	22	3594.161312	163.3709687			
Total	23	3865.625				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	18.68590165	6.775835897	2.757726418	0.0115	4.633678075	32.73812523
Group 1	-0.143998243	0.111709154	-1.289046044	0.2108	-0.375668849	0.087672363

Figure K.4. ANOVA Analysis for Rater 4.

Rater 5

OVERALL FIT			
Multiple R	0.517530427	AIC	101.3
R Square	0.267837743	AICc	102.5
Adjusted R Square	0.234557641	SBC	103.65
Standard Error	7.928727962		
Observations	24		

ANOVA					Alpha 0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	505.9343373	505.9343373	8.048	0.009595216	yes
Residual	22	1383.023996	62.86472709			
Total	23	1888.958333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	21.33837596	4.216887475	5.060219435	5E-05	12.59308659	30.08366532
Group 1	-0.197925362	0.069768259	-2.836896951	0.0096	-0.342615877	-0.053234848

Figure K.5. ANOVA Analysis for Rater 5.

Rater 6

OVERALL FIT			
Multiple R	0.578826713	AIC	108.9
R Square	0.335040364	AICc	110.1
Adjusted R Square	0.304814926	SBC	111.26
Standard Error	9.290791859		
Observations	24		

ANOVA					Alpha 0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	956.8194393	956.8194393	11.085	0.003042333	yes
Residual	22	1899.013894	86.31881336			
Total	23	2855.833333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	11.44055084	4.619416251	2.476622634	0.0214	1.860467889	21.0206338
Group 1	-0.284028034	0.085309804	-3.329371552	0.003	-0.460949739	-0.107106329

Figure K.6. ANOVA Analysis for Rater 6.

Rater 7

OVERALL FIT			
Multiple R	0.330990946	AIC	131.46
R Square	0.109555006	AICc	132.66
Adjusted R Square	0.069080234	SBC	133.82
Standard Error	14.86376619		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	598.0060016	598.0060016	2.7067	0.114138475	no
Residual	22	4860.493998	220.9315454			
Total	23	5458.5				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	11.07523466	7.522705985	1.472240798	0.1551	-4.52590268	26.676372
Group 1	-0.22407719	0.136198945	-1.645219719	0.1141	-0.506536515	0.058382134

Figure K.7. ANOVA Analysis for Rater 7.

Rater 8

OVERALL FIT			
Multiple R	0.456568872	AIC	104.04
R Square	0.208455135	AICc	105.24
Adjusted R Square	0.172475823	SBC	106.4
Standard Error	8.395354259		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	408.3549243	408.3549243	5.7937	0.024916828	yes
Residual	22	1550.603409	70.48197314			
Total	23	1958.958333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	13.60276279	4.230903951	3.215096099	0.004	4.828405029	22.37712054
Group 1	-0.176304779	0.07324605	-2.407020958	0.0249	-0.32820779	-0.024401767

Figure K.8. ANOVA Analysis for Rater 8.

Rater 9

OVERALL FIT			
Multiple R	0.15542401	AIC	111.93
R Square	0.024156623	AICc	113.13
Adjusted R Square	-0.020199894	SBC	114.29
Standard Error	9.896049571		
Observations	24		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	53.33379704	53.33379704	0.5446	0.468331369	no
Residual	22	2154.499536	97.9317971			
Total	23	2207.833333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	7.622150632	4.79037935	1.591137168	0.1258	-2.312488088	17.55678935
Group 1	-0.06062381	0.082149295	-0.73797116	0.4683	-0.23099102	0.109743399

Figure K.9. ANOVA Analysis for Rater 9.

Rater 10

OVERALL FIT			
Multiple R	0.539840422	AIC	107.99
R Square	0.291427682	AICc	109.19
Adjusted R Square	0.259219849	SBC	110.34
Standard Error	9.11504966		
Observations	24		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	751.7741333	751.7741333	9.0483	0.006472729	yes
Residual	22	1827.850867	83.0841303			
Total	23	2579.625				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	20.83334424	4.845567974	4.299463829	0.0003	10.78425131	30.88243716
Group 1	-0.247604647	0.082314082	-3.008047225	0.0065	-0.418313605	-0.076895688

Figure K.10. ANOVA Analysis for Rater 10.

