ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

Comparison of Boxplots for Outlier Detection in Performance Modelling

[1] Joice Mary Philip, [2] Abraham George
[1] Research Scholar, School of Engineering, Cochin University of Science and Technology, Kerala, India
[2] Deputy General Manager, NLC India Ltd, Neyveli, Tamilnadu, India
Corresponding Author email: jophabres@gmail.com

***Abstract:*-- Distorted values creeping into a data due to sampling, experimental, instrumental, manual , data handling or data processing errors can mislead the prediction of performance. Misfits in an observational data has to be diagnosed which need to be treated before modelling. Quality of data on the material characteristics, determines the accuracy in the performance prediction of a product. In this paper, the reported incompetence of models in a research data and the reason for model inaccuracy is considered . Examination of the data under study using Tukey's traditional boxplot, and two other medcouple based adjusted boxplots indicated presence of outliers in the data on characteristics of different types of fly ash. Skewness in the data on fly ash characteristics revealed through histogram and density plots were dealt by transformations done to the data. Impact of data transformation in outlier detection is studied for the 3 boxplots. Suitability of each method for the detection of outliers is assessed using sensitivity and specificity calculations. Sensitivity or True Positive Rate is found to be maximum in modified adjusted boxplots while specificity or True Negative Rate is found to be maximum in traditional boxplots. Adjusted boxplots showed least variation in the results with transformed and nontransformed data which suggests it to be suitable for a nontransformed data. Performance models could predict well for the winsorised data based on adjusted box plots.**

Index Terms:- BoxPlot, Data Transformation , Interquartile Range, Medcouple, Sensitivity, Specificity, True Positive Rate, False Positive Rate, True Negative Rate

I. INTRODUCTION

Inconsistent observations in a data, if not properly treated, leads to distorted models with wrong predictions for product performance resulting in faulty products. Hence detection and treatment of outliers in the data is crucial in modelling any test results. Hawkins D.M [2] defines outlier intuitively as, "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Boxplot is a versatile exploratory data analysis (EDA) tool, which helps to visualize the location, spread, and skewness of data distributions, along with unusual values or outliers. The most common version of the boxplot by Tukey is termed as traditional boxplot in which an observation is considered as a mild outlier when its value is not within the inner fence and as an extreme outlier when it is not within the outer fence. Insia Hussain [3] suggests that boxplot which uses 'mode and density' approach is more reliable compared to the bagplot while obtaining the outliers.

Masking-swamping effect, skewness of data set and number of records affects the detection of outliers. Masking is the phenomenon which makes each outlier difficult to be detected due to the presence of some outliers. M. K. James et al.,[4] informs that in cases of non-normality, for simplicity and robustness, Tukey box plot stands-out for sample sizes greater than 10. Y. H. Dovoedo [5] reports that 'traditional boxplot does not use the extreme observations in the construction of the fences, which makes the boxplot not suffer from "masking". Tukey's traditional boxplot is being widely used in outlier detection for normal symmetric data.

Medcouple (MC) a measure of skewness, that is robust to outliers was introduced by Brys, Hubert, and Struyf [6]. Traditional boxplot combined with medcouple for analyzing skewed distributions is used by Hubert and Vandervieren [7] to make an adjusted boxplot. Fences in an Adjusted boxplot is fixed by considering semiinterquartile range. The medcouple (MC), incorporated in the Adjusted boxplot to accommodate skewness gives fences of the form [Q1-h_{l*}(MC)*IQR, Q3+h_{u*}(MC)*IQR], where h_{l*}(MC) and h_{u*}(MC) are fence "constants" adjusted for skewness. Y. H. Dovoedo [5] proposed modification for the adjusted boxplot in the multiples of semiinterquartile range as a function of medcouple.

Suitability of boxplots for a data, has to be decided based on its sensitivity and specificity to the presence of an outlier. Singh A, Masuku M [8] describes that the proportion of actual positives which are correctly detected as positives is called sensitivity while the proportion of negatives which are correctly identified is termed specificity. Sensitivity and specificity are usually expressed in percentage. In this paper,

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

the boxplots are assessed for its capability in outlier detection based on its 'sensitivity' and 'specificity'.

II. OBJECTIVES AND METHODOLOGY

Inaccurate predictions with the models for paste system properties is reported by Tanikella, P., & Olek, J. [1] in the research on type C and type F fly ash from different sources. With a focus on evaluating the reason for reported model inaccuracy, the extracted data on the characteristics of fly ash used for modelling is checked for outliers with the help of boxplots. Objective of this paper is to compare the capability of 3 boxplots namely Tukey's boxplot, adjusted boxplot and modified adjusted boxplot to identify the outliers in the extracted data. Coding for boxplots is done using R programming. Histogram and densityplots uncovered the skewness in data which are handled with proper transformations on the data. Sensitivity and Specificity calculations suggest the best among the three box plots for detection and treatment of outliers in the material characteristics data.

