IMPROVED MONTE CARLO SIMULATION IN THE PRESENCE OF OUTLIERS USING LABELING AND BAYESIAN AVERAGING

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Kwabena Yeboah Dadson

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Agribusiness and Applied Economics

April 2023

Fargo, North Dakota

North Dakota State University
Graduate School

Title

IMPROVED MONTE CARLO SIMULATION IN THE PRESENCE OF OUTLIERS USING LABELING AND BAYESIAN AVERAGING

By

Kwabena Yeboah Dadson

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. David Bullock
Chair

Dr. William W. Wilson

Dr. Ruilin Tian

Approved:

7/8/2023
Date

Dr. William Nganje
Department Chair

ABSTRACT

The problem of outliers is an old phenomenon in statistics, and it appears with surprising frequency in many datasets in both the natural and social sciences and can have both positive and negative effects on statistical analysis. Unlike the traditional approach to dealing with outliers in a dataset, this study considers both the base and contaminating distributions that generate outliers and estimates the best-fitting distribution for each separately. Using the natural conjugate prior distribution for the probability of occurrence, the 'Bayesian averaging' technique is used in a way that preserves most of the information in the total dataset. The KS-test and AD-test statistics were computed by contrasting the simulated to the actual data distribution to obtain the comparative metric. Analysis of seven sample datasets (each containing outliers) indicated that these alternate simulation procedures provided a stronger goodness-of-fit to the historical data when compared to other, more traditional approaches.

ACKNOWLEDGMENTS

I would like to acknowledge the support and contributions of several individuals without whom this dissertation would not have been possible. First and foremost, I am grateful to my supervisors, Dr. David Bullock, Dr. William Wilson, and Dr. Ruilin Tian, for their guidance, expertise, and constant support throughout the research process. I am also thankful to the participants who generously volunteered their time and provided valuable insights for this study. Additionally, I extend my appreciation to my family, friends, and colleagues for their unwavering encouragement and understanding. Lastly, I express my gratitude to the academic community for their contributions to the field, which has greatly influenced this research.

DEDICATION

I dedicate this dissertation to my loving family, whose unwavering support and encouragement have been the driving force behind my journey. To my parents, whose sacrifices and belief in my abilities have shaped me into the person I am today, thank you for instilling in me the value of education and the pursuit of knowledge. To my siblings, for their constant encouragement and understanding during the long hours spent immersed in research. To my dear friends, who have been pillars of strength and a source of inspiration throughout this challenging endeavor. This work is dedicated to all those who have believed in me and stood by my side, illuminating the path to achievement.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS iv

DEDICATION v

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF APPENDIX TABLES xii

LIST OF APPENDIX FIGURES xiii

1. INTRODUCTION 1

1.1. Problem Statement 2

1.2. Research Goals 3

1.3. Relevance of the Study 4

1.4. Organization 5

2. LITERATURE REVIEW 6

2.1. Genesis of Outliers 6

2.2. Outlier Categorization 8

2.3. Data Modeling and Detection of Outliers 8

2.4. Classification of Outlier Detection Methods 9

2.4.1. Clustering 10

2.4.2. Density Estimators 10

2.4.3. Classifiers 10

2.4.4. Fixed Size Windows 11

2.4.5. Distance Measures 11

2.5. Outlier Labeling Techniques 12

2.5.1. Formal Tests 12

2.5.2. Informal Tests 15
3. THEORY AND CONCEPTUAL FRAMEWORK 19
3.1. Research Procedure 19
3.2. Labeling Procedures Utilized in the Study 20
3.2.1. Modified Z-Score 20
3.2.2. Tukey's Interquartile Range (IQR) 22
3.2.3. K-means Clustering 24
3.3. Distribution Fitting Procedures Used in the Study 26
3.3.1. Palisade Bestfit Procedure 26
3.3.2. Review of Kolmogorov-Smirnov Test 28
3.3.3. Review of Anderson-Darling Test 33
3.3.4. Akaike Information Criterion 36
3.3.5. Bayesian Information Criterion 38
3.4. Bayesian Averaging Procedure Used in the Study 39
3.4.1. Review of Bayesian Averaging 39
3.5. Procedure for Comparing Goodness-of-Fit to Original Sample Data 41
4. DATA SERIES DESCRIPTION AND CHARACTERISTICS 43
4.1. Daily Car Values (DCV) for Secondary Rail Shipping 43
4.2. Electricity Wholesale Price Daily Changes 45
4.2.1. General Description 45
4.2.2. Indiana Hub 49
4.2.3. PJM West Hub 51
4.2.4. NEPOOL Hub 54
4.3. Futures Price and Spread Daily Changes 56
4.3.1. General Description 56

4.3.2. Nearby Chicago Oats Futures 58
4.3.3. Chicago Nearby Wheat Futures Spread 60
4.3.4. Chicago Nearby Intermonth Wheat Spread 63
5. EMPIRICAL RESULTS AND DISCUSSION 65
5.1. Analysis and Results 65
5.1.1. Daily Car Values (DCV) 65
5.1.2. Indiana Hub Electricity Prices 68
5.1.3. NEPOOL Hub Electricity Prices 71
5.1.4. PJM West Hub Electricity Prices 74
5.1.5. Chicago Oats Futures Prices 78
5.1.6. KC - Chicago Wheat Intermarket Spread 81
5.1.7. Chicago Wheat Intermonth Spread 83
5.2. General Observations 86
6. SUMMARY AND CONCLUSIONS 87
6.1. Summary of Problem 87
6.2. Summary of Methodology 87
6.3. Study Results 88
6.4. Major Observations 90
6.5. Contributions to Existing Literature 92
6.6. Study Implications for Risk Researchers and Practitioners 92
6.7. Avenues for Future Study 93
REFERENCES 96
APPENDIX A. SUPPLEMENTARY TABLES 99
APPENDIX B. SUPPLEMENTARY FIGURES 101

LIST OF TABLES

Table		Page
1.	DCV ($/car) Descriptive Statistic Summary	45
2.	Indiana Hub Price Change ($/MWh) Descriptive Statistic Summary	50
3.	PJM West Hub Price Change ($/MWh) Descriptive Statistic Summary	53
4.	NEPOOL Hub Price Change ($/MWh) Descriptive Summary Statistics	55
5.	Nearby Chicago Oats Futures Descriptive Statistic Summary	59
6.	Chicago Nearby Wheat Futures Spread Summary Descriptive Statistics	62
7.	Chicago Nearby Intermonth Wheat Spread Descriptive Summary Statistics	64
8.	Summary Results of DCV Goodness-of-Fit Test	67
9.	Summary Results of Indiana Hub Electricity Prices Goodness-of-Fit Test	71
10.	Summary Results of NEPOOL Hub Electricity Goodness-of-Fit Test	74
11.	Summary Results of PJM West Hub Electricity Goodness-of-Fit Test	77
12.	Results of Oats Futures Price Goodness-of-Fit Test	80
13.	Results of KC - Chicago Wheat Intermarket Spread Goodness-of-Fit Test	83
14.	Results of KC - Chicago Wheat Intermonth Spread Goodness-of-Fit Test	86

LIST OF FIGURES

Figure		Page
1.	Panel A. Scatterplot of individual patient data illustrating the relationship between body surface area and diameter of the Sinuses of Valsalva. Solid line regression = equation; dashed line = confidence Intervals. Panel B. Nomogram generated from the scatterplot in Panel A for determining individual Z-score according to body surface area and diameter of the Sinuses of Valsalva (Curtis et al. 2016).	17
2.	Components of a classic Tukey boxplot with Interquartile Range (IQR) illustrated.	23
3.	Application of Tukey boxplot to Indiana Electricity price change data.	23
4.	Cumulative Probability	29
5.	Illustration of the Kolmogorov-Smirnov Distribution PDF	30
6.	Illustration of the two-sample Kolmogorov-Smirnov statistics. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.	33
7.	Time series (DCV ($/car))	44
8.	(DCV ($/car)) Boxplot	45
9.	Regional Transmission Organizations/ Electricity Wholesale Market	46
10.	Time series (Indiana Hub Price Change ($/MWh))	49
11.	Box plot (Indiana Hub Price Change ($/MWh))	51
12.	Time series (PJM West Hub Price Change ($/MWh))	52
13.	Box plot (PJM West Hub Price Change ($/MWh))	53
14.	Time series (NEPOOL Hub Price Change ($/MWh))	54
15.	Box plot (NEPOOL Hub Price Change ($/MWh))	56
16.	Time series (Change in Nearby Oats Futures (cents/bu)):	58
17.	Box plot Change in Nearby Oats Futures (cents/bu)	60
18.	KC-Chicago Nearby Wheat Futures Spread	61
19.	Box plot (Change in KC - Chicago Wheat Spread (cents/bu))	62

20. Chicago Nearby Intermonth Wheat Spread Time Series 63

21. Box plot (Change in Chicago Wheat Futures Nearby Intermonth Spread (cents/bu)) 64

LIST OF APPENDIX TABLES

Table Page

A1. Two-sample KS test - Ignore Outlier 99

A2. Two-sample KS test - Drop Outlier 99

A3. Two-sample KS test - Z-Score Bayesian Averaging 99

A4. Two-sample KS test Tukey IQR Bayesian Averaging 100

A5. Two-sample KS test K-means IQR Bayesian Averaging 100

LIST OF APPENDIX FIGURES

Figure — Page

A1. DCV KS-Ignore Outlier Cumulative Distribution Comparison. 101

A2. DCV KS-Drop Outlier Cumulative Distribution Comparison..... 101

A3. DCV KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison..... 102

A4. DCV KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison. 102

A5. DCV KS-k-Means Bayesian Averaging Cumulative Distribution Comparison. 103

A6. Indiana Hub KS-Ignore Outlier Bayesian Averaging Cumulative Distribution Comparison. 103

A7. Indiana Hub KS-Drop Outlier Bayesian Averaging Cumulative Distribution Comparison. 104

A8. Indiana Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison 104

A9. Indiana Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 105

A10. Indiana Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison 105

A11. NEPOOL Hub KS-Ignore Outlier Cumulative Distribution Comparison. 106

A12. NEPOOL Hub KS-Drop Outlier Cumulative Distribution Comparison. 106

A13. NEPOOL Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison. 107

A14. NEPOOL Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 107

A15. NEPOOL Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison 108

A16. PJM West Hub KS-Ignore Outliers Cumulative Distribution Comparison 108

A17. PJM West Hub KS-Drop Outliers Cumulative Distribution Comparison 109

A18. PJM West Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison 109

A19. PJM West Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 110

A20. PJM West Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison 110

A21. Chng_O KS-Ignore Outliers Cumulative Distribution Comparison..... 111

A22. Chng_O KS-Drop Outliers Cumulative Distribution Comparison 111

A23. Chng_O KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison..... 112

A24. Chng_O KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 112

A25. Chng_O KS-k-Means Bayesian Averaging Cumulative Distribution Comparison 113

A26. KW-W KS-Ignore Outliers Cumulative Distribution Comparison 113

A27. KW-W KS-Drop Outliers Cumulative Distribution Comparison..... 114

A28. KW-W KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison 114

A29. KW-W KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 115

A30. KW-W KS-k-Means Bayesian Averaging Cumulative Distribution Comparison 115

A31. W KS- Ignore Outliers Cumulative Distribution Comparison 116

A32. W KS- Drop Outliers Cumulative Distribution Comparison..... 116

A33. W KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison 117

A34. W KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison 117

A35. W KS-k-Means Bayesian Averaging Cumulative Distribution Comparison. 118

1. INTRODUCTION

A significant amount of research has been dedicated to developing efficient outlier detection techniques that consider efficiency, data dimensionality, and accuracy, among many other factors. Traditionally, data cleaning was at the core of outlier detection so that parametric statistical models would fit the training data more smoothly (Boukerche, Zheng, and Alfandi 2021). In recent times, more attention has been shifted to the outliers themselves because they represent information of interest. Outlier detection and labeling have been applied to a variety of real-life scenarios, including system diagnosis related to host-based intrusion detection systems, mechanical-based systems, and UNIX systems, among others.

Most of these systems, such as medical diagnosis, intrusion detection, and biological applications, generate discrete-valued temporal sequences. In biological applications, the data may contain sequences of amino acids in which anomalous sub-sequences can present unusual properties of genome sequences. On the other hand, in medical anomaly diagnosis, medical equipment is used to collect data on potential disease risks (Akram et al. 2021), while in intrusion detection systems, malicious activities are detected through collected data. These discrete sequences are caused by temporal ordering in certain fields, such as intrusion detection and systems diagnosis, whereas physical ordering is the reason in others, such as biological data (Aggarwal 2017).

In time series, outliers are either contextual or collective anomalies. Outliers are contextual when large subsequences within a time series have unusual shapes. Contextual outliers have values at specific time stamps that suddenly change with respect to their temporary adjacent values.

Aggarwal (2017) defines a sequence as an ordered set of symbols, $a_1, a_2, \ldots a_r$ drawn from the symbol set $\Sigma = \{\sigma_1, \ldots, \sigma|a_\Sigma|\}$. In the most general form, a sequence can also be defined as an ordered set of sets $S_1, S_2, \ldots S_r$, where each S_i is a subset of Σ. Consequently, the common regression-modeling methods used for identifying anomalies in continuous data are challenging to apply in this context. However, outliers can be defined either by their divergence from expected values at certain timestamps or by unusual sequential arrangements of sequence values. The goal is to create a discrete predictive or regularity model that mirrors its continuous counterpart. Just like with continuous data, outliers are categorized into two types based on whether individual locations are viewed as outliers or if combinations of symbols are seen as outliers. In a position-based model, values at specific locations are forecasted to measure the deviation from a model and identify certain positions as outliers. Predictive outlier detection employs Markovian techniques. When dealing with outliers, an entire symbol combination is used to assess if the complete test sequence is atypical.

1.1. Problem Statement

The presence of outliers in a data set can introduce severe bias when utilizing distribution fitting procedures such as the Bestfit procedure in Palisade's @Risk Monte Carlo simulation software. The usual approach of ignoring or discarding outlier observations ignores useful information that may be present in these observations. Outliers are generally produced from a separate "contaminating" distribution (Hawkins 1981) that may assert their presence from time to time. Therefore, the true underlying stochastic process is likely to be a probability-weighted combination of the "base" and "contaminating" distributions. Outliers in a sample will show such characteristics as large gaps between 'outlying' and 'inlying' observations and deviations between the outliers and the group of inliers as measured on a standard scale. (Hawkins, 1980).

1.2. Research Goals

The primary goal of this study is to develop and evaluate an improved procedure for fitting statistical distributions to data in the presence of "contaminating" outlier distributions. This alternative procedure will be applied to a variety of real-world datasets and compared to traditional methods using non-parametric statistical goodness-of-fit tests (Kolmogorov-Smirnov and Anderson-Darling) as the primary comparison metrics.

This alternative procedure begins with dividing the candidate dataset into "base" and "contaminating" distributions using one of three alternative outlier labeling methods: 1) the modified Z-statistic proposed by Iglewicz and Hoaglin (1993), 2) an approach using the interquartile range (IQR) proposed by Tukey (1977), and 3) an approach using k-means clustering (MacQueen, 1967) to divide the dataset. The three approaches were all compared using the Kolmogorov-Smirnov and Anderson-Darling statistics for goodness-of-fit for each dataset.

Once the dataset was divided into the "base" and "contaminating" subsets, the Bestfit procedure contained in the @Risk (Palisade, 2022) computer software was used to find the best-fitting statistical distributions for each subset individually. For simulation purposes, these individual subset distributions were combined using a technique known as "Bayesian averaging," where the frequency of draws from each individual distribution is based upon the prior probability distribution using a beta or multivariate beta (i.e., Dirichlet) distribution to generate random probabilities for each subset based upon their observed frequency in the actual dataset.

To calculate the comparison metric, the Kolmogorov-Smirnov and Anderson-Darling test statistics were estimated by comparing the actual sample data to the Monte Carlo simulated data from each alternative approach (simulated at the same exact sample size as the original data).

Lower values of the test statistics would indicate a stronger goodness-of-fit between the actual and candidate data distributions. Two traditional procedures where Bestfit is applied to the entire original dataset (i.e., ignore the presence of outliers) and where identified outliers are discarded from the estimation dataset were also included in the comparisons.

1.3. Relevance of the Study

The problem of outliers is one of the oldest in statistics, and it appears with surprising frequency in many datasets in both the natural and social sciences. Most previous studies focused upon the identification of outliers using either labeling or statistical procedures but generally handle them using one of three methods: (1) just ignore their existence and estimate a best-fitting distribution to the total dataset using a "fat-tails" distribution such as the Pareto; (2) remove the offending observations from the dataset and estimate a best-fitting distribution upon the remaining values; or (3) use a non-parametric distribution based upon order-statistics (such as the median as a measure of centrality) to minimize the influence of the offending values.

This study is unique in that it directly attacks the outlier problem by considering both the "base" (or natural) distribution generating the data and the "contaminating" distributions that generate outliers and estimating best fitting distributions separately to each. Using the natural conjugate prior (beta or Dirichlet) distribution for the probability of occurrence, the distributions are combined in a manner that effectively preserves most of the information in the total dataset. This improved method can be used by practitioners of Monte Carlo simulation modeling to more accurately reflect historical data distributions that have a significant number of outliers present in the historical data.

1.4. Organization

The next section (Chapter 2) contains a review of the relevant literature organized in subsections by topic. Chapter 3 presents a discussion on the theory and conceptual framework of the study. It discusses the procedure and methodological approaches employed in the study, including background on the outlier labeling procedures, the statistical tests employed, and a discussion of the seven sample data sets used for the analysis. Chapter 4 contains a description and characteristics of the data series. Chapter 5 presents the empirical results and discussion of the labeling procedures, distribution fitting, and statistical goodness-of-fit comparisons using the Kolmogorov-Smirnov and Anderson-Darling statistics. Chapter 6 summarizes the major observations from the empirical analysis and recommendations for future studies.

2. LITERATURE REVIEW

Outliers have varied interpretations despite its general understanding. It has been defined and interpreted by many professionals (Iglewicz & Hoaglin, 1993a). In 1960, Grubbs likely provided the first definition of outliers (Boukerche et al., 2021a): "An outlier is a deviation that appears to deviate markedly from other members of the sample in which it occurs." Hawkins (1980) defines outliers similarly "as an observation that deviates so much from other observations as to arouse the suspicions that it was generated by a different mechanism." Alternatively, outliers can also be noise that harms a data process, distorts, and affect statistical conclusions (Yang et al., 2019). It makes their use inaccurate because of the presence of biasedness. Similarly, Aggarwal (2005) referred to outliers as deviants, discordance, abnormalities, or anomalies in data mining and statistics literature.

2.1. Genesis of Outliers

Outliers and anomalies are common words with almost similar meanings used interchangeably for outlier detection. However, Ajith Kumar et al. (2019) argued that the terms have different meanings. The motive behind outlier detection is to find the type of abnormality in any data set that can be identified as an outlier. Boukerche et al. (2021) placed a strong emphasis on the subtle difference between them. While anomalies suggest a different understanding of the underlying generative mechanism, outliers highlight statistical rarity and deviation, and whether they are generated by a different mechanism is not directly addressed. Anomalies are the preferred term in supervised learning since there is solid advice to represent the aberrant generating mechanism. Unsupervised learning approaches, on the other hand, rely on the inherent distribution or structure of the data to quantify the deviation from the norm, hoping that the detected outliers reflect the anomalies of interest owing to a lack of solid guidance. As far as

statistics and machine learning are concerned, outliers are those instances of data (sometimes erroneous data points) that complicate fitting a particular model One can either remove outliers or use robust models to minimize their impact.

Many factors can contribute to outliers, including data errors, variable construction, omitted variables, sampling errors, and non-normality. What constitutes a sufficient deviation for a point to be considered an outlier is often a subjective judgment (Lightfoot and O'Connell 2016). Usually, deviations of interest are those that have a significant impact on the data. Various mechanisms give rise to samples that show outliers in them. In one mechanism, data proceeds from a heavy-tailed distribution. Hawkins (1980) classifies statistical distributions into two families, including outlier-prone and outlier-resistant families, respectively. Outlier-prone families tend to have tails that approach zero slowly. An outlier-prone family is said to be absolute if there exists ε, δ > c > 1, δ > 0, and n0 > 0 such that:

$$\Pr\left[X_{n,n} - X_{n,n-1} - 1 > \varepsilon\right] \geq \delta \text{ for all } n > n_0 \quad (1)$$

$$\Pr\left[X_{n,n}/X_{n,n-1} - 1 > c\right] \geq \delta \text{ for all } n > n_0 \quad (2)$$

where X_n is the statistic based on sample size n and n_0 is an integer.

There is a propensity for the most significant order statistic to be suspiciously large compared to its predecessors if either of these conditions holds. Absolute and relative outlier-resistant distributions are not absolute and relative outlier-prone, respectively. The former class includes the normal family distributions, and the latter the gamma family (Hawkins, 1980). In the second mechanism, data proceeds from two distributions: primary and contaminating. Basic distribution generates good observations, while contaminating distribution generates contaminants. Contaminating distributions with heavier tails are more likely to be outliers than

basic distributions because they can be separated from the good observations, which become inliers.

2.2. Outlier Categorization

Outliers can be organized categorically in different ways. First, outliers can be categorized based on the number of data instances involved to comprise an abnormal pattern, such as point outliers and collective outliers (Smith & Bryant, 1975). A point outlier can be defined as an individual that deviates mainly from the rest of a dataset. Point outliers can be further subclassified as local outliers and global outliers based on the scope of comparison. Whereas the detection of local outliers relies on the differences in characteristics between an outlier and its nearest neighbors, global outliers differ significantly from the entire dataset. The concept of local outliers was first introduced by Breunig et al. (1999). Collective outliers are a group of occurrences that appear anomalous from the rest of an entire dataset (Boukerche et al., 2021b).

Outliers can also be categorized based on the type of input data, that is, vector outliers and graph outliers (Boukerche et al., 2021b; Zhang, 2013). Graph outliers are associated with graph data, while vector outliers are associated with vectorlike multidimensional data. Multiple attributes are related to a vectorlike data point, each with a numerical categorical value. The interdependencies between data objects are best represented by the nodes and edges that make up graph data. Examples include node outliers, edge outliers, and subgraph outliers (Zhang, 2013).

2.3. Data Modeling and Detection of Outliers

In all outlier detection algorithms, the normal patterns in the data are modeled first, then the deviations from these patterns determine the outlier score for a given data point. Following Aggarwal (2017), a data point's outlier score is computed by evaluating the quality of its fit to

the model. Many times, the model is algorithmically defined. By analyzing the distribution of its k-nearest neighbor distance, nearest neighbor-based algorithms can predict the outlier tendency of a data point. In this case, outliers are assumed to be located at a distance from the majority of the data. It is thus assumed that outliers are located far from the majority of data in this case. The underlying presumption is that outliers are spread widely apart from most of the data. There is no doubt that the choice of the data model is crucial. Poor results may be produced from an incorrect choice of data. For example, Aggarwal (2017) shows that if underlying data is clustered arbitrarily, a linear model may not work.

Comparably, the Gaussian mixture model may not work if sufficient data points are not available to learn the parameters of the model. As a result, outliers are sometimes reported incorrectly due to poor fit of the erroneous assumptions of the model. Distinguished from supervised data mining problems where there are labeled examples to learn the best model, outlier detection is largely an unsupervised problem with few examples to learn the best model for a particular data set. Hence, it is important to carefully evaluate the relevant modeling properties of a data domain before constructing an effective model for it. Again, there are trade-offs that exist with model choice. A simple model created with an understanding of the data will likely produce much better results compared to a highly complex model with too many parameters which overfits the data and outliers. However, an oversimplified model is likely to declare normal patterns as outliers.

2.4. Classification of Outlier Detection Methods

There are different methods associated with outlier detection depending on the type of data. For time series data, Ajith Kumar et al. (2019) identified five methods:

2.4.1. Clustering

The dataset is separated into several clusters based on mathematical distance or another metric. A point is then termed an outlier if the point does not belong to any of the clusters (i.e., a single point cluster). This method is used for supervised detection and sometimes for semi-supervised detection).

2.4.2. Density Estimators

A point is considered to be an outlier based on assigned weights to points in a dataset. This method does not give a binary result, and it is based primarily on the relative neighborhood theory. Thus, whether an observation is considered an outlier or not depends upon the mathematical distance of the observation to its neighborhood centroid value.

2.4.3. Classifiers

Classifiers are widely used in statistical and machine-learning applications to assign a label or category to a data point based on its attributes. There are various types of classifiers, including logistic regression, naïve Bayes, K-nearest neighbors (KNN), support vector machines (SVM), decision trees, random forests, and neural networks. The choice of the classifier depends on the specific task and characteristics, such as the number of classes, the distribution of the features, and the size of the training dataset. In a review of various classifiers used in machine learning, Kotsiantis et al. (2007) noted that the performance of a classifier is dependent on factors such as the number of classes, the dimensionality of the feature space, the size and quality of the training dataset, and the nature of the decision boundary. Under this method, a model is established for historical time series data using either a regression model or a vector regression model. One of the novel applications of this approach is the graph-based outlier detection technique. The Classifier-based outlier detection technique is used in machine learning to detect

anomalous data points in a dataset. In this approach, a classifier is trained on most of the data points. It is then used to predict the class label of each data point. Data points assigned different class labels than the majority are considered outliers. The classifier-based outlier detection technique was applied to Wireless Sensor Networks (WSN) data with success, and the results were promising. Janakiram et al. (2006) proposed the use of an outlier detection scheme based on Bayesian belief networks, which captured the conditional dependencies among the observations of the attributes to detect outliers in the sensor streamed data in WSN. Their approach improves accuracy in detecting the outliers and missing values, demonstrating its effectiveness and reliability.

2.4.4. Fixed Size Windows

The fixed size windows are a subset of consecutive observations within observations within a dataset that have a pre-determined size or length. They are used in time series analysis and signal processing to compute statistical measures within a specific time interval or window. This methodology is dependent on the observation that a longer time-variant series can be separated into fixed smaller time series windows. Outliers can then be searched for after the division. There is the possibility of an outlier showing in the window if there was an outlier in the original time series. Compared with factor-based approaches, the method shows a significant level of performance.

2.4.5. Distance Measures

Knox and Ng (1998) first introduced the distance-based outlier detection techniques. According to them, "An object p in a dataset DS is a DB (q, distance) outlier if at least fraction q of the object in DS lies at a greater distance than the distance from p". The simple approach was improved by adding a rank based on the distance and using the rank as an outlier score by

(Ramaswamy et al., 2000.). A new concept, hubness-awareness was introduced to increase the efficiency of distance-based outlier analysis and the results were shown to be promising when applied on multidimensional datasets. In predicting the efficiency of the algorithm performance, the structural characteristics of the problem play an important role. The metadata of the problem can be extracted and used as an important parameter towards the outlier analysis and detection approach. (Ajith Kumar et al. 2019).

2.5. Outlier Labeling Techniques

There are many statistical techniques used for outlier identification. This paper discusses two kinds of outlier detection methods: formal tests and informal tests. Formal tests are usually referred to as tests of discordancy while informal tests are referred to as outlier labeling methods.

2.5.1. Formal Tests

Formal tests are usually based on the assumption of well-behaving distribution, and if the extreme value is an outlier of the distribution regardless of whether it deviates from the distribution. There are tests for single outliers and tests for multiple outliers. The choice of these tests largely depends on number and type of target outliers, and type of data distribution (Acuna and Rodriguez 2004). Iglewicz & Hoaglin, (1993b) and Barnett and Lewis (1994) have discussed various tests according to the choice of distributions.

2.5.1.1. Grubbs Test

The Grubbs test is also known as the Pearson-Hartley or the Extreme Studentized Deviate tests for a single outlier. Although Grubbs (1950) is the primary source, it might be challenging to find the actual source for the G statistic as it is described by sources like Wikipedia and NIST. Grubbs' test is defined as $G = \max|x_i - u|/s$ where u is the sample mean and s is the sample

standard deviation. It is used to detect a single outlier in a univariate data set that follows an approximately normal distribution (Grubbs 1969 and Stefansky 1972).

2.5.1.2. Dixon Test

While the Grubbs test is used for relatively large sample size (N>30), the Dixon test (Dixon, 1950) is used for small sample size (N<30). The Dixon test also tests for single outliers and is considered superior to the Grubbs test. The idea is to compare the gap between the smallest or largest value and its adjacent value to the range, given a test statistic.

2.5.1.3. T-tests

The t-test is a statistical test used to determine if there is a significant difference between the means of two groups. It is commonly used in hypothesis testing to compare the means of two samples, such as the mean test scores of two groups of students. The t-test is based on the t-distribution, which is a probability distribution that considers the sample size and standard deviation of the samples being compared . There are two types of t-tests: the independent-samples t-test and the paired-end t-test. The independent samples t-test is used to compare the means of two independent groups, while the paired samples t-test is used to compare the means of two related groups, such as before and after treatment.

2.5.1.4. ANOVA Test

Compared to T-test, ANOVA is used when comparing means of three or more independent groups or samples. ANOVA assumes homogeneity of variances between all groups being compared. In other words, T-test is a special case of ANOVA, where there are only two groups. If the assumption of variance homogeneity is violated, a modified version of ANOVA called the Welch's ANOVA can be used. There are several different types of ANOVA tests, including one-way ANOVA, two-way ANOVA, and repeated measures ANOVA.

2.5.1.5. Wilcoxon rank-sum test

The Wilcoxon rank-sum test, also known as Mann-Whitney U test, is a nonparametric statistical test used to compare two independent samples. It is often used as an alternative to the T-test when the assumption of normal distribution is violated, or when the data is ordinal or skewed. The Wilcoxon rank-sum test involves ranking all the observations from both groups and calculating the sum of ranks for each group. The test statistic is then calculated as the smaller of the sum of ranks for the two groups, and the expected value of this statistic under the null hypothesis is calculated using a reference distribution (Mann, H. B., & Whitney, D. R., 1947)

2.5.1.6. Kruskal-Wallis H test

The Kruskal-Wallis H (1952) test is a nonparametric statistical test used to compare the means of two or more groups. It is similar to the ANOVA, but it is more robust and can be used when the assumptions of the ANOVA are not met, that is assumption of normality is violated, or when the data is ordinal or skewed. The Kruskal-Wallis H test involves ranking all the observations from all groups combined and calculating the sum of ranks for each group. The test statistic is then calculated as a function of the sum of ranks, and the expected value of this statistic under the null hypothesis is calculated using a reference distribution.

2.5.1.7. Chi-square test of independence

The chi-square test of independence is a non-parametric test used to determine whether there is a significant association between two categorical variables. The test statistic is calculated by summing the squared differences between the observed and expected frequencies and dividing by the expected frequencies. The resulting chi-square statistic follows a chi-square distribution with (r-1) (c-1) degrees of freedom, where r is the number of rows in the contingency table and c is the number of columns. This statistic is then compared to a critical

value to determine whether the two variables are independent. If the chi-square statistic is greater than the critical value, then the two variables are dependent. Chi-square test of independence can be used to test for independence in a variety of situations, including comparing proportions, testing for association, and testing for goodness of fit.

2.5.2. Informal Tests

Conversely, informal labeling tests involve flagging potential outlier erroneous data, indicative of an inappropriate distributional model for further investigation. Most informal outlier labeling approaches create an interval or criterion for outlier detection rather than doing hypothesis tests, and any observations that fall outside of the interval or criterion are regarded as outliers. Even though the labeling approach is typically straightforward to apply, if the outliers are specified as just data that differ from the assumed distribution, then some observations outside the interval may be incorrectly detected after a formal test.

Unlike the formal test, z-score, a labeling method measures how far the mean is from a data point. It can be place on a normal distribution curve and ranges from -3 standard deviations up to +3 standard deviations. The modified z-score, more robust to the z-score, measures outlier strength or how much more the scores differ from the typical score.

Tukey's (1977) robust IQR method(boxplot) for labeling outliers is a graphical tool for displaying information about continuous univariate data, such as the median, lower quartile, upper quartile, lower extreme, and upper extreme of a data set.

2.5.2.1. Z-Score

The z-score or standard score of an observation is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a gaussian distribution. Z-scores are a means of expressing the deviation of a given anatomic or physical measurement

from a size- or age-specific population mean (Curtis et al. 2016) This makes z-score a parametric method. the z-score of any data point can be calculated with the following expression:

$$Z = \frac{(x-\mu)}{\sigma}$$

where χ = the observed measurement,

μ = the expected measurement (population mean), and

σ = the population standard deviation

When the Z-score is above the population mean it will have a positive value, while when it is below the population mean it will have a negative value. The greater the deviation from zero, the greater the magnitude of the deviation (Curtis et al. 2016).

Z-score can be used in different fields, such as agriculture, health, and business, among many others, to determine an outlier. Curtis et al., (2016) demonstrated using Z-scores in measuring the thoracic to determine treatment strategies in aneurysmal disease. The Z-scores allowed them to determine whether actual pathology exists, which can be challenging in growing children. In addition, Z-scores allow for thoughtful interpretation of aortic size in different genders, ethnicities, and geographical regions. The advantage of the Z-score is its inclusion of body surface area (BSA) in determining whether an aorta is within normal size limits. An aorta outside the regular size limits presents itself as an outlier and, hence of interest for treatment.

For a Z-score to be calculated, the mean and standard deviation for that body structure (e.g., aortic root diameter) was determined in the population. The mean and standard deviation were calculated in many individual studies of varying sample sizes. The individual studies were used to generate nomograms. A parameter of interest (e.g., aortic root diameter) was then recorded for each individual, allowing the generation of a scatterplot (Figure 1) and the calculation and plotting of a regression equation and confidence intervals. The scatterplot was

transformed into a nomogram (Figure 1B), allowing them to determine the Z-score for an individual patient given their BSA and parameter of interest (e.g., aortic root diameter). As a result of their extensive evidence base, Z-scores are a convenient tool for diagnosing and monitoring cardiovascular disease and for determining treatment efficacy in aortic aneurysms.

