Construction and Applications of Stairboxplot for Exploratory Data Analysis

Babangida Ibrahim Babura¹, Mohd Bakri Adam², Muhammad Sani³, Usman Waziri⁴, Felix Yakubu Eguda¹

¹Department of Mathematics, Federal University Dutse, Jigawa State, Nigeria, ²Department of Mathematics, University Putra Malaysia, 43300 Serdang Selangor, ³Department of Mathematical Sciences, Federal University Dutin-Ma Katsina State, Nigeria ⁴Department of Mathematical Sciences, Bauchi State University, Bauchi State, Nigeria.

E-mail: bibabura@gmail.com

Abstract. The classical construction of boxplot requires estimates of five robust statistics of interest namely; the first quartile, the median, the third quartile, the lower fence and the upper fence. The fence estimate is usually dependent on the three quartiles and is utilized to visually identify outliers in a batch of univariate dataset. Some scholars are critical of the limitation of boxplox to display individual data points, density of observations and distributional shape in multiple batch comparison among to mention. In this paper display enhancement to address the limitations of classical boxplot is proposed according to a new construction method called stairboxplot. The construction begins with display of four stairs of consecutive boxes according to quadbins to replace box and whiskers in the classical boxplot construction and an inscription of individual observations using scale adjusted outlyingness estimate of each data point. The advantage of stairboxplot as a data display toolkit was explored using simulation and real life dataset.

1. Introduction
Visualization techniques is regarded as the most popular exploratory data analysis (EDA) tools for data analytic in the new field of Data Science. Stairboxplot was recently introduced as a data visualization tool that enhance and extend the features of classical Boxplot method in EDA. Boxplot’s main features is define according to a resistant rule for identifying outliers in a univariate dataset. The first quartile \(Q_1\) and third quartiles \(Q_3\) are approximate estimates of the lower and upper fourth of a dataset in which \(Q_2\) is the median of the dataset. The resistance rule label observation as ”outside” if the observation fall below \(Q_1 - 1.5IQR\), or above \(Q_3 + 1.5IQR\), where \(IQR = Q_1 - Q_3\), the Tukey’s classical boxplot rule. Further studies on the classical boxplot construction methods redefine the resistance rule where extensive done done.

Literature on boxplot construction variants where proposed to incorporate additional display information and requirement. For example the notch boxplot in which the incorporated notch representing confidence interval around the median estimate, the curving (circular) boxplot for circular data and the K-boxplot specifically developed for mixture data such that a multi-modality display enhancement was incorporated within the k-boxplot construction method.
A recent development that provide easy accessibility of computational power undermine the initial construction philosophy of classical boxplot of simplify (with paper and pencil) [2] and consequently bring in new developments. Many literature advances consider transformation of the entire boxplot display characters to capture some distributional attributes of interest on a dataset. Such advances includes; a vase plot [16], a violin plot [17], a beanplot [18], the raindrop plot [19] and other plots [20, 21, 22] are boxplot inspired EDA facilities. Unfortunately, not all visualization enhancement and transformation of boxplot are consistent with the resistance rule a feature regarded as the most important component of a boxplot method. Other EDA visualization tools which are inspired from boxplot method can be found in the literature [20, 21, 22].

Stairboxplot is proposed with computationally advanced features than classical boxplot but specifically maintain the robust display summaries with additional features of histogram and dotplot [1]. The plot consider a display of each individual observation in a dataset according to outlyingness estimates within the respective quadbin levels [1]. A connected stairs of figures (quadbins) were drawn as a boxes whose width are proportional to the number of observations and height equals to a quarter of the data range. In this article we revisit and redefine some terms in the construction of stairboxplot to allow its implementation to a real life dataset. This article describe in detail statistical estimators of range, quadbin, scale outlyingness and a description of how to incorporate them in the construction of the proposed enhancement to stairboxplot. We give an illustration of the implementation of stairboxplot for real life data such as monthly maximum rainfall data, 46 years monthly maximum precipitation data and annual maximum river flow discharge data.