2.1 DATA SELECTION

Binary paste system properties with fly ash and cement is described in the report *Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2017/11* [1] available for download. Tanikella, P., & Olek, J.[1] aimed to predict the properties of binary (cement+ fly ash) and ternary (cement + 2 fly ashes) paste systems, with 20% replacement of cement. Relevant characteristics of thirteen class C and seven class F fly ash from different powerplant sources as well as the properties of pastes namely initial setting time, heat of hydration, rate of strength gain, non-evaporable water content and calcium hydroxide content are available in the published report [1]. Experimental research data for the 20 sources considered in the present study include 206 responses on 12 dependent variables and 437 observations for the 23 predictor variables. Characteristics extracted from the source [1] are taken as the independent variables with notations X_1 to X_{23} and the properties studied are the dependent variables namely Initial setting time (Initial_Set), Peak Peak Heat of Hydration (HOH), Calcium Hydroxide content and Non evaporable water content at 1, 3,7, and 28 days, (represented as CH_1, CH_3, CH_7, CH_28, NEwater_1, NEwater_3, NEwater_7, NEwater_28), and Stength Activity Index at 7 and 28 days. Dependent variables are denoted as Y_1 to Y_{12} .

2.2 . DATA ANALYSIS

Exploratory data analysis tool namely boxplots are used for detecting presence of outliers in the observational data. Using a Tukey's traditional box plot, outliers are determined based on inner and outer fencings given at 1.5 and 3 times respectively of the Inter Quartile Range (IQR). Interquartile range is the difference between the 1st and 3rd quartiles for the predictors. In adjusted and modified adjusted box plots, outliers are determined with respect to the fencing given based on medcouple, semiinterquartile range and inter quartile range. Fences of the boxplot by Hubert and Vandervieren [7] are given by equations (1) and (2),

$[Q1-1.5e^{-4*MC}\ IQR,\ Q3+1.5e^{3*MC}\ IQR]$ for $MC >= 0$ …… (1)

$[Q1-1.5\ e^{-3*MC}\ IQR, Q3+1.5\ e^{4*MC}\ IQR]$ for $MC < 0$……(2)

Y. H. Dovoedo [5] proposed some modifications in the adjusted boxplot by Hubert and Vandervieren, and advocates outliers as the observations falling outside the fences given by equation (3)

$[Q_2-4e^{-2MC}SIQR_L,\ Q_2+4e^{2MC}SIQR_U]$……..(3)

where $SIQR_L = Q2-Q1$ and $SIQR_U = Q3-Q2$.

Y. H. Dovoedo [5] proposed some modifications in the adjusted boxplot by Hubert and Vandervieren, and advocates outliers as the observations falling outside the fences given by equation (3)

$[Q_2-4e^{-2MC}SIQR_L,\ Q_2+4e^{2MC}SIQR_U]$……..(3)

where $SIQR_L = Q2-Q1$ and $SIQR_U = Q3-Q2$.

2.3. DATA TRANSFORMATION AND SKEWNESS REDUCTION

Logarithmic, square root, and exponential transformations on the actual data imparted normality to those distributions which presented skewness.

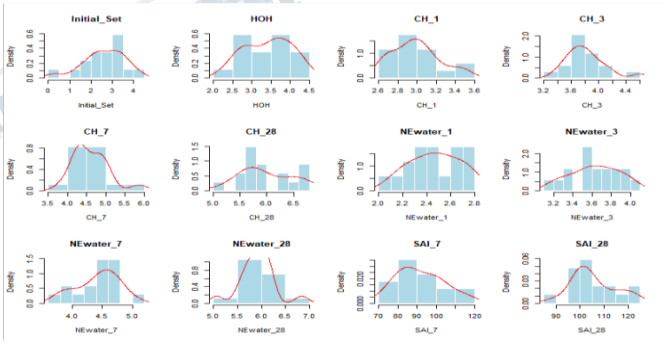

Fig.1.. Histogram & Density Plot for 12 Parameters

While detecting the outliers in the parameters, Initial_Set, HOH, NEwater_1, and NEwater_3, being normally or nearly normally distributed were not transformed for skewness, whereas other parameters were given transformations according to the skewness displayed as in **Fig.1.** The transformation adopted for each variable to detect outliers is provided in **Table 1.**

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

Table 1. Transformations for Skewness Reduction

Variable	Adopted Transformation			
	Actual data	Logarithmic	Square root	Exponential
Independent Variables	X_1, X_2, X_3, X_{11}, X_{12}, X_{13}, X_{23}	X_7, X_8, X_9, X_{10}, X_{14}, X_{15}, X_{16}, X_{18}, X_{20}, X_{21}, X_{22}	X_4, X_6, X_{17}, X_{19}	
Dependent variables	Y_1, Y_2, Y_7, Y_8	Y_3, Y_4, Y_5, Y_6, Y_{11}, Y_{12}	Y_{10}	Y_9