Rater 11

OVERALL FIT			
Multiple R	0.253138072	AIC	114.35
R Square	0.064078883	AICc	115.55
Adjusted R Square	0.021537014	SBC	116.71
Standard Error	10.40694518		
Observations	24		

ANOVA					Alpha 0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	163.1341569	163.1341569	1.5063	0.232683022	no
Residual	22	2382.699176	108.304508			
Total	23	2545.833333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	14.90885084	5.324547335	2.800022219	0.0104	3.866415525	25.95128616
Group 1	-0.108701754	0.088570147	-1.227295621	0.2327	-0.292384997	0.074981489

Figure K.11. ANOVA Analysis for Rater 11.

Rater 12

OVERALL FIT			
Multiple R	0.446710883	AIC	122
R Square	0.199550613	AICc	123.2
Adjusted R Square	0.16316655	SBC	124.36
Standard Error	12.20551294		
Observations	24		

ANOVA					Alpha 0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	817.0599853	817.0599853	5.4846	0.028644292	yes
Residual	22	3277.440015	148.9745461			
Total	23	4094.5				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	13.60795877	6.224253971	2.186279486	0.0397	0.699646095	26.51627145
Group 1	-0.262995087	0.11229921	-2.34191396	0.0286	-0.495889393	-0.03010078

Figure K.12. ANOVA Analysis for Rater 12.

Rater 13

OVERALL FIT			
Multiple R	0.230167708	AIC	120.45
R Square	0.052977174	AICc	121.65
Adjusted R Square	0.009930682	SBC	122.81
Standard Error	11.81705633		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	171.8579515	171.8579515	1.2307	0.279251467	no
Residual	22	3072.142049	139.6428204			
Total	23	3244				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	4.638679245	5.626115093	0.824490642	0.4185	-7.029169322	16.30652781
Group 1	-0.112398922	0.101318006	-1.109367681	0.2793	-0.322519607	0.097721763

Figure K.13. ANOVA Analysis for Rater 13.

Rater 14

OVERALL FIT			
Multiple R	0.079104663	AIC	97.989
R Square	0.006257548	AICc	99.189
Adjusted R Square	-0.038912564	SBC	100.35
Standard Error	7.400973343		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	7.588058711	7.588058711	0.1385	0.713305136	no
Residual	22	1205.036941	54.77440642			
Total	23	1212.625				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	3.71811492	3.455920438	1.075868206	0.2936	-3.449025401	10.88525524
Group 1	0.021785204	0.058530889	0.372200114	0.7133	-0.099600431	0.143170838

Figure K.14. ANOVA Analysis for Rater 14.

Rater 15

OVERALL FIT			
Multiple R	0.165181829	AIC	92.968
R Square	0.027285037	AICc	94.168
Adjusted R Square	-0.01692928	SBC	95.324
Standard Error	6.665849221		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	27.42032489	27.42032489	0.6171	0.440502702	no
Residual	22	977.5380084	44.43354584			
Total	23	1004.958333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-2.033343456	2.971289261	-0.684330362	0.5009	-8.195420232	4.12873332
Group 1	0.040937314	0.052112095	0.785562624	0.4405	-0.067136557	0.149011185

Figure K.15. ANOVA Analysis for Rater 15.

Rater 16

OVERALL FIT			
Multiple R	0.371895284	AIC	114.11
R Square	0.138306102	AICc	115.31
Adjusted R Square	0.099138198	SBC	116.46
Standard Error	10.35449711		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	378.5899042	378.5899042	3.5311	0.073540518	no
Residual	22	2358.743429	107.2156104			
Total	23	2737.333333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	8.45314098	5.050661134	1.67367019	0.1084	-2.021289121	18.92757108
Group 1	-0.170408059	0.090684827	-1.879124275	0.0735	-0.358476878	0.01766076

Figure K.16. ANOVA Analysis for Rater 16.