2. MATERIALS AND METHODS

2.1. Definition of Terms

Definition 2.1. The range of an ordered univariate sample \( X = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\} \) of size \( n \), is given by \( r = x_{(n)} - x_{(1)} \) where \( x_{(i)} \) and \( x_{(n)} \) are the minimum and maximum observations in \( X \) respectively.

Definition 2.2. Define

\[
\begin{align*}
q_1 &= \{x \in X : x \leq r_1\}, \\
q_2 &= \{x \in X : r_1 < x \leq r_2\}, \\
q_3 &= \{x \in X : r_2 < x \leq r_3\}, \\
q_4 &= \{x \in X : x > r_3\},
\end{align*}
\]

where \( r_1 = x_{(1)} + \frac{1}{4} r, \quad r_2 = x_{(1)} + \frac{1}{2} r, \quad r_3 = x_{(1)} + \frac{3}{4} r \), then \( q_i \) for \( i = 1, 2, 3, 4 \), are called the Quadbin level \( i \) of \( X \) for \( i = 1, 2, 3, 4 \) respectively.

From Definition 2.2 we can simply observe that Quadbin are synonymous to histogram bins only that we restrict the number of bins to exactly four.

Definition 2.3. The outlyingness of a data point \( x_{(i)} \in X \) is a measure of how far an observation \( x_{(i)} \) lies from an estimate of a central point of the data, usually standardized by means of a robust scale (e.g. the median).

According to [23], in Definition 2.3 it does not matter if a data point is smaller or larger than the median. However, when the distribution is skewed, a scale on each side of the median is proposed [23] such that adjusted outlyingness \( AO \) is mathematically given by

\[
AO_i = \begin{cases} 
\frac{Q_3 - x_{(i)}}{w_u - w_l} & \text{if } x_{(i)} < Q_2, \\
\frac{x_{(i)} - Q_2}{w_u - w_l} & \text{if } x_{(i)} > Q_2,
\end{cases}
\]

where \( w_l \) and \( w_u \) are the lower and upper whisker of the adjusted boxplot of [9] applied to the dataset \( X \).
Figure 1. A Graphical description of measurements, $c_l, c_u, d_l$ and $d_u$ required to estimate outlyingness from a boxplot.

The analogous definition to $AO$ is given as follows

**Definition 2.4.** The *stairboxplot outlyingness* $SO_i$ of $x_i \in X$ is distance between an observation $x_i$ and median ($Q_2$) of the dataset $X$ scaled with the distance between corresponding fence cutoff value $m_j$, ($j = 1, 2$) to the median $Q_2$ and is given by

$$SO_i = \begin{cases} Q_2 - x_i & \text{if } x_i < Q_2, \\ \frac{Q_2 - m_2}{x_i - Q_2} & \text{if } x_i > Q_2. \end{cases}$$  \hspace{1cm} (3)$$

**Definition 2.5.** The *scale stairboxplot outlyingness* is given by

$$SO_\alpha = \alpha SO_i \hspace{1cm} (0 < \alpha \leq 1)$$

The main reason we have Definition 2.5 is to have a control on width of $SO_i$ as in some instances of data visualization especially skewed data the line representing $SO_i$ might overlap during batch comparison using the proposed stairboxplot.

Figure 1 illustrate with clarity how to estimate $SO_i$ for data point $x_{(i)}$. An arbitrary observation $x_l$ is chosen and lies below the lower fence as outlier whose value is less than median, it’s stairboxplot outlyingness will be estimated as $SO = \frac{Q_2 - x_l}{m_1 - Q_2}$. Also for an arbitrary point $x_u$ from the sample $X$ whose value is more than the median and below the upper fence position, it’s stairboxplot outlyingness is given by $SO_u = \frac{x_u - Q_2}{Q_2 - m_2}$. So, in general if the value $c_i > d_i$ for an arbitrary sample point $x_i$, then $SO_i$ will assume a value greater than 1 and thus such sample point becomes a suspected outlier. Alternatively, when $c_i \leq d_i$ for an sample point $x_i$, such point will assume the value of $SO_i \leq 0$ and is marked as regular sample poin (observation) by $SO$.