Initial_Set, HOH, NEwater_1, NEwater_3 had lowest skewness for original data than transformed data and hence detected outliers with original variables were chosen. NEwater_28 gets near normality in the distribution for square root transformation, while CH_1, CH_3, CH_7, CH_28, SAI_7 and SAI_28 gains near normality in distribution for $\log_{10}$ transformation and NEwater_7 gets normality through exponential transformation. Outliers were detected for the variables accordingly with the respective transformations

III. RESULTS AND DISCUSSION

Detection of outliers for the actual and transformed data by Tukey's traditional boxplot , adjusted boxplot and modified adjusted boxplots are analysed for the values, number, and percentage.

3.1 OUTLIERS WITH ACTUAL DATA

Details of outliers detected using the 3 boxplots for the independent variables before data transformation are tabulated in **Table 2**. **Fig.2** depicts the number of outliers for independent variables in non transformed data.

Table 2 Outliers in the dataset identified before transformation

V.No:	**Outliers**		
	Traditional box plot	**Adjusted box plot**	**Modified Adjusted box plot**
X_2	No outlier	22075	22075
X_3	0.25	1.27,1.67,0.25	1.78,1.27,1.67,0.25
X_6	37.72	No outlier	0, 0.07, 0, 0.31, 0, 0, 0.12, 0
X_8	29.45, 26.02	16.91, 29.45	16.91, 29.45
X_9	No outlier	4.96, 4.62	6.25, 5.91, 6.03, 6.19, 4.96, 4.62
X_{10}	No outlier	56.41	56.41, 58, 57.32
X_{11}	No outlier	No outlier	26.98, 26.23, 27.66, 26.97
X_{12}	No outlier	5.9, 5.83, 5.78	5.9, 5.83, 5.64, 5.51, 5.78
X_{14}	3.7	3.7	3.7
X_{15}	No outlier	0.34, 0.36	0.47, 0.38, 0.34, 0.49, 0.36, 0.37
X_{16}	3.92	3.92, 3.06	3.92, 3.06
X_{17}	2.06	0.07, 0.05, 0.22	0.07, 0.05, 0.22
X_{18}	No outlier	0.35, 0.25, 0.17, 0.35	0.43, 0.35, 0.25, 0.44, 0.17, 0.44, 0.43, 0.35
X_{19}	200, 169, 161	4,2	4, 2, 16, 16
X_{20}	0.32, 0.18	0.32, 0.18	0.32, 0.18
X_{21}	8.2	1.6, 1.6, 1.4	1.6, 1.6, 1.4
X_{22}	2.881, 2.485	2.881, 2.485, 2.13	2.881, 2.485, 2.13