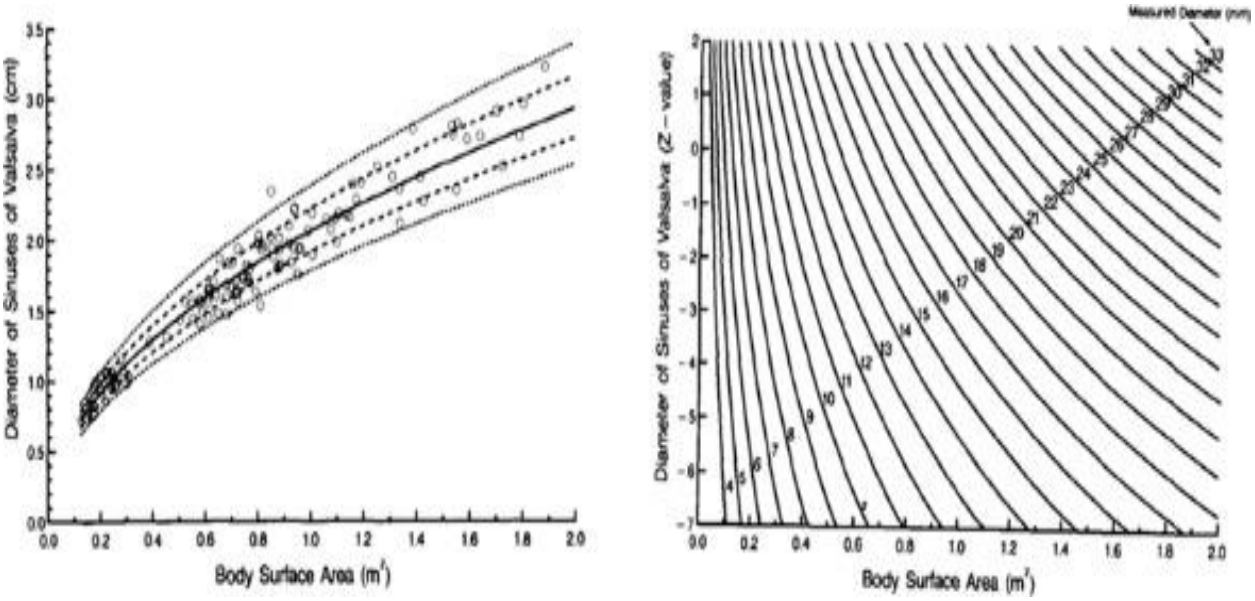

Figure 1. Panel A. Scatterplot of individual patient data illustrating the relationship between body surface area and diameter of the Sinuses of Valsalva. Solid line regression = equation; dashed line = confidence Intervals. Panel B. Nomogram generated from the scatterplot in Panel A for determining individual Z-score according to body surface area and diameter of the Sinuses of Valsalva (Curtis et al. 2016).

2.5.2.2. Cook's Distance

There are many techniques to remove outliers from a dataset. One method that is often used in regression settings is Cook's Distance (D). It has been studied by several authors, such as Cook (1977), Besley et al., (1980 p. 201) under Gaussian errors, the identification problem of outliers or influential data in univariate or multivariate linear regression. In regression analysis, Cook's distance, Di, is used to identify outliers in a set of predictor variables, which can negatively impact your model. In other words, it's a way to identify outliers. Cook's distance is

determined by combining the leverage and residual values of each observation, the higher the leverage and residuals, the greater the distance. Technically, Cook's D is calculated by removing the ith data point from the model and recalculating the regression. It summarizes how much all the values in the regression model change when the ith observation is removed. The formula for Cook's distance is:

$$D\mathrm{i} = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\hat{\sigma}^2} \quad (3)$$

For example, Jagadeeswari et al. (2013) used Cook's distance to identify outliers in agriculture datasets. They used the Principal Component Analysis-Minimum Volume Ellipsoid (PCA-MVE), which is one of the data techniques along with Cook's distance. To do this, he applied the technique of Gentleman and Wilk (1975) and John and Draper (1978), who investigated the problem of detecting outliers in a two-way table and provided a statistic, Qk, which is difference between the sum of square of residuals from original data and sum of squares revised residuals resulting from fitting the basic model after deleting K-influential observations. The Cook statistics also detect the outlying observations in experimental data. The results showed that such detection of the outliers helps in identifying errors in agriculture data set.

3. THEORY AND CONCEPTUAL FRAMEWORK

3.1. Research Procedure

This section briefly describes the procedure used in the study. To start the study, we asked faculty members of the Department of Agribusiness and Applied Economics for ideas about data containing potential outliers. We got some suggestions like the whole electricity data and spreads data. We had seven (7) candidate data series for the study (refer to the chapter for data description). We then applied Grubb's test in XLSTAT (Addinsoft., 2023), an add-in software for Microsoft Excel, to confirm the presence of outliers in the candidate series. The results from Grubb's test confirmed the presence of outliers in all the candidate series.

Once we confirmed the presence of outliers in all the candidate series, we applied the outlier labeling procedures: Z-Score, Tukey IQR, and K-means clustering to each series. Each dataset for each candidate series was divided into base datasets and upper datasets or lower datasets if indicated by the labeling procedures. We fitted distribution to each candidate labeled set using @Risk (Palisade Software., 2023), an add-in software for Microsoft Excel. We ignored the outliers for full datasets, which is one candidate dataset. We removed outliers for truncated datasets identified by the modified Z-Score procedure. For each labeling procedure used, we labeled the separated datasets as base datasets, and upper outlier dataset if indicated or lower outlier dataset if indicated. The lower datasets were transformed by taking the negative transform so that we could fit more long-tailed skewed distributions. Each of the labeled dataset was then combined using the Bayesian Averaging Procedure (described in section 3.4.1). In the case of two outlier distributions, we use the beta distribution to generate random probabilities for each subset based on their observed frequency in the dataset. Similarly, we used multivariate beta

(i.e., Dirichlet) for three outlier distributions to generate random probabilities for each subset based on their observed frequency in the actual dataset.

Finally, we compared the Goodness-of-Fit to the actual dataset using the Two-Sample KS and AD Tests. That is, we compared the CDF of the datasets to actual datasets. For fitting the distributions, we used the one-sample test where we had a candidate and tested where the candidate fit the data.

3.2. Labeling Procedures Utilized in the Study

Labeling outliers involves identifying extreme values in a data set and assigning them a label or category. Labeling is essential to data analysis and interpretation, allowing us to organize and understand complex data sets. One approach to labeling outliers is using statistical methods to identify data points significantly different from the rest of the data. For example, the modified z-score and Tukey IQR methods are commonly used techniques for identifying outliers based on their distance from the mean or quartile ranges respectively (Aggarwal, 2017). The objective of the procedure is to flag potential outlier erroneous data indicating an inappropriate distributional model (Iglewicz and Hoaglin, 1993) of the datasets and to compare them using the Kolmogorov-Smirnov statistic goodness of fit test and the Anderson-Darling test. These two tests are considered the basis of comparison to ensure consistency, coherence, and correspondence.

3.2.1. Modified Z-Score

The standard Z-score is sensitive to the influence of values and can be affected by outliers in a dataset. Two estimators of Z-score: sample mean ($\tilde{x}$) and sample standard deviation can be affected by extreme values or even a single value. To avoid this problem, the median and the median of the absolute deviation (MAD) are employed in the modified Z-score (Iglewicz and Hoaglin, 1993). MAD is one of the basic robust methods which is not affected by the presence of

extreme values in a dataset (Burke, 1998). The median and MAD must first be estimated to calculate the modified Z-score for a data point. The median is the middle value of data sorted into ascending order. The sample mean has a breakdown point of approximately 50%, but the exact percentage depends on the sample size, even or odd. The breakdown point of the median for n odd is $\tilde{x}$ = xm or 100((n-1)/2)/n% = 50((1-1)/n)% and the corresponding n even is $\tilde{x}$ = (xm+xm+1)/2 or 100((n-2)/2)/n= 50((1-2)/n). MAD is defined as follows:

2MAD Method: Median ± 2MAD

2MAD Method: Median ± 2MAD, where MAD = 1.483 x MAD for large normal data and is an estimator f the spread in a data.

MAD = median $\{|x_i - \tilde{x}|\}$, where $\tilde{x}$ is the sample mean.

The modified z-score method assumes that if the data is usually distributed, approximately 50% of the values should have a z-score between -1.96 and 1.96, while 95% of the values should have a z-score between -2.58 and 2.58 (Iglewicz & Hoaglin, 1993). However, in practice, many data sets may not be normally distributed, and extreme values or outliers can significantly affect the mean and standard deviation. The modified z-score method is designed to be more robust in these situations and is less likely to incorrectly identify legitimate data points as outliers. The modified Z-score can be computed as

$$\text{Mi} = \left(\frac{0.6745(x_i - \tilde{x})}{MAD}\right) \quad (4)$$

where E(MAD) = 0.675σ for large normal data.

Iglwicz and Hoaglin, (1993) suggested that observations are labeled outliers when $|M_i| >$ 3.5 through the simulation based on pseudo-normal observations for sample sizes 10,20 and 40.

We selected the Modified Z-Score option in the Grubb's procedure in XLSTATS, an add-in software for Microsoft Excel for the candidate datasets. All scores greater than or equal to 3.5

were labeled upper datasets for each of the series. All scores less than or equal to -3.5 were labeled lower datasets for each series while scores outside of these ranges labeled as base datasets for each of the series.

3.2.2. Tukey's Interquartile Range (IQR)

Tukey's interquartile range (IQR) is a statistical technique for identifying outliers in a data set. It is based on the difference between the upper and lower quartiles of the data and is less sensitive to extreme values than other outlier detection methods that use the mean and standard deviation. The interquartile range is a valuable tool for describing the spread of a given set of data or distribution. It is typically used when there are outliers present in the distribution in the distribution or the distribution is skewed. The interquartile range value is estimated by subtracting the first quartile (Q1) from the third quartile value (Q3). The population IQR for a continuous distribution is defined to be:

$$IQR = Q3 – Q1 \quad (5)$$

Tukey's (1977) method of constructing a boxplot is a graphical tool to display information about the continuous data, such as the median, lower quartile, upper quartile, lower extreme, and upper extreme of a data set. The method has the following rules:

$$\text{Low outliers} = Q1 - 1.5IQR \quad (6)$$

$$\text{Upper outliers} = Q1 + 1.5IQR \quad (7)$$

IQR is the difference between inner fences and outer fences. The interval with 1.5IQR (inner fences) is situated below the Q1 and Q3 at 1.5 IQR3 distance. The interval with 3IQR (outer fences) is situated below Q1 and above Q3 at 3IQR distance. The observations among the inner fences and outer fences are considered potential outliers (Saleem, Aslam, & Shaukat, 2021). This approach can detect more observations as outliers as the measure of skewness in the

data increases (Seo, 2021). Figure 2: components of a classic boxplot show the components of the boxplots. In figure 3, the Tukey boxplot shows the analysis of Indiana Electricity Peak data.

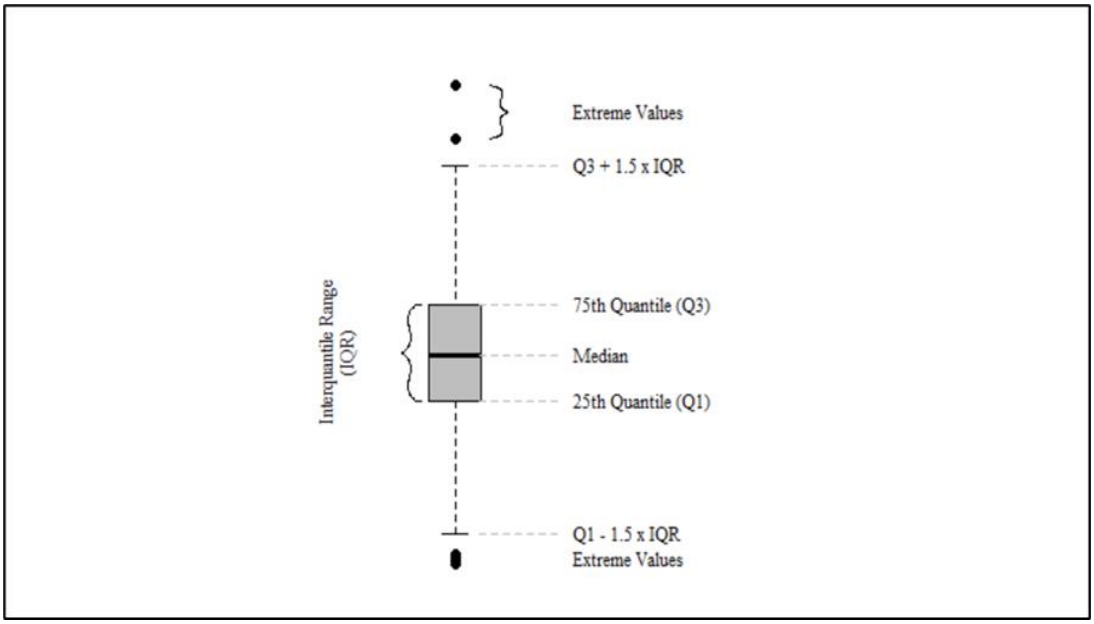

Figure 2. Components of a classic Tukey boxplot with Interquartile Range (IQR) illustrated.

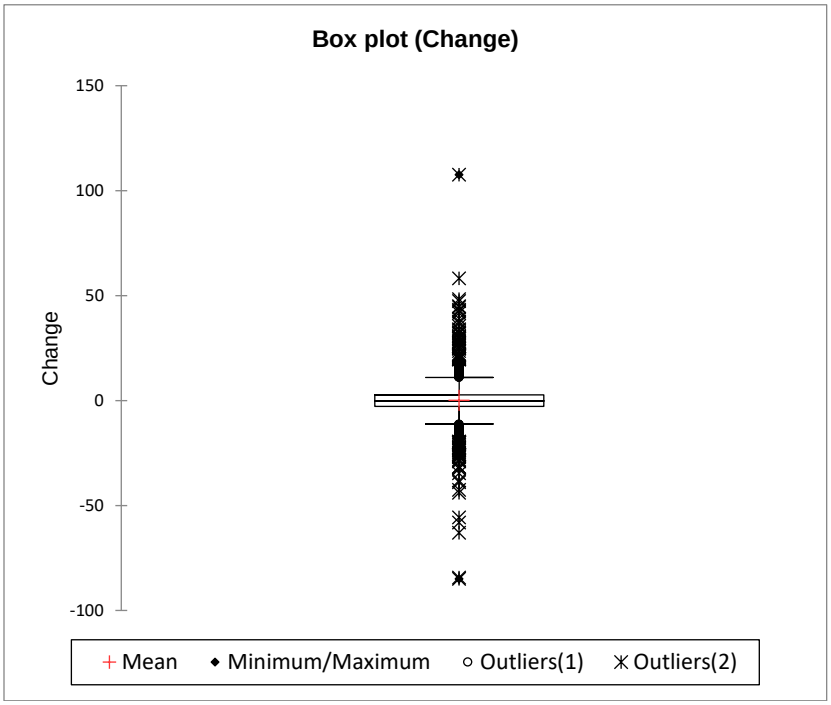

Figure 3. Application of Tukey boxplot to Indiana Electricity price change data.

We followed Tukey IQR's procedure described above and manually calculated the IQR from the descriptive statistics of each of the candidate series. We applied the Q1- 1.5IQR for lower outliers and Q1+1.5IQR for upper outliers. Aside from the DCV data series, the lower outliers and upper outliers were computed for all the candidate series for each dataset.

3.2.3. K-means Clustering

Data clusters are frequently encountered when sampling from a population because several distinct subpopulations may exist within the population. It is much easier to visually detect clusters in univariate or bivariate data, however, the task becomes complex as the dimensionality of the data increases. K-means is a clustering algorithm based on a partition where the data is only entered into one K cluster, the algorithm determines the number of grub in the beginning and defines the K centroid. (Nayak et al. 2015). A data cluster can be treated collectively as one group and may be considered as a form of data compression (Han & Kamber, 2001). James MacQueen introduced K-means in 1967 (MacQueen 1967). K-means clustering uses various distance functions to measure the similarity among the objects. To measure the similarity among the datasets, distance plays a significant role. It identifies the way datasets are interrelated, how various data are similar and which measures are used for comparison. As a result, distance metrics functions are calculated based on which data are clustered. The method of K-means begins with the random selection of k number of objects and is represented as cluster means. For each of the residual objects, a similar object is assigned which helps to complete a new cluster mean depending on the distance metric between the object and the cluster mean. For this work, the k-means algorithm using Euclidean distance function was used.

Euclidean distance is a widely used measure of distance between two points in a multidimensional space which is a space that obeys the axioms of Euclidean geometry (Hastie,

Tibshirani, & Friedman, 2009). The Euclidean distance can be calculated using the Pythagorean theorem: $a^2 + b^2 = c^2$, and the formula can be extended to higher-dimensional spaces. In machine learning, the Euclidean distance is widely used to measure the similarity or dissimilarity between data points in a dataset. For example, in k-means clustering, the distance between each data point and the centroid of each cluster is calculated using the Euclidean distance (Tan, Steinbach, & Kumar, 2019). The Euclidean distance is also commonly used in other machine learning algorithms, such as k-nearest neighbors, support vector machines, and principal component analysis. The Euclidean distance is the prototypical example of the distance in a metric space and defines all the properties of a metric space (Ivanov, 2013):

1. It is symmetric; that is, for all points *p* and *q*, *d(p,q) = d(q,p)*
2. It is positive; that is, the distance between every two distinct points is a positive number, while the distance from any point to itself is zero.
3. It obeys the triangular inequality. That is, for every three (3) points p, q and r, *d(d,p) + d(q,r) ≥ d(p,r).*

It is used in many applications such as clustering, classification, and data visualization. To calculate the K-means by the Euclidean distance method for each of the candidate series of each datasets (electricity, oats futures, DCV data, wheat daily intermarket spread and daily intermonth spread) we applied the K-means function in *XLSTAT*, an add-in software for Microsoft Excel, we followed the procedure: let $X=\{x_1, x_2, x_3, \ldots, x_n\}$ be the set of data points and $B=\{b_1, b_2, b_3, \ldots, b_n\}$ be the set of clusters.

1. Define 'N' be the number of clusters.
2. Suppose 'C' randomly selected as cluster center.

3. Calculate the distance of each data point from cluster centers' C' using Euclidean distance.

$$\text{distxy} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \tag{8}$$

4. Data point is assigned to the cluster center whose distance "dist" from cluster center is minimum.
5. The cluster center is recalculated using,

$$b_t = \left(\frac{1}{c_i}\right)\sum_{1}^{c_i} x_i \tag{9}$$

6. The distance between the new cluster center and data points is calculated and data point with minimum distance from a particular cluster center is assigned to that cluster center.
7. If no data point was reassigned then stop, otherwise repeat steps from 3 to 6.

We then used the shadow scores to pick the optimal number of clusters from the data. There were three (3) categories identified for KW-W Intermarket Spread, W Intermonth Spread, Indiana Hub, and NEPOOL. PJM had four (4) categories while two (2) categories were identified for DCV data. For three (3) clusters, we labeled the third cluster as the lower outlier while the first cluster was labeled base and the second or middle cluser labeled as upper outlier. The DCV dataset had two clusters and were labeled base and upper outlier. For four clusters as was the case of PJM Hub, the two clusters, that is, cluster three (3) and cluster four (4) were grouped into the same cluster. They were then labeled in the same order as three (3) clusters.

3.3. Distribution Fitting Procedures Used in the Study

3.3.1. Palisade Bestfit Procedure

The Palisade BestFit procedure is a statistical software tool available in the @Risk software, which is widely used in risk analysis and decision-making in various industries such as

finance, engineering, and healthcare. The BestFit procedure in @Risk is designed to fit input data to a range of statistical distributions, including normal, lognormal, Weibull, exponential, and other distributions. The procedure uses the maximum likelihood method (MOM) to determine the distribution that best fits the input data. For the distribution artist and distributions with poorly defined derivatives, Bestfit uses a hybrid between MOM and ordinary least squares (OLS) regression to calibrate the shapes of the data and the theoretical probability density function (PDF), cumulative distribution function (CDF) or probability mass function (PMF).

Maximum likelihood estimator (MLE) is a statistical method used to estimate the parameters of a probability distribution based on a set of observed data. The basic idea of the MLE is to find the values of the distribution parameters that maximize the likelihood of observing the given data. The likelihood function is a function of the parameters that describe the probability of observing the data given a specific set of parameter values. The MLE estimates the parameter values that maximize the likelihood function, that is, the parameter values that make the observed data most probable. For the distribution of the candidate series, the log-likelihood function was derived using:

$$L(x;\Theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\ f(x_i|\theta) \tag{10}$$

where the xi are the n observed values, and Θ is the vector of parameter values for the distribution in question. The log-likelihood function with respect to the parameter vector Θ was maximized by taking the partial derivate of L with respect to Θ and set to zero.

To use the Palisade Bestfit procedure contained in @Risk software, we fitted candidate distributions to data using MLE. The procedure compared the likelihood values for each candidate distribution and ranked the distributions based on parametric statistics tests (Chi-squared), non-parametric statistical tests (Kolmogorov-Smirnov (KS) and Anderson-Darling

(AD) and information criterion (Akaike Information Criterion (AIC) and Bayesian (Schwarz) Information Criterion (BIC). However, we did not use the Chi-squared because it is better for discrete than continuous fitting due to binning of data. If the four criteria used (KS, AD, AIC, BIC) were not unanimous in their recommendations, then we used visual examination of the Probability-Probability (P-P) plots to break the tie. If we could not visually distinguish a dominant distribution, then we went with the recommendation from the BIC criterion. One of the significant benefits of using the BestFit procedure in @Risk is that it provides a comprehensive list of statistical distributions to choose from, allowing users to find the best fit for their data.

3.3.2. Review of Kolmogorov-Smirnov Test

The goodness of fit refers to a statistical concept that determines how well sample data fits a distribution from a population with a normal distribution or one with a Weibull distribution (Jaccard & Becker, 2019). Measures of goodness of fit summarize the difference between observed values and the expected values under a model to see if there is a significant difference between the two. Such measures can be used in statistical testing to test for the normality of residuals, test whether two samples are drawn from the identical distribution, or whether the outcome follows a specified distribution. There are several methods for assessing goodness of fit, including graphical methods and statistical tests: the chi-squared test, the Kolmogorov-Smirnov test, and the Anderson-Darling test. The choice of test depends on the specific characteristics of the data and the theoretical distribution being tested. For the purpose of this paper, the Kolmogorov-Smirnov test and the Anderson-Darling test are applied.

The goodness of fit statistics measures the compatibility of random samples against some theoretical probability distribution function. The Kolmogorov-Smirnov test (K-S test or KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions

that are used to compare a sample with a reference probability distribution (one-sample K-S test) or to compare two samples (two-sample K-S test). It assumes that a list of data points can be easily converted to a cumulative distribution function. The test uses the maximum absolute difference between two cumulative distribution functions. The K-S statistic, when comparing one data set F(x) against a known cumulative distribution function P(x), is

$$DKS = \max|F(x) - P(x)| \tag{11}$$

When comparing two samples with cumulative distribution functions F(x) and G(x), the statistic is defined as

$$DKS = \max|F(x) - G(x)| \tag{12}$$

The Kolmogorov-Smirnov Test can be graphically represented in Figure 4.

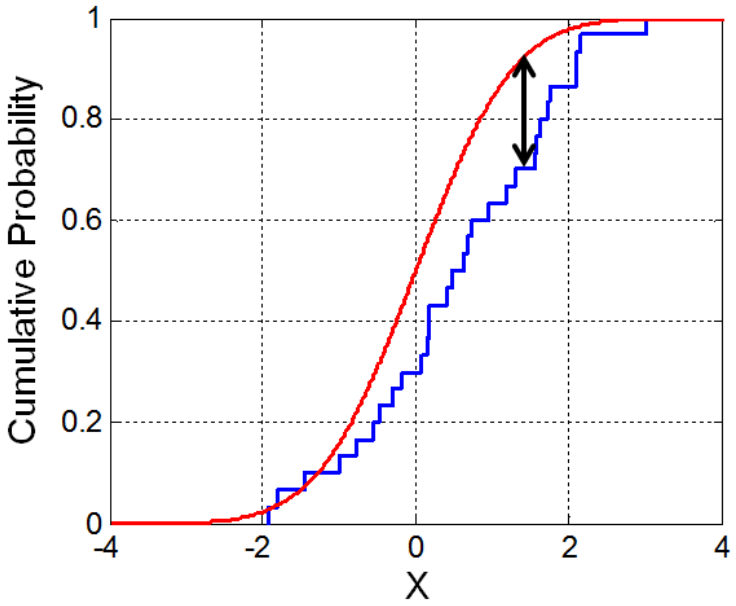

Figure 4. Cumulative Probability

The Kolmogorov distribution, also known as the Kolmogorov-Smirnov distribution, is a probability distribution that describes the maximum deviation between a cumulative distribution function (CDF) and a hypothetical reference distribution. The Kolmogorov distribution is used in statistical tests, such as the Kolmogorov-Smirnov test, to determine whether a given dataset

follows a specified distribution. The Kolmogorov distribution is a continuous distribution with support on the interval [0,1]. The probability distribution is graphically represented in figure 4. Its PDF is given by:

$$f(x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} (e)^{\frac{-(2k-1)^2 x^2}{8x^2}}, \quad k=1,2,3,... \tag{13}$$

where x is a value between 0 and 1, and the summation goes over all positive integers k. The Kolmogorov-Smirnov statistic for a given cumulative distribution function F(x) is

$$D_n = sup_X |F_n(x) - F(x)| \tag{14}$$

where sup_X is the supremum of the set of distances. Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values.

The Kolmogorov distribution is important in statistical inference because it provides a way to quantify the goodness-of-fit of a distribution to a dataset. Specifically, the Kolmogorov-Smirnov test uses the Kolmogorov distribution to determine whether a given dataset is consistent with a specified distribution (refer to figure 5).

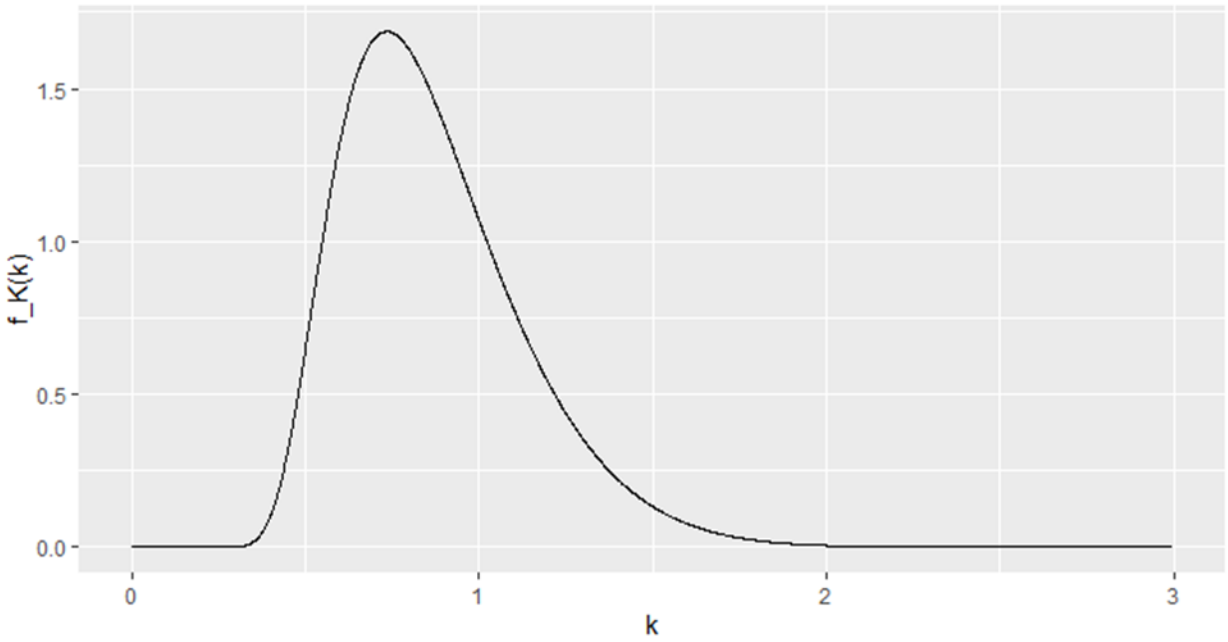

Figure 5. Illustration of the Kolmogorov-Smirnov Distribution PDF

3.3.2.1. One-Sample Kolmogorov-Smirnov Test Statistic

The one-sample Kolmogorov-Smirnov (KS) test is a nonparametric statistical test that is used to determine whether a sample follows a specific distribution or not. The test compares the cumulative distribution function (CDF) of the sample with the CDF of a specified theoretical distribution, such as the normal distribution or the uniform distribution. The null hypothesis of the one-sample KS test is that the sample comes from the specified distribution, while the alternative hypothesis is that the sample does not come from the specified distribution. The one-sample KS test is often used when the underlying distribution of the sample is unknown, and it is used to test whether the sample can be considered representative of the population. The test is based on the maximum absolute difference between the CDF of the sample and the CDF of the specified distribution. This maximum difference is called the KS statistic. The critical value of the KS statistic depends on the sample size and the significance level of the test. If the p-value is less than the significance level, typically 0.05, then the null hypothesis is rejected, and it is concluded that the sample data does not come from the theoretical distribution being tested. The empirical distribution function Fn for n independent and identically distributed (i.i.d.) ordered observations Xi is defined as

$$F_n = \frac{number\ of\ (elements\ in\ the\ sample \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}\,(X_i), \qquad (15)$$

Where $1_{(-\infty,\mathrm{x}]}(X_i)$ is the indicator function and equal to 1 if $X_i \leq x$ and 0 otherwise.

3.3.2.2. Two-Sample Kolmogorov-Smirnov Test Statistic

The K-S test may also be used to test whether two underlying one-dimensional probability distributions differ. It is a non-parametric statistical test used to compare two samples to determine if they come from the same distribution. It is based on the maximum difference

between the empirical distribution functions (EDF) of the two samples. The K-S statistic is given by:

$$D_{\mathrm{n,m}} = sup_x \left| F_{1,n}(x) \; - \; F_{2,m}(x) \right| \tag{16}$$

Where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution of functions of the first and the second sample respectively and sup is the supremum function. For large samples, the null hypothesis is rejected at level of α if

$$D_{\mathrm{n,m}} = c(\alpha) \sqrt{\frac{n+m}{nm}} \tag{17}$$

Where n and m are the sizes of first and second sample respectively. The steps involved in performing the two-sample KS test are as follows:

Calculate the empirical distribution function (EDF) for each sample, which is a step function that assigns a probability of (1/n) for each observed data point, where n is the sample size.

Compute the KS test statistic, which is the maximum absolute difference (D) between the two EDFs.

Determine the critical value of the KS statistic using a significance level (α) and the sample sizes, either from a table or software.

Compare the computed KS test statistic with the critical value. If the computed KS statistic is less than the critical value, then we fail to reject the null hypothesis that the two samples are drawn from the same distribution. Otherwise, we reject the null hypothesis and conclude that the two samples come from different distributions.

The null hypothesis (Ho) is that the two dataset values are from the same continuous distribution. The alternative hypothesis (Ha) is that these two datasets are from different continuous distributions. It is a two-tailed test and can detect differences in both the location and

shape of the distributions. The two-sample KS test is a useful non-parametric statistical test that can be used to compare two samples without any assumptions about the underlying distribution of the data. It is particularly useful for small sample sizes, data with unknown or non-normal distributions, and data with outliers or extreme values.

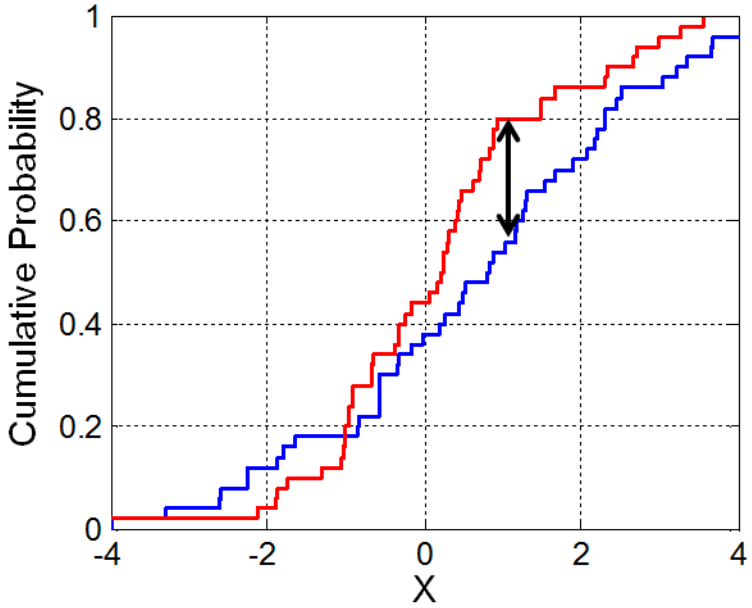

Figure 6. Illustration of the two-sample Kolmogorov-Smirnov statistics. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

3.3.3. Review of Anderson-Darling Test

The Anderson-Darling (AD) test (Stephens 1974) is a statistical test used to determine whether a given sample of data comes from a given probability distribution, such as the normal distribution. The test compares the sample's empirical distribution function (EDF) with the cumulative distribution function (CDF) of the theoretical distribution being tested. It assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values are distribution-free. If the hypothesized distribution is F, and the

empirical cumulative distribution function is Fn, then the quadratic function EDF statistics measure the distance between F and Fn by

$$n\int_{-\infty}^{\infty}|F_n(x)-F(x)|^2\Psi(x)f(x)dx, \tag{18}$$

Where n is the number of elements in the sample, and $\boldsymbol{\Psi(x)}$ is a weighting function. The test was first proposed by Anderson and Darling in 1952 as an improvement over the more commonly used Kolmogorov-Smirnov (KS) test. Compared to the Kolmogorov-Smirnov test, the AD test considers the entire distribution of the sample data, including the tails, and assigns more weight to deviations. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. The Anderson-Darling test statistics is defined as

$$A_n^2=\int_{-\infty}^{\infty}|F_n(x)-F(x)|^2\Psi(x)f(x)dx, \tag{19}$$

which is obtained when the weight function is

$$\Psi(x)=\frac{n}{F(x)\{1-F(x)\}} \tag{20}$$

If the AD test hypothesizes the underlying distribution and assumes that the data does not arise from the distribution, the CDF of the data can be assumed to follow a uniform distribution. The data can then be tested for uniformity with a distance test (Shapiro 1980). The formula for the test statistic $\boldsymbol{A}$ to assess if data $\{Y_1<\ldots<Y_n\}$is obtained from a CDF **F** is

$$A^2=-n\ -\ S, \tag{21}$$

where

$$S=\sum_{i=1}^{n}\frac{2i-1}{n}\left[\ln\left(F(Y_I)\right)+\ln\left(1-F(Y_{n+1-i})\right)\right] \tag{22}$$

The test statistic is then compared to the critical values of the theoretical distribution. In this case, no parameters are estimated in relation to the cumulative distribution function F. If the AD statistic is greater than the critical value, we reject the null hypothesis. Otherwise, we fail to

reject the null hypothesis. The AD test is a powerful tool for testing whether a sample of data comes from a specific probability distribution. It is particularly useful for testing normality or other specific distributions. However, like all statistical tests, it is important to interpret the results in the context of the specific problem being addressed.