### 3. RESULT AND DISCUSSION

#### 3.1. Implementation of Stairboxplot Construction

The construction of stairboxplot is at first instance synonymous to histogram which is by first dividing the data according to quadbins. The histogram bins are usually side (left,right,up
or down) align, while stairboxplot quadbins are centre align with either horizontal or vertical orientation. The vertical orientation is our default choice for implementation in this article. That is an ascending joint boxes that are drawn according the quadbins levels order of equal hight of $\frac{1}{4}r$ and width proportional to number of observations will be drawn. Each individual point $x_i \in X$ is represented as a horizontal lines and plotted with width proportional to the value its corresponding outlyingness $SO_\alpha$ and we set $\alpha = 1/4$ for our implementation. In this case a regular observation will be represented with a full line while a potential outlier will we represented with a dash line.

However, the quartiles $Q_1, Q_2$ and $Q_3$ of $X$ will be represented with avertical thicker full line of length equals to the width of the quadbin that the quartile falls within. To enhance the visual analysis, we plot each observation less than the median with blue colour and more than the median with black colour while outliers with red colours. Furthermore, the median is plotted with blue line colour when the data is positive skewed and black line colour when the data is negative skewed.

3.2. Display and Application of Stairboxplot

![Stairboxplot Diagram](image)

**Figure 2.** To the left is regular boxplot as compared with the right stairboxplot display of outliers

Figure 2 illustrate the visual comparison of the regular boxplot method to the left against the proposed stairboxplot method. The display in Figure 2 shows that the centre, spread and outliers of the two plots of a dataset. We can observe the striking difference among the two plot methods in which the stairboxplot displays each individual observation relative to its
outlyingness $SO_\alpha$. So also the identification of skewness of the data is better illustrated with the stairboxplot than the classical boxplot in which no matter how small the skewness is can be seen in the stairboxplot.

Figure 3. Comparison of regular boxplot with stairboxplot display shape of distribution

Figure 4 shows further advantage of stairboxplot according to the display of shape of a distribution. We have two batch of datasets generated from different distributions and displayed by the two plots. The first batch (left) is generated from the normal distribution while the second batch (right) is generated from uniform distribution. The display by the classical boxplot indicate only a difference in data spread of that is middle boxes but no much information about the distributional difference. But as observed according to stairboxplot method that the distribution of the two batches of the data are different. The uniform distributed data shows approximately symmetric boxes at each of quadbins level while for the Gaussian sample, it indicates a bigger but symmetrical boxes for second and third quadbin levels compare with the smaller symmetrical boxes at the first and fourth quadbins levels.

3.3. Exploratory Analysis of Real life Dataset

Among the interesting functionality of a boxplot is observing trend in batches of datasets. Stairboxplot had maintain this property as in Figures 4, 5 and 6 with additional informations to the conventional boxplot display as describe below.

3.3.1. The Rainfall Intensity Dataset

Figures 4 and 5 are stairboxplots visualization of records of 31 years monthly maximum rainfall data from Petaling Jaya record station, source from the
Figure 4. Trend in monthly maximum rainfall data for 31 years (1974 - 1988) at Patalin Jaya, Malaysia.

UKM earth observatory center. Each batch is an annual monthly maxima collections. The most immediate information visualise in this trend is that, the rainfall records exhibit an inconsistent trend throughout the period 31 year. Both centre and spread deviates as compared according to the displayed sequence. The median records are mostly below 50mm for first 8 years (1974-1981) while above 50mm there after (1982 - 2003). The highest record is observed in 2001 a period which is associated with devastating flood issues in Malaysia.