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

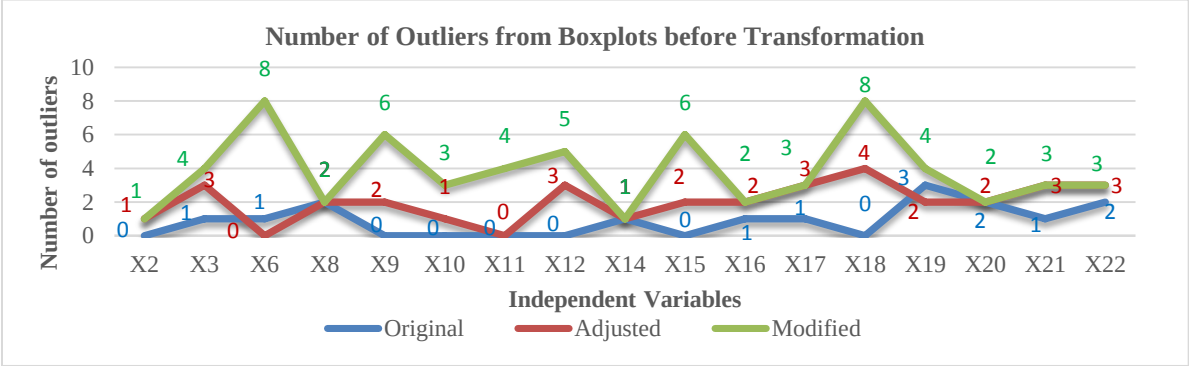

Fig.2 Number of Outliers for non transformed variables

Outlier detection using the 3 boxplots on a randomly chosen dependent variable CH_28 is given in **Fig.3**. Boxplots for all the remaining dependent and independent variables are not presented here. Selection of type of boxplot is based on the analysis of anomaly detection for independent variables. Same outlier values occurred only in the case of X_{14} and X_{20} whereas same number of outliers were detected for X_8, X_{14} & X_{20} by the three boxplots. Traditional boxplots showed outliers only in 10 variables while Adjusted boxplots indicated outliers in 15 variables and Modified Adjusted boxplot revealed outliers in 22 variables. The traditional box plot gives a maximum of 3 outliers for variable X_{19} while the maximum number (4) of observations from Adjusted boxplots occurred for X_{18} which gave the maximum outliers in Modified Adjusted boxplots as well. X_6 too marked the maximum number with Modified Adjusted boxplots. Many common outliers occur between adjusted and modified adjusted boxplots whereas few entirely different outliers occur in the Tukey's boxplot.

Outliers for 8 variables exposed by Adjusted boxplots and Modified Adjusted boxplots had a complete match in number and values of the outliers detected. Six variables had all the detected outliers from Tukey's boxplot occurring under the other two boxplots as well, though modified adjusted boxplot gives more number of outliers for these variables . Those common observations were considered as definite outliers. Three hundred and sixty five observations are not detected as outliers from all the 3 boxplots and hence those observations are considered as definite inliers.

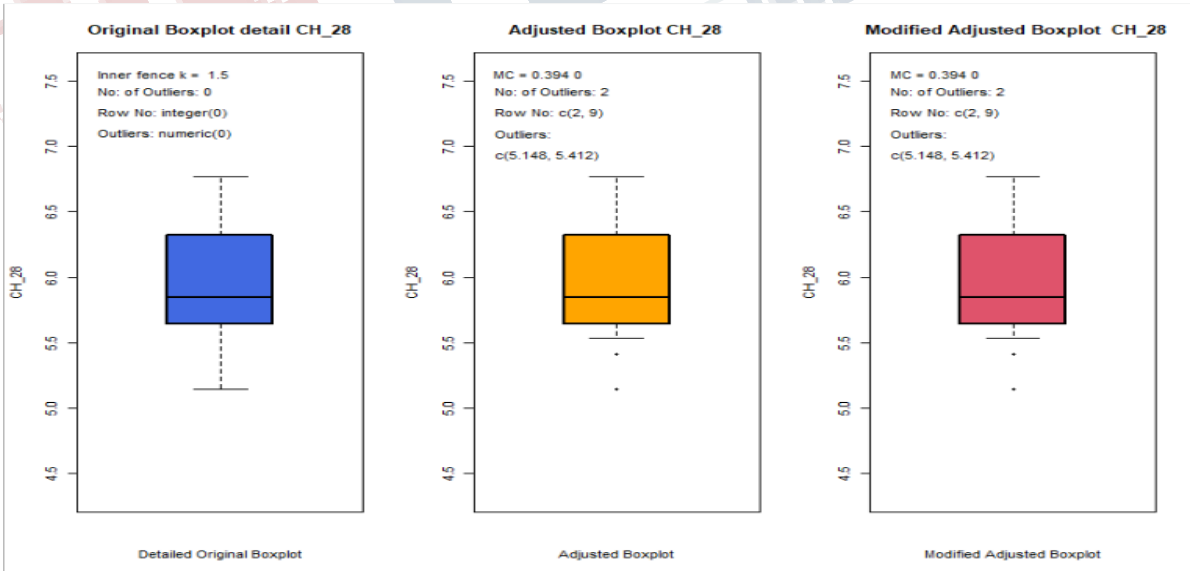

Fig.3 Original, Adjusted, and Modified Adjusted Boxplots for parameter CH_28

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

3.2 DATA TRANSFORMATION AND OUTLIER DETECTION

Logarithmic, square root and exponential transformations on the data as given in **Table 2,** could improve the normality of the density plots**.** Detected outlier and Inlier particulars after data transformation of independent variables are presented in **Fig.4 & Fig.5**.