Rater 17

OVERALL FIT			
Multiple R	0.064258429	AIC	126.03
R Square	0.004129146	AICc	127.23
Adjusted R Square	-0.041137711	SBC	128.38
Standard Error	13.27294478		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	16.06994698	16.06994698	0.0912	0.765471294	no
Residual	22	3875.763386	176.171063			
Total	23	3891.833333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-13.8496128	5.463578245	-2.53489786	0.0189	-25.18038058	-2.518845024
Group 1	0.032231216	0.106717773	0.302022943	0.7655	-0.189087899	0.25355033

Figure K.17. ANOVA Analysis for Rater 17.

Rater 18

OVERALL FIT			
Multiple R	0.042074901	AIC	118.53
R Square	0.001770297	AICc	119.73
Adjusted R Square	-0.04360378	SBC	120.89
Standard Error	11.3544211		
Observations	24		

ANOVA					Alpha	0.05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	5.030004799	5.030004799	0.039	0.845231362	no
Residual	22	2836.303329	128.9228786			
Total	23	2841.333333				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-5.537013559	4.978673258	-1.112146404	0.2781	-15.86214994	4.788122825
Group 1	0.018007177	0.091164642	0.197523696	0.8452	-0.17105672	0.207071074

Figure K.18. ANOVA Analysis for Rater 18.

Appendix L

Normality Tests

Case Study 2

Real Statistics Report

Shapiro-Wilk Test

	7D
W-stat	0.894487
p-value	0
alpha	0.05
normal	no

QQ Plot - 7D

Count	2078	4156
Mean	70.93215	
Std Dev	16.34736	

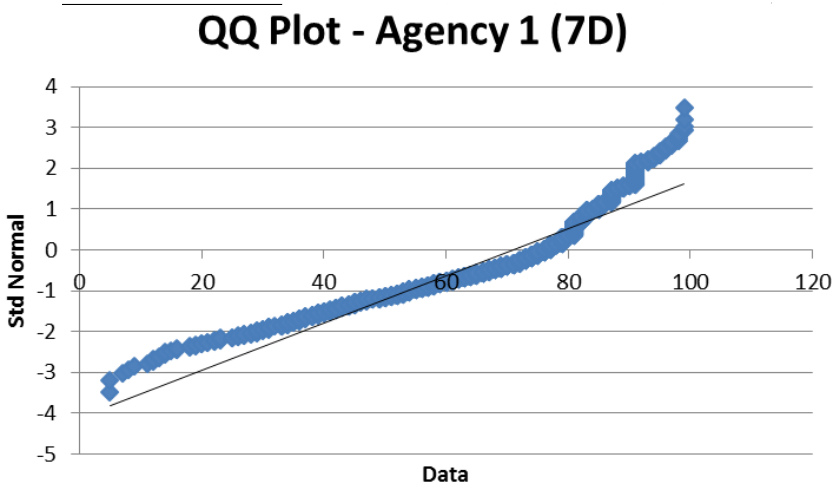


Figure L.1. Shapiro-Wilk Test and QQ Plot for Agency 1 (seven distresses).

Shapiro-Wilk Test

	8D
W-stat	0.905439
p-value	0
alpha	0.05
normal	no

QQ Plot - 8D

Count	2078	4156
Mean	73.72089	
Std Dev	17.30065	

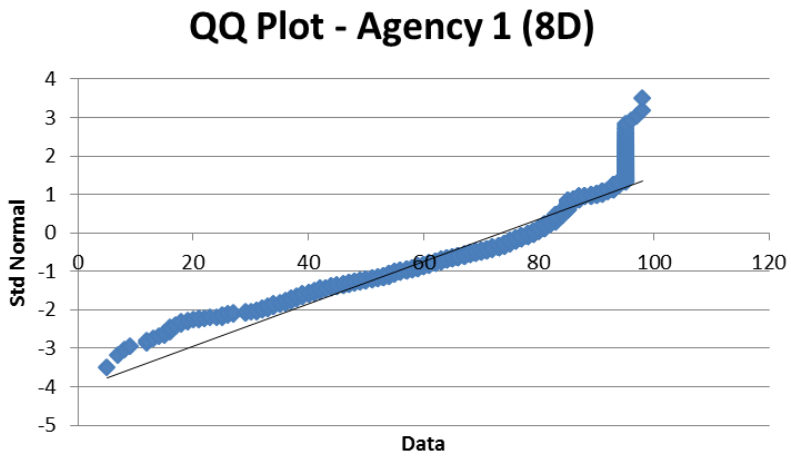


Figure L.2. Shapiro-Wilk Test and QQ Plot for Agency 1 (eight distresses).