3.3.3.1. Review of One-Sample Test

The one-sample Anderson-Darling test is a statistical test used to determine if a given sample of data comes from a particular probability distribution, such as the normal distribution. It is a modification of the Kolmogorov-Smirnov test that places more emphasis on the tails of the distribution. The test calculates a test statistic, the Anderson-Darling statistic, which measures the difference between the observed distribution and the expected distribution, and a p-value that determines the level of significance. Also, more weight is given to the tails of the distribution being fitted. To check the CDF F(x) of sample X to evaluate how well it fits a continuous distribution, sort the sample X in ascending order: x1 ≤ x2 ≤ …xn and perform the one-sample AD test:

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\mathrm{LN}\left(F(x_i)\right) + \mathrm{LN}\left(1 - F(x_{n-i+1})\right)\right] \quad (23)$$

The null-hypothesis that { x1 ≤ x2 ≤ …xn) comes from the underlying distribution F(x) is rejected if AD is larger than the critical value AD. In Stephens (1974), the one-sample AD test was relatively easy to calculate and provided a powerful test for assessing the fit of a given sample of data to a particular probability distribution. The study also provided tables of critical values for the Anderson-Darling statistic under various distributions, including the normal, exponential, and Weibull distributions. Best and Rayner (2006) suggested that a well-performed single test of fit statistic is the AD test statistic. The one-sample Anderson-Darling test is a valuable statistical tool for evaluating the normality of data and the goodness of fit of statistical

models. Its relatively simple calculation and high power make it a popular choice for many applications.

3.3.3.2. Review of Two-Sample Test

The two-sample Anderson-Darling test is a statistical test used to determine whether two independent samples of data come from the same underlying distribution. The two-sample AD test, introduced by Darling (1957) and Pettitt (1976), generalizes to the formula:

$$AD = \frac{1}{mn} \sum_{i=1}^{n+m} \left(N_i Z_{(n+m-ni)}\right)^2 \frac{1}{iZ_{(n+m-i)}} \quad (24)$$

where $Z_{(n+m)}$ is the combined and ordered samples $X_{(n)}$ and $Y_{(m)}$, of size n and m, respectively, and $N_{(i)}$ is the number of observations in $X_{(n)}$ that is equal to or smaller than the *ith* observation in $Z_{(n+m)}$. The null hypothesis that samples $X_{(n)}$ and $Y_{(m)}$ comes from the same continuous distribution is rejected if AD is larger than the correspondent critical value.

3.3.4. Akaike Information Criterion

Akaike's (1974) Information Criterion (AIC) is a statistical method used for model selection and comparison. It is an estimator of prediction error and, thus, a relative quality of statistical models for a given dataset (Stoica and Selen 2004; McElreath 2018; Taddy 2019). Thus, AIC provides a means for model selection. AIC is based on the principle of minimizing the Kullback-Leibler (KL) divergence between the data's true underlying distribution and the model's probability distribution.

Kullback-Leibler (1951) (KL) is a type of statistical distance that measures how one probability distribution P is different from a second, reference probability distribution Q. If P is the data, or a measured probability distribution and Q represents a theory, model, or an approximation of P, then the KL divergence is the average difference of the number of bits required for encoding samples of P using a code optimized for Q than one optimized for P. For

discrete probabilities, P and Q defined on a sample space, X, the relative entropy from Q to P is defined as

$$D_{KL}(P|Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{25}$$

For a distribution $\boldsymbol{P}$ and $\boldsymbol{Q}$ of a continuous random variable, the relative entropy is defined to be the integral as

$$D_{KL}(P|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \tag{26}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are the probabilities densities $\boldsymbol{P}$ and $\boldsymbol{Q}$

For the stated statistical model for the datasets, let $\boldsymbol{k}$ be the number of the estimated parameters in the model. Let $\hat{\boldsymbol{L}}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is

$$AIC = 2k - 2\ln \hat{L} \tag{27}$$

The AIC penalizes models with more parameters since such models are more complex and may overfit the data. Lower AIC values indicate a better trade-off between model complexity and goodness of fit. Hence the model with the least AIC value is considered the best model among the competing models. The AIC is a simple and flexible method for model selection that can be applied to various models. It is more reliable than other methods because it is a principled method that considers both model complexity and goodness of fit rather than one of these factors. However, AIC does not measure uncertainty related to model selection. It is only applicable to models that are based on MLE. Finally, the AIC values may not be meaningful if the models are wrongly specified. Using XLSTAT, an add-in Microsoft Excel AIC was applied among competing models and compared to the BIC's model selection.

3.3.5. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) has a theoretical foundation in Bayesian statistical analysis, especially the Bayes Factor ((Kass and Raftery 1995; Kass and Wasserman 1995; Kass and Vaidyanathan 1992; Kuha 2004). BIC is a statistical measure used to compare different models and select the one that best fits the data while considering the complexity of each model. The computation of BIC is based on the log-likelihood, does not require the specification priors, and is closely related to the AIC (Schwarz 1978). Thus, BIC appeals in many Bayesian modeling problems where priors are hard to set precisely. It is possible to increase the likelihood by adding parameters when fitting models. However, this may lead to overfitting. The BIC is defined as

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{28}$$

where:

$\hat{L}$ = the maximized value of the likelihood function of the model ***M,*** that is $\hat{L} = p(x|\hat{\theta}, M)$

where $\widehat{\boldsymbol{\theta}}$ are the parameter values that maximize the likelihood of

$\boldsymbol{x}$ = the observed data

n = the number of data points in ***x,*** the number of observations, the sample size.

k = the number of parameters estimated by the model.

In a model selection, the minimum value of BIC is chosen as the best model. BIC has the property of consistency for model selection. Theoretically, consistency is arguably the strongest optimality property of BIC. Consistency requires the correct specification of one of the candidate models. This condition is not required by the Bayesian justification of BIC. It is necessary to modify the idea of consistency for BIC if the true model is not included in the candidate collection. Using the BIC selection model, the selected model is likely to converge with

probability one to a model that can be termed quasi-true. In a candidate collection, the quasi-true model is the most parsimonious model closest to the true model in terms of Kullback-Leibler information (Claeskens and Hjort 2008). BIC selects models more parsimonious than those AIC favors as a model selection criterion. Simulation studies show that BIC outperforms other popular model selection criteria, such as AIC, in small to moderate sample size settings, as measured by the proportion of times a criterion selects the correct model structure (Nylund, Asparouhov, and Muthén 2007; Chih-Chien Yang and Chih-Chiang Yang 2007; Shao 1997). However, the BIC has some limitations, such as the assumption of independent and identically distributed data and sensitivity to the choice of prior distributions for the parameters of the model.

3.4. Bayesian Averaging Procedure Used in the Study

3.4.1. Review of Bayesian Averaging

Bayesian Averaging (BA) is a statistical technique used to estimate the parameters of a model by averaging the estimates obtained from multiple models. It is based on the Bayesian theory, which involves assigning probabilities to hypotheses based on available evidence. The first mention of the model combination was provided by Barnard (1963) in a paper studying airline passenger data. However, most of the early work in model combination was not in statistical journals. Early work on model averaging in the statistical literature includes Roberts (1965), who proposed a distribution that integrates the views of two experts (or models). This distribution is similar to BMA since it is simply a weighted average of the posterior distributions of two models. Leamer (1978) developed this concept and provided the fundamental Bayesian Model Averaging (BMA) paradigm. Hoeting et al. (1999), Draper 1995, Chatfield 1995, and Kass and Raftery 1995, provided further review on BMA.

Following the procedure in this study (section 3.1) including the labeling procedures (section 3.2) and estimating the parameters of each distribution of the dataset using the maximum likelihood estimator, the labeled separated datasets or models were combined using the BA simulation procedure. We used the Monte Carlo simulation in @risk for the simulation process.

In a case where the labeling outlier procedure identified only one outlier distribution, that is, upper or lower, we simulated probability weights for the two samples (base and outlier) for observation i as:

$$\{\tilde{p}_{base_i}, \tilde{p}_{outlier_i}\} = \{\text{Beta}(n_{base}+1, n_{outlier}+1), 1-\text{beta}(n_{base}+1, n_{outlier}+1)\},$$

where:

$\tilde{p}_{base_i}$ = simulated frequency probability for base dataset,

$\tilde{p}_{outlier_i}$ = simulated frequency probability for outlier dataset,

n_{base} = number of observations in base dataset,

$n_{outlier}$ = number of observations in outlier dataset.

and '~' over-score indicates a random variable and {.} indicates a vector. We then simulated one random variable from each of the Bestfit fitted distributions ($\tilde{o}_{basei}, \tilde{o}_{outlieri}$). Finally, we combined the two simulated observations using a discrete distribution to simulate an observation (i) of the simulated random variable ($\tilde{x}_i$) as follows:

$$\tilde{x}_i = \text{Discrete}(\{\tilde{o}_{base_i}, \tilde{o}_{outlier_i}\}, \{\tilde{p}_{base_i}, \tilde{p}_{outlier_i}\}).$$

In the case where the labeling procedure identifies two outlier distributions: upper and lower, we simulated the probability weights for the three samples, that is, base, lower outlier, and upper outlier for observations i as:

$$\{\tilde{p}_{base_i}, \tilde{p}_{upper_i}, \tilde{p}_{lower_i}\} = \text{Dirichlet}(n_{base}+1, n_{upper}+1, n_{lower}+1),$$

where:

$\tilde{p}_{base_i}$ = simulated frequency probability for base dataset,

$\tilde{p}_{upper_i}$ = simulated frequency probability for upper outlier dataset,

$\tilde{p}_{lower_i}$ = simulated frequency probability for lower outlier dataset,

n_{base} = number of observations in base dataset,

n_{upper} = number of observations in upper outlier dataset,

n_{lower} = number of observations in lower outlier dataset.

Following the simulation of the probability weights for the three samples, we simulated one random variable from each of the three fitted distributions ($\tilde{o}_{base_i}, \tilde{o}_{upper_i}, \tilde{o}_{lower_i}$). Lastly, we combined the three simulated observations using a discrete distribution to simulate an observation (i) of the simulated random variable ($\tilde{x}_i$) as below:

$$\tilde{x}_i = \text{Discrete}(\{\tilde{o}_{base_i}, \tilde{o}_{upper_i}, \tilde{o}_{lower_i}\}, \{\tilde{p}_{base_i}, \tilde{p}_{lower_i}, \tilde{p}_{upper_i}\}).$$

3.5. Procedure for Comparing Goodness-of-Fit to Original Sample Data

The procedure for comparing goodness-of-fit to original datasets was used to evaluate the accuracy of the models and whether the original dataset (sample 1) best fit the alternate fitting procedures (sample 2). The alternate fitting procedures are simulated to the same number of observations as the original dataset. We used the two (2)-sample KS test and AD test statistics for comparing the goodness-of-fit of the alternate fitting procedures to the original datasets. The KS test involves comparing the CDF of the original dataset to the alternate fitting procedures. The AD test is based on the comparison of the CDF of the original dataset and the EDF of the alternate fitting procedure. The mathematical process for the KS test and AD test has been described in sections 3.3.2 and 3.3.3, respectively.

In Fit1, the labeled ss, ‘Ignore Outlier,’ is fitted to the full sample. For Fit 2, the labeled candidate, ‘Truncate Outlier,’ is fit to sample minus modified Z-Score identified outliers. Fit 3 is the Bayesian Averaging applied to the partition of datasets based upon modified Z-score labeling. Fit 4 is the Bayesian Averaging applied to the partition of datasets based upon Tukey IQR labeling, while Fit 5 is Bayesian Averaging applied to the partition of datasets based upon k-Means cluster labeling. Both the KS and AD tests were used to compare the goodness-of-fit of each of the original datasets to the alternative procedure for all the fitting procedures by defining the null hypothesis Ho as the two samples follow the same distribution and the alternate hypothesis Ha as the distribution follows a different distribution. Rejection of the null hypothesis indicates a poor fit to the original dataset, and failure to reject the null hypothesis indicates an acceptable fit to the distribution. The level of significance alpha (D*) was set to 0.05. The KS and AD test statistic (D) functions in XLSTAT were used to determine the critical value and test statistic. We then compared the test statistics to the critical value. If $D > D^*$, we reject the null hypothesis Ho. If $D \leq D^*$, we fail to reject the null hypothesis Ho. The results of the goodness-of-fit test are explained in Chapter 5.

4. DATA SERIES DESCRIPTION AND CHARACTERISTICS

4.1. Daily Car Values (DCV) for Secondary Rail Shipping

Daily Car Values (DCV) is a method used in the secondary rail shipping industry to calculate the value of railroad cars daily. The value of a car is calculated by taking the average daily lease rate for a specific type of railcar and adjusting it based on several factors, such as its age, condition, and type of cargo it is designed to carry. The DCV data for secondary rail shipping is a weekly market report sourced from Trade West Brokerage Company from 1/2/2004 through 9/1/2022 (figure 8). The missing values in the series were filled using linear interpolation. This series is interesting in that DCV market can be volatile at times because of changes in demand and supply, fluctuations in the overall economy, fuel prices, weather events among many others. These factors can cause sudden changes in demand for certain types of railcars or disruptions to supply chains, which can impact the value of railcars. Additionally, DCV price is quoted as a premium (positive) or discount (discount) to the standard railcar tariffs which makes it even more interesting. The series is important because it helps to identify potential risks and opportunities. Additionally, it is a valuable tool for understanding the value of railroad cars in the secondary rail shipping industry and is used by a variety of stakeholders to make informed decisions about their logistics operations. The time series graph is presented in Figure 7.

The series was tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis Ho, and accept the alternative

hypothesis, Ha. We conclude that there is no unit root, and the series is stationary at level. Hence, there is no need to difference the series for stationarity.

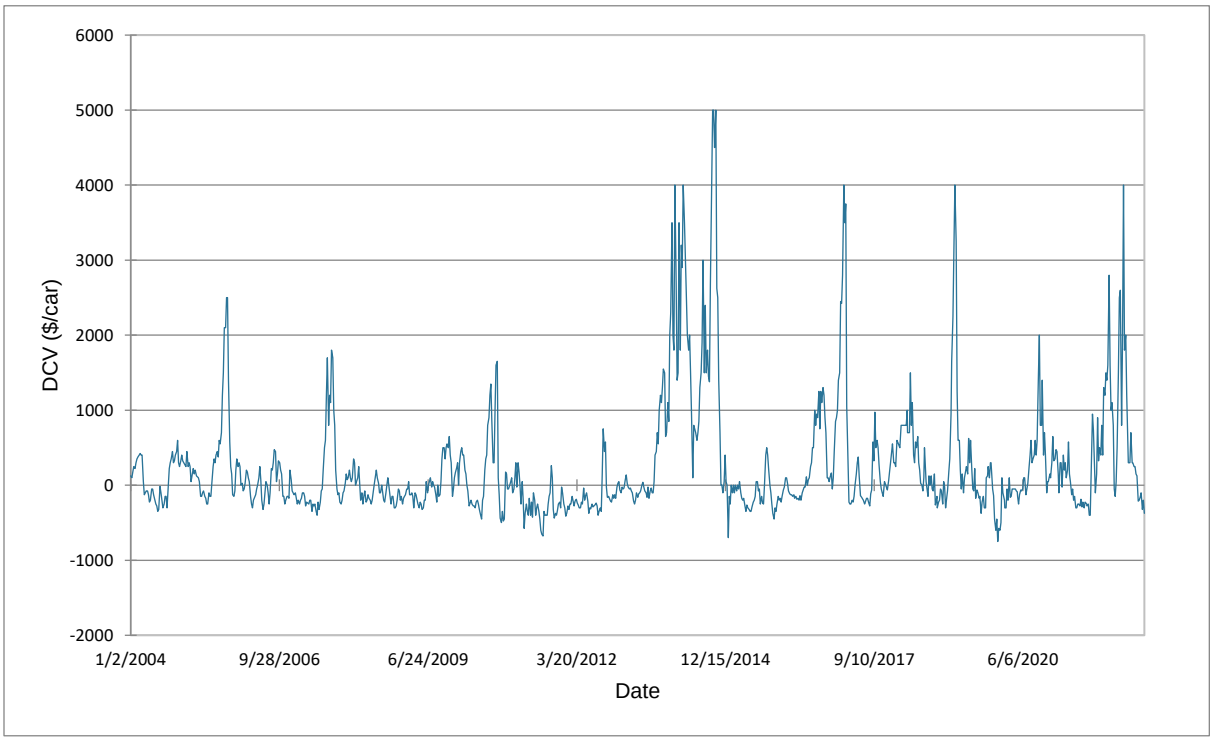

Figure 7. Time series (DCV ($/car))

DCV series is quoted in U.S. dollars per car ($/car). There were 964 DCV time series observations for the period. The minimum and maximum value are $750 and $5000 respectively with a mean value of $260.82 and standard deviation of $788.24. The 1st quartile value is $175 while the 3rd quartile value is $356.25. Pearson's skewness and kurtosis are 2.7976 and 9.5957 respectively. The descriptive statistic summary is presented in Table 1 below:

Table 1. DCV ($/car) Descriptive Statistic Summary

Statistic	Nbr. of obs	Minimum	Maximum	1st Quartile	3rd Quartile	Mean	Standard deviation (n-1)
DCV ($/car)	964	(750.00)	5,000.00	(175.00)	356.25	260.82	788.24

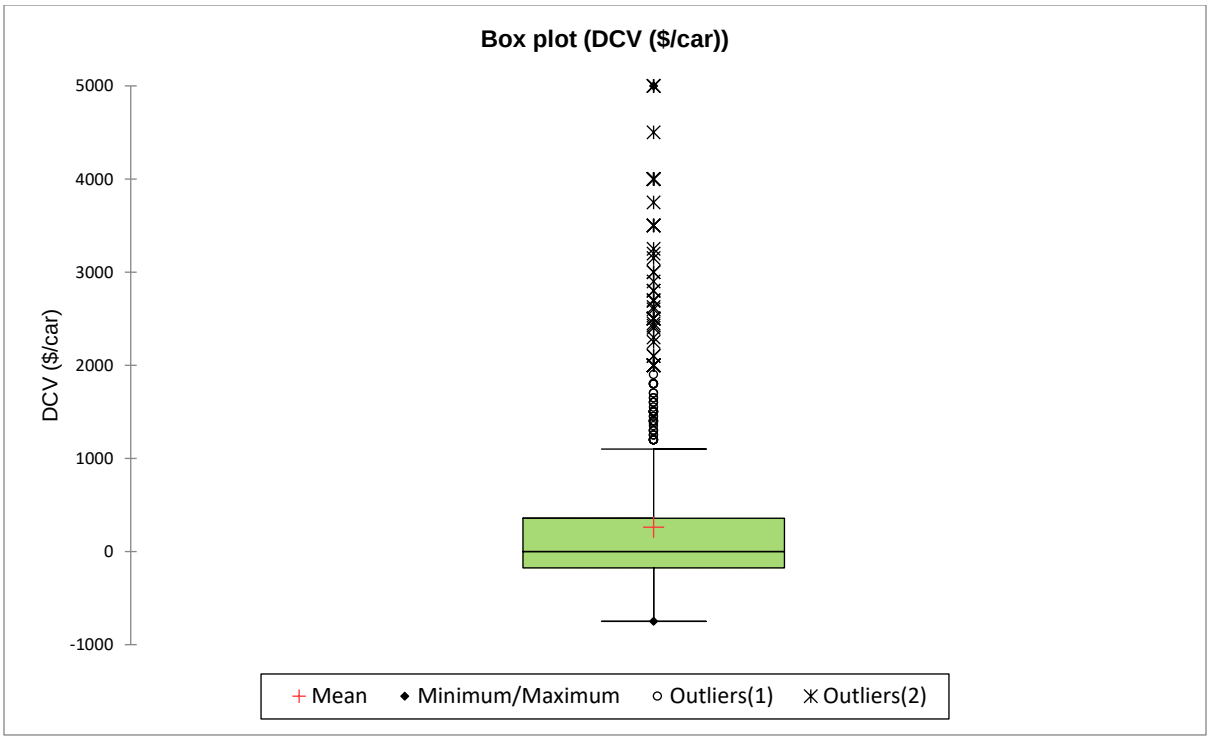

Figure 8. (DCV ($/car)) Boxplot

Figure 9 shows the DCV $/car Boxplot with outliers. The outliers presented in the boxplot are upwards. It identified one outlier representing one and half IQR and two outliers representing 3IQRs.

4.2. Electricity Wholesale Price Daily Changes

4.2.1. General Description

The electricity wholesale market in the United States is regulated by Federal Energy Regulatory Commission (FERC). The wholesale electricity market operates on a regional basis, with different regions having their own wholesale electricity markets. The management of the

transmission grid and the operation of the wholesale energy markets are the responsibilities of the regional transmission organizations (RTOs) and independent system operators (ISOs). There are currently seven RTOs/ISOs operating in the United States including California ISO (CAISO), Electric Reliability Council of Texas (ERCOT), Midcontinent Independent System Operator (MISO), New York Independent System Operator (NYISO), PJM Interconnection (PJM), Southwest Power Pool (SPP), and the ISO New England (ISO-NE). Figure 10 shows the wholesale electricity markets or the RTO map.

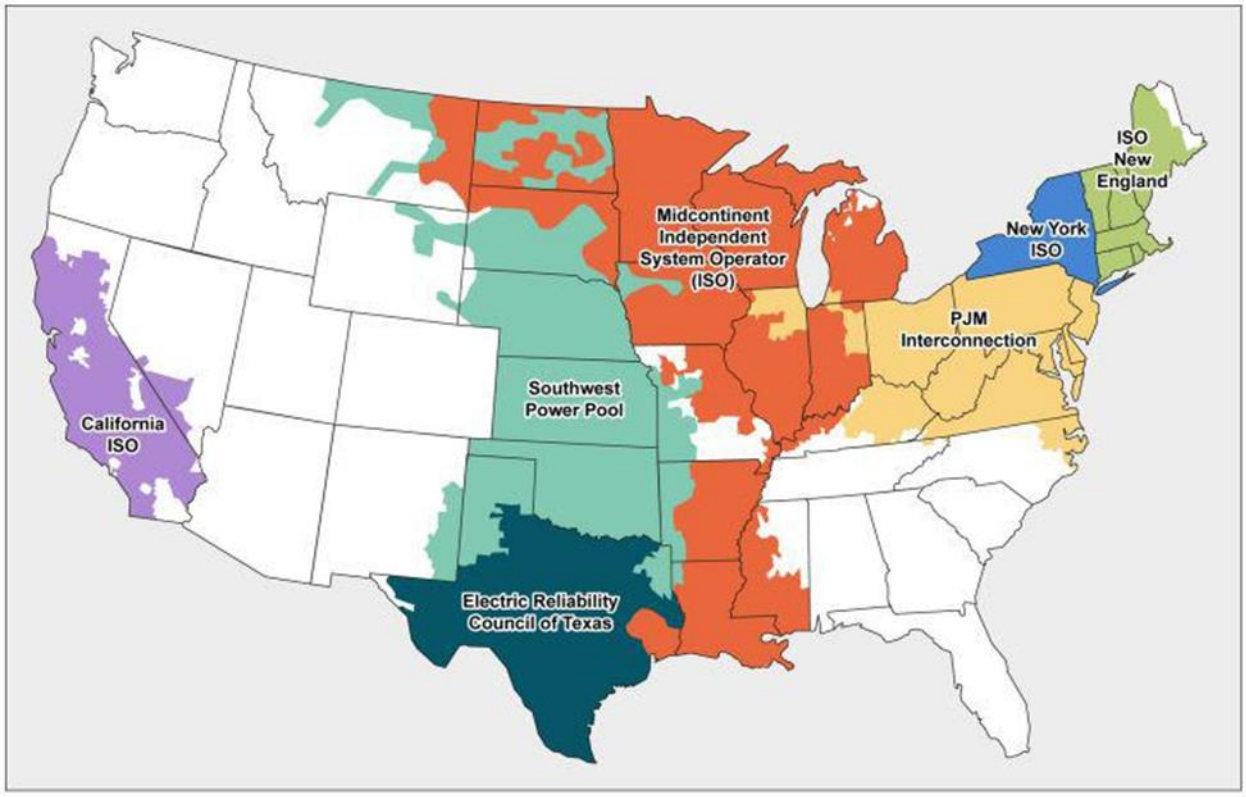

Figure 9. Regional Transmission Organizations/ Electricity Wholesale Market

Like how wholesale and retail markets operate for other products, electricity is bought, sold, and traded in wholesale and retail markets. The purchase and sale of electricity to resellers is done in the wholesale market, while the purchase and sale of electricity to consumers is done in the retail market. The wholesale electricity price is predetermined by a buyer and seller

through bilateral contracts or set of organized wholesale markets. The clearing price for electricity in the wholesale markets is determined by an auction in which generation offer in a price at which they can supply a specific number of megawatt-hours of power. A successful bid which is said to "clear" the market. The cheapest options "clear" the market first followed by the cheapest option until demand is met. The market is "cleared" when supply equals demand and the price of the last resource to offer becomes the wholesale price of power. The nature of wholesale market price makes it an ideal candidate for the study.

The wholesale price of power is calculated by dividing the sum of each transaction's price multiplied by its volume by the total number of all qualifying transactions. Mathematically, it is defined as

$$I = \sum(P * V) / T \tag{29}$$

where:

I = Volumetric Weighted Average Index Price

P = price or premium of individual transaction

V = volume of individual transaction

$\sum(P * V)$ = sum of each transaction's price multiplied by its volume

T = total volume of all qualifying transactions

We randomly selected three (3) RTOs including Midcontinent Independent System Operator (MISO), PJM Interconnection (PJM), and the ISO New England (ISO-NE) for the study. Wholesale electricity prices for these regions were sourced from Energy Information Administration: Wholesale Electricity and Natural Gas Market Data. The ICE Electricity product names for these regions is Indiana Hub RT Peak, PJM WH Real Time Peak, and NEPOOL MH

DA LMP Peak respectively. However, we used the variable names; Indiana Hub, PJM West and NEPOOL to represent the daily change in wholesale electricity prices.

NEPOOL and PJM West hub daily price series are from January 2001 to October 2022 while Indiana hub daily price series is from January 2006 to September 2022. The wholesale daily electricity prices are sourced from the U.S. Energy Information Administration (EIA)[1]. The market data provided by the EIA are republished from data collected by the Intercontinental Exchange (ICE) and updated biweekly. The ICE wholesale electricity markets include more than two dozen hubs and delivery points in North America.

One of the main reasons why this data series is interesting is that the wholesale electricity market is volatile since electricity cannot be easily stored. As a result, the supply of electricity must be constantly matched to the demand in real time. This means that even small changes in supply or demand can have a significant impact on the market price. In addition, the wholesale electricity market is heavily influenced by regulatory policies, which can be subject to change. Overall, the electricity market is subject to a range of factors that can cause volatility such as unexpected power plant outages, shifts in energy policy, fluctuations in fuel prices etc., making it important for energy traders, utilities, and policy makers to closely monitor market trends and anticipate potential changes. Wholesale electricity market is important because it provides insights into market trends, enables more efficient energy procurement, helps to mitigate risks, and identify potential such as changes in demand for electricity of shifts or shifts in electricity policy.

[1] ("U.S. Energy Information Administration - EIA - Independent Statistics and Analysis" 2023.)
Energy Information Agency, U.S. Department of Energy. Website: https://www.eia.gov/electricity/wholesale/

4.2.2. Indiana Hub

The Indiana Hub Peak Electricity (IPE) is a market index that tracks the price of electricity at the Indiana Hub during peak hours. The Indiana Hub is a major electricity trading hub located in the Midwestern region of the United States, serving a large portion of the Midwest and Mid-Atlantic regions. The IPE index is used as a benchmark for pricing electricity futures and options contracts, and it is also used by market participants for risk management and hedging purposes. The series was sourced from Energy Information Agency, U.S. Department of Energy. It is reported by Intercontinental Exchange as Indiana Hub RT Peak Price in $ per megawatt hour (MWh), volume weighted average for next day delivery. It is a daily price change from 1/5/2006 to 9/30/2022. The time series graph for Indiana Hub price change ($/MWh) is presented in Figure 11. The series is stable across the period with extreme prices during certain peak periods. For example, price increased to $107.60, and price was as low as $ (85) during the height of covid-19 in 2020.

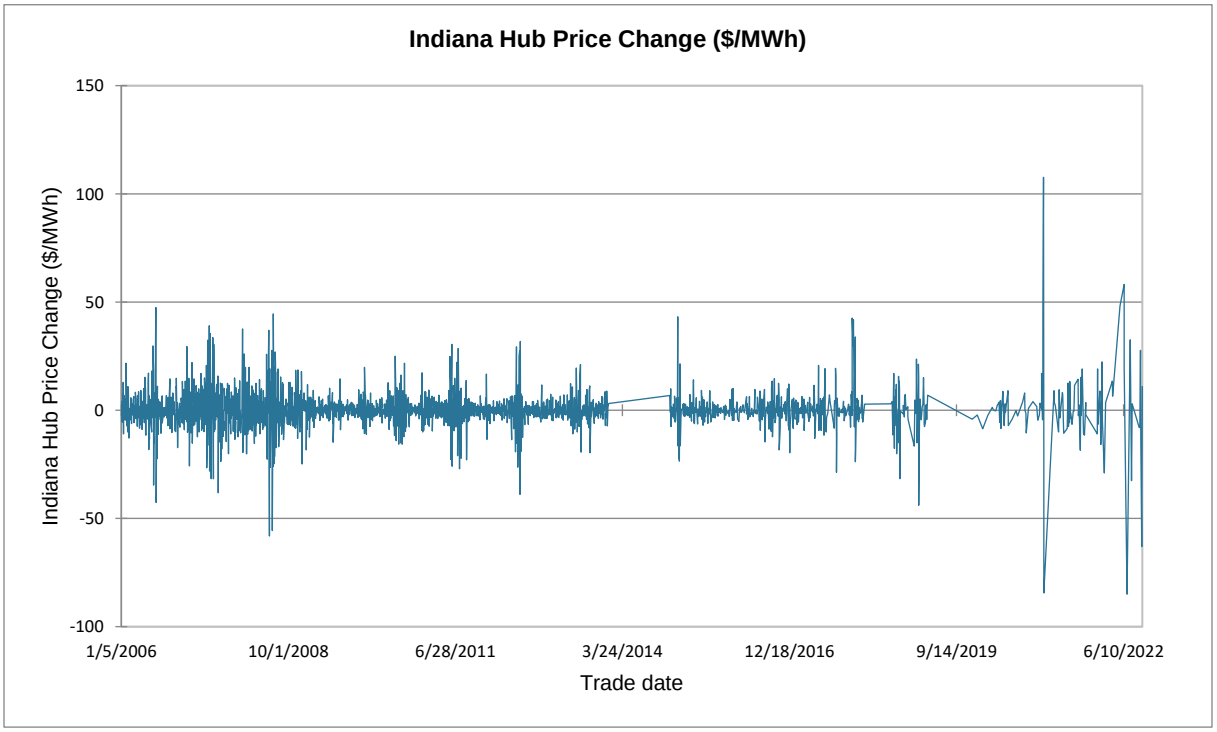

Figure 10. Time series (Indiana Hub Price Change ($/MWh))

The series was tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis Ho, and accept the alternative hypothesis, Ha. We conclude that there is no unit root, and Indian Hub price change ($/MWh) series is stationary at first differencing. There were 2692 Indian Hub price change ($/MWh) time series observations for the period. The minimum and maximum value are $(85) and $107.60 respectively with a mean value of $0.08 and standard deviation of $8.54. The descriptive statistic summary for Indian Hub price change ($/MWh) is presented in Table 2.

Table 2. Indiana Hub Price Change ($/MWh) Descriptive Statistic Summary

Statistic	**Nbr. of Obs**	**Min**	**Max**	**1st Quartile**	**Median**	**3rd Quartile**	**Mean**	**Standard deviation (n-1)**
Indiana Hub Price Change ($/MWh)	2692	(85.00)	107.60	(2.82)	(0.15)	2.75	0.08	8.54

A Boxplot with outliers for Indiana Hub Price Change ($/MWh) is illustrated in Figure 12. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Indiana Hub Price Change ($/MWh). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and half IQR and outliers (2) indicates 3IQRs.