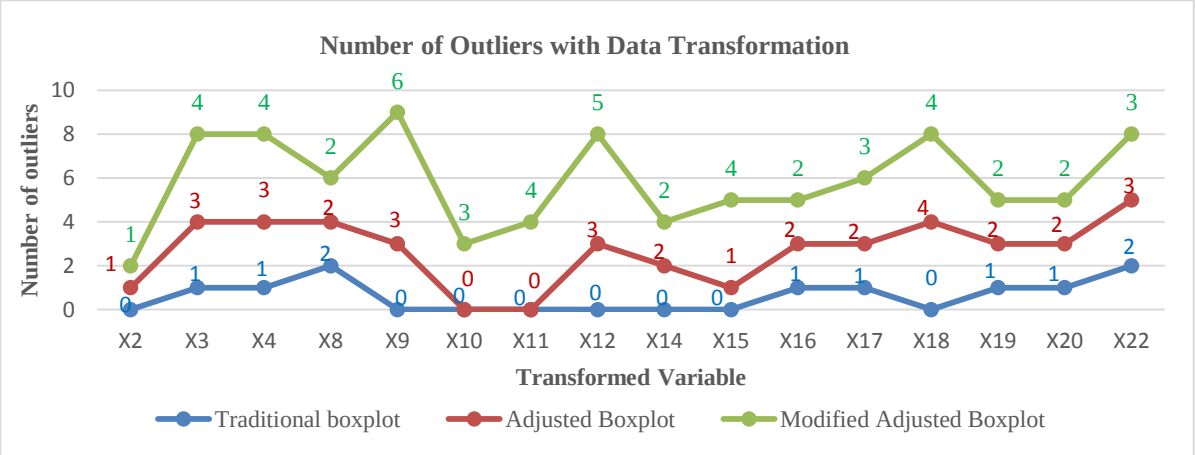

Fig.5 Number of Outliers for transformed variables

Seventeen variables contained outliers before transformation, but only sixteen variables contained outliers for the transformed data. X_4 which was free of outliers before transformation displayed outliers after transformation whereas X_6 and X_{21} which had outliers before transformation became outlier free after the transformation. 83.5% observations are detected as definite inliers before transformation while 87.64 % observations fall in as inliers after the transformations. The number of definite outliers turns down from 8 to seven after the transformation. The percentage of outliers and non outliers are presented in Fig.6. The percentage of outliers detected using Tukey's Boxplot showed 32.7% difference in the result with transformation. Modified Adjusted Boxplot detected 11.67% data as outliers after transformation which is 21.67% shift from outliers before transformation.

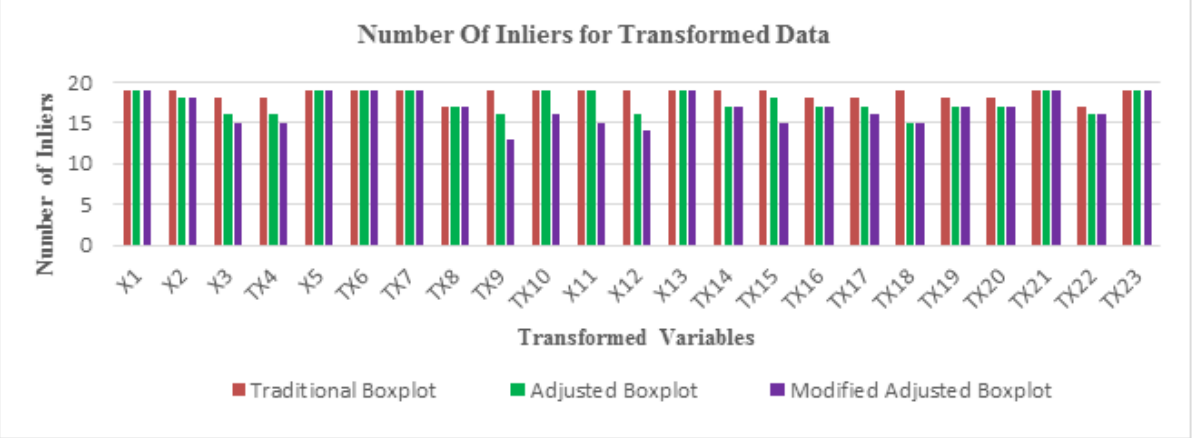

Fig.5 Number of Inliers for transformed variables

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

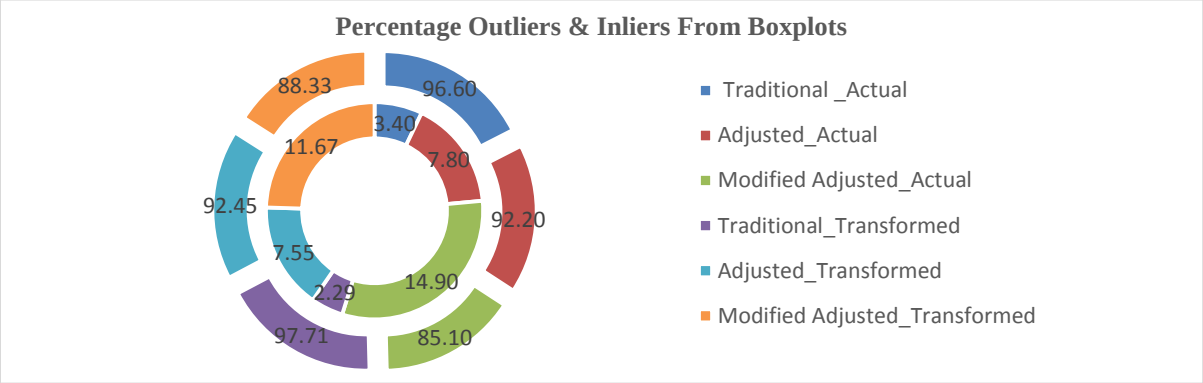

Fig.6 Percentage of detected outliers and Inliers in the dataset

Adjusted Boxplot on the other hand showed mere 3.19% change in the percentage outliers with transformation. Similarly the Inliers detected before transformation varied from that after transformations by 0.27% ,1.15%, and 3.79 % for the Adjusted, Tukey's traditional, and Modified Adjusted Boxplots respectively.