One of the popular assumption that characterise a display of similar set of data with the regular boxplot is, the data is Gaussian [24]. It is quite interesting to observe that the stairboxplot give more detail to avoid such mis-information. For example, the trends in 1977, 1978, 1979, 1980, 1995, 1997, and 2001 demonstrate other distributional pattern as against that of normal distribution. Consider the skewed data in 1999 which shows symmetric pattern within the inter quartile range, one might consider such to follow Gaussian distribution when visualized with the regular boxplot method. But the stairboxplot gives more detail about the shape of the distribution that may fit and can suggest it to be fitted with Gumbel type GEV distribution a visual character demonstrated by [1].

So many batches of the dataset e.g. 1977, 1980, 1995, 1996, 1999 indicate shape of an extreme data and can be associated with family of GEV distribution [11]. Another information that can go unnoticed with regular boxplot display is multi-modality in some batches such as 1977, 1979, 1983, 1992, and 2001.
3.3.2. The Maximum Precipitation Dataset: Precipitation is simply any form of water particles regardless of their state that forms up in the atmosphere and falls to the ground under some influence of gravity [25]. Meteorologies observe and study different forms of precipitation such as drizzle, rain, sleet, snow, graupel and hail.

More than 300 storms were identified by [26] technical report since the late 1800s that have produced very heavy precipitation either over sizeable areas in and near the mountains of Colorado. Among the large set of heavy precipitation events in [26] are 36 from more severely extreme storms that stand out for some selected geographical regions of the Colorado state for analysis on safety policy considerations [26]. We observe the trend of the dataset for over 46 years, i.e., from 1947 to 1993 as source from [26] using stairboxplot. The trend was specifically compared between months of the year throughout the 46 years maximum precipitation events.

Figure 5 is the trend between batches of monthly maximum precipitation events for 46 years (1947 - 1993) in Colorado (US) precipitation data record in [26]. We compare the trend in monthly batches of 46 years each rather than the annual batches due to repetitive tie values in the data which makes a batch of annual events less informative. However, the monthly maximum can give a seasonal pattern of the maximum precipitation events, unlike the annual maximum batch. The trend further explores how the monthly maximum events paired with the overall annual maximum events. As can be observed in the figure, the monthly maximum pattern of spread can be put into two categories, a narrow spread which can be observed among, January to May and October to December, then a wide spread in June to September for which the most extremal events occurred. When we compare the batches of monthly maximum events with 46
years annual maximum events we can observe also that almost all the batches in the narrow spread monthly maximum events fall below the medium annual maximum events while the wide spread maximum events has about 25% of the records above the medium annual maximum events which signifies that the severely extreme storms events happen from among the months of June to September.

3.3.3. The Annual Maximum River Flow Discharge Dataset: The 60 years annual maximum of flood discharges, in cubic meters per second, at unknown location of a river, are obtainable at [27]. Castillo [27] use the data for analysis and design of a flood protection device at that particular location using extreme value statistical analysis. Since the data is presented in a single batch with less information as the other datasets in Subsections 3.3.1 and 3.3.2 we concentrate to analyse the data based on single batch stairboxplot visualization.

Figure 7 is the stairboxplot display of the annual maximum flood discharge dataset records using modified boxplot outlyingness method. The stairboxplot has pointed the 2 observations as potential outlier with demonstration of right skewness an indication that the dataset can fit a Fréchet type GEV distribution for modelling purposes. We can observe asymmetric behaviour of the dataset with stairboxplot basically in two ways;

- the non symmetry of boxes at first and fourth quadbin levels or boxes at the second and third quadbin levels,
Figure 7. Stairboxplot display of Colorado Annual Maximum River Flow Discharge Data [26]

- the displayed data points according to outlyingness has upto the third quartile (75% data points) within the first and second range levels.