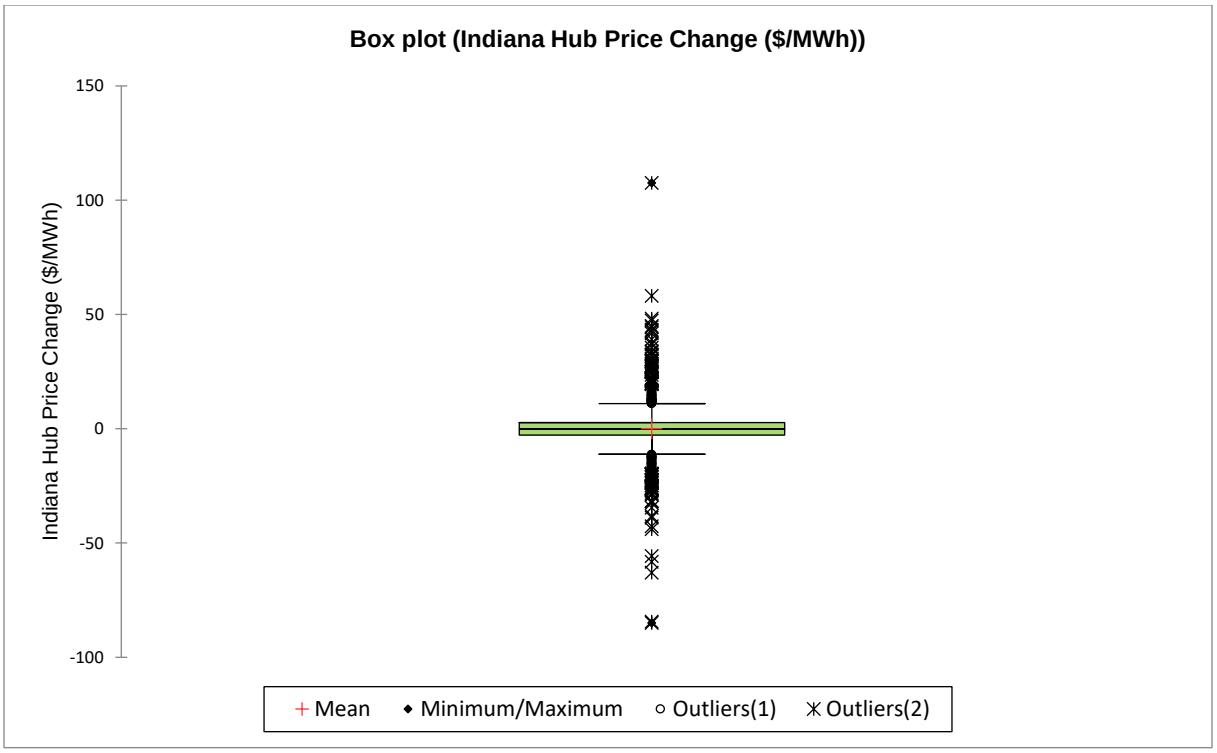

Figure 11. Box plot (Indiana Hub Price Change ($/MWh))

4.2.3. PJM West Hub

The PJM Interconnection is a regional transmission organization (RTO) that manages the transmission grid and operates the wholesale electricity market in 13 states and the District of Columbia in the Eastern United States. The PJM Interconnection is the largest power grid in North America, serving more than 65 million people and covering an area of over 243,000 square miles. PJM operates several wholesale electricity markets, including a day-ahead market, a real-time market, and a capacity market. The day-ahead market allows market participants to purchase electricity for delivery the following day, while the real-time market allows for the purchase and sale of electricity in real-time to address changes in supply and demand. The series was sourced from Energy Information Agency, U.S. Department of Energy. It is reported by PJM WH Real Time Peak Price in $ per megawatt hour (MWh), volume weighted average for next day delivery. It is a daily price change from 1/3/2001 to 10/4/2022. The time series graph

for PJM West Hub price change ($/MWh) is presented in Figure 13. The series is stable across the period with extreme prices during certain peak periods. For example, price increased to $194.10, and price was as low as $ (287.48) during the height of covid-19 in 2020.

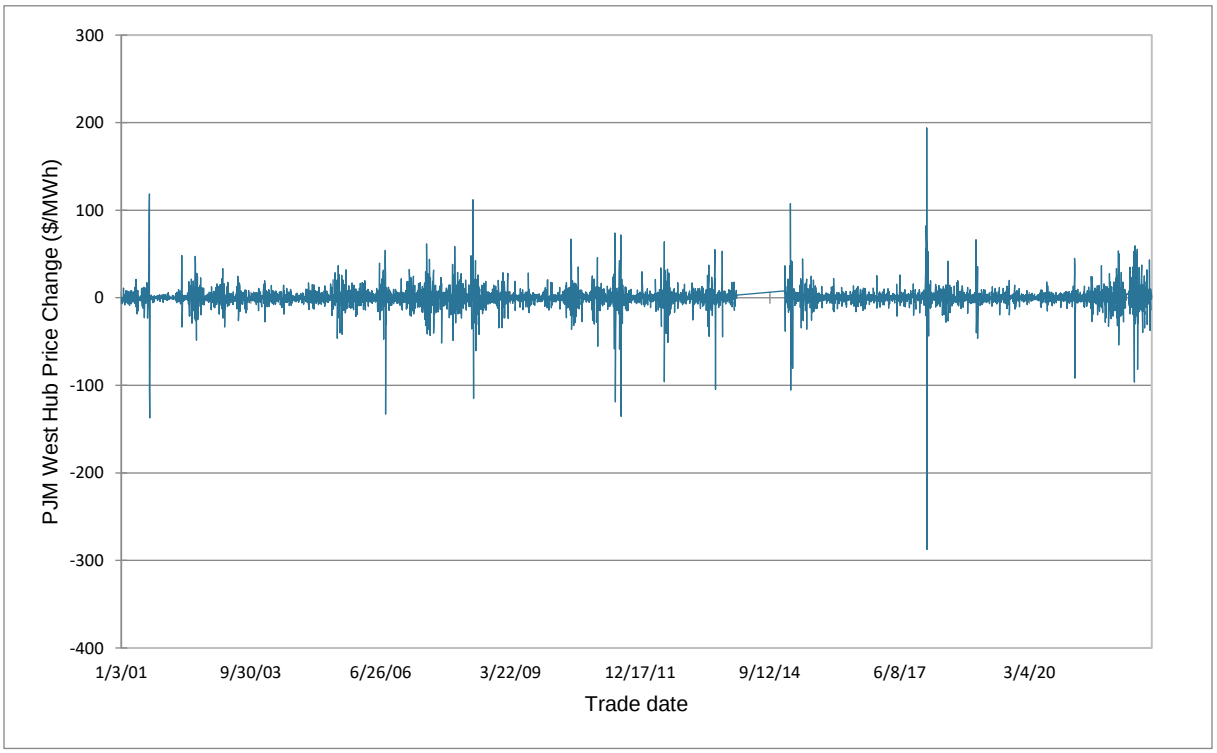

Figure 12. Time series (PJM West Hub Price Change ($/MWh))

The series was also tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis Ho, and accept the alternative hypothesis, Ha. We conclude that there is no unit root, and PJM West Hub price change ($/MWh) series is stationary at first differencing. There were 5249 PJM West Hub price change ($/MWh) time series observations for the period. The minimum and maximum value are $(287.48) and $194.10 respectively with a mean value of $0.11 and standard deviation of $12.76.

The descriptive statistic summary for PJM West Hub price change ($/MWh) is presented in Table 3.

Table 3. PJM West Hub Price Change ($/MWh) Descriptive Statistic Summary

Statistic	Nbr. of obs	Min	Max	1st Quartile	Median	3rd Quartile	Mean	Standard deviation (n-1)
PJM West Hub Price Change ($/MWh)	5249	(287.48)	194.1	(3.42)	(0.11)	3.56	0.03	12.76

A Boxplot with outliers for PJM West Hub Price Change ($/MW h) is illustrated in Figure 14. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Indiana Hub Price Change ($/MWh). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and half IQR and outliers (2) indicates 3IQRs.

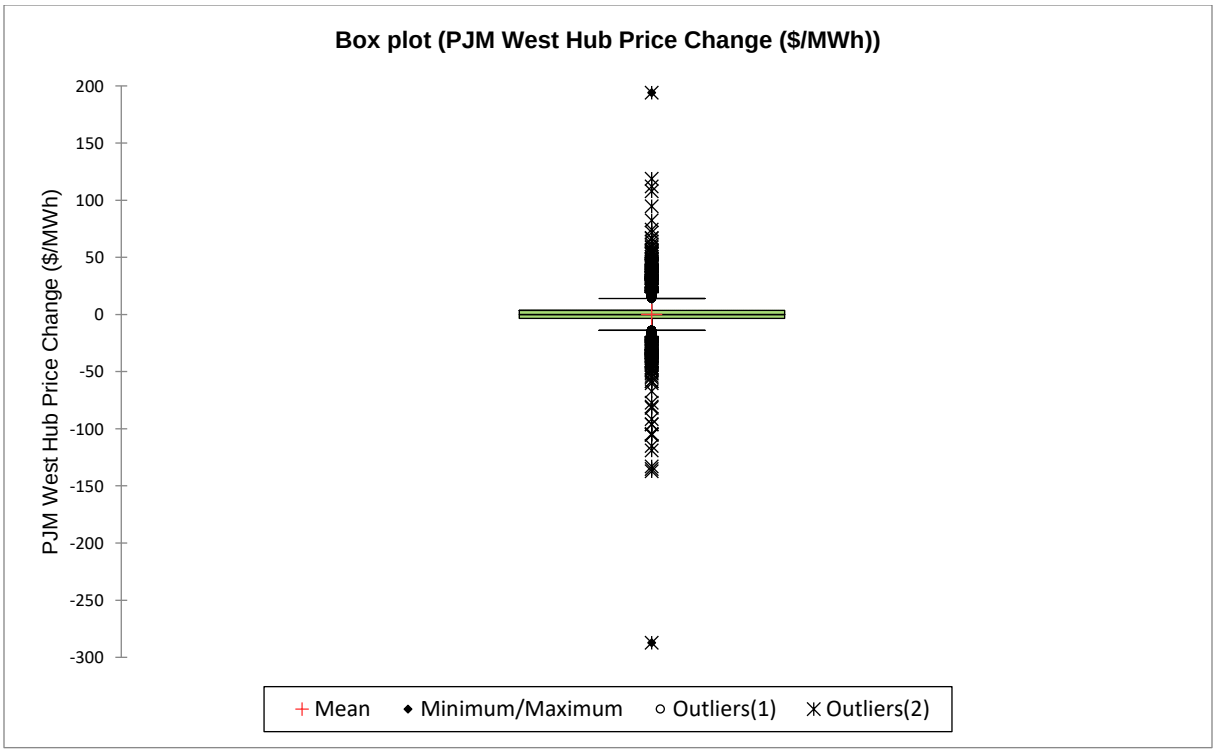

Figure 13. Box plot (PJM West Hub Price Change ($/MWh))

4.2.4. NEPOOL Hub

The NEPOOL is a regional transmission organization that manages the transmission grid and operates the wholesale electricity market in the New England region of the United States. The NEPOOL Peak Electricity is a market index that tracks the price of electricity at the NEPOOL peak load hours. Market players use the NEPOOL Peak Electricity index as a benchmark for pricing electricity futures and options contracts as well as for risk management and hedging. The series was sourced from the Energy Information Agency, U.S. Department of Energy. It is reported by NEPOOL MH DA LMP Peak Price in $ per megawatt-hour (MWh), volume weighted average for next-day delivery. It is a daily price change from 1/8/2001 to 10/4/2022. The time series graph for NEPOOL Hub price change ($/MWh) is presented in Figure 15. The series is stable across the period, with extreme prices during certain peak periods. For example, the price increased to $206.94, and the price was as low as $ (162.06) during the height of covid-19 in 2020.

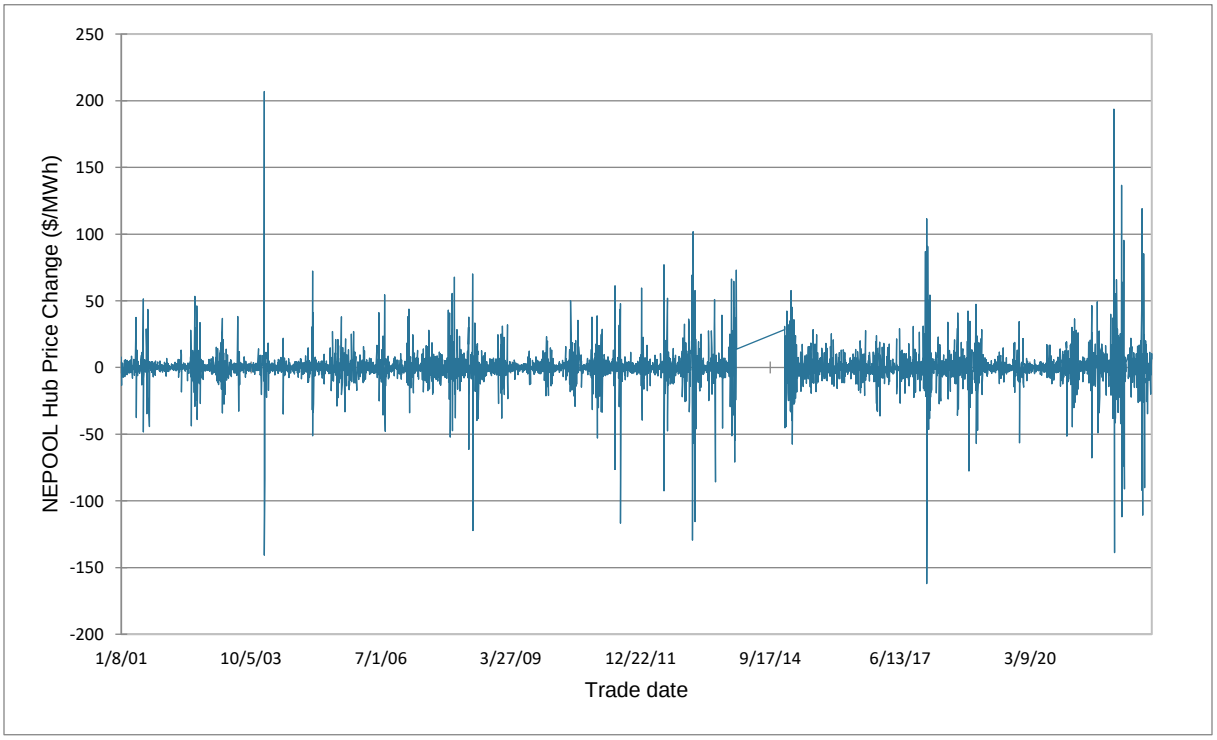

Figure 14. Time series (NEPOOL Hub Price Change ($/MWh))

The series was also tested for stationarity using the ADF stationarity test statistics. The null hypothesis, Ho, states that there is a unit root for the series, and the alternative hypothesis, Ha is that there is no unit root for the series at a 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis, Ho, and accept the alternative hypothesis, Ha. We conclude that there is no unit root, and NEPOOL Hub Price Change ($/MWh) series is stationary at first differencing. There were 4855 NEPOOL Hub Price Change ($/MWh) time series observations for the period. The minimum and maximum values are $(162.06) and $206.94, respectively, with a mean value of $0.03 and a standard deviation of $14.38. The descriptive statistic summary for NEPOOL Hub Price Change ($/MWh) is presented in Table 4.

Table 4. NEPOOL Hub Price Change ($/MWh) Descriptive Summary Statistics

Statistic	**Nbr of obs**	**Min**	**Max**	**1st Quartile**	**Median**	**3rd Quartile**	**Mean**	**Standard deviation (n-1)**
NEPOOL Hub Price Change ($/MWh)	4855	(162.06)	206.94	(3.28)	(0.19)	3.14	0.03	14.38

A Boxplot with outliers for NEPOOL Hub Price Change ($/MWh) is illustrated in Figure 16. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Indiana Hub Price Change ($/MWh). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and a half IQR, and outliers (2) indicate 3IQRs.

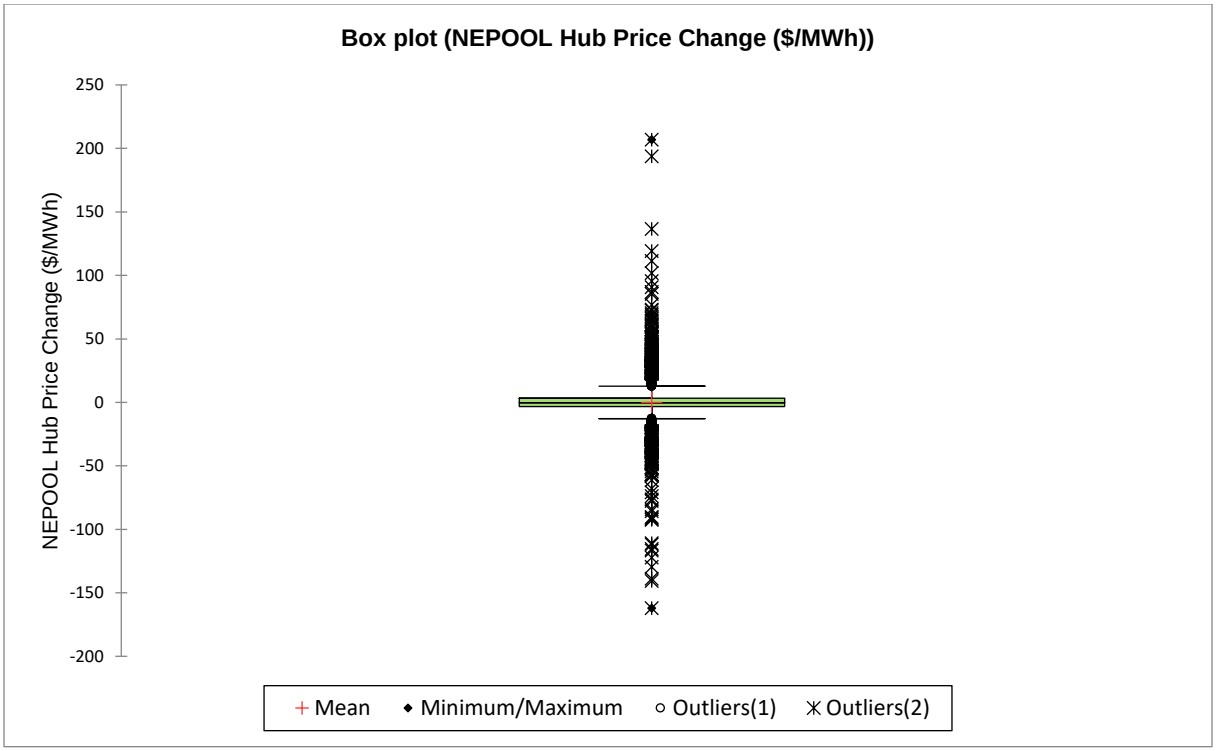

Figure 15. Box plot (NEPOOL Hub Price Change ($/MWh))

4.3. Futures Price and Spread Daily Changes

4.3.1. General Description

Agriculture futures price is a financial instrument used to speculate on the future price of agricultural commodities such as wheat, corn, soybeans, coffee, sugar, and livestock. These futures contracts enable farmers, producers, processors, and traders to manage the price risk associated with the production, transportation, and storage of agricultural products. Spread daily change is the difference between two different futures contracts for the same commodity with different delivery dates. The spread is the cost of carrying agriculture commodities over time and can be affected by a variety of factors such as storage costs, interest rates, and supply and demand dynamics. In futures trading, a spread is created by simultaneously buying and selling two different futures contracts, with the expectation that the price difference between the contracts will widen or narrow over time.

For futures price and spread daily changes, we sourced Nearby Chicago Oats Futures, KC-Chicago Nearby Wheat Futures Spread and Chicago Nearby Intermonth Wheat Spread time series data from DTN ProphetX. Nearby Chicago Oats Futures is a daily change in Chicago oats futures price quoted in cents per bushel. The code that was used to extract the Oats futures daily price is @O@C. Daily differencing was applied to the extracted price. Oats futures daily price is based on the price difference per bushel for delivery in the future. That is, the value of an oats futures is calculated by multiplying the price of oats per bushel by the number of bushels in the contract. The daily oats futures prices are daily observations from January 2010 to September 2022. KC-Chicago Nearby Wheat Futures Spread is a daily intermarket nearby price spread between Kansas City and Chicago wheat futures quoted in cents per bushel (cents/bu). The code that was used to extract Wheat Futures Spread is @W@C2 - @W@C1. KC-Chicago Nearby Wheat Futures Spread is reported daily from January 2010 to September 2022. The spread is calculated by subtracting the price of the KC wheat futures contract from the price of the Chicago wheat futures contract. A positive spread indicates that the price of the Chicago contract is higher than the price of the KC contract, while a negative spread indicates the opposite. Grain Futures and Transportation Market Prices and Spreads. Chicago Nearby Intermonth Wheat Spread is a daily Chicago wheat nearby intermonth futures spread priced in cents per bushel (cents/bu). It is reported daily from January 2010 to September 2022 and is quoted in U.S. cents per bushel (cents/bu). The code that was used to extract Chicago Nearby Intermonth Wheat Spread is @KW@C - @W@C. The spread is calculated by subtracting the price of the wheat futures contract for the nearby contract from the price of the wheat futures deferred contract. A positive spread indicates that the price of the farther-out delivery month is higher than the price of the nearby delivery month, while a negative spread indicates the opposite.

Futures Price and Spread Daily Changes are interesting because of the volatility in commodity trading. The price of futures can fluctuate daily based on a variety of factors, including demand and supply, geopolitical events, and market sentiment. The Chicago Nearby Intermonth Wheat Spread is an important futures trading strategy in the wheat market, providing a tool to stakeholders, farmers, traders, investors, and analyst for managing risk and potentially profiting from price differentials between related contracts with different delivery months. In addition, traders can adjust their trading strategies to take advantage of potential profit or manage risk exposure.

4.3.2. Nearby Chicago Oats Futures

The time series graph for Nearby Chicago Oats Futures is presented in Figure 17. The series is stable across the period with extreme values presenting potential outliers.

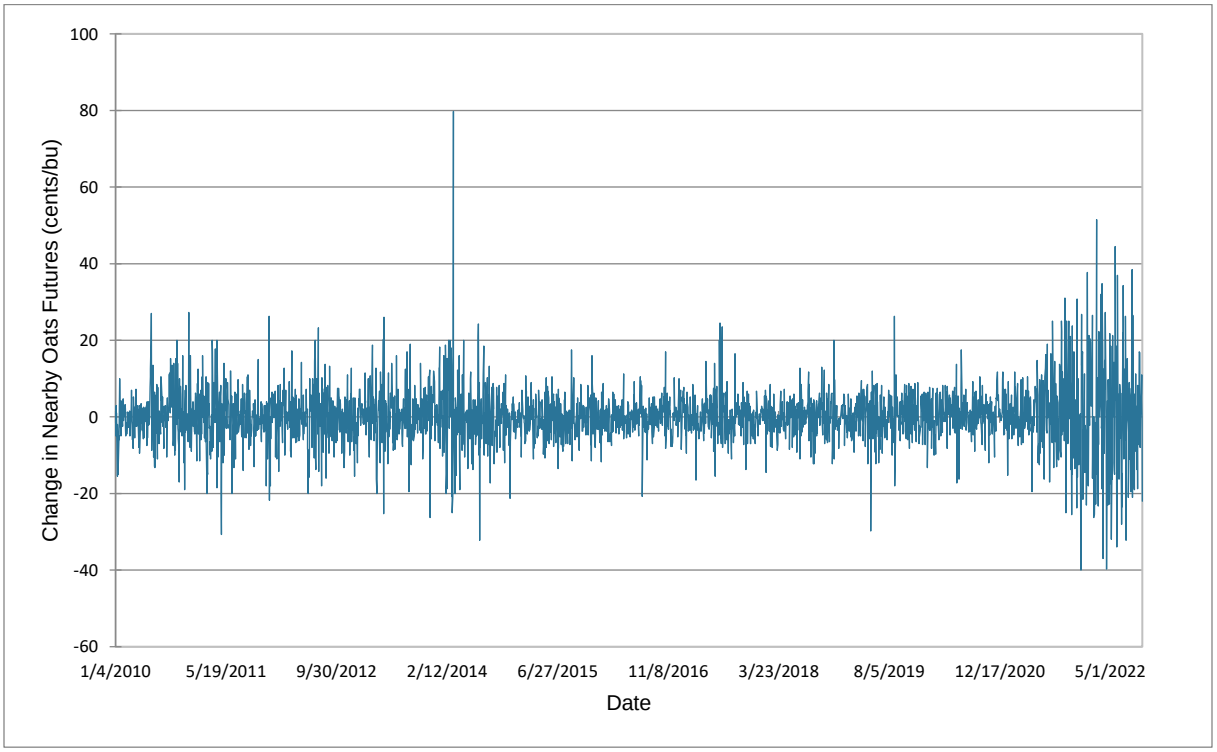

Figure 16. Time series (Change in Nearby Oats Futures (cents/bu)):

The series was also tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis Ho, and accept the alternative hypothesis, Ha. We conclude that there is no unit root, and Nearby Chicago Oats Futures price (cents/bu)) series is stationary at first differencing. There were 3143 Nearby Chicago Oats Futures time series observations from 1/4/2010 to 9/23/2022. The minimum and maximum value are (-40) and 79.75 respectively with a mean value of 0.25 and standard deviation of 7.53. The descriptive statistic summary for Nearby Chicago Oats Futures (cents/bu) is presented in Table 5.

Table 5. Nearby Chicago Oats Futures Descriptive Statistic Summary

Statistic	**Nbr. of obs**	**Min**	**Max**	**1st Quartile**	**Median**	**3rd Quartile**	**Mean**	**Standard deviation (n-1)**
Change in Nearby Oats Futures (cents/bu)	3143	(40.00)	79.75	(3.25)	-	3.75	0.25	7.53

A Boxplot with outliers for Nearby Chicago Oats Futures is illustrated in Figure 18. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Nearby Chicago Oats Futures (cents/bu). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and half IQR and outliers (2) indicates 3IQRs.

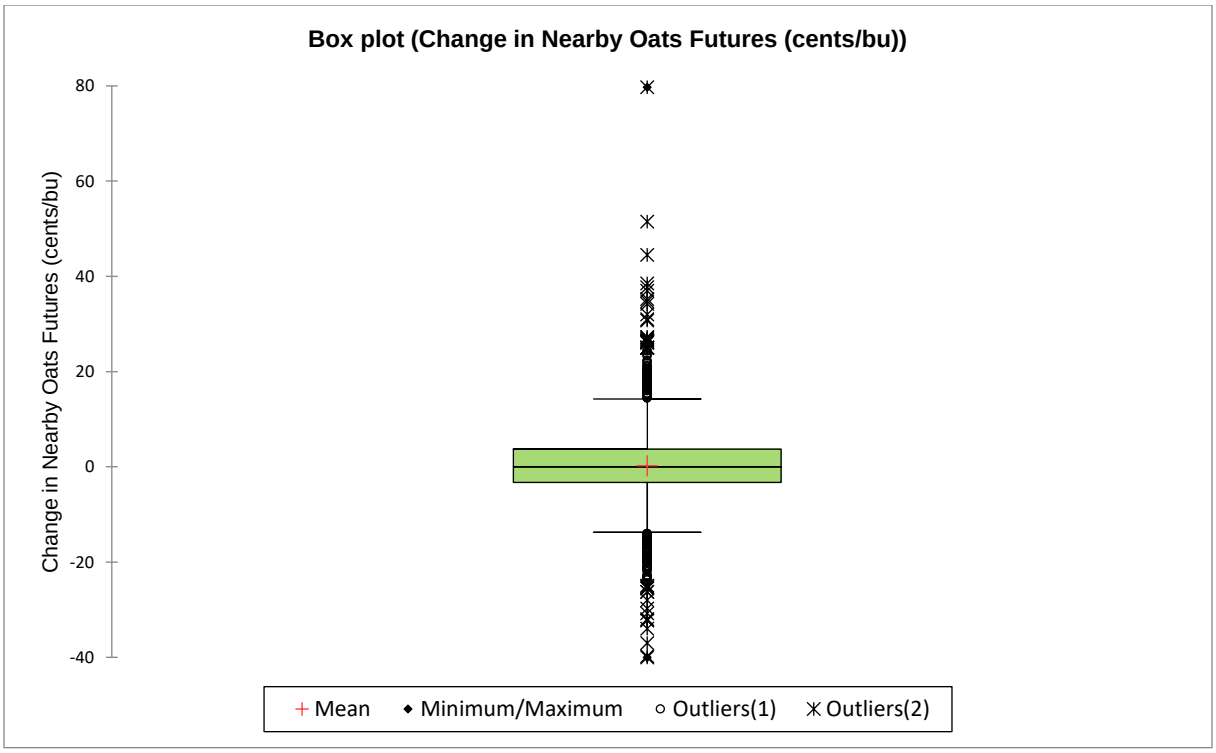

Figure 17. Box plot Change in Nearby Oats Futures (cents/bu)

4.3.3. Chicago Nearby Wheat Futures Spread

Figure 19 shows the time series graph for Chicago Nearby Wheat Futures Spread is presented in.

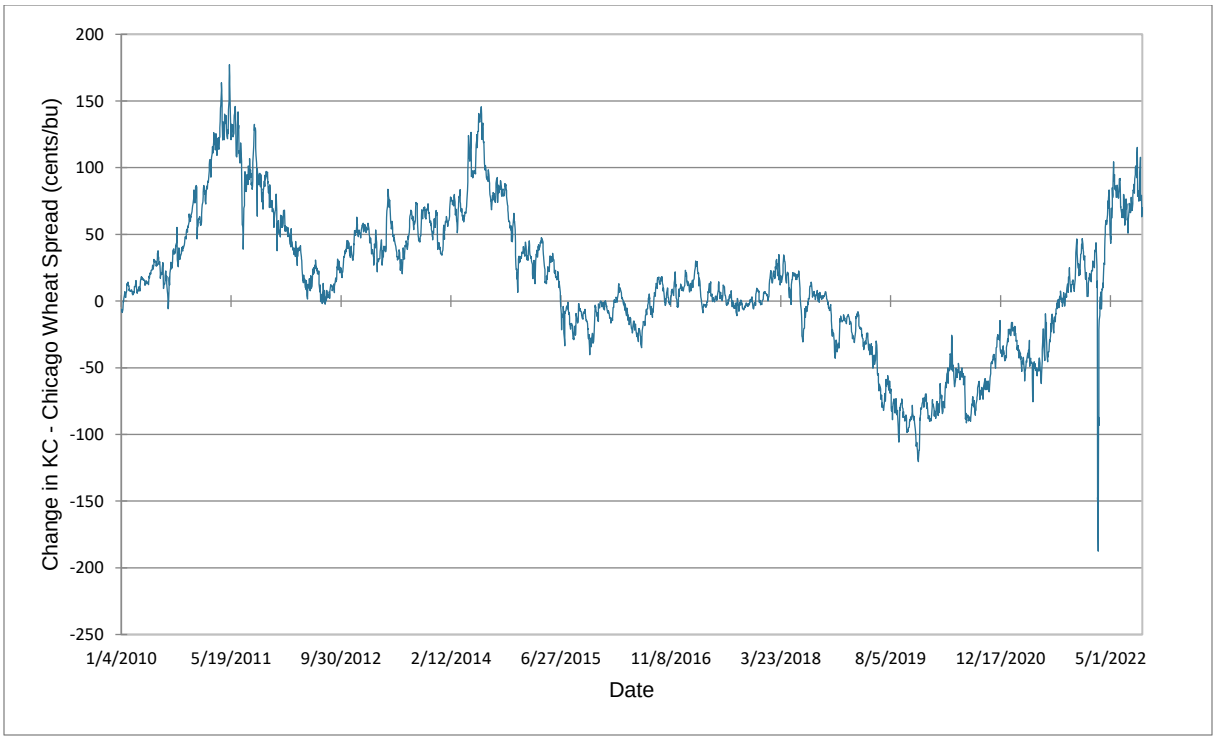

Figure 18. KC-Chicago Nearby Wheat Futures Spread

The series was also tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is greater than the significance level, we cannot reject the null hypothesis Ho. We cannot conclude that there is no unit root, and Chicago Nearby Wheat Futures Spread (cents/bu)) series is not stationary at first differencing. There were 3207 Chicago Nearby Wheat Futures Spread time series observations from 1/4/2010 to 9/23/22. The minimum and maximum value are (187.5) and 177.25 respectively with a mean value of 12.25 and standard deviation of 51.71. The descriptive statistic summary for Chicago Nearby Wheat Futures Spread (cents/bu) is presented in Table 6.

Table 6. Chicago Nearby Wheat Futures Spread Summary Descriptive Statistics

Statistic	Nbr. of obs	Min	Max	1st Quartile	Median	3rd Quartile	Mean	Standard deviation (n-1)
@W@C2 - @W@C1	3207	(187.50)	177.25	(13.50)	12.25	48.88	15.07	51.71

A Boxplot with outliers for Chicago Wheat Spread is illustrated in Figure 20. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Nearby Chicago Oats Futures (cents/bu). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and half IQR and outliers (2) indicates 3IQRs.