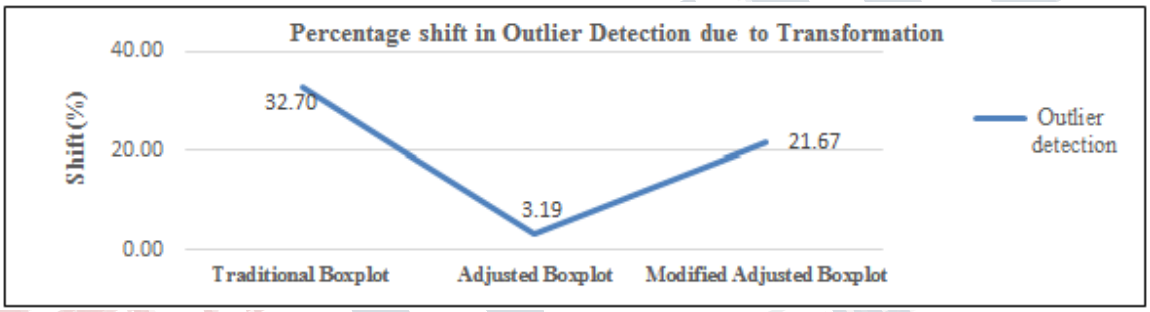

Fig.7a. Shift in Outlier detection due to data transformation

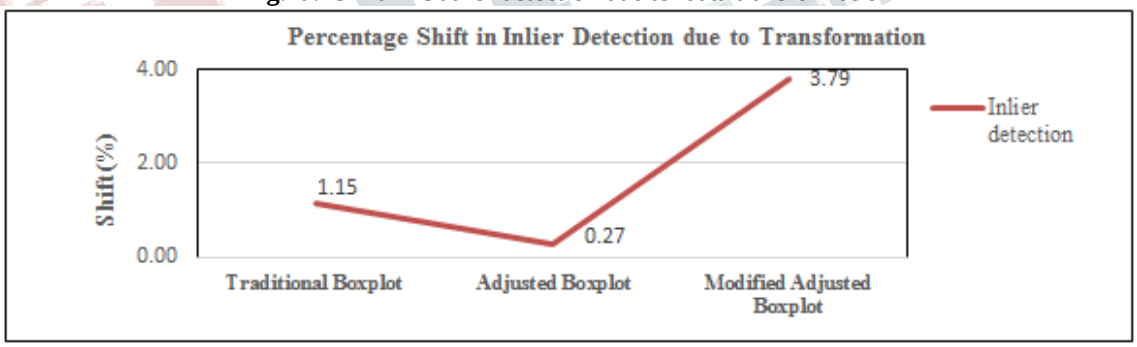

Fig.7b. Shift in Inlier detection due to data transformation

Fig.7a & Fig.7b depicts the percentage shift in outlier and inlier detection due to data transformation. Among the three boxplots studied, Adjusted Boxplots are found to have minimum change due to data transformation in the percentage of outlier detected.

3.3 EVALUATION OF THE BOXPLOTS FOR OUTLIER DETECTION

A detection correctly obtained is termed as 'true positive' while a precise non detection is termed as 'true negative' for any test. False positives and false negatives in a test upsets the dependability of the method and hence efficacy of an outlier detection method should be evaluated based on the rate of detection and non detection of the outliers. Rate of detection (true positive rate) of a test is the ratio of number of true positives obtained from the test to the real number of

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

positives whereas false positive rate is the ratio of number of false positives obtained to the real number of negatives.

$$\textbf{True Positive rate}(\textbf{TPR})\% = \frac{\textbf{True Positives}}{(\textbf{True Positives}+\textbf{False Negatives})} * \mathbf{100} \quad \text{.....(4)}$$

$$\textbf{False Positive rate}\ (\textbf{FPR})\% = \frac{\textbf{False Positives}}{(\textbf{False Positives}+\textbf{True Negatives})} * \mathbf{100} \quad \text{.....(5)}$$

$$\textbf{True Negative Rate}(\textbf{TNR})\% = \frac{\textbf{True Negatives}}{(\textbf{False Positives}+\textbf{True Negatives})} * \mathbf{100} \quad \text{.....(6)}$$

3.3.1 SENSISTIVITY AND SPECIFICITY OF BOXPLOTS

'Sensitivity' in outlier detection is the percentage of outliers correctly identified as outliers whereas percentage of non-outliers (inliers) correctly detected as non-outliers is termed as the 'Specificity'. Sensitivity is also termed True positive rate (TPR) while specificity denotes the true negative rate(TNR).