Shapiro-Wilk Test

	7D
W-stat	0.921127
p-value	0
alpha	0.05
normal	no

QQ Plot - 7D

Count	1597	3194
Mean	71.65247	
Std Dev	17.65228	

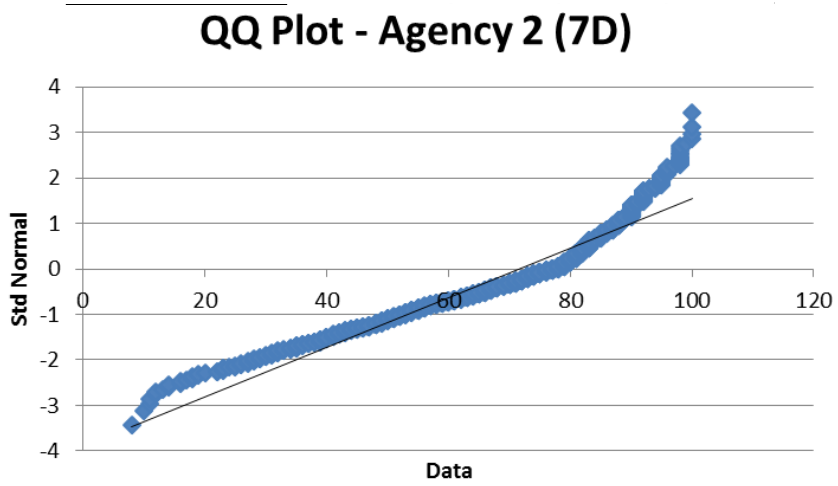


Figure L.3. Shapiro-Wilk Test and QQ Plot for Agency 2 (seven distresses).

Shapiro-Wilk Test

	8D
W-stat	0.91798
p-value	0
alpha	0.05
normal	no

QQ Plot - 8D

Count	1597	3194
Mean	72.89543	
Std Dev	19.23619	

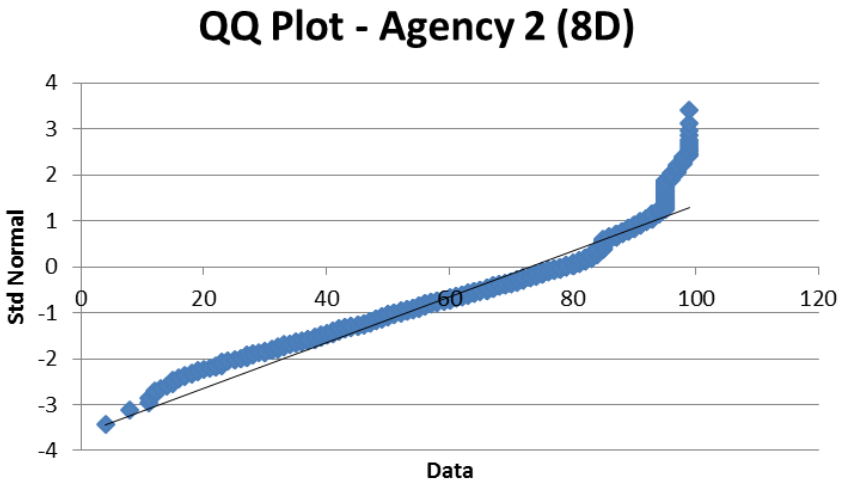


Figure L.4. Shapiro-Wilk Test and QQ Plot for Agency 2 (eight distresses).