4. CONCLUSION

The proposed quadbins or quadbins level is an improvement of the concept of range level introduced earlier in the literature [1]. The concept of quadbins is synonymous with bins in histogram construction except that the number of bin to be implemented in the stairboxplot are exactly four in number. Furthermore, The histogram bins are usually side (left, right, up or down) align, while stairboxplot quadbins are centre align with either horizontal or vertical orientation. The vertical orientation is our default choice for implementation in this article. That is an ascending joint boxes that are drawn according the quadbins levels order of equal height of $\frac{1}{4}r$ and width proportional to number of observations will be drawn. Each individual point $x_i \in X$ is represented as a horizontal lines and plotted with width proportional to the value its corresponding outlyingness $SO_\alpha$ and we set $\alpha = 1/4$ for our implementation. In this case a regular observation will be represented with a full line while a potential outlier will we represented with a dash line.

References

[1] Babura B I, Adam M B, Fitrianto A and Mohammad A T 2019 Stairboxplot and visual analysis of univariate dataset p 050041 URL [http://aip.scitation.org/doi/abs/10.1063/1.5136429](http://aip.scitation.org/doi/abs/10.1063/1.5136429)

[2] Tukey J W 1977 Exploratory Data Analysis (Reading, Mass: Addison-Wesley Pub. Co)
[3] Hoaglin D, Mosteller F and Tukey J 1983 Understanding Robust and Exploratory Data Analysis Wiley series in probability and mathematical statistics: Applied probability and statistics (New York: Wiley) URL https://books.google.com.my/books?id=FRnvAAAAAIAJ
[4] Frigge M, Hoaglin D C and Iglewicz B 1989 The American Statistician 43 50–54
[5] Kimber A 1990 Applied Statistics 21–30
[6] Carling K 2000 Computational Statistics & Data Analysis 33 249–258
[7] Schwertman N C, Owens M A and Adnan R 2004 Computational Statistics and Data Analysis 47 165–174 ISSN 01679473
[8] Schwertman N C and de Silva R 2007 Computational Statistics and Data Analysis 51 3800–3810 ISSN 01679473
[9] Hubert M and Vandervieren E 2008 Computational Statistics & Data Analysis 52 5186–5201
[10] Bruffaerts C, Verardi V and Vermandele C 2014 Statistics & Probability Letters 95 110–117
[11] Adil I H and Irshad A U R 2015 Pakistan Journal of Statistics and Operation Research ISSN 22205810
[12] Babura B I, Adam M B, Fitrianto A and Rahim A S A 2017 Modified boxplot for extreme data vol 030034 p 030034 ISBN 9780735415126 URL http://aip.scitation.org/doi/abs/10.1063/1.4982872
[13] McGill R, Tukey J W and Larsen W A 1978 The American Statistician 32 12–16
[14] Abuzaid A H, Mohamed I B and Hussin A G 2012 Computational Statistics 1–12
[15] Qarmalah N M, Einbeck J and Coolen F P 2016 Statistical Papers 1–16
[16] Benjamini Y 1988 The American Statistician 42 257–262
[17] Hintze J L and Nelson R D 1998 The American Statistician 52 181–184
[18] Kampstra P et al. 2008 Journal of Statistical Software 28 1–9
[19] Barrowman N J and Myers R A 2003 The American Statistician 57 268–274
[20] Zani S, Riani M and Corbellini A 1998 Computational Statistics & Data Analysis 28 257–270
[21] Hyndman R J and Shang H L 2010 Journal of Computational and Graphical Statistics 19 29–45
[22] Adam M B, Babura B I and Gopal K 2018 MATEMATIKA 34 187–204
[23] Hubert M and Van der Veeken S 2008 Journal of Chemometrics 22 235–246
[24] Brillinger D R 2001 Time Series: Data Analysis and Theory (Philadelphia: SIAM)
[25] Potter T D and Colman B R 2003 Handbook of weather, climate, and water: atmospheric chemistry, hydrology, and societal impacts (Hoboken, N.J.: Wiley-Interscience) ISBN 978-0-471-21489-2 oCLC: 123158897
[26] Thomas B M and Nolan J