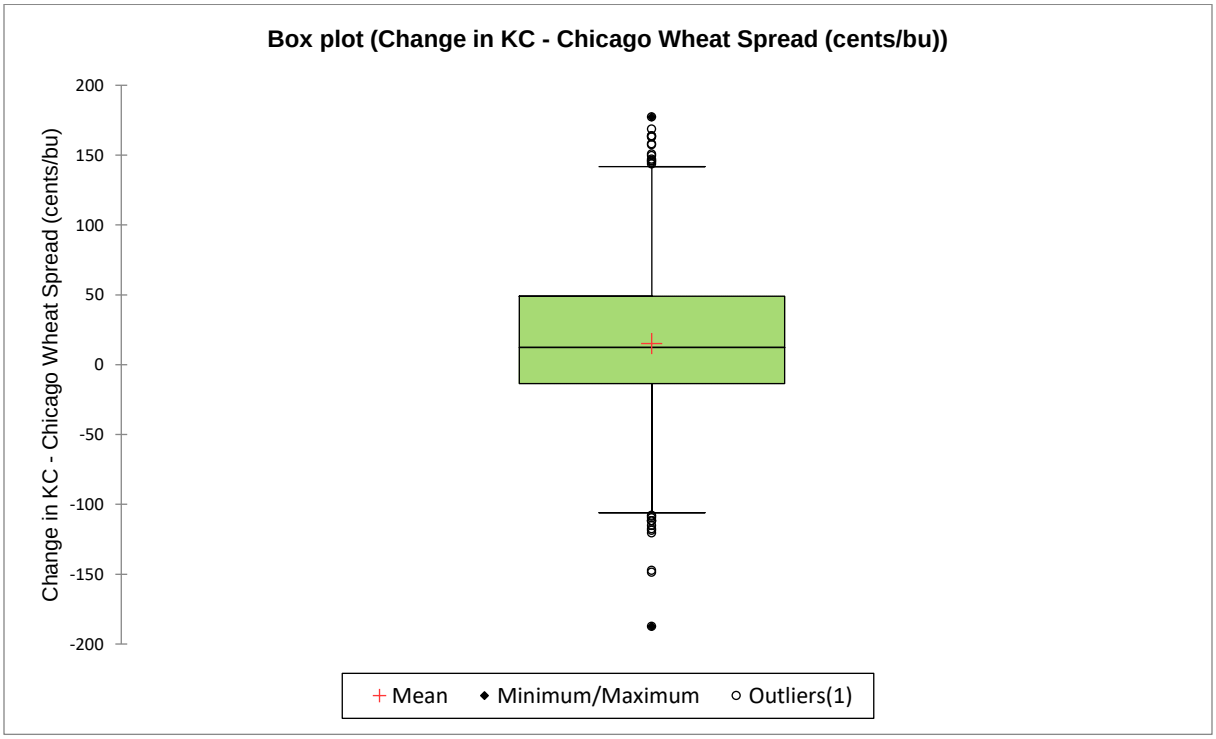

Figure 19. Box plot (Change in KC - Chicago Wheat Spread (cents/bu))

4.3.4. Chicago Nearby Intermonth Wheat Spread

The time series graph for Chicago Nearby Intermonth Wheat Spread in Figure 21. The series is stable across the period with extreme values presenting potential outliers.

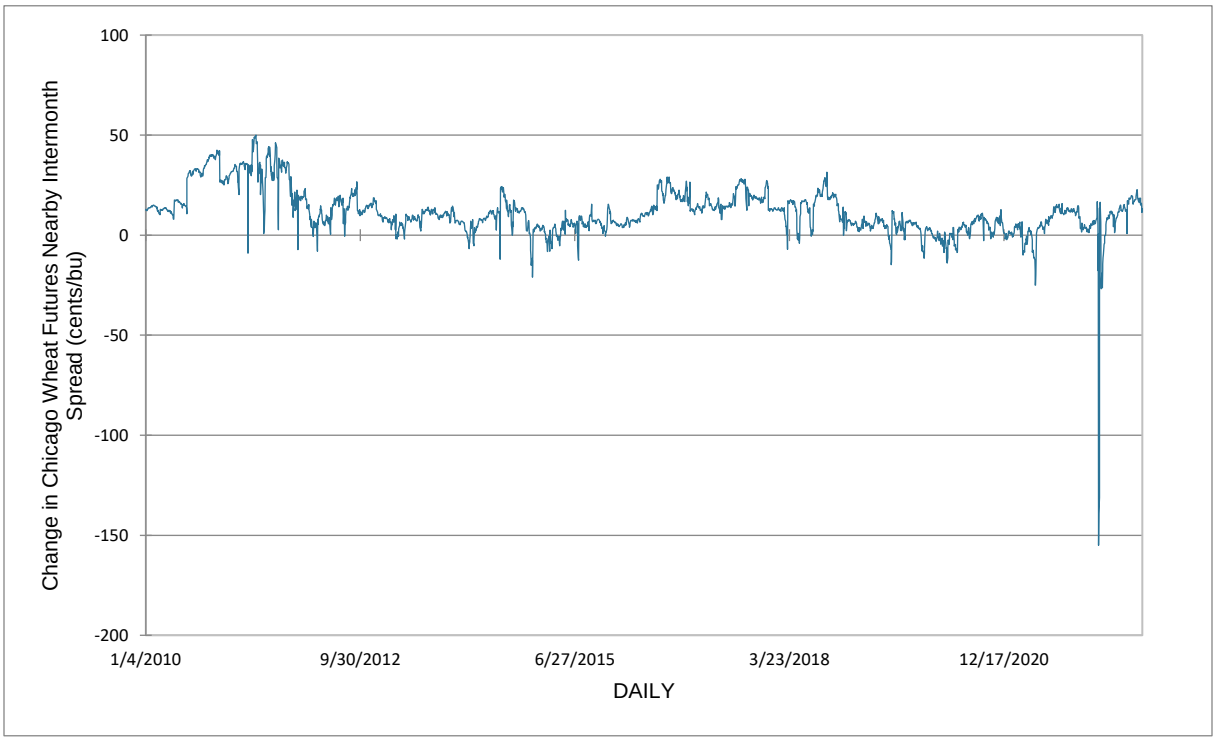

Figure 20. Chicago Nearby Intermonth Wheat Spread Time Series

The series was also tested for stationarity using the ADF stationarity test statistics. The null hypothesis Ho states that there is a unit root for the series and the alternative hypothesis Ha is that there is no unit root for the series at 5% significance level. As the computed p-value is lower than the significance level, we reject the null hypothesis H0, and accept the alternative hypothesis, Ha. We cannot conclude that there is no unit root, and Chicago Nearby Intermonth Wheat Spread (cents/bu)) series is not stationary at first differencing. There were 3207 Chicago Nearby Intermonth Wheat Spread series observations from 1/4/2010 to 9/23/22. The minimum and maximum values are (155) and 50 respectively with a mean value of 11.79 and standard

deviation of 11.54. The descriptive statistic summary for Chicago Nearby Intermonth Wheat Spread (cents/bu) is presented in Table 7.

Table 7. Chicago Nearby Intermonth Wheat Spread Descriptive Summary Statistics

Statistic	Nbr. of obs	Min	Max	1st Quartile	Median	3rd Quartile	Mean	Standard deviation (n-1)
@KW@C - @W@C	3207	(155.00)	50.00	5.25	10.75	16.75	11.79	11.54

A Boxplot with outliers for Chicago Wheat Futures Nearby Intermonth is illustrated in Figure 22. The Boxplot shows the mean, minimum/maximum, outliers, and number of IQR for Nearby Chicago Oats Futures (cents/bu). The outliers fall below the minimum and above the maximum. Outlier (1) indicates one and half IQR and outliers (2) indicates 3IQRs.

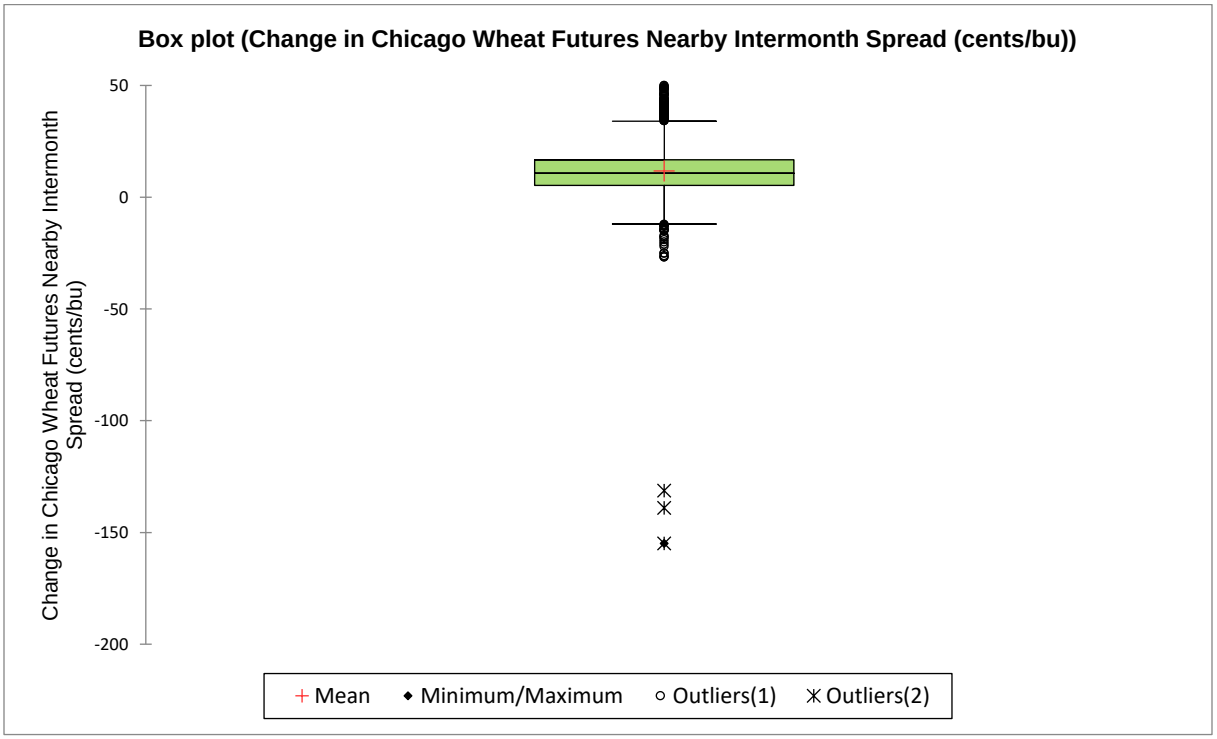

Figure 21. Box plot (Change in Chicago Wheat Futures Nearby Intermonth Spread (cents/bu))

5. EMPIRICAL RESULTS AND DISCUSSION

This chapter presents the empirical results and discussion of the result. The chapter begins with the discussion analysis and results of each of the alternative procedures. The chapter concludes with general observations.

5.1. Analysis and Results

5.1.1. Daily Car Values (DCV)

The first step of the research procedure was to confirm the presence of outliers in the dataset series. Grubb's test for DCV ($/car) showed that the G-Scores on maximum values are 6.012, 6.130, and 6.254 while the G-Scores on minimum values are 4.053 and 4.116 respectively. Again, Grubb's test identified the presence of 37 outliers in DCV series. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in DCV dataset.

There are two labeled subgroups for DCV ($/car), that is, base and upper labeled dataset. The result for Modified Z-Score showed that there are 878 and 89 number of observations in base and upper labeled subgroups, and a mean of 260.82 ($/MWh) and standard deviation of 788.24 ($/MWh) respectively. The minimum observed value in base subgroup is -750 ($/MWh) and the maximum observed value is 1200 ($/MWh). The mean and the standard deviation for base using the Modified Z-Score technique are 53.32 ($/MWh) and 354.30 ($/MWh). For upper subgroup, the minimum observed value is 1250 ($/MWh) and the maximum observed value is 5000 ($/MWh). The mean and the standard deviation for upper using the Modified Z-Score technique are 2288.99 ($/MWh) and 984.56 ($/MWh). The result for Tukey IQR labeling showed that there are 867 and 97 number of observations in base and upper labeled subgroups with a mean of 260.82 ($/MWh) and standard deviation of 788.24 ($/MWh) respectively. The

minimum observed value in base subgroup is -750 ($/MWh) and the maximum observed value is 1100. The mean and the standard deviation for base using the Tukey IQR technique are 43.96 and 11.49. For upper subgroup, the minimum observed value is 1200 and the maximum observed value is $5000. The mean and the standard deviation for upper using the Tukey IQR technique are 2199.18 and 100.48. For k-Means clustering, the shadow scores for cluster 1 and 2 are 0.8131 and 0.4570. Also, there are 861 and 103 number of observations in base and upside labeled subgroups, and a mean of 260.82 and standard deviation of 788.24 respectively. The minimum observed value in base subgroup is -750 and the maximum observed value is 1050. The mean and the standard deviation for base using the k-Means clustering technique are 36.60 and 11.17. For upper subgroup, the minimum observed value is 1100 and the maximum observed value is $5000. The mean and the standard deviation for upper using the k-Means clustering technique are 2160.90 and 108.67.

The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, Ha, is that the distributions of the two samples are different.

For DCV full dataset, the computed p-value is greater than the significance level at 5%, we cannot reject the null hypothesis H0 (refer to Table A1 and Figure A1) Hence, we conclude that the two samples follow the same distribution. However, we reject the null hypothesis H0 and accept the alternate hypothesis H0 since the computed p-value is lower than the significance

level at 5% (Refer to Table A1 and Figure A2). Therefore, we conclude that the distributions of the two samples are different. The KS statistical test results for Z-Score Bayesian Averaging showed that the computed p-value is lower than the significance level at 5%, hence we reject the null hypothesis H0, and accept the alternative hypothesis Ha. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A3). The KS statistical test results for Tukey IQR Bayesian Averaging showed that the computed p-value is lower than the significance level at 5%, hence we reject the null hypothesis H0, and accept the alternative hypothesis Ha. We conclude that the distributions of two samples are different (refer to Table A2 and Figure A4). Similarly, The KS statistical test results for k-Means Bayesian Averaging showed that the computed p-value is lower than the significance level at 5%, hence we the reject null hypothesis H0, and accept the alternative hypothesis Ha. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A5).

In the final step of the research procedure, we compared the Goodness-of-Fit to the actual dataset by using the Two-Sample KS and AD Tests. Both the KS and AD test statistic indicated the p-value is greater than the significance level 5%, hence, we accept the null hypothesis and reject the alternative Ha (refer to Table 8). Hence, we conclude that Full dataset model is a perfect fit to the actual data.

Table 8. Summary Results of DCV Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-Means**
Kolgomorov-Smirnov (KS)	Statistic	0.0488	0.3911	0.3994	0.3932	0.3932
	p-value (Two-tailed)	0.2020	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Anderson-Darling (AD)	Statistic	1.0806	135.1767	144.3072	139.0969	138.4067
	p-value (Two-tailed)	0.3176	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

5.1.2. Indiana Hub Electricity Prices

The first step of the research procedure was to confirm the presence of outliers in the dataset series. Grubb's test for Indiana Hub Electricity Prices showed that the G-Scores on maximum value is 12.597 while the G-Score on minimum value is 10.27. Again, Grubb's test identified the presence of 45 outliers in Indiana Hub Electricity Prices. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in Indiana Hub Electricity Prices ($/MWh).

The three labeling techniques employed in this study were used to divide the datasets. There are two labeled subgroups for Indiana Hub Electricity Prices ($/MWh) series was divided into the base and contaminants labeled upper and lower datasets. The result for Modified Z-Score showed that there are 2513, 97 and 82 number of observations in base, upper and lower labeled subgroups. The mean of the original dataset is 0.80 and standard deviation 8.54 respectively. The minimum observed daily price for Indiana Hub Electricity in the **base** dataset is -14.64 ($/MWh) and the maximum observed daily price is 14.28 ($/MWh). The mean and the standard deviation for **base** using the Modified Z-Score technique are -0.06 ($/MWh) and 4.74 ($/MWh). For **upper**, the minimum observed daily price is 14.4 ($/MWh) and the maximum observed value is 107.6 ($/MW/h). The mean and the standard deviation for **upper** using the Modified Z-Score technique are 24.72 and 12.54 ($/MWh) respectively. In the **lower** dataset, the minimum observed daily price is -85 and the maximum observed daily price is -14.72 ($/MWh). The mean and the standard deviation for **lower labeled dataset** are -24.83 and 13.45 ($/MWh). The result for the Tukey IQR procedure showed that there are 2425, 149 and 118 number of observations in **base, upper** and **lower** labeled datasets. The mean of the original Indiana Hub Electricity dataset is 0.80 and standard deviation 8.54 respectively. The minimum observed daily

price for **base** labeled dataset is -11.16 and the maximum observed value is 11 ($/MWh). The mean and the standard deviation for **base** dataset using the Tukey IQR technique are -0.14 and 4.18 ($/MWh).

For **upper** labeled dataset, the minimum observed daily price is 12.76 and the maximum observed daily price is 206.94 ($/MWh). The mean and the standard deviation for **upper** labeled data using the Tukey IQR technique are 28.70 and 21.99 ($/MWh). In the **lower** 12.91 ($/MWh). The mean and the standard deviation for **lower labeled** dataset are -29.91 and 22.76 ($/MWh). Using the **k-Means** clustering procedure, we identified 5 clusters. The shadow scores for these clusters 1, 2, 3, 4 and 5 are 0.457, 0.690, 0.469, 0.395 and 0.574. Also, there are 2211, 213 and 268 number of observations in **base, upper** and **lower** labeled datasets. The minimum observed daily price in **base** labeled dataset is -19.49 and the maximum observed daily price is 17.74 ($/MWh). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are -0.19 and 5.87. For **upper** labeled dataset, the minimum observed daily price is 17.9 and the maximum daily price is 206.94 ($/MWh). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 38.88 and 24.15 ($/MWh). In the **lower** labeled dataset, the minimum observed dataset is -162.06 and the maximum observed daily price is -19.67 ($/MWh). The mean and standard deviation are -39.14 and 25.34 ($/MWh) respectively.

We then combined each of the labeled dataset using the Bayesian Averaging Procedure. Following the Bayesian Averaging Procedure, we used the KS and AD test statistics to compare the Bestfit for each of the labeled dataset. The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging

with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H_0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, H_a, is that the distributions of the two samples are different.

For Indiana Hub Electricity full dataset, the computed p-value is greater than the significance level at 5%, we cannot reject the null hypothesis H_0 (refer to Table A1 and Figure A6). Hence, we conclude that the two samples follow the same distribution. However, we reject the null hypothesis H_0 and accept the alternate hypothesis H_a since the computed p-value is lower than the significance level at 5% for the truncated labeled dataset. (refer to Table A1 and Figure A7). Therefore, we conclude that the distributions of the two samples are different. The KS statistical test results for Z-Score Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail to reject the null hypothesis H_0, and reject the alternative hypothesis, H_a. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A8). The KS statistical test results for Tukey IQR Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail reject the null hypothesis H_0, and reject the alternative hypothesis, H_a, and conclude that the distributions of two samples follow the same distribution (refer to Table A2 and Figure A9). Similarly, The KS statistical test results for k-Means Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail to reject null hypothesis H_0, and reject the alternative hypothesis, H_a. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A10).

In the final step of the research procedure, we compared the Goodness-of-Fit of each procedure or model to the actual dataset by using the Two-Sample KS and AD Tests. Both the

KS and AD test statistics indicated that the p-value is greater than the significance level at 5% for k-Means clustering model. Hence, we accept the null hypothesis and reject the alternative hypothesis, H_a (refer to Table 9) and conclude that k-Means is a perfect fit to the actual data.

Table 9. Summary Results of Indiana Hub Electricity Prices Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-means**
Kolmogorov-Smirnov (KS)	Statistic	0.0331	0.0468	0.0178	0.0160	0.0115
	p-value (Two-tailed)	0.1055	**0.0055**	0.7855	0.8822	0.9941
Anderson-Darling (AD)	Statistic	4.7167	9.4931	0.4942	0.5202	0.2389
	p-value (Two-tailed)	**0.0039**	**0.0000**	0.7529	0.7264	0.9761

5.1.3. NEPOOL Hub Electricity Prices

The first step of the research procedure was to confirm the presence of outliers in the dataset series. Grubb's test for NEPOOL Hub Electricity prices showed that the G-Scores on maximum value is 11.74 while the G-Score on minimum value is 14.385. Again, Grubb's test identified the presence of 117 outliers in NEPOOL Hub Electricity Prices. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in NEPOOL Hub Electricity Prices ($/MWh).

The three labeling techniques employed in this study were used to divide the datasets. There are two labeled subgroups for NEPOOL Hub Electricity Prices ($/MWh) series was divided into the **base** and contaminants labeled **upper** and **lower** datasets. The mean of the original dataset is 0.034 and standard deviation standard is 14.383. The result for Modified Z-Score showed that there are 4316, 259 and 235 number of observations in **base**, **upper** and **lower** labeled subgroups. The minimum observed daily price for NEPOOL Hub Electricity in the **base** dataset is -16.87 ($/MWh) and the maximum observed daily price is 16.51 ($/MWh). The mean and the standard deviation for **base** using the Modified Z-Score technique are -0.09 ($/MWh)

and 5.47 ($/MWh). For upper, the minimum observed daily price is 16.55 ($/MWh) and the maximum observed value is 206.94 ($/MW/h). The mean and the standard deviation for **upper** using the Modified Z-Score technique are 34.15 and 23.63 ($/MWh) respectively. In the **lower** dataset, the minimum observed daily price is -162.06 and the maximum observed daily price is -16.92 ($/MWh). The mean and the standard deviation for **lower** labeled dataset are -35.13 and 24.23 ($/MWh). The result for the Tukey IQR procedure showed that there are 4183, 357 and 315 number of observations in **base**, **upper** and **lower** labeled datasets. The minimum observed daily price for **base** labeled dataset is -12.82 and the maximum observed value is 12.75 ($/MWh). The mean and the standard deviation for **base** dataset using the Tukey IQR technique are -0.16 and 4.72 ($/MWh). For **upper** labeled dataset, the minimum observed daily price is 12.76 and the maximum observed daily price is 206.94 ($/MWh). The mean and the standard deviation for **upper** labeled data using the Tukey IQR technique are 28.70 and 21.99 ($/MWh). In the **lower** dataset, the minimum observed daily price is -162.06 and the maximum observed daily price is -12.91 ($/MWh). The mean and the standard deviation for **lower** labeled dataset are -29.91 and 22.76 ($/MWh). Using the **k-Means** clustering procedure, we identified 5 clusters. The shadow scores for these clusters 1, 2, 3, 4 and 5 are 0.457, 0.690, 0.469, 0.395 and 0.574. Also, there are 4430, 235 and 190 number of observations in **base, upper** and **lower** labeled datasets. The minimum observed daily price in **base** labeled dataset is -19.49 and the maximum observed daily price is 17.74 ($/MWh). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are -0.19 and 5.87. For **upper** labeled dataset, the minimum observed daily price is 17.9 and the maximum daily price is 206.94 ($/MWh). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 35.88 and 24.15 ($/MWh). In the **lower** labeled dataset, the minimum observed

dataset is -162.06 and the maximum observed daily price is -19.67 ($/MWh). The mean and standard deviation are -39.14 and 25.34 ($/MWh) respectively.

We then combined each of the labeled dataset using the Bayesian Averaging Procedure. Following the Bayesian Averaging Procedure, we used the KS and AD test statistics to compare the Bestfit for each of the labeled dataset. The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H_0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, H_a, is that the distributions of the two samples are different.

For NEPOOL Hub Electricity full dataset, the computed p-value is greater than the significance level at 5%, we cannot reject the null hypothesis H_0 (refer to Table A1 and Figure A11). Hence, we conclude that the two samples follow the same distribution. However, we reject the null hypothesis H_0 and accept the alternate hypothesis H_a since the computed p-value is lower than the significance level at 5% for the truncated labeled dataset. (refer to Table A1 and Figure A12). Therefore, we conclude that the distributions of the two samples are different. The KS statistical test results for Z-Score Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail to reject the null hypothesis H0, and reject the alternative hypothesis, Ha. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A13). The KS statistical test results for Tukey IQR Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail reject the null hypothesis H_0, and reject the alternative hypothesis, H_a, and

conclude that the distributions of two samples follow the same distribution (refer to Table A2 and Figure A14). Similarly, The KS statistical test results for k-Means Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we fail to reject null hypothesis H_0, and reject the alternative hypothesis, H_a. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A15).

In the final step of the research procedure, we compared the Goodness-of-Fit of each procedure or model to the actual dataset by using the Two-Sample KS and AD Tests. Both the KS and AD test statistics indicated that the p-value is greater than the significance level at 5% for k-Means clustering model. Hence, we accept the null hypothesis and reject the alternative hypothesis, H_a (refer to Table 10) and conclude that Z-Score model is a perfect fit to the actual data.

Table 10. Summary Results of NEPOOL Hub Electricity Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-means**
Kolmogorov-Smirnov (KS)	Statistic	0.0152	0.0614	0.0109	0.0177	0.0142
	p-value (Two-tailed)	0.6255	<0.0001	0.9345	0.4314	0.7109
Anderson-Darling (AD)	Statistic	2.7422	32.2602	0.4281	0.8452	0.5776
	p-value (Two-tailed)	0.0371	<0.0001	0.8205	0.4496	0.6695

5.1.4. PJM West Hub Electricity Prices

Grubb's test for PJM West Hub Electricity prices showed that the G-Scores on maximum value is 16 while the G-Score on minimum value is 22.53 Again, Grubb's test identified the presence of 107 outliers in PJM West Hub Electricity Prices. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in NEPOOL Hub Electricity Prices ($/MWh).

The three labeling techniques employed in this study were used to divide the datasets into three labeled subgroups for PJM West Hub Electricity Prices ($/MWh) series was namely **base**

and contaminants labeled **upper** and **lower** datasets. The mean of the original dataset is 0.03 and standard deviation standard is 12.76. The result for Modified Z-Score showed that there are 4887, 191 and 171 number of observations in **base**, **upper** and **lower** labeled subgroups. The minimum observed daily price for PJM West Hub Electricity in the **base** dataset is -18.05 ($/MWh) and the maximum observed daily price is 17.94 ($/MWh). The mean and the standard deviation for **base** using the Modified Z-Score technique are 0.01 ($/MWh) and 5.93 ($/MWh). For **upper**, the minimum observed daily price is 18.01 ($/MWh) and the maximum observed value is 194.1 ($/MW/h). The mean and the standard deviation for **upper** using the Modified Z-Score technique are 33.44 and 20.47 ($/MWh) respectively. In the **lower** dataset, the minimum observed daily price is -287.48 and the maximum observed daily price is -18.3 ($/MWh). The mean and the standard deviation for **lower** labeled dataset are -36.58 and 30.76 ($/MWh). The result for the Tukey IQR procedure showed that there are 4701, 281 and 264 number of observations in base, upper and lower labeled datasets. The minimum observed daily price for **base** labeled dataset is -12.82 and the maximum observed value is 12.75 ($/MWh). The mean and the standard deviation for **base** dataset using the Tukey IQR technique are -0.13 and 5.18 ($/MWh). For **upper** labeled dataset, the minimum observed daily price is 14.08 and the maximum observed daily price is 194.1 ($/MWh). The mean and the standard deviation for **upper** labeled data using the Tukey IQR technique are 27.77 and 18.79 ($/MWh). In the **lower** dataset, the minimum observed daily price is -287.48 and the maximum observed daily price is -13.97 ($/MWh) while the mean and standard deviation are -29.19 and 26.70 ($/MWh).

Using the **k-Means** clustering procedure, we identified 5 clusters. The shadow scores for these clusters 1, 2, 3, 4 and 5 are 0.626, 0.452, 0.416, 0.454 and 0.524. Also, there are 4740, 390 and 119 number of observations in **base, upper** and **lower** labeled datasets. The minimum

observed daily price in **base** labeled dataset is -21.96 and the maximum observed daily price is 11.29 ($/MWh). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are -0.8 and 5.61. For **upper** labeled dataset, the minimum observed daily price is 11.42 and the maximum daily price is 194.1 ($/MWh). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 23.54 and 17.34 ($/MWh). In the **lower** labeled dataset, the minimum observed dataset is -287.48 and the maximum observed daily price is -22.38 ($/MWh). The mean and standard deviation are -43.90 and 34.42 ($/MWh) respectively.

We then combined each of the labeled dataset using the Bayesian Averaging Procedure. Following the Bayesian Averaging Procedure, we used the KS and AD test statistics to compare the Bestfit for each of the labeled dataset. The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H_0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, H_a, is that the distributions of the two samples are different.

For PJM Hub Electricity full dataset, the computed p-value is lower than the significance level at 5%, we reject the null hypothesis H_0 (refer to Table A1 and Figure A16). Hence, we conclude that the two samples do not follow the same distribution. Also, we reject the null hypothesis H_0 and accept the alternate hypothesis H_a since the computed p-value is lower than the significance level at 5% for the truncated labeled dataset. (refer to Table A1 and Figure A17). Therefore, we conclude that the distributions of the two samples are different. The KS - Z-Score

Bayesian Averaging results showed that the computed p-value is greater than the significance level at 5%. Hence we fail to reject the null hypothesis H_0 and reject the alternative hypothesis, Ha. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A18). The KS - Tukey IQR Bayesian Averaging results showed that the computed p-value is greater than the significance level at 5%, hence we accept the null hypothesis H_0, and reject the alternative hypothesis, H_a, and conclude that the distributions of two samples follow the same distribution (refer to Table A2 and Figure A19). Similarly, The KS - k-Means Bayesian Averaging showed that the computed p-value is less than the significance level at 5%, hence we reject the null hypothesis H_0, and accept the alternative hypothesis, H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A20).

In the final step of the research procedure, we compared the Goodness-of-Fit of each procedure or model to the actual dataset by using the Two-Sample KS and AD Tests. Both the KS and AD test statistics indicated that the p-value is greater than the significance level at 5% for k-Means clustering model. Hence, we accept the null hypothesis and reject the alternative hypothesis, H_a (refer to Table 11) and conclude that Z-Score model is a perfect fit to the actual data.

Table 11. Summary Results of PJM West Hub Electricity Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-means**
Kolmogorov-Smirnov (KS)	Statistic	0.0533	0.0455	0.0173	0.0164	0.0402
	p-value (Two-tailed)	**<0.0001**	**<0.0001**	0.4093	0.4816	**0.0004**
Anderson-Darling (AD)	Statistic	14.9585	19.7970	1.1995	0.9388	4.0837
	p-value (Two-tailed)	**<0.0001**	**<0.0001**	0.2677	0.3910	**0.0079**

5.1.5. Chicago Oats Futures Prices

Grubb's test for Chicago Oats Futures prices (Chng_O (cents/bu)) showed that the G-Scores on maximum value is 10.564 while the G-Scores on minimum value is 5.516 respectively. Again, Grubb's test identified the presence of 21 outliers in Chng_O. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in Chng_O.

Chng_O, was divided into base, upper and lower labeled dataset. The result for Modified Z-Score analysis showed that there are 3023, 66, and 54 number of observations in **base**, **upper** and **lower** labeled subgroups, and a mean of 0.25 and standard deviation of 7.53 (cents/bu) respectively. The minimum observed value in **base** subgroup is -18 (cents/bu) and the maximum observed value is 18 (cents/bu). The mean and the standard deviation for **base** using the Modified Z-Score technique are 0.12 (cents/bu) and 5.68 (cents/bu). For **upper**, the minimum value is 18.25 (cents/bu) and the maximum observed value is 79.75 (cents/bu). The mean and the standard deviation for **upper** using are 25.42 (cents/bu) and 9.51 (cents/bu). In **lower** subgroup, the minimum value is -40 while the maximum value is -18.5. The mean is -23.73 with a standard deviation of 5.46 (cents/bu). The results of Tukey IQR labeling analysis showed that there are 2940, 104, and 99 number of observations in **base**, **upper and lower** labeled subgroups respectively. The minimum observed value in **base** is -18 (cents/bu) and the maximum observed value is 18 (cents/bu). The mean and the standard deviation for base using the Tukey IQR technique are 0.12 and 5.68 (cents/bu). For **upper**, the minimum observed value is 18.25 and the maximum observed value is $79.75 (cents/bu). In **lower** subgroup, the minimum value is -40 while the maximum value is -18.5 The mean and the standard deviation for upper using the Tukey IQR technique are 25.42 and 9.51 (cents/bu). In **lower** subgroup, the minimum value is -

40 while the maximum value is -18.5. The mean is -23.73 with a standard deviation of 5.46 (cents/bu). The mean is -23.73 with a standard deviation of 5.46 (cents/bu). For k-Means clustering analysis, there were four clusters. The shadow scores for these clusters are 0.527, 0.553, 0.440 and 0.483. There are 2089, 518 and 536 number of observations in **base, upper** and **lower** labeled datasets. The minimum observed daily price in **base** labeled dataset is -4.75 and the maximum observed daily price is 5.5 (cents/bu). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are 0.20 and 2.71. For **upper** labeled dataset, the minimum observed daily price is 5.75 and the maximum daily price is 79.75 (cents/bu). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 11.15 and 7.02 (cents/bu). In the **lower** labeled dataset, the minimum observed dataset is -40 and the maximum observed daily price is -5 (cents/bu). The mean and standard deviation are -10.13 and 5.76 (cents/bu) respectively.

The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H_0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, H_a, is that the distributions of the two samples are different.

For Chng_O full dataset, the computed p-value is greater than the significance level at 5%, we cannot reject the null hypothesis H_0 (refer to Table A1 and Figure A21) Hence, we conclude that the two samples follow the same distribution. Also, we accept the null hypothesis H_0 and reject the alternate hypothesis H_0 since the computed p-value is greater than the

significance level at 5% (Refer to Table A1 and Figure A22) for Chng_O truncated dataset. Therefore, we conclude that the two samples follow the same distribution. The Z-Score Bayesian Averaging results showed that the computed p-value is greater than the significance level at 5%, hence we accept the null hypothesis H_0, and reject the alternative hypothesis H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A23). KS Tukey IQR Bayesian Averaging results showed that the computed p-value is lower than the significance level at 5%, hence we reject the null hypothesis H_0, and accept the alternative hypothesis H_a. We conclude that the distributions of two samples are different (refer to Table A2 and Figure A24). Similarly, The KS k-Means Bayesian Averaging showed that the computed p-value is lower than the significance level at 5%, hence we the reject null hypothesis H_0, and accept the alternative hypothesis H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A25).

In the final step of the research procedure, we compared the Goodness-of-Fit to the actual dataset by using the Two-Sample KS and AD Tests. Both the KS and AD test statistic indicated the p-value is greater than the significance level 5%, hence, we accept the null hypothesis and reject the alternative H_a (refer to Table 12). Hence, we conclude that alternate procedure, Modified Z-Score model is a perfect fit to the actual data.

Table 12. Results of Oats Futures Price Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-Means**
Kolmogorov-Smirnov (KS)	Statistic	0.0286	0.0283	0.0175	0.0655	0.1995
	p-value (Two-tailed)	0.1519	0.1608	0.7217	<0.0001	<0.0001
Anderson-Darling (AD)	Statistic	1.1884	4.1352	0.3837	20.9727	207.7736
	p-value (Two-tailed)	0.2720	0.0075	0.8649	<0.0001	<0.0001

5.1.6. KC - Chicago Wheat Intermarket Spread

Grubb's test for Chicago Wheat Intermarket Spread (KW-W (cents/bu)) showed that p-value is greater than the significance level alpha=0.05 and hence we cannot reject the null hypothesis. Thus, we conclude that Grubb's test did not confirm the presence of outliers in KW-W dataset.

However, the results of the Modified Z-Score analysis showed that there are outliers in the series. There are 3204 price observations in **base**, while there is 1 observed outlier in the **upper** and 2 observed outliers in the **lower** labeled subgroups respectively. The mean for KW-W series is 15.07 and standard deviation of 51.71 (cents/bu) respectively. The minimum observed value in **base** subgroup is -147.25 (cents/bu) and the maximum observed value is 168.5 (cents/bu). The mean and the standard deviation for **base** using the Modified Z-Score technique are 15.13 (cents/bu) and 51.54 (cents/bu). In **lower** subgroup, the minimum value is -40 while the maximum value is -18.5. The mean is -23.73 with a standard deviation of 5.46 (cents/bu). The results of the Tukey IQR labeling analysis showed that there are 3180, 14, and 13 number of observations in **base**, **upper and lower** labeled subgroups respectively. The minimum observed value in **base** is -106 (cents/bu) and the maximum observed value is 141.25 (cents/bu). The mean and the standard deviation for **base** using the Tukey IQR technique are 15.08 and 50.28 (cents/bu). For **upper**, the minimum observed value is 143.5 and the maximum observed value is 177.25 (cents/bu). In **lower** subgroup, the minimum value is -187.5 while the maximum value is -108. The mean and the standard deviation for upper using the Tukey IQR technique are -124.40 and 23.24 (cents/bu). For k-Means clustering analysis, there were three clusters. The shadow scores for these clusters are 0.599, 0.565, and 00.565. There are 1889, 1318 and 657 number of observations in **base, upper** and **lower** labeled datasets. The minimum observed daily price in

base labeled dataset is -187.5 and the maximum observed daily price is 45.5 (cents/bu). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are 8.32 and 21.03 (cents/bu). For **upper** labeled dataset, the minimum observed daily price is 22.5 and the maximum daily price is 28.58 (cents/bu). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 73.98 and 28.58 (cents/bu). In the **lower** labeled dataset, the minimum observed dataset is -187.5 and the maximum observed daily price is -108 (cents/bu). The mean and standard deviation are -124.40 and 23.24 (cents/bu) respectively.

For KW-W full dataset, the computed p-value is lower than the significance level at 5%, and reject the null hypothesis H_0 (refer to Table A1 and Figure A26) Hence, we conclude that the two samples do not follow the same distribution. Also, we reject the null hypothesis H_0 and accept the alternate hypothesis H_0 since the computed p-value is greater than the significance level at 5% (Refer to Table A1 and Figure A27) for KW-W truncated dataset. Therefore, we conclude that the two samples do not follow the same distribution. The Z-Score Bayesian Averaging results showed that the computed p-value is greater than the significance level at 5%, hence we accept the null hypothesis H_0, and reject the alternative hypothesis H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A28). KS Tukey IQR Bayesian Averaging results showed that the computed p-value is lower than the significance level at 5%, hence we reject the null hypothesis H_0, and accept the alternative hypothesis H_a. We conclude that the distributions of two samples are different (refer to Table A2 and Figure A29). Similarly, The KS k-Means Bayesian Averaging showed that the computed p-value is lower than the significance level at 5%, hence we the reject null hypothesis H_0, and

accept the alternative hypothesis H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A30).

In the final step of the research procedure, we compared the Goodness-of-Fit to the actual dataset by using the Two-Sample KS and AD Tests. The KS test statistic indicated the alternate procedure, Tukey IQR as the perfect fit to the actual data while the AD test statistic indicated the traditional approach of ignoring outliers to be the best fit model to the actual dataset (refer to Table 13).

Table 13. Results of KC - Chicago Wheat Intermarket Spread Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-Means**
Kolmogorov-Smirnov (KS)	Statistic	0.0362	0.0355	0.0362	0.0340	0.0493
	p-value (Two-tailed)	**0.0301**	**0.0348**	**0.0301**	**0.0492**	**0.0008**
Anderson-Darling (AD)	Statistic	3.4524	3.4827	3.4576	3.6372	7.7783
	p-value (Two-tailed)	**0.0162**	**0.0157**	**0.0161**	**0.0131**	**0.0001**

5.1.7. Chicago Wheat Intermonth Spread

Grubb's test for Chicago Oats Futures prices (W (cents/bu)) showed that the G-Scores on maximum value is -131.25 while the G-Scores on minimum value is -155 respectively. Again, Grubb's test identified the presence of 3 outliers in W. Thus, we conclude that the G-Scores on maximum and minimum values confirmed the presence of outliers in W.

W, was divided into base, upper and lower labeled dataset. The result for Modified Z-Score analysis showed that there are 3162, 64, and 11 number of observations in **base**, **upper** and **lower** labeled subgroups, The actual W dataset had a mean of 0.25 and standard deviation of 7.53 (cents/bu). The minimum observed value in **base** subgroup is -19 (cents/bu) and the maximum observed value is 40.5 (cents/bu). The mean and the standard deviation for **base** using the Modified Z-Score technique are 11.67 (cents/bu) and 9.87 (cents/bu) respectively. For **upper**,

the minimum value is 40.75 (cents/bu) and the maximum observed value is 50 (cents/bu). The mean and the standard deviation for **upper** are 44.87 (cents/bu) and .85 (cents/bu). In **lower** subgroup, the minimum value is -155 while the maximum value is -20. The mean is -56.18 with a standard deviation of 55.27 (cents/bu). The results of Tukey IQR labeling analysis showed that there are 3027, 155, and 25 number of observations in **base**, **upper and lower** labeled subgroups respectively. The minimum observed value in **base** is -12 (cents/bu) and the maximum observed value is 34 (cents/bu). The mean and the standard deviation for base using the Tukey IQR technique are 10.78 and 8.48 (cents/bu). For **upper**, the minimum observed value is 34.25 and the maximum observed value is $50 (cents/bu). The mean and the standard deviation for **upper** using the Tukey IQR technique are 38.67 and 3.93 (cents/bu). In **lower** subgroup, the minimum value is -155 while the maximum value is -12.25. The mean is -33.06 while the standard deviation 41.39 (cents/bu). For k-Means clustering analysis, there are five clusters. The shadow scores for these clusters are 0.576, 0.470, 0.604, 0.286, and 0.530. There are 1434, 430 and 1343 number of observations in **base, upper** and **lower** labeled datasets. The minimum observed daily price in **base** labeled dataset is 8.5 and the maximum observed daily price is 23 (cents/bu). The mean and the standard deviation for **base** labeled dataset using the k-Means clustering technique are 14.14 and 3.61. For **upper** labeled dataset, the minimum observed daily price is 23.25 and the maximum daily price is 50 (cents/bu). The mean and the standard deviation for **upper** labeled dataset using the k-Means clustering technique are 32.20 and 5.93 (cents/bu). In the **lower** labeled dataset, the minimum observed dataset is -155 and the maximum observed daily price is 8.25 (cents/bu). The mean and standard deviation are 2.74 and 8.44 (cents/bu) respectively.

The distribution fitting results here show the output analysis for the full dataset (ignore outliers), truncated dataset (remove outliers), Bayesian Averaging with modified Z-Score, Bayesian Averaging with Tukey IQR, Bayesian Averaging with k-Means Clustering, and comparison of goodness-of-fit with original data. The distribution fitting results followed the same hypothesis. The null hypothesis, H_0, of the two-sample Kolmogorov test is that the two samples follow the same distribution. The alternative hypothesis, H_a, is that the distributions of the two samples are different.

For W full dataset, the computed p-value is lower than the significance level at 5%, we reject the null hypothesis H_0 (refer to Table A1 and Figure A31) Hence, we conclude that the two samples do not follow the same distribution. Also, we reject the null hypothesis H_0 and accept the alternate hypothesis H_a since the computed p-value is lower than the significance level at 5% (Refer to Table A1 and Figure A32) for W truncated dataset. Therefore, we conclude that the two samples do not follow the same distribution. The Z-Score Bayesian Averaging results showed that the computed p-value is lower than the significance level at 5%, hence we reject the null hypothesis H_0, and accept the alternative hypothesis H_a. We conclude that the two samples do not follow the same distribution (refer to Table A2 and Figure A33). KS Tukey IQR Bayesian Averaging results showed that the computed p-value is greater than the significance level at 5%, hence we accept the null hypothesis H_0, and reject the alternative hypothesis H_a. We conclude that the distributions of two samples follow the same distribution (refer to Table A2 and Figure A34). Similarly, The KS k-Means Bayesian Averaging showed that the computed p-value is greater than the significance level at 5%, hence we the accept null hypothesis H_0, and reject the alternative hypothesis H_a. We conclude that the two samples follow the same distribution (refer to Table A2 and Figure A35).

In the final step of the research procedure, we compared the Goodness-of-Fit to the actual dataset by using the Two-Sample KS and AD Tests. Both the KS and AD test statistics indicated the k-means model as the perfect fit to the actual data (refer to Table 14).

Table 14. Results of KC - Chicago Wheat Intermonth Spread Goodness-of-Fit Test

Statistical Test		**Ignore Outliers**	**Drop Outliers**	**Z-Score**	**Tukey IQR**	**k-Means**
Kolmogorov-Smirnov (KS)	Statistic	0.2831	0.0477	0.0418	0.0324	0.0187
	p-value (Two-tailed)	**<0.0001**	**0.0014**	**0.0074**	0.0686	0.6285
Anderson-Darling (AD)	Statistic	301.3086	4.2461	2.6923	1.5325	0.5314
	p-value (Two-tailed)	**<0.0001**	**0.0066**	**0.0393**	0.1689	0.7152

5.2. General Observations

The alternative procedures proved to be better fit to the historical data compared to traditional method. Among the alternative procedures, the k-Means model was a better fit when applied to Indiana Hub Electricity Prices, NEPOOL Hub Electricity Prices, and KC Intermonth Spread. This was followed by the Tukey IQR model in the case of PJM West Hub and KC Intermarket Spreads. However, the KS test statistic results show that Tukey IQR model provided a better fit to KC Intermarket Spread while the AD test statistic results showed that ignoring the outlier proved to be a better fit to KC Intermarket Spread. Interestingly, ignoring the outliers in DCV data series proved to be a better fit compared to the alternative models. This could be due to the fact the outliers were only upward bound.

6. SUMMARY AND CONCLUSIONS

6.1. Summary of Problem

Outliers are problematic for many statistical analyses especially when using distribution fitting procedures in @risk Monte Carlo Simulation software because they can cause tests to miss significant findings or distort real results. This usually occurs with the traditional approach of ignoring or removing outlier observations which overlooks important information in an observation. Thus, the actual stochastic process is undermined. This study seeks to develop and evaluate a more effective approach for fitting statistical distributions to real-world data in the presence of outlier observations and compared to traditional approach using the Kolmogorov-Smirnov and Anderson-Darling test statistics.

6.2. Summary of Methodology

In this study, we used the Grubbs test to detect the outliers in all seven (7) candidate data series; DCV, wholesale electricity prices, oats futures prices and wheat spreads. After detecting outlier, we used nonparametric labeling procedures that is, Z-Score, Tukey IQR and k-Means clustering to divide the datasets into base and contaminating subsets. The contaminating subsets were labeled upper and lower outlier datasets. The Bestfit procedure in @Risk was then used to determine the best-fitting statistical distributions for each subset. The best-fitting procedures for each of the subsets were then combined for simulation by using the Bayesian Averaging technique to generate stochastic probabilities for each subset based on their observed frequency in the actual datasets. We used beta distributions in the case of two distributions to generate random probabilities and multivariate beta in the case of three distributions. Finally, we compared the actual sample data to the Monte Carlo Simulated data for each alternate approach by estimating the Kolmogorov-Smirnov and Anderson-Darling test statistics. The lower values

of the KS and AD test statistics indicated a stronger goodness-of-fit between the actual and candidate data distributions. We also compared and evaluated the overall utility of the procedure by applying the Bestfit to the total original dataset. Finally, we compared the traditional approaches to the labeling approaches for each of the datasets to determine the best the bestfit.

6.3. Study Results

The Grubbs test was estimated to identify outliers in the datasets. The results for the estimated Grubb's test confirmed the presence of outliers in all the datasets used in the study. The labeling procedures; modified Z-Score, Tukey IQR, and K-means further confirmed outliers in the datasets. We labeled the contaminants as lower and upper outliers. In the case of DCV, the modified Z-Score, Tukey IQR and k-Means labeled outliers as upper outliers only. In the case of wholesale electricity prices, oats futures price, and wheat spreads, the modified Z-Score, Tukey IQR and k-Means categorized outliers into upper and lower outliers.

After combining the labeled dataset using the Bayesian Averaging Procedure, the KS test statistic was estimated to find the best fit for each subset. For DCV series, the KS test statistic results showed the two samples followed the same distribution for full dataset. The two samples did not follow the same distribution for truncated data series. The KS test results showed that the samples did not follow the same distribution for Z-Score Bayesian Averaging, Tukey IQR Bayesian Averaging and k-Means Bayesian Averaging. Ignoring the outliers was identified as the better fit to the actual dataset compared to the alternative procedures.

For Indiana Hub Electricity Price series, the KS test statistic results showed the two samples followed the same distribution for full datasets series, truncated data series, Z-Score Bayesian Averaging, Tukey IQR Bayesian Averaging and k-Means Bayesian Averaging. Both the KS and AD goodness-of-fit test selected k-Means procedure as the best. This was followed

by Tukey IQR and Z-Score. This indicated that the alternative procedures outperformed traditional procedures in the case of Indiana Hub Electricity prices.

For NEPOOL Hub Electricity Price series, the KS test statistic results showed the two samples followed the same distribution for full datasets series, Z-Score Bayesian Averaging, Tukey IQR Bayesian Averaging and k-Means Bayesian Averaging except truncated data series Both the KS and AD goodness-of-fit test selected k-Means procedure as the bestfit. This was followed by Z-Score and Tukey IQR. This indicated that the alternative procedures outperformed traditional procedures in the case of NEPOOL Hub Electricity prices.

For PJM West Hub Electricity Price series, the KS test statistic results showed the two samples did not follow the same distribution for full datasets series, truncated data series, and k-Means Bayesian Averaging. On the hand, the two samples followed the same distribution for Z-Score and Tukey IQR Bayesian Averaging. Both the KS and AD goodness-of-fit test selected Tukey IQR procedure as the bestfit. This was followed by Z-Score and k-Means. This indicated that the alternative procedures outperformed traditional procedures in the case of PJM West Hub Electricity prices.

For Chng_0, the KS test statistic results showed the two samples followed the same distribution for full datasets series, truncated data series, and Z-Score Bayesian Averaging. On the hand, the two samples followed the same distribution for Tukey IQR and k-Means Bayesian Averaging. Both the KS and AD goodness-of-fit test selected Z-Score procedure as the bestfit. This indicated that the alternative procedures outperformed traditional procedures in the case of Chicago Oats Futures Prices.

For KW-W, the KS test statistic results showed the two samples followed the same distribution for truncated data series, and Z-Score Bayesian Averaging. On the hand, the two

samples followed the same distribution for full data series, Tukey IQR Bayesian Averaging and k-Means Bayesian Averaging. The KS goodness-of-fit test selected Tukey IQR procedure as the bestfit. Comparably, the AD goodness-of-fit test indicated that ignoring outliers in Chicago Wheat Intermarket Spread was a better fit.

For W, the KS test statistic results showed the two samples followed the same distribution for full data series, truncated data series, and Z-Score Bayesian Averaging. On the hand, the two samples followed the same distribution for Tukey IQR Bayesian Averaging and k-Means Bayesian Averaging. Both the KS and AD goodness-of-fit tests selected the k-Means procedure as the bestfit. This indicated that the alternative procedures outperformed traditional procedures in the case of Chicago Intermonth Spread.

6.4. Major Observations

Results from the sample datasets indicate, for the most part, that our procedures have improved fits than traditional approaches of handling outliers with two exceptions. Firstly, the labeling procedures tend to fit better when the distributions are two-tails rather than one-tail. Secondly, labeling procedures provide significant goodness-of-fit results compared to traditional approaches. Lastly, no one improved procedure was dominant in all the samples.

One of the major observations in this study was that the labeling procedures worked best when they were two-tailed outliers, that is, upper and lower outliers. Most distributions are adequate for fitting upper tails or where there is a wide range of upper skewed distributions. There is only one low-tailed distribution, that is, the extreme minimum. Thus, by flipping the lower tail and fitting a wide range of upper tails, we provide a better fit. The benefit of using the labeling approaches is particularly in the case where we have upper and lower tails. For example, the DCV was one tail and was more amendable to just the traditional approaches. While the AD

test indicated that the traditional approach (ignoring outliers) provided a better fit for KW-W, the labeling approach was also significant. For example, the Z-score identified only one (lower) and two (upper) outliers for KW-W, respectively, which were not enough outliers to fit well. So, we used real sample distribution to fit them. Tukey IQR and k-Means, on the other hand, identified adequate outliers. While Tukey IQR identified 13 (lower) and 14 (upper) outliers, k-Means identified 657 (downside) and 1318 (upper) outliers. However, Tukey IQR produced better results and better fit compared to k-Means.

We can therefore observe that the two tests sometimes do not agree. While the Anderson-Darling test fits the whole CDF, that is, gives more weight to the tails of the distribution by considering the squared difference between the observed EDF and expected CDF, the KS test looks at the maximum absolute difference between the EDF sample and CDF of the hypothesized distribution. Additionally, it is important to note that both tests require the selection of a significant level to determine the threshold for rejecting the null hypothesis. Finally, the performance and suitability of each test depend on the unique characteristics of the data and the research question. Thus, these differences between AD and KS tests may sometimes cause discrepancies in the goodness-of-fit test.

Additionally, we observed that the labeling procedures provided better fitting results for the stationary series compared to the traditional approaches. For example, since the KW-W series was non-stationary from a difference one different perspective, the labeling procedures were ineffective. As evidenced by the one non-stationary series, the fitting procedure did not work. The fitting procedures tend to work best with stationary series, as evidenced by all the series except the KW-W series. Hence, the labeling procedures provide improved fitting compared to the traditional approaches.

Finally, none of the improved procedures was dominant across all the sample datasets. For example, K-Means provided better fitting results for the samples, NEPOOL Hub, Indiana Hub, and W. Tukey IQR procedure was a better fit for PJM Hub, while Z-Score procedure was a better fit for Chng_O. This shows that having multiple labeling procedures will provide better results than only one.

6.5. Contributions to Existing Literature

Outlier detection methods are important for many applications. Traditional methods of ignoring or discarding outliers are mostly used to detect and treat outliers. In this study, we developed and evaluated improved procedures, labeling techniques, and Bayesian Averaging Modeling for fitting distribution to data in the presence of outliers. These improved procedures help to select the best model for fitting statistical distribution parameters to data in the presence of outliers by combining multiple models, each with different assumptions and parameter estimates. For example, one model can assume a heavy-tailed distribution to accommodate extreme outliers, while another model may assume a more normal distribution for most of the data. Our procedure, in many cases, was superior to traditional approaches. Thus, our procedure improved performance in the presence of outliers. That is, it helped to capture different types of outliers by considering models with different assumptions and allowing for the possibility of including models that specifically account for such types of outliers, leading to improved fitting performance. Therefore, the improved procedure we used in the study improves estimation and interpretation while enhancing the utilization of resources.

6.6. Study Implications for Risk Researchers and Practitioners

This study will improve the work of researchers and practitioners, particularly those who employ Monte-Carlo modeling, and they can find better-improved fitting of distributions,

particularly when outliers are present using labeling and Bayesian averaging. This can be used in risk modeling where there is high volatility, especially wholesale electricity prices, which are problematic in predicting day-to-day fluctuations in electricity prices. It gives you an idea of how bad things could get and how frequently, thereby enhancing decision-making. Bayesian averaging allows researchers to integrate diverse estimates or predictions, giving each estimate an appropriate weight based on its performance. The aggregate estimates provided a more robust and reliable basis for decision-making, as was the case in this study. By using the labeling and Bayesian averaging technique, we were able to select the best distribution-fitting model for each case study. In addition, this study will improve the work of researchers and practitioners by providing them the flexibility in modeling outliers. The flexibility of Bayesian averaging gives room for understanding outliers and their influence on risk assessment by capturing uncertainty associated with extreme values upward bound outliers, resulting in more accurate risk and parameter estimates for DCV prices. It gives risk practitioners who use Monte Carlo an idea about potential outcomes of DCV prices that helps to assess the likelihood of different DCV price levels, evaluate downside risks, and determine appropriate risk management strategies.

6.7. Avenues for Future Study

Although this thesis developed and evaluated improved procedures for fitting statistical distributions to historical data in the presence of "contaminating" outlier distributions, there are many topics that remain unexplored. The following sections address the selected topics for further studies.

First, this study focused on evaluating improved procedures for fitting distributions to only time series data in the presence of outliers. These alternative procedures were specifically developed to deal with outliers present in time series data but not cross-sectional data. Cross-

sectional data will be useful to understand the impact of Covid-19, the Russia-Ukraine war, price volatility, weather, demand, and supply, among other factors, on distribution fitting, risk analysis and assessment, and sensitivity analysis at a particular point in time. For example, given a 20% outlier contamination rate in the DCV and wheat futures market because of the Russian-Ukraine war, we can develop and evaluate improved procedures for fitting distributions to the datasets. More improved procedures can be developed for cross-sectional data analysis.

Second, other labeling procedures such as Dixon, Local Outlier Factor (LOF), k-Nearest Neighbor, Isolation Forest, and Generalized ESD remain unexplored in this study. The idea is to develop a more robust model tailored for each specific database, which would imply that the best-ranked procedure will provide a better distributional fit for further analysis. Given the multiple procedures and the different types of datasets, the BMA will help to select the most suitable procedure or model datasets in the presence of outliers. Thus, the question is, "whether the many labeling procedures are more convenient and more efficient than analyzing outliers using selected or few procedures?"

Third, applications in time series regression modeling and non-stationary series will be novel for future research. Regression models can assist in identifying outliers in time series. By modeling the relationship between the dependent variable and independent variables, regression models can identify observations that do not conform to the established relationship, thus aiding in outlier detection and data quality assessment. Also, regression models, such as linear or polynomial regression, can help to estimate and quantify the underlying behavior and facilitate trend analysis and forecasting. For example, the impact of covid-19 and the Russia-Ukraine war on wholesale electricity, futures, and DCV can be modeled to validate and evaluate the improved alternative procedures, forecasting, and sensibility analysis.

Lastly, outlier detection becomes challenging when dealing with huge volumes of data. Traditional methods of detecting outliers do not work effectively on big data. Machine learning algorithms offer advanced techniques to detect outliers by learning patterns and relationships from data. Techniques like clustering, density estimation, and deep learning-based anomaly detection can effectively identify anomalies in various data types and dimensions. In addition, ensemble methods, such as voting, averaging, or stacking, provide a robust framework for integrating multiple perspectives and enhancing overall outlier detection performance. Hence combining these methods will help improve the alternative procedures.

REFERENCES

Acuna, Edgar, and Caroline Rodriguez. 2004. “A Meta Analysis Study of Outlier Detection Methods in Classification” 1 (November):25.

Aggarwal, Charu C. 2017. *Outlier Analysis*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3.

Ajith Kumar, S.P., Priyank Pandey, Kaushal Mehta, and Manoj Kumar. 2019. “Identifying Outliers for Climatology Time Variant Series with Sliding Window.” In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 113–16.

Akram, Faiza, Dongsheng Liu, Peibiao Zhao, Natalia Kryvinska, Sidra Abbas, and Muhammad Rizwan. 2021. “Trustworthy Intrusion Detection in E-Healthcare Systems.” *Frontiers in Public Health* 9 (December):788347. https://doi.org/10.3389/fpubh.2021.788347.

Anderson, T. W., and D. A. Darling. 1952. “Asymptotic Theory of Certain ‘Goodness of Fit’ Criteria Based on Stochastic Processes.” *The Annals of Mathematical Statistics* 23 (2): 193–212. https://www.jstor.org/stable/2236446.

Barnard, G. A. 1963. “New Methods of Quality Control.” *Journal of the Royal Statistical Society. Series A (General)* 126 (2): 255–58. https://doi.org/10.2307/2982365.

Best, D. J., and J. C. W. Rayner. 2006. “Improved Testing for the Binomial Distribution Using Chi-Squared Components with Data-Dependent Cells.” *Journal of Statistical Computation and Simulation* 76 (1): 75–81. https://doi.org/10.1080/00949650412331320891.

Boukerche, Azzedine, Lining Zheng, and Omar Alfandi. 2021. “Outlier Detection: Methods, Models, and Classification.” *ACM Computing Surveys* 53 (3): 1–37. https://doi.org/10.1145/3381028.

Chatfield, Chris. 1995. “Model Uncertainty, Data Mining and Statistical Inference.” *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158 (3): 419–66. https://doi.org/10.2307/2983440.

Chih-Chien Yang and Chih-Chiang Yang. 2007. “Separating Latent Classes by Information Criteria.” *Journal of Classification* 24 (2): 183–203. https://doi.org/10.1007/s00357-007-0010-1.

Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge ; New York: Cambridge University Press.

Curtis, Alexander E., Tanya A. Smith, Bulat A. Ziganshin, and John A. Elefteriades. 2016. “The Mystery of the Z-Score.” *AORTA* 04 (4): 124–30. https://doi.org/10.12945/j.aorta.2016.16.014.

Darling, D. A. 1957. “The Kolmogorov-Smirnov, Cramer-von Mises Tests.” *The Annals of Mathematical Statistics* 28 (4): 823–38. https://doi.org/10.1214/aoms/1177706788.

Draper, David. 1995. “Assessment and Propagation of Model Uncertainty.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 45–97. https://www.jstor.org/stable/2346087.

“Euclidean Distance.” 2023. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Euclidean_distance&oldid=1142600543#cite_note-9.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14 (4): 382–401. https://www.jstor.org/stable/2676803.

Iglewicz, Boris, and David C. Hoaglin. 1993. *How to Detect and Handle Outliers*. Edited by Edward F. Mykytka. Vol. 16. 16 vols. Basic References in Quality Control: Statistical Techniques. Milwaukee, WI: American Society of Quality Control.

Jagadeeswari, T, and N Harini. 2013. "Identification of Outliers by Cook's Distance in Agriculture Datasets" 2 (6): 4.

Janakiram, D., A.V.U.P. Kumar, and Adi Mallikarjuna Reddy V. 2006. "Outlier Detection in Wireless Sensor Networks Using Bayesian Belief Networks." In *2006 1st International Conference on Communication Systems Software & Middleware*, 1–6. https://doi.org/10.1109/COMSWA.2006.1665221.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–95. https://doi.org/10.2307/2291091.

Kass, Robert E., and Suresh K. Vaidyanathan. 1992. "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions." *Journal of the Royal Statistical Society. Series B (Methodological)* 54 (1): 129–44. https://www.jstor.org/stable/2345950.

Kass, Robert E., and Larry Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90 (431): 928–34. https://doi.org/10.2307/2291327.

Knox, Edwin M., and Raymond T. Ng. 1998. "Algorithms for Mining Distancebased Outliers in Large Datasets." In *Proceedings of the International Conference on Very Large Data Bases*, 392–403. Citeseer.

Kotsiantis, S B, Ioannis Zaharakis, and P Pintelas. 2007. "Supervised Machine Learning: A Review of Classification Techniques," July, 249–68.

Kuha, Jouni. 2004. "AIC and BIC: Comparisons of Assumptions and Performance." *Sociological Methods & Research* 33 (2): 188–229. https://doi.org/10.1177/0049124103262065.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1): 79–86. https://doi.org/10.1214/aoms/1177729694.

Lightfoot, Emma, and Tamsin C. O'Connell. 2016. "On the Use of Biomineral Oxygen Isotope Data to Identify Human Migrants in the Archaeological Record: Intra-Sample Variation, Statistical Methods and Geographical Considerations." *PloS One* 11 (4): e0153850. https://doi.org/10.1371/journal.pone.0153850.

MacQueen, James. 1967. "Classification and Analysis of Multivariate Observations." In *5th Berkeley Symp. Math. Statist. Probability*, 281–97.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Nayak, Janmenjoy, Bighnaraj Naik, D. P. Kanungo, and H. S. Behera. 2015. "An Improved Swarm Based Hybrid K-Means Clustering for Optimal Cluster Centers." In *Information Systems Design and Intelligent Applications*, edited by J. K. Mandal, Suresh Chandra Satapathy, Manas Kumar Sanyal, Partha Pratim Sarkar, and Anirban Mukhopadhyay, 339:545–53. Advances in Intelligent Systems and Computing. New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2250-7_54.

Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén. 2007. "Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo

Simulation Study." *Structural Equation Modeling: A Multidisciplinary Journal* 14 (4): 535–69. https://doi.org/10.1080/10705510701575396.
Pettitt, A. N. 1976. "A Two-Sample Anderson-Darling Rank Statistic." *Biometrika* 63 (1): 161–68. https://doi.org/10.1093/biomet/63.1.161.
Ramaswamy, Sridhar, Rajeev Rastogi, and Kyuseok Shim. 2000. "Efficient Algorithms for Mining Outliers from Large Data Sets," May, 427–38.
Roberts, Harry V. 1965. "Probabilistic Prediction." *Journal of the American Statistical Association* 60 (309): 50–62. https://doi.org/10.2307/2283136.
Saleem, Sehar, Maria Aslam, and Mah Rukh Shaukat. n.d. "A Review and Empirical Comparison of Univariate Outlier Detection Methods."
Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://www.jstor.org/stable/2958889.
Shao, Jun. 1997. "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica* 7 (2): 221–42. https://www.jstor.org/stable/24306073.
Stephens, M. A. 1974. "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association* 69 (347): 730–37. https://doi.org/10.2307/2286009.
Stoica, P., and Y. Selen. 2004. "Model-Order Selection: A Review of Information Criterion Rules." *IEEE Signal Processing Magazine* 21 (4): 36–47. https://doi.org/10.1109/MSP.2004.1311138.
Taddy, Matt. 2019. *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*. McGraw Hill Professional.
Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, MA: Addison-Wesley.
"U.S. Energy Information Administration - EIA - Independent Statistics and Analysis." n.d. Accessed April 19, 2023. https://www.eia.gov/electricity/wholesale/#history.

APPENDIX A. SUPPLEMENTARY TABLES

Table A1. Two-sample KS test - Ignore Outlier

Variable	p-value	alpha	D
PJM	0.0001***	0.05	0.0533
Indiana	0.1055	0.05	0.0331
NEPOOL	0.6255	0.05	0.0614
Oats Futures	0.1519	0.05	0.0286
W Nearby Daily Intermonth Spread	0.0001***	0.05	0.2831
KW-W Daily Intermarket Spread	0.0301*	0.05	0.0362
DCV($/car)	0.2020	0.05	0.0488

Signification codes: 0 < "***" < 0.001 < "**" < 0.01 < "*" < 0.05 < "." < 0.1 < " " < 1

Table A2. Two-sample KS test - Drop Outlier

Variable	p-value	alpha	D
PJM	0.0001***	0.05	0.0455
Indiana	0.0055**	0.05	0.0468
NEPOOL	0.0001***	0.05	0.0614
Oats Futures	0.1608	0.05	0.0283
W Nearby Daily Intermonth Spread	0.0014**	0.05	0.0477
KW-W Daily Intermarket Spread	0.0348*	0.05	0.0355
DCV($/car)	0.0001***	0.05	0.3911

Signification codes: 0 < "***" < 0.001 < "**" < 0.01 < "*" < 0.05 < "." < 0.1 < " " < 1

Table A3. Two-sample KS test - Z-Score Bayesian Averaging

Variable	p-value	alpha	D
PJM	0.4093	0.05	0.0173
Indiana	0.7855	0.05	0.0178
NEPOOL	0.9345	0.05	0.0109
Oats Futures	0.7217	0.05	0.0175
W Nearby Daily Intermonth Spread	0.0074**	0.05	0.0418
KW-W Daily Intermarket Spread	0.0301*	0.05	0.0362
DCV($/car)	0.0001***	0.05	0.3994

Signification codes: 0 < "***" < 0.001 < "**" < 0.01 < "*" < 0.05 < "." < 0.1 < " " < 1

Table A4. Two-sample KS test Tukey IQR Bayesian Averaging

Variable	p-value	alpha	D
PJM	0.4816	0.05	0.0164
Indiana	0.8822	0.05	0.0160
NEPOOL	0.9345	0.05	0.0109
Oats Futures	0.0001***	0.05	0.0655
W Nearby Daily Intermonth Spread	0.0686	0.05	0.0324
KW-W Daily Intermarket Spread	0.0492*	0.05	0.0340
DCV($/car)	0.0001***	0.05	0.3932

Signification codes: 0 < "***" < 0.001 < "**" < 0.01 < "*" < 0.05 < "." < 0.1 < " " < 1

Table A5. Two-sample KS test K-means IQR Bayesian Averaging

Variable	p-value	alpha	D
PJM	0.0004***	0.05	0.0402
Indiana	0.9941	0.05	0.0115
NEPOOL	0.7109	0.05	0.0142
Oats Futures	0.0001***	0.05	0.1995
W Nearby Daily Intermonth Spread	0.6285	0.05	0.0187
KW-W Daily Intermarket Spread	0.0008***	0.05	0.0493
DCV($/car)	0.0001***	0.05	0.3932

Signification codes: 0 < "***" < 0.001 < "**" < 0.01 < "*" < 0.05 < "." < 0.1 < " " < 1

APPENDIX B. SUPPLEMENTARY FIGURES

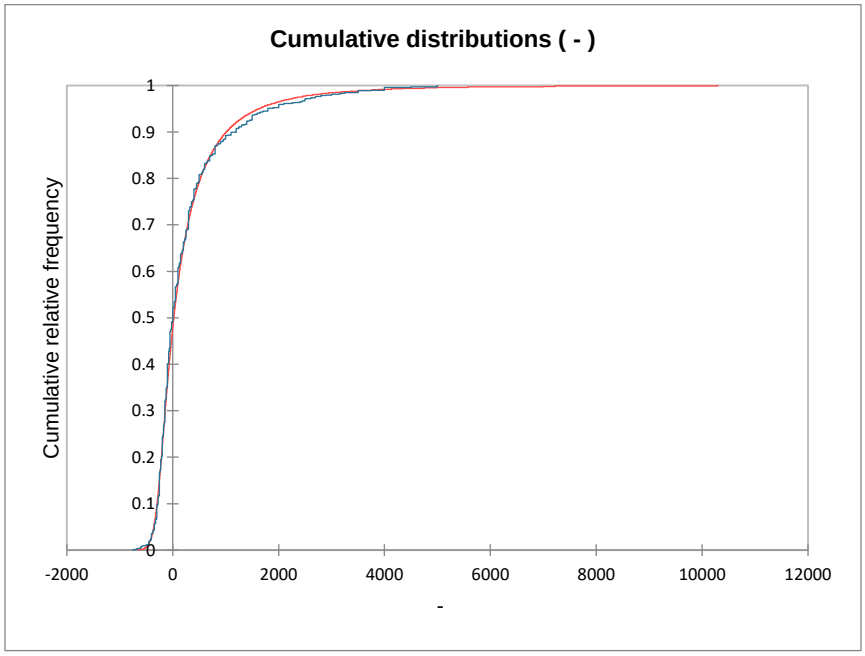

Figure A1. DCV KS-Ignore Outlier Cumulative Distribution Comparison.

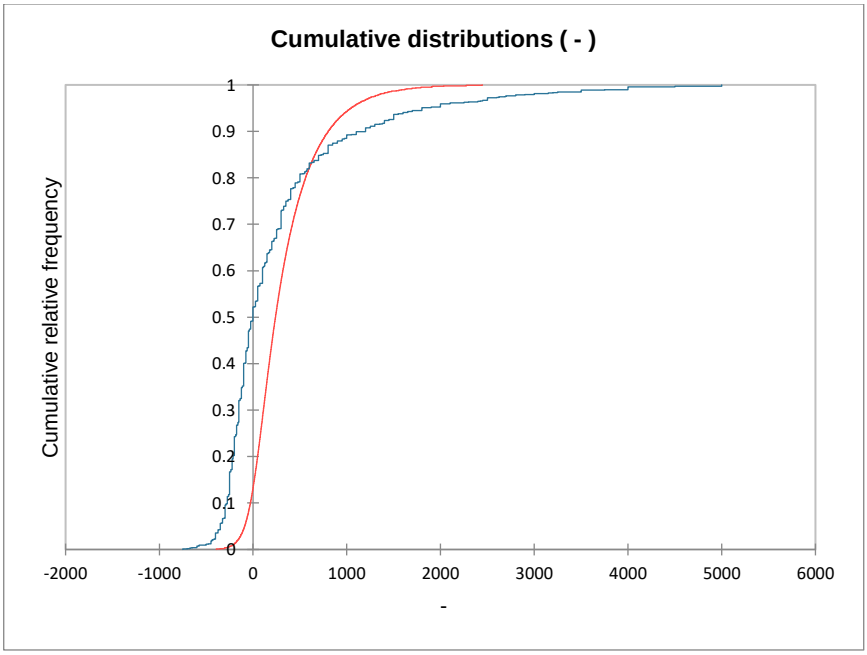

Figure A2. DCV KS-Drop Outlier Cumulative Distribution Comparison.

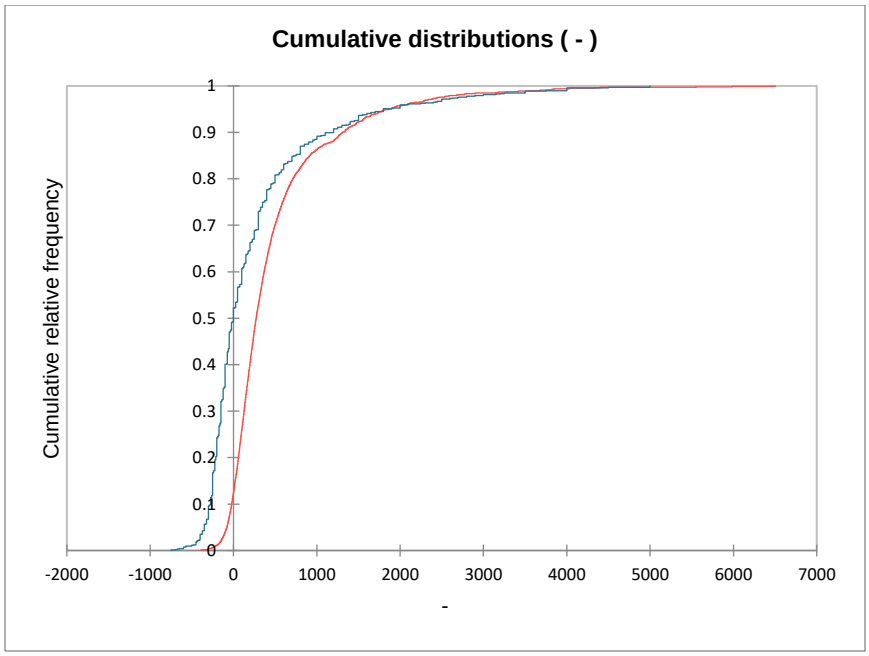

Figure A3. DCV KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison.

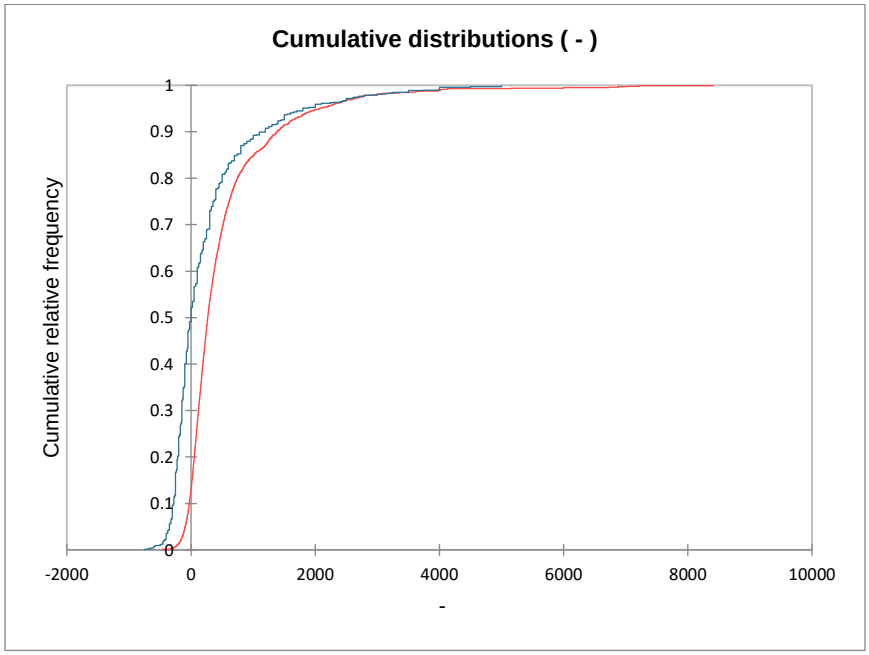

Figure A4. DCV KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison.

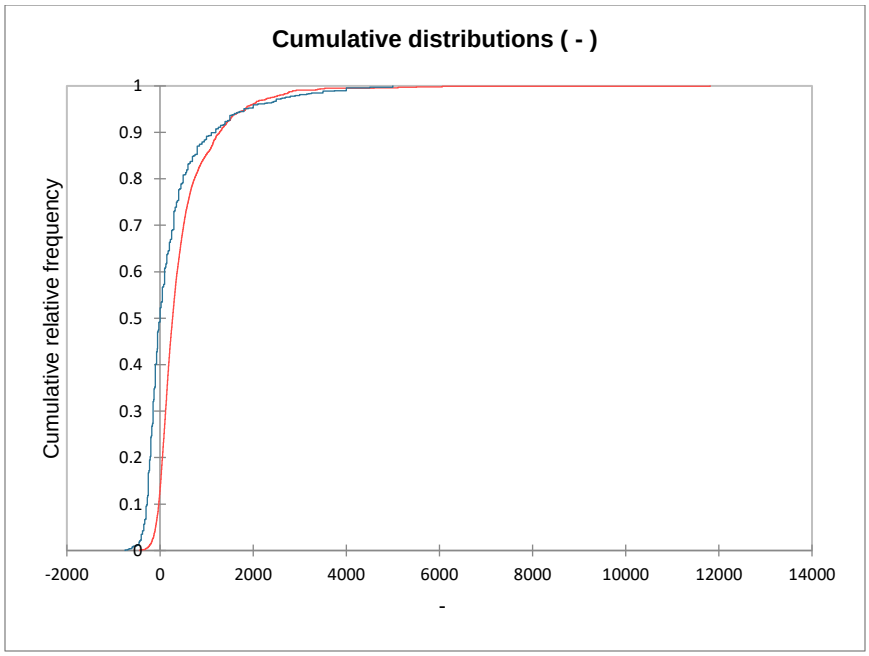

Figure A5. DCV KS-k-Means Bayesian Averaging Cumulative Distribution Comparison.

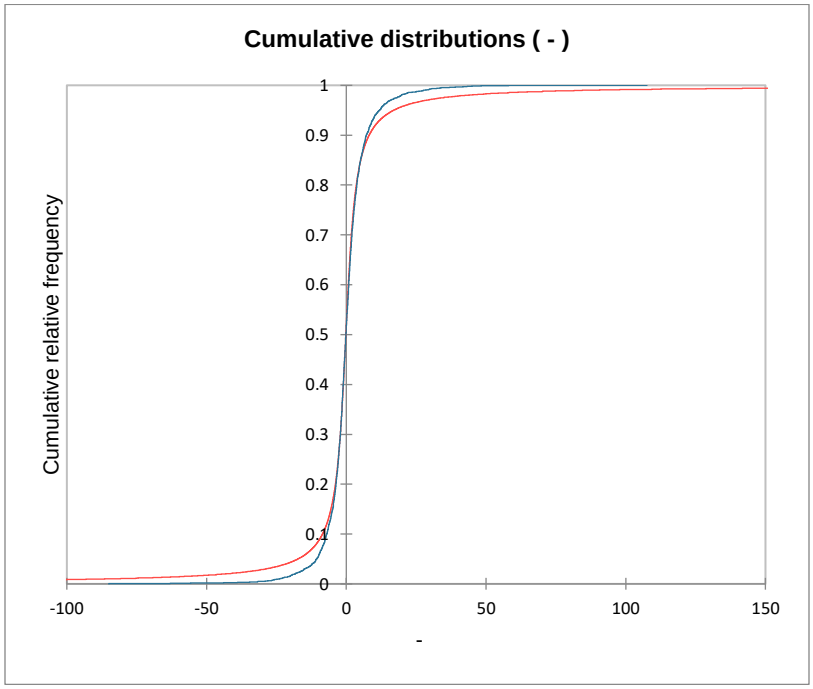

Figure A6. Indiana Hub KS-Ignore Outlier Bayesian Averaging Cumulative Distribution Comparison.

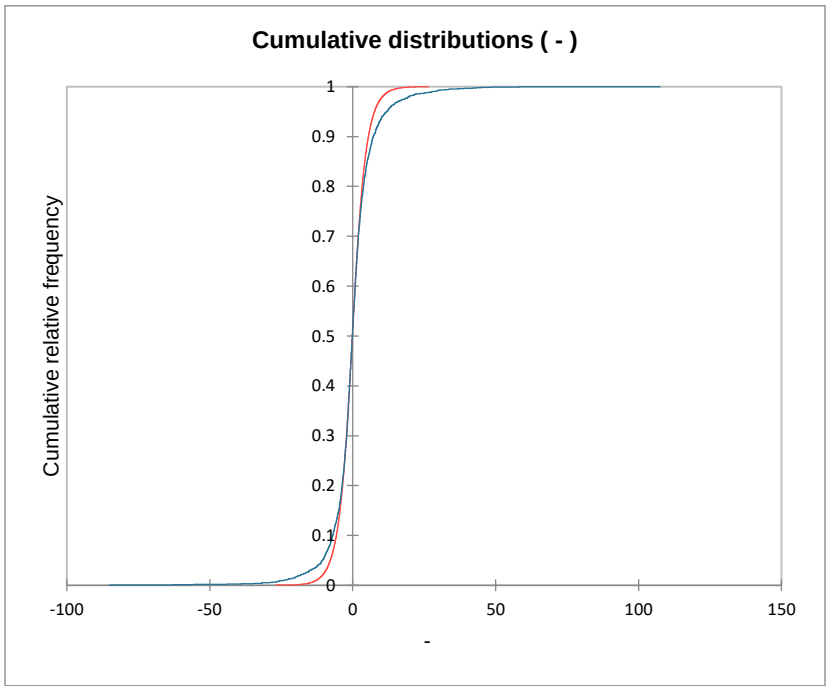

Figure A7. Indiana Hub KS-Drop Outlier Bayesian Averaging Cumulative Distribution Comparison.

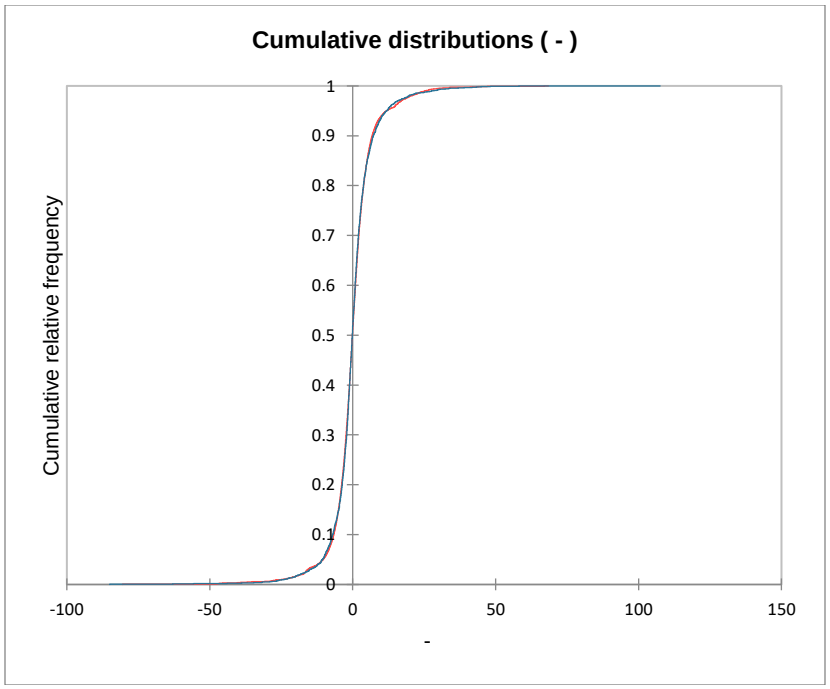

Figure A8. Indiana Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison

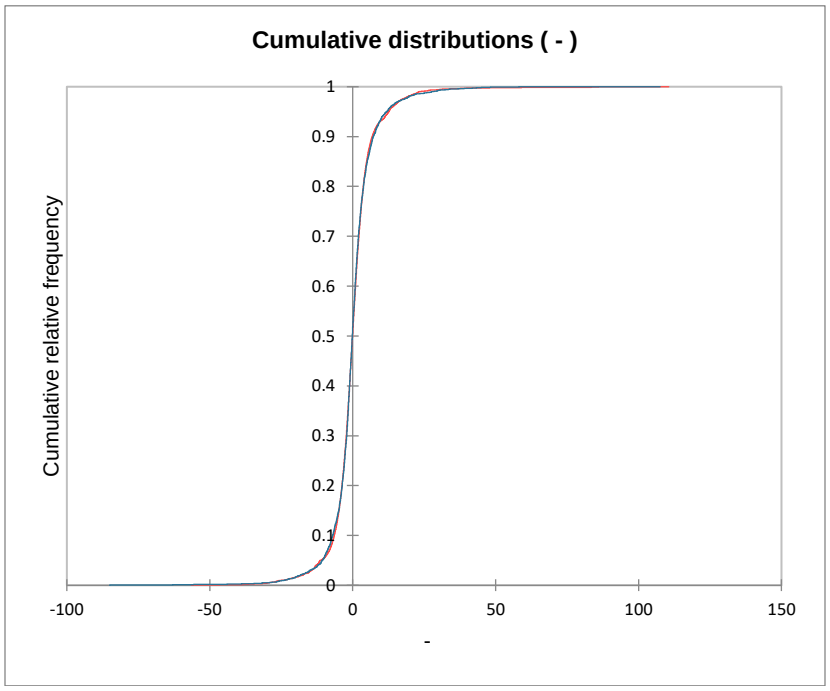

Figure A9. Indiana Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

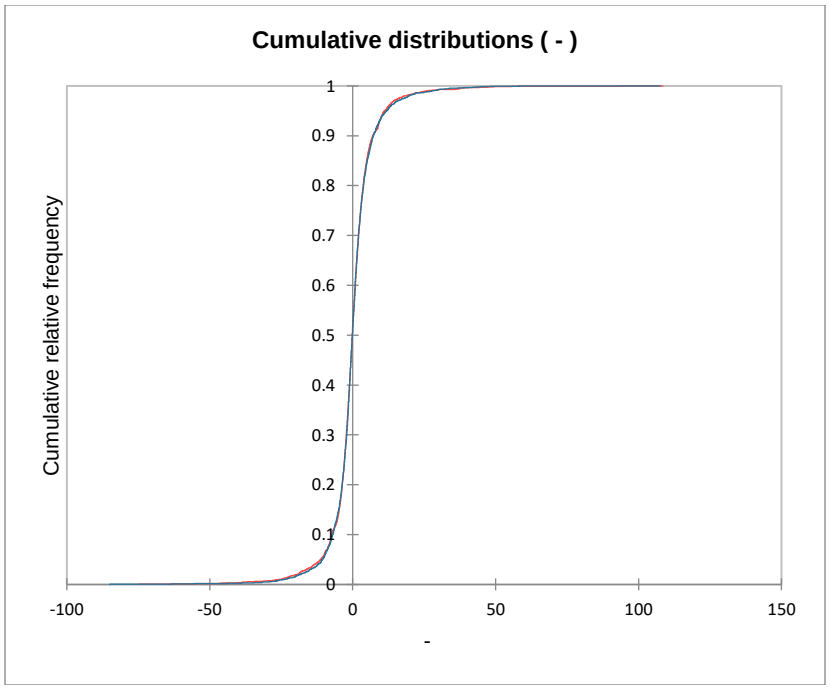

Figure A10. Indiana Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison

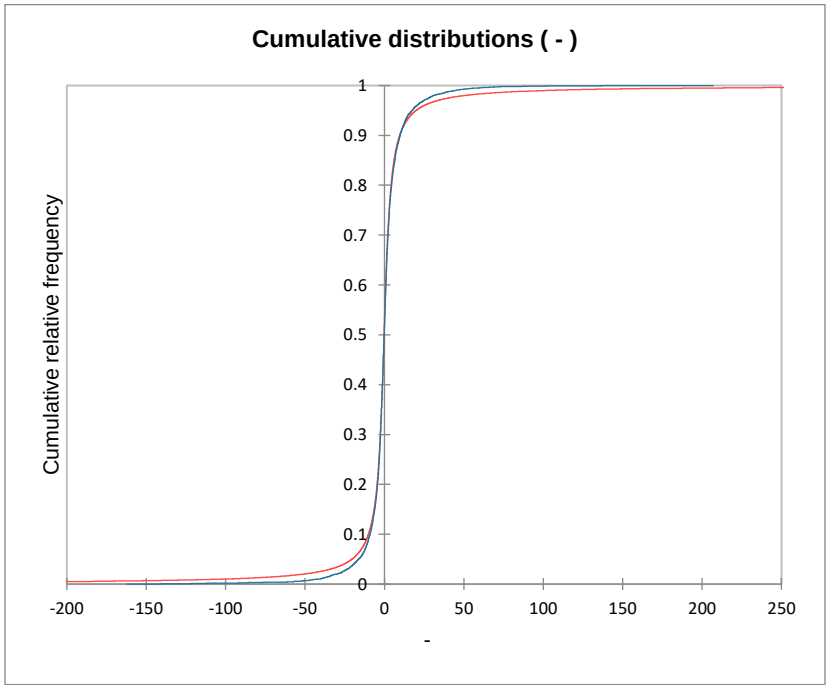

Figure A11. NEPOOL Hub KS-Ignore Outlier Cumulative Distribution Comparison.

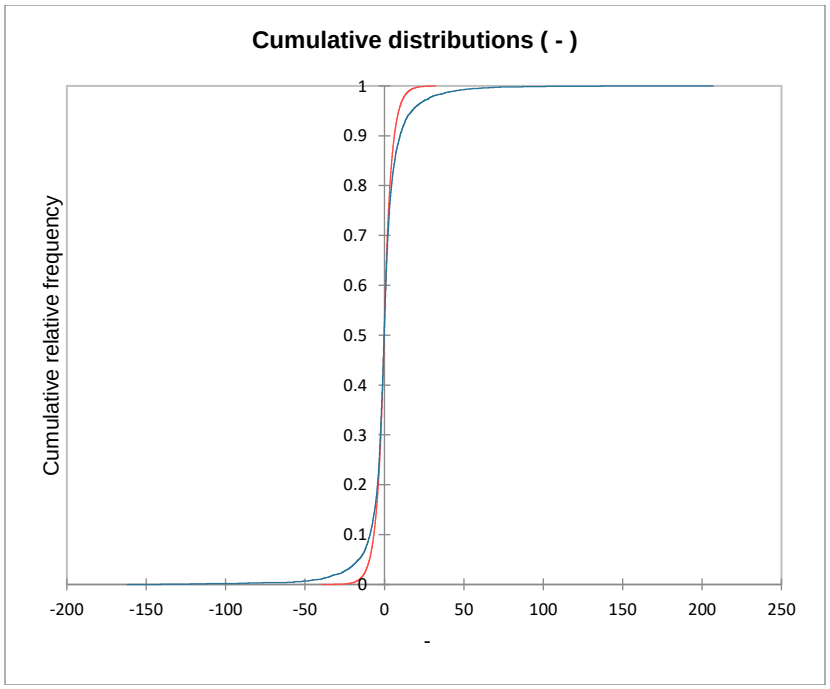

Figure A12. NEPOOL Hub KS-Drop Outlier Cumulative Distribution Comparison.

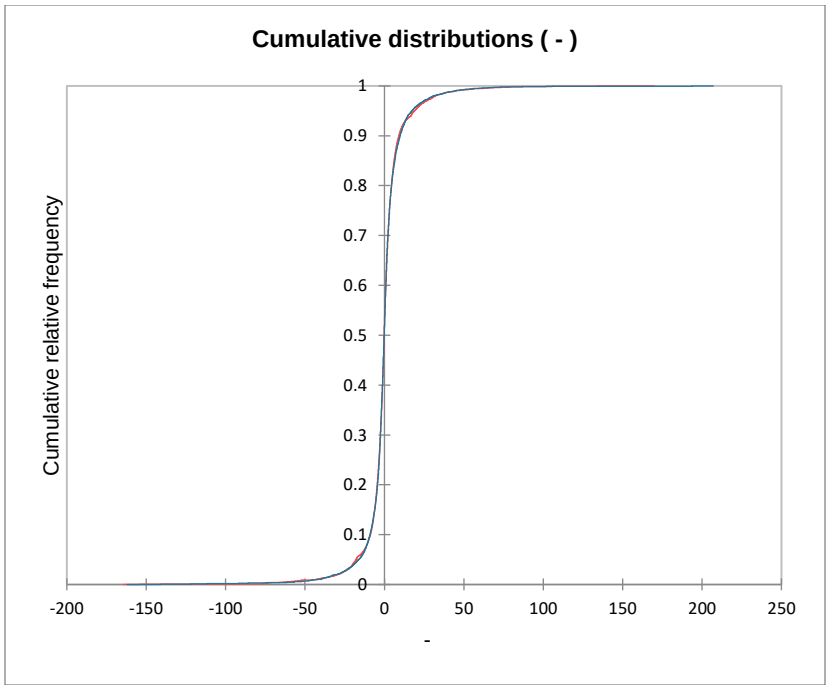

Figure A13. NEPOOL Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison.

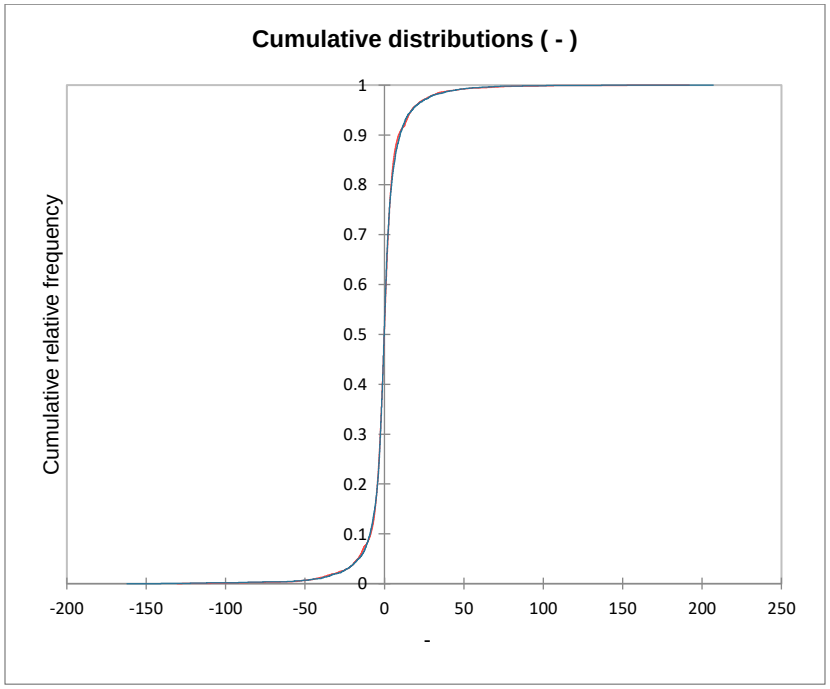

Figure A14. NEPOOL Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

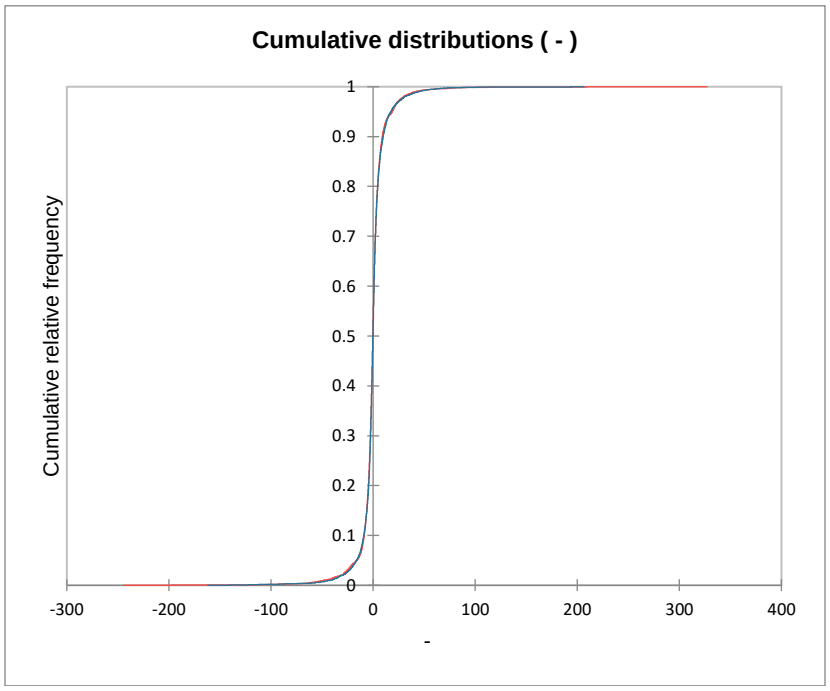

Figure A15. NEPOOL Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison

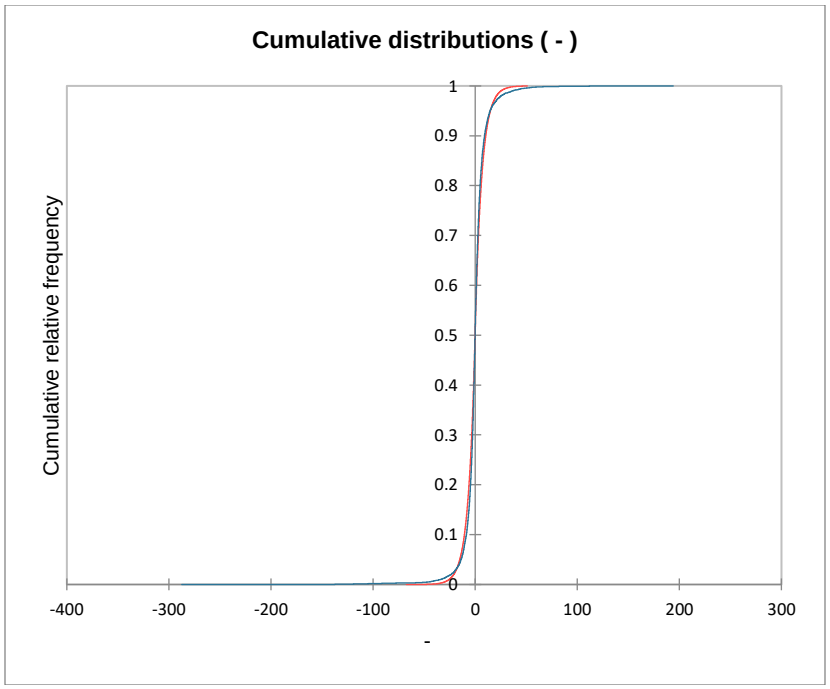

Figure A16. PJM West Hub KS-Ignore Outliers Cumulative Distribution Comparison

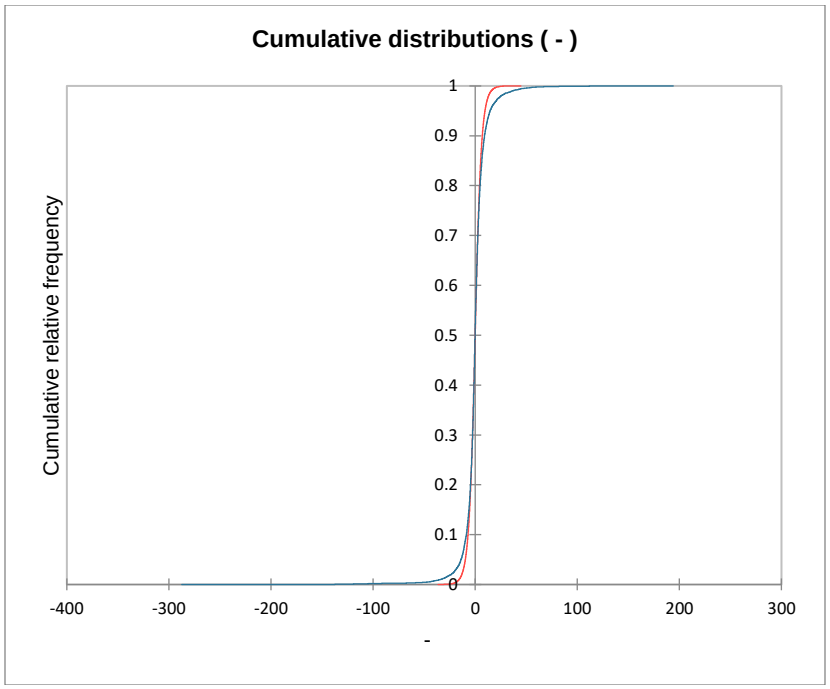

Figure A17. PJM West Hub KS-Drop Outliers Cumulative Distribution Comparison

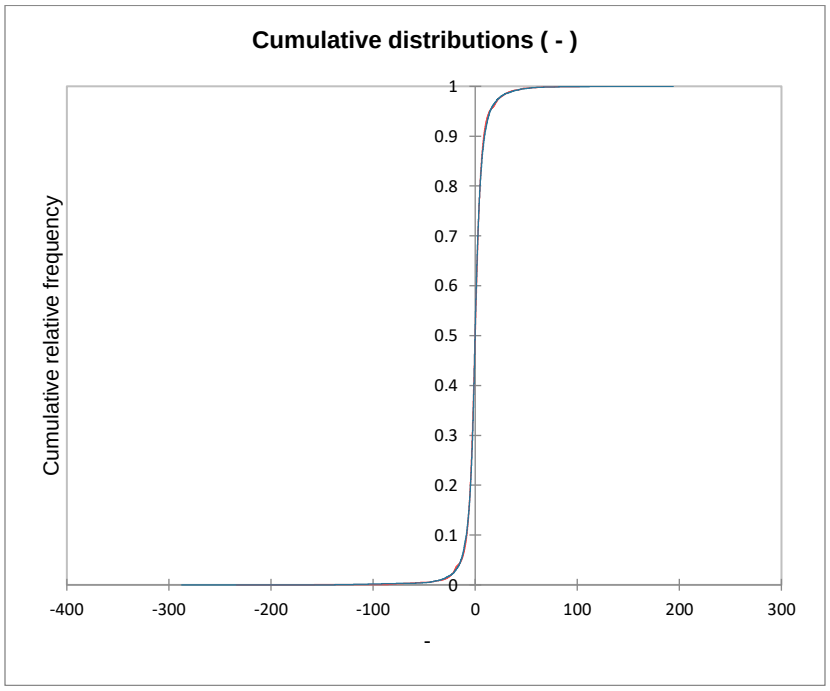

Figure A18. PJM West Hub KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison

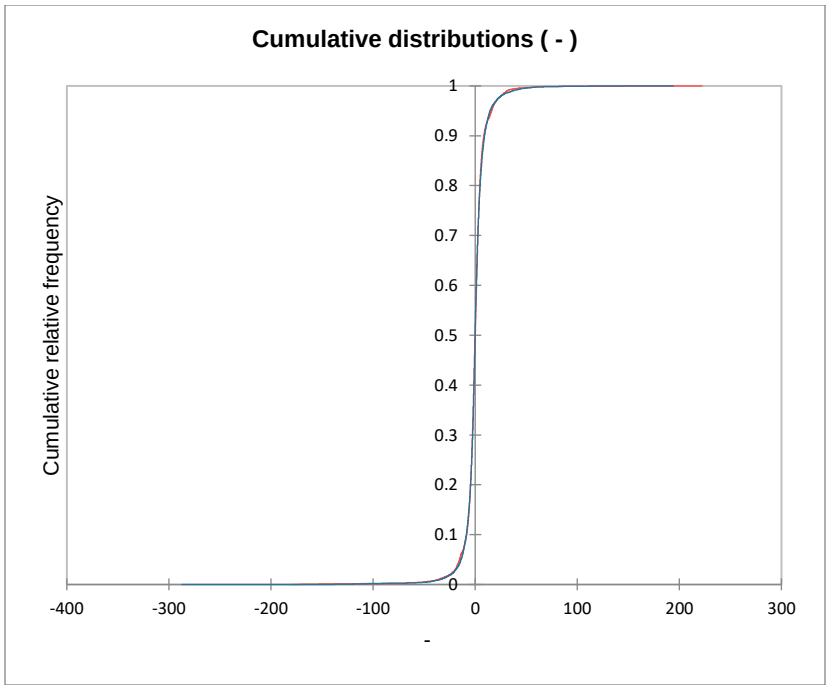

Figure A19. PJM West Hub KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

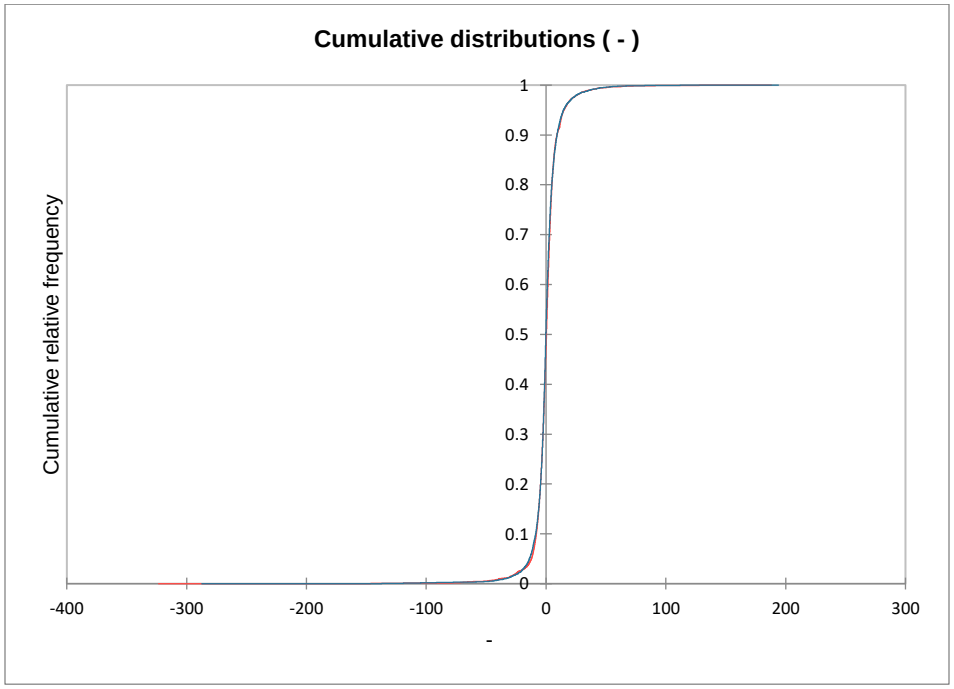

Figure A20. PJM West Hub KS-k-Means Bayesian Averaging Cumulative Distribution Comparison

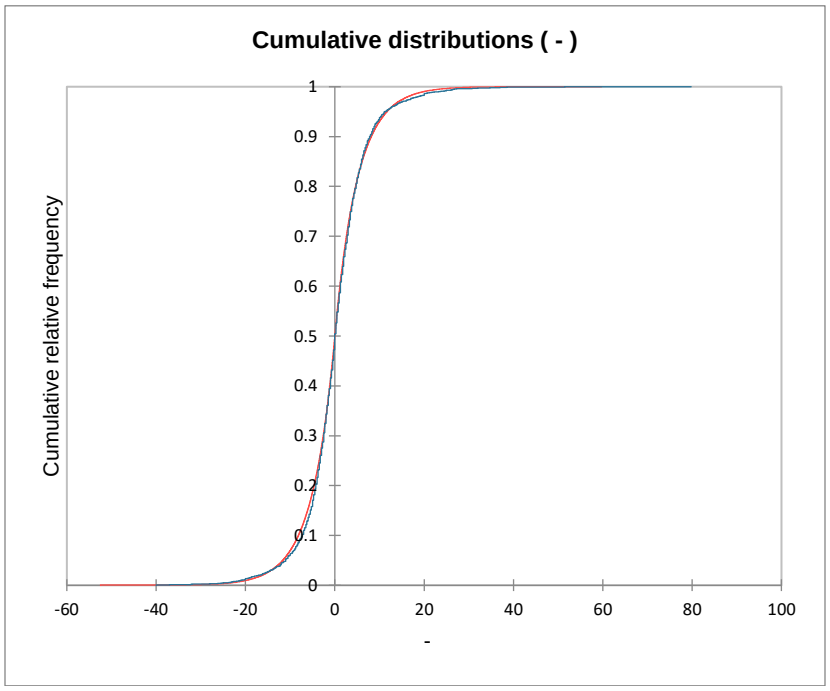

Figure A21. Chng_O KS-Ignore Outliers Cumulative Distribution Comparison

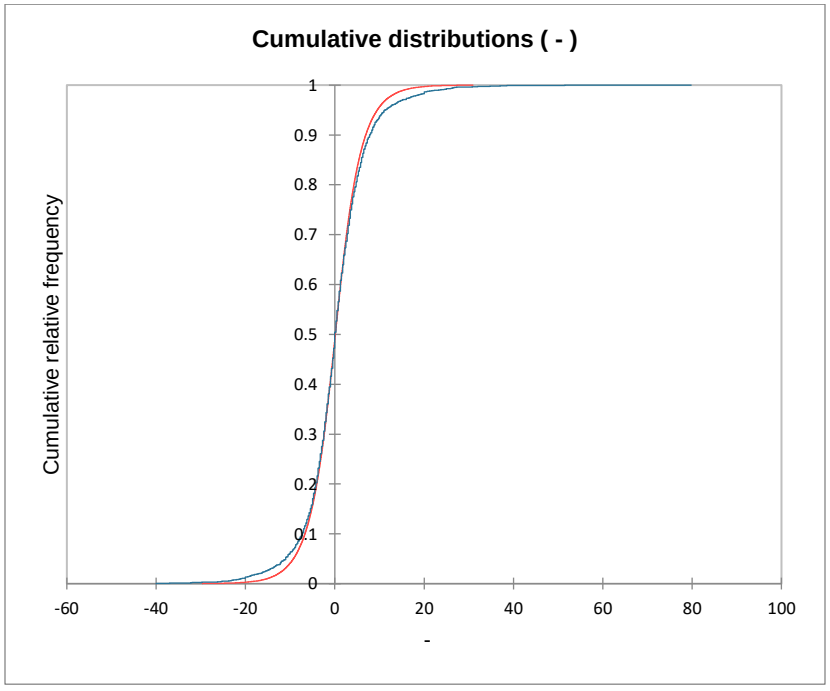

Figure A22. Chng_O KS-Drop Outliers Cumulative Distribution Comparison

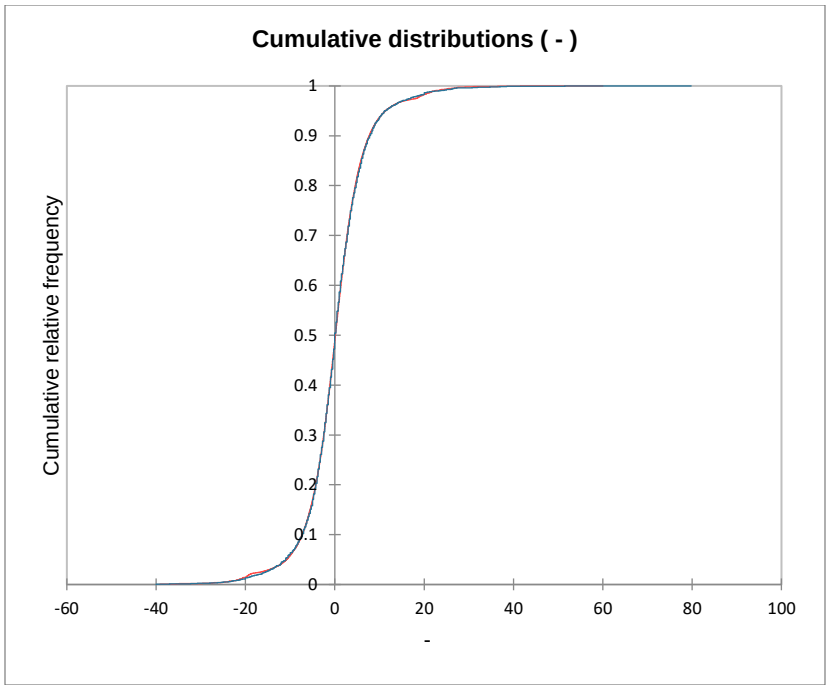

Figure A23. Chng_O KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison

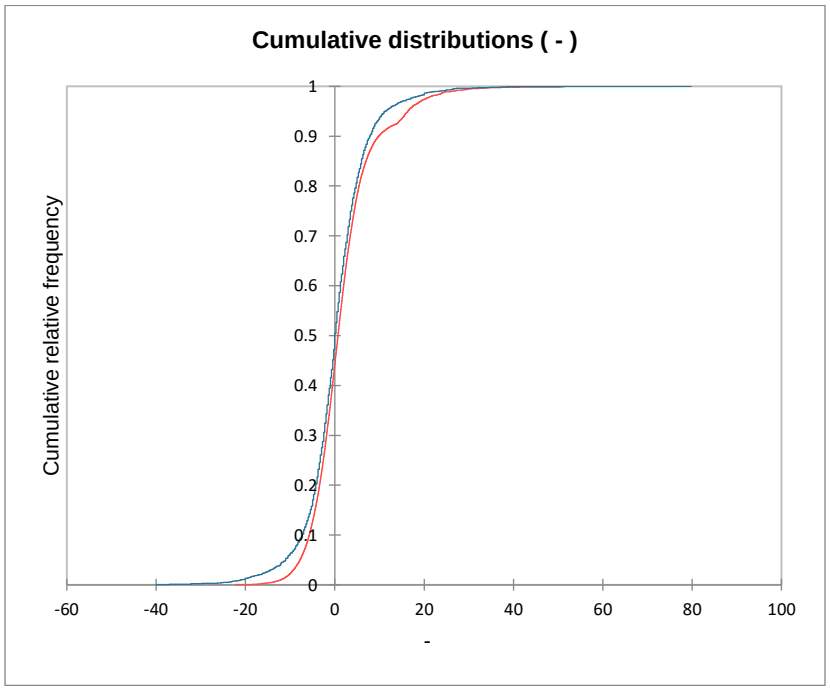

Figure A24. Chng_O KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

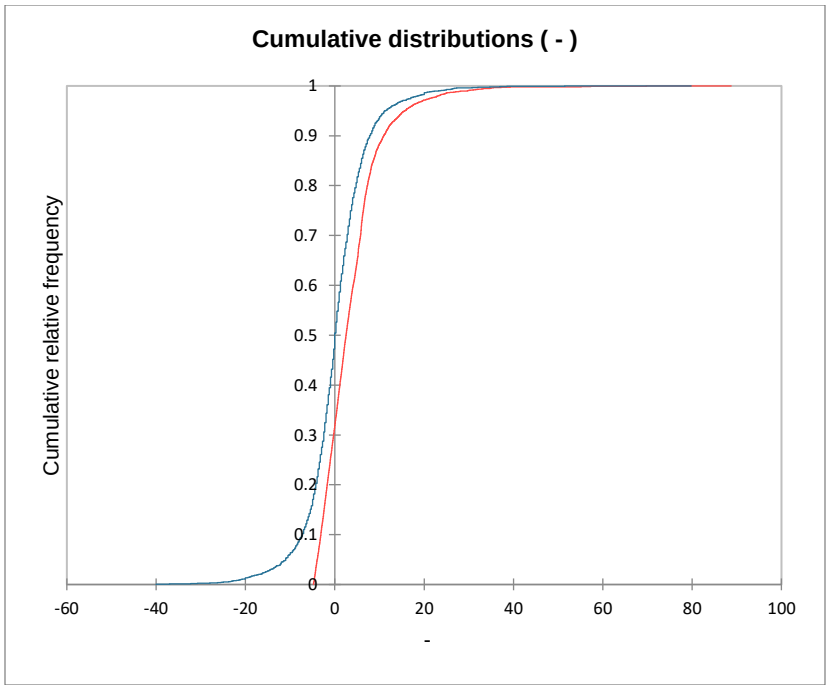

Figure A25. Chng_O KS-k-Means Bayesian Averaging Cumulative Distribution Comparison

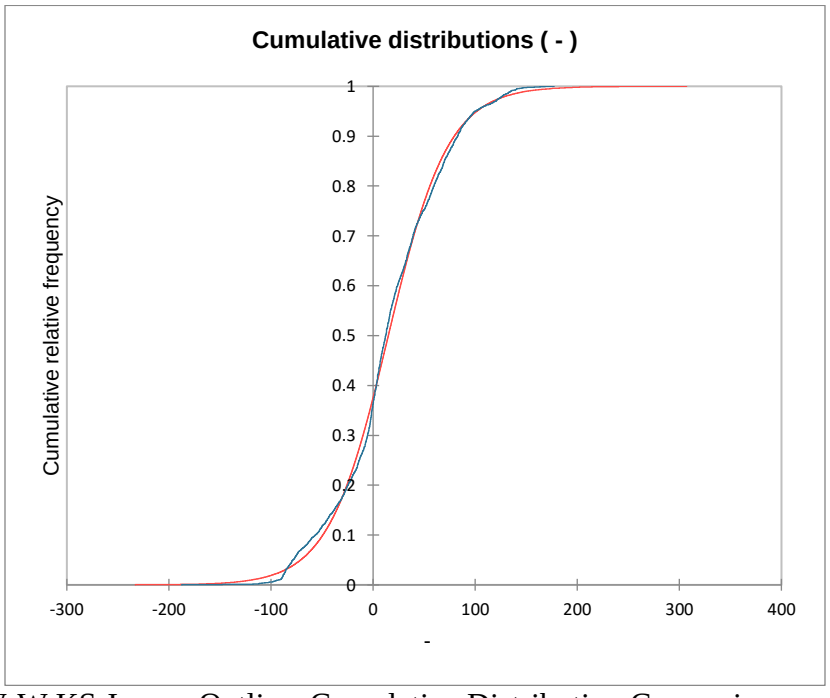

Figure A26. KW-W KS-Ignore Outliers Cumulative Distribution Comparison

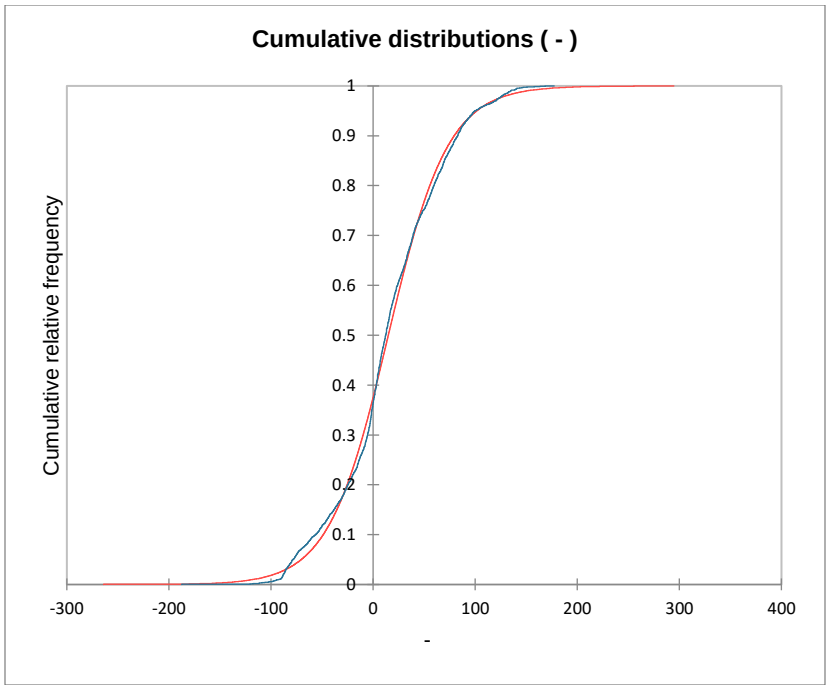

Figure A27. KW-W KS-Drop Outliers Cumulative Distribution Comparison

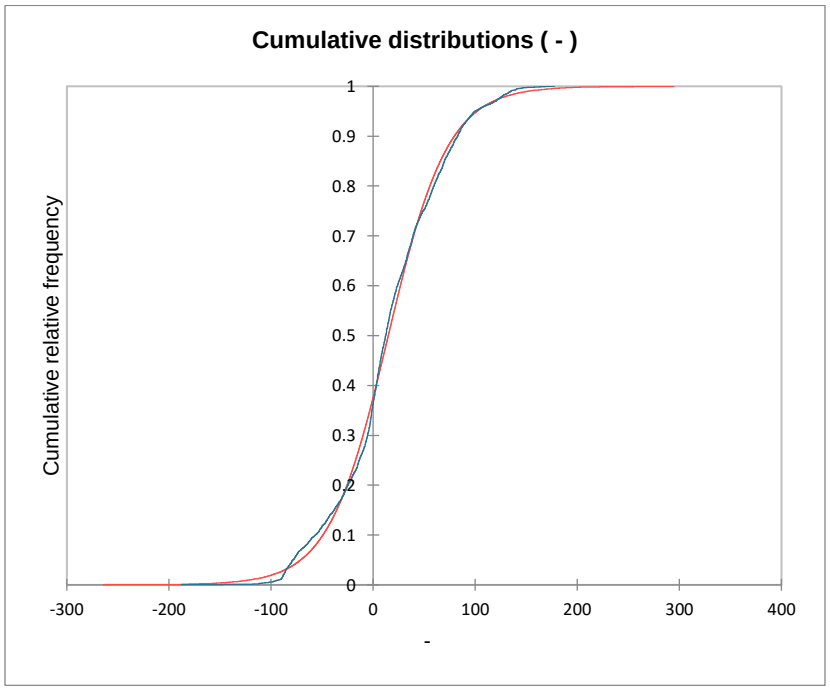

Figure A28. KW-W KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison

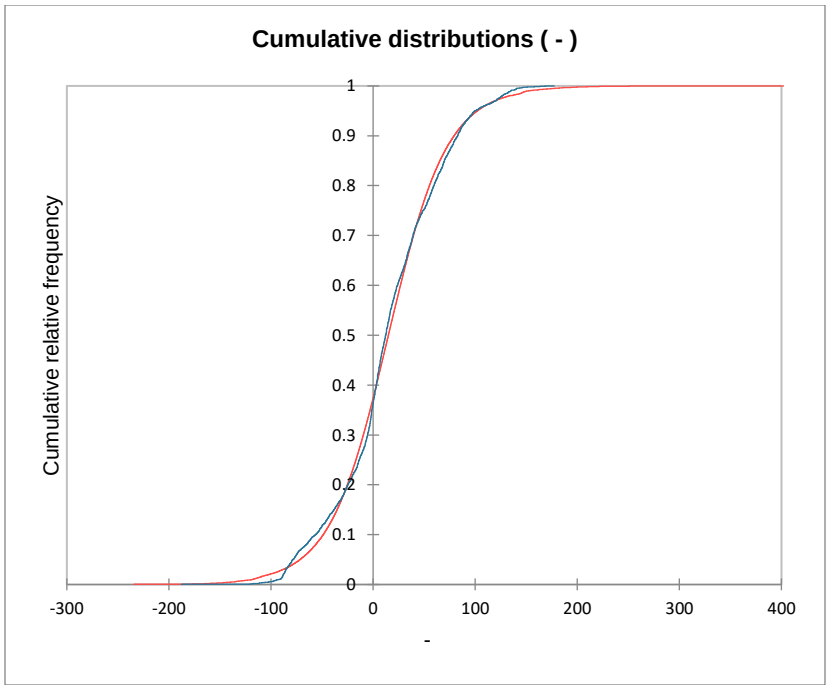

Figure A29. KW-W KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

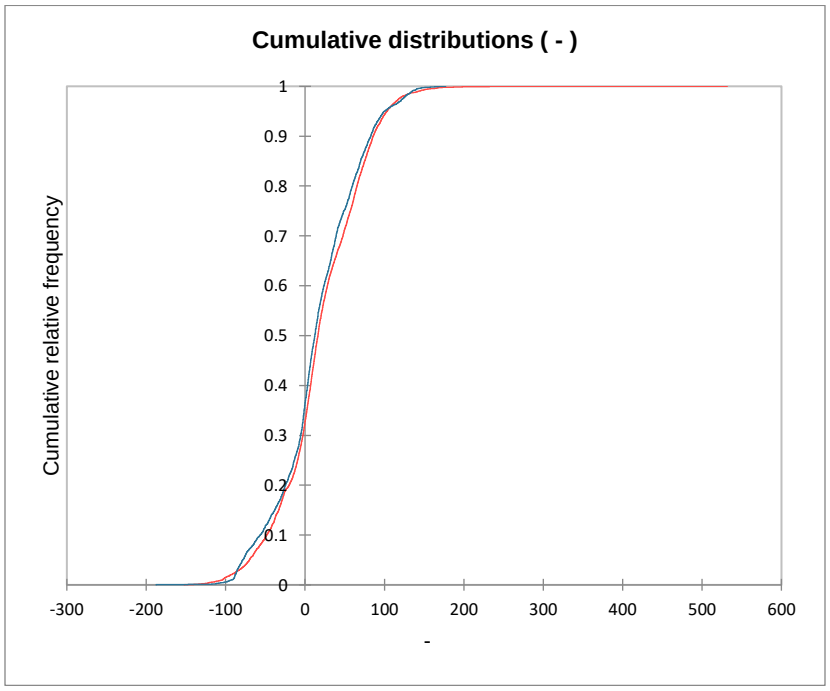

Figure A30. KW-W KS-k-Means Bayesian Averaging Cumulative Distribution Comparison

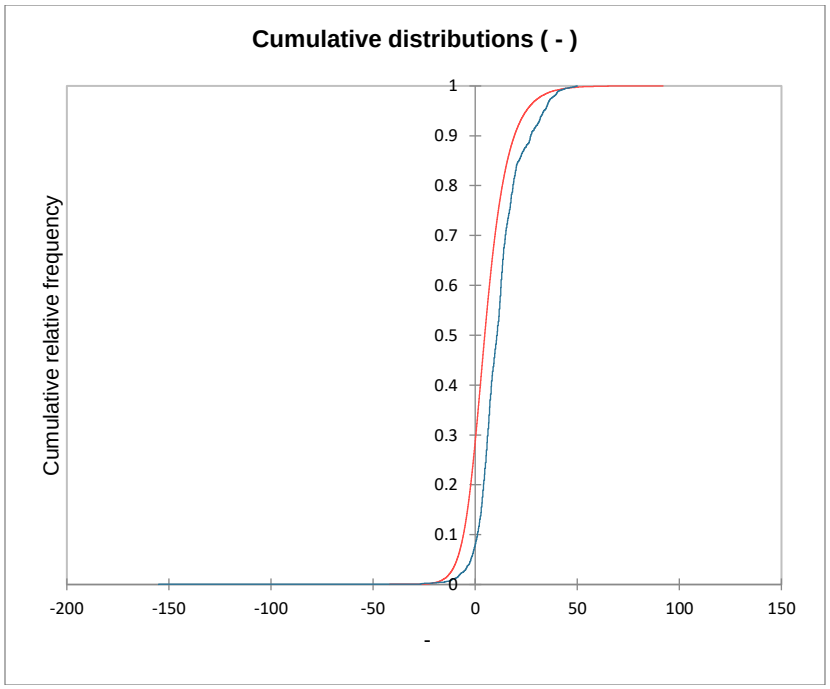

Figure A31. W KS- Ignore Outliers Cumulative Distribution Comparison

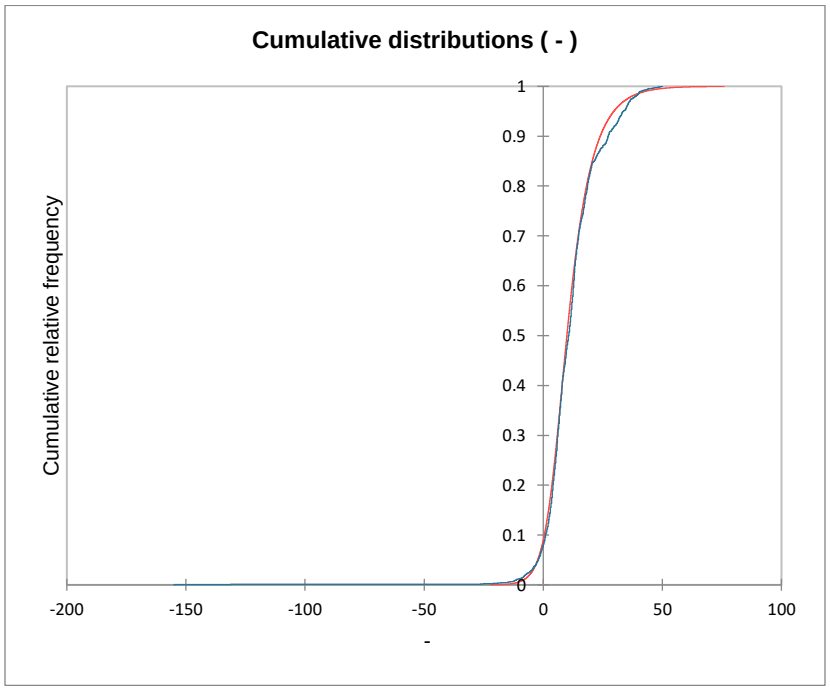

Figure A32. W KS- Drop Outliers Cumulative Distribution Comparison

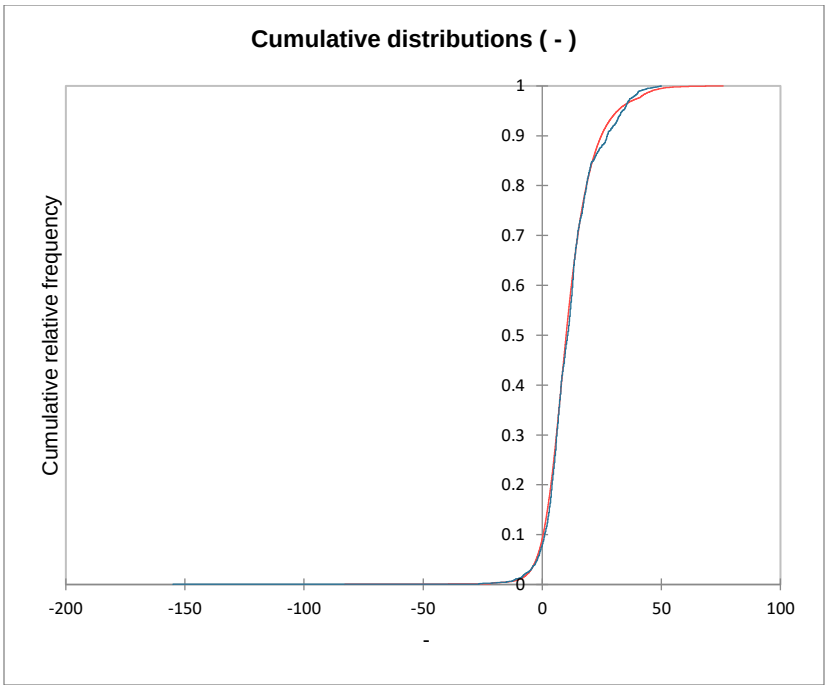

Figure A33. W KS-Z-Score Bayesian Averaging Cumulative Distribution Comparison

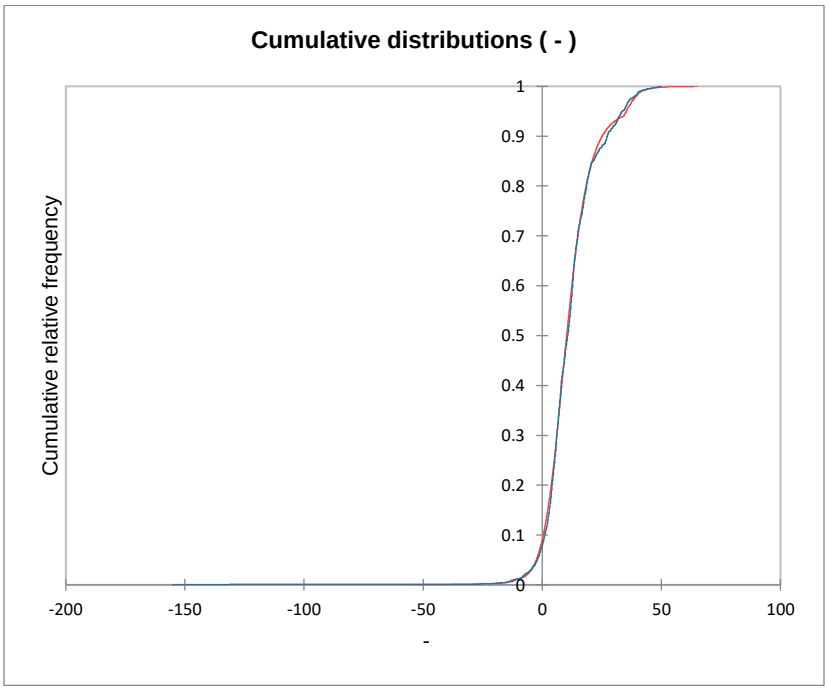

Figure A34. W KS-Tukey IQR Bayesian Averaging Cumulative Distribution Comparison

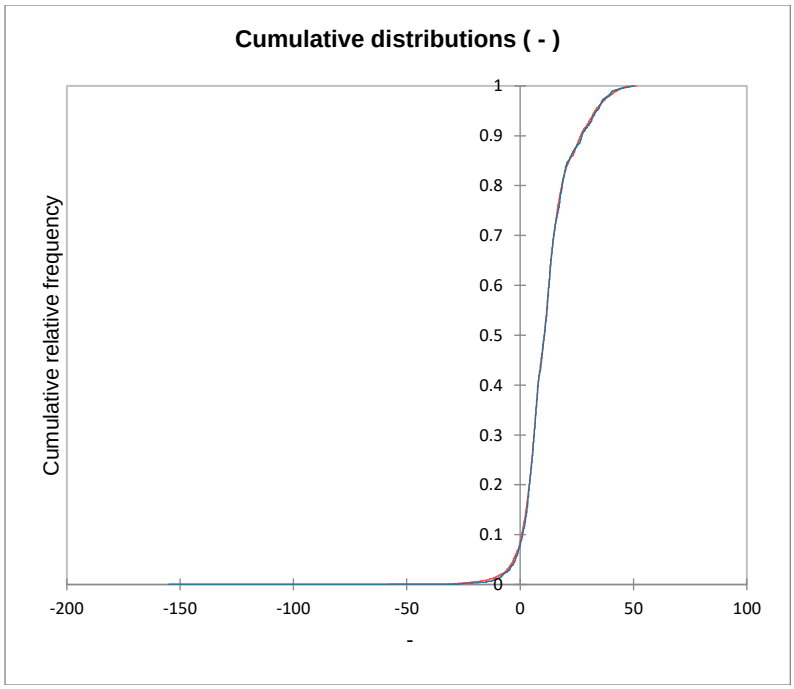

Figure A35. W KS-k-Means Bayesian Averaging Cumulative Distribution Comparison.