Shapiro-Wilk Test

	7D
W-stat	0.947385
p-value	0
alpha	0.05
normal	no

QQ Plot - 7D

Count	1809	3618
Mean	67.89773	
Std Dev	17.3715	

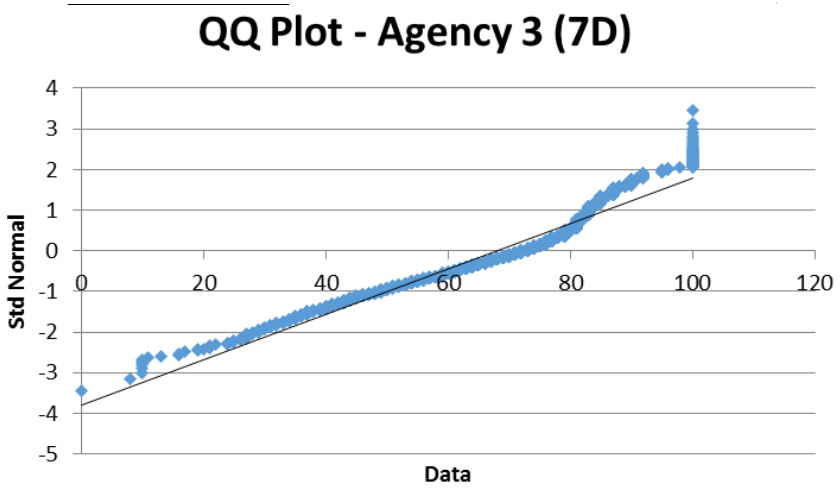


Figure L.5. Shapiro-Wilk Test and QQ Plot for Agency 3 (seven distresses).

Shapiro-Wilk Test

	8D
W-stat	0.940218
p-value	0
alpha	0.05
normal	no

QQ Plot - 8D

Count	1809	3618
Mean	71.8021	
Std Dev	18.29791	

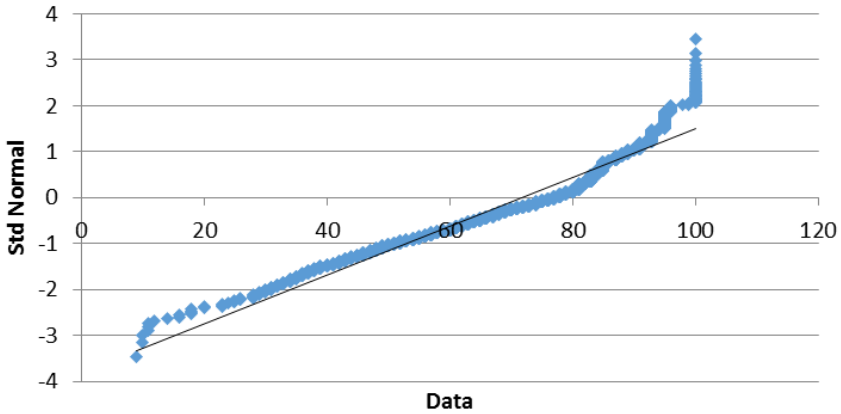


Figure L.6. Shapiro-Wilk Test and QQ Plot for Agency 3 (eight distresses).

Appendix M

Mann-Whitney Test

Case Study 2

Real Statistics Report

Mann-Whitney Test for Two Independent Samples

	InspectedPCI (7)	InspectedPCI (8)
count	2078	2078
median	76	79
rank sum	4029578.5	4608667.5
U	2448586.5	1869497.5

	one tail	two tail
alpha	0.05	
U	1869497.5	
mean	2159042	
std dev	38676.30109	
z-score	7.486341555	
effect r	0.116126642	
U-crit	2095424.646	2083237.343
p-value	3.54161E-14	7.08322E-14
sig (norm)	yes	yes

Figure M.1. Mann-Whitney Test for Agency 1.

Mann-Whitney Test for Two Independent Samples

	InspectedPCI (7)	InspectedPCI (8)
count	1597	1597
median	77	79
rank sum	2454301.5	2648113.5
U	1372110.5	1178298.5

	one tail	two tail
alpha	0.05	
U	1178298.5	
mean	1275204.5	
std dev	26058.51869	
z-score	3.718764722	
effect r	0.065800811	
U-crit	1232341.551	1224130.242
p-value	0.0001001	0.000200199
sig (norm)	yes	yes

Figure M.2. Mann-Whitney Test for Agency 2.