Table 4. Comparison of Outlier detection rates

Particulars	Actual Data			Transformed Data		
	Traditional Boxplot	Adjusted Boxplot	Modified Adjusted Boxplot	Traditional Boxplot	Adjusted Boxplot	Modified Adjusted Boxplot
Definite Outliers (True Positives)	8	8	8	7	7	7
Definite Inliers (True negatives)	365	365	365	383	383	383
False Positives	7	26	57	3	26	44
False Negatives	57	38	7	44	21	3
TNR (Specificity)*	98.1	93.4	86.5	99.2	93.6	89.7
TPR (Sensitivity) *	12.3	17.4	53.3	13.7	25.0	70.0
FPR (1-Specificity)*	1.9	6.6	13.5	0.8	6.4	10.3
Mean TPR	32.0	32.0	32.0	32.0	32.0	32.0
Mean TNR	93.4	93.4	93.4	93.4	93.4	93.4
Mean FPR	6.6	6.6	6.6	6.6	6.6	6.6
Standard Deviation in Sensitivity	13.9	10.3	15.1	12.9	4.9	26.9
Standard Deviation in Specificity	3.3	0.0	4.9	4.1	0.2	2.6
* TPR= True Positive Rate, FPR= False Positive Rate, TNR= True Negative Rate						

False Positive Rate (FPR) is calculated by subtracting Specificity(%) from 100(%).In this paper, the definite outliers are considered as the 'true positives' and definite inliers are considered as the 'true negatives' which remains constant in all the 3 boxplots. Sensitivity and specificity are calculated on constant true positives and true negatives but varying false positives and false negatives as shown in **Table 4**. The percentage of outliers and non-outliers correctly detected are referred to as 'Sensitivity' and 'Specificity' respectively. The efficiency of the boxplots under study are represented by True Positive Rates and False Positive rates as shown in **Fig.8**

TPR as well as FPR is found to be maximum in the case of modified adjusted boxplot while traditional boxplot had minimum TPR as well as FPR. On transformation of data, the TPR values got increased while FPR values decreased for the three boxplots. The mean Sensitivity and Specificity for the three boxplots before transformation were found to be 27.7% and 92.7% while it changed to 36.2% and 94.2% respectively after transformations. Thus the sensitivity as well as specificity of the test methods got improved with transformation. The transformed data boxplots assessed for minimum deviation from the mean showed the deviations as shown in **Fig.9**.