Mann-Whitney Test for Two Independent Samples

	Sample 1	Sample 2
count	1809	1809
median	72	77
rank sum	3013018.5	3533752.5
U	1896607.5	1375873.5

	one tail	two tail
alpha	0.05	
U	1375873.5	
mean	1636240.5	
std dev	31415.37407	
z-score	8.287868843	
effect r	0.137787109	
U-crit	1584566.308	1574666.998
p-value	0	0
sig (norm)	yes	yes

Figure M.3. Mann-Whitney Test for Agency 3.

Appendix N

ANOVA Analysis

Case Study 2

Real Statistics Report

Regression Analysis

OVERALL FIT			
Multiple R	0.3309587	AIC	4172.386644
R Square	0.109533661	AICc	4172.398216
Adjusted R Square	0.109104728	SBC	4183.664967
Standard Error	2.727708524		
Observations	2078		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	1899.998985	1899.998985	255.3626915	2.65944E-54	yes
Residual	2076	15446.25751	7.440393792			
Total	2077	17346.2565				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-1.336591865	0.264998902	-5.043763786	4.96116E-07	-1.856283159	-0.81690057
Group 1	0.057037602	0.003569296	15.9800717	2.65944E-54	0.05003783	0.064037374

Figure N.1. ANOVA analysis for Agency 1.

Regression Analysis

OVERALL FIT			
Multiple R	0.42884107	AIC	3864.885223
R Square	0.183904664	AICc	3864.900289
Adjusted R Square	0.183393005	SBC	3875.636987
Standard Error	3.351539142		
Observations	1597		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	4037.39393	4037.39393	359.4285196	1.90705E-72	yes
Residual	1595	17916.33932	11.23281462			
Total	1596	21953.73325				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	-5.015360541	0.340591499	-14.72544254	4.08164E-46	-5.683414557	-4.347306525
Group 1	0.086591586	0.004567404	18.95860015	1.90705E-72	0.077632841	0.095550332

Figure N.2. ANOVA analysis for Agency 2.

Regression Analysis

OVERALL FIT			
Multiple R	0.276302182	AIC	4251.47014
R Square	0.076342896	AICc	4251.483436
Adjusted R Square	0.075831741	SBC	4262.471199
Standard Error	3.236640793		
Observations	1809		

ANOVA				Alpha	0.05	
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	1	1564.606078	1564.606078	149.3537069	4.65741E-33	yes
Residual	1807	18929.84942	10.47584362			
Total	1808	20494.4555				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	0.245902769	0.308878858	0.796113953	0.426070377	-0.359894439	0.851699977
Group 1	0.052376072	0.004285731	12.22103543	4.65741E-33	0.043970563	0.060781581

Figure N.3. ANOVA analysis for Agency 3.

Vita

Edgar Rodriguez was born in Piura, Peru. In 2004, he started his undergraduate studies in Civil Engineering at the University of Piura (UDEP). He earned his Bachelors of Science in 2008 and, the next year, obtained the professional licensure as a civil engineer by presenting the thesis titled "Calculation of Pavement Condition Index in Luis Montero Avenue, District of Castilla". In 2009, Edgar became a faculty at UDEP teaching the undergraduate classes of Statics, Geology, Soil Mechanics, and Pavement Design. In 2011, he participated at the "Inca Engineering Expedition: The Inca Road" in Cuzco (Peru), a research project financed by the National Science Foundation (NSF) and the Pan-American Advanced Studies Institute (PASI). In 2012, Edgar completed an academic internship in the field of structural evaluation of asphalt pavements in the Road Research Center of the National Technological University in La Plata (Argentina). As a professor at UDEP, Edgar published articles for international conferences related to education in engineering, construction engineering, and rehabilitation technology. At present, Edgar is part of the team responsible for the development of the project "Stochastic-Mathematical Model that Incorporates Local Characteristics to the Measurements of Urban Pavements Condition, Based on Data Generated with Existing Computer Applications," held in Piura. In 2016, Edgar began his Masters at the University of Texas at El Paso (UTEP) supported by UDEP. During his graduate studies, he participated in the development of articles for international conferences related to infrastructure management. He also participated in the development of research projects prepared for the National Cooperative Highway Research Program (NCHRP), Texas Department of Transportation (TxDOT), and the Metropolitan Transportation Commission (MTC).