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 6, Issue 10, October 2021

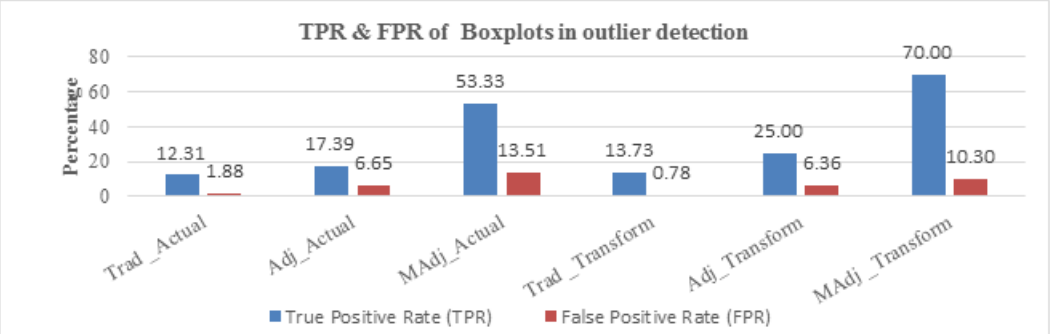

Fig.8 Efficiency of Boxplots in Outlier Detection

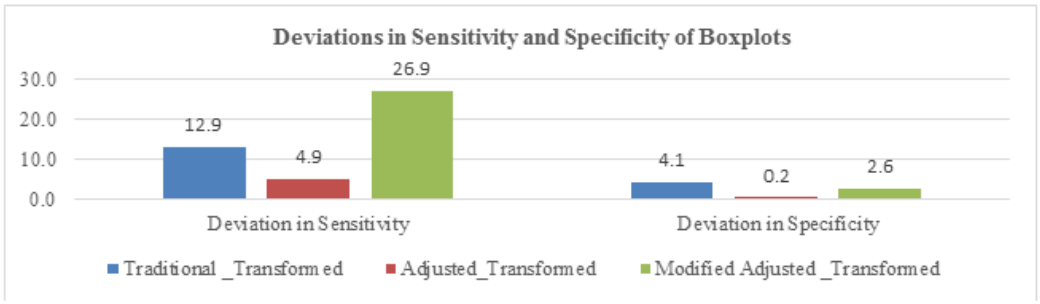

Fig 9. Effect of transformation on Sensitivity and Specificity of Boxplots

IV. CONCLUSIONS

Minimum deviation in Sensitivity and Specificity is located in Adjusted Boxplots compared to the other two Boxplots. All the models with the data winsorised for outliers based on adjusted boxplots could give '*Significance* F' less than 0.05, which indicated the improved models' efficiency in prediction.

Examining the data used for predictions in the source paper as well as the transformed data, using Tukey's, Adjusted and Modified adjusted boxplots bring out the following conclusions .

- Tukey's boxplot detected 3.4% of the observations as outliers while adjusted boxplot detected 7.8% and modified adjusted boxplot detected 14.9% of the observations as outliers. The identified definite inliers from the actual data boxplots, figures to mere 83.5% of the observations. Thus based on boxplots analysis of the data without transformation, 16.5% of the observations considered for modelling in the source data are doubtful to perform as inliers.
- With reduced skewness in the distributions with transformed data, 87.6% of the observations could definitely be used for modelling of the test results. 12.4 % of the observations used in the source data, are still doubtful to perform as inliers. Incorporation of those doubtful inliers (12.4% observations) in the modelling procedures of the source paper could have caused the reported prediction inaccuracy.
- Minimum difference in outlier and inlier detection (3.19% & 0.27% respectively), between the actual and transformed data occured for Adjusted boxplot compared to Traditional and Modified adjusted Boxplots, which indicates that Adjusted Boxplots co produce better results in outlier detection for the data, even without transformations. Traditional Boxplots and Modified Adjusted Boxplots should adjust the skewness by way of transformations in the data.
- Modified Adjusted Boxplot had the highest (70%) sensitivity among the three boxplots on the transformed data, which indicates that 70% of outliers detected by Modified Adjusted boxplot on transformed data is correctly identified. Maximum specificity occured for Tukey's Traditional Boxplot on the transformed and actual data (99.2% & 98.1% respectively) which indicates that Tukey's Boxplot is more dependable in inlier detection for the selected data than the other two boxplots.
- The deviation in sensitivity between actual and transformed data are found to be minimum in the case of Adjusted Boxplots compared to traditional and Modified Adjusted Boxplots. Likewise specificity also showed minimum deviation between actual and transformed data in the case of Adjusted Boxplots. Hence Adjusted

boxplots can perform anomaly detections, regardless of transformation, resulting in no substantial shift in the detections.

- Adjusted Boxplots could mend the data resulting in performance models with accurate predictions for the wisorised data.

REFERENCES

[1] Tanikella, P., & Olek, J. (2017). "Updating physical and chemical characteristics of fly ash for use in concrete" (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2017/11). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284315213)

[2] Hawkins D.M., 1980, "Identification of Outliers", Chapman & Hall, ISBN 978-94-015-3996-8, ISBN 978-94-015-3994-4(ebook), DOI 10.1007/978-94-015-3994-4

[3] Insia Hussain, "Outlier Detection using Graphical and Nongraphical Functional Methods in Hydrology", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019

[4] Mark Kasunic James, McCurley, Dennis Goldenson, David Zubrow December 2011, "An Investigation of Techniques for Detecting Data Anomalies in Earned Value Management Data", Technical Report Cmu/Sei-2011-Tr-027 Esc-Tr-2011-027

[5] Yinaze Herve Dovoedo(2011), "Contributions To Outlier Detection Methods: Some Theory And Applications", Tuscaloosa, Alabama

[6] Brys G., Hubert M., Struyf A. (2003) "A Comparison of Some New Measures of Skewness". In: Dutter R., Filzmoser P., Gather U., Rousseeuw P.J. (eds) Developments in Robust Statistics. Physica, Heidelberg. https://doi.org/10.1007/978-3-642-57338-5_8

[7] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Computational Statistics & Data Analysis, Volume 52, Issue 12, 2008,Pages 5186-5201,ISSN 0167-9473, https://doi.org/10.1016/j.csda.2007.11.008. http://www.sciencedirect.com/science/article/pii/S0167947307004434)

[8] Singh A, Masuku M. "Understanding and applications of test characteristics and basics inferential statistics in hypothesis testing.", European Journal of Applied Sciences (ISSN 2079-2077) 4 (2): 90-97, 2012, DOI: 10.5829/idosi.ejas.2012.4.2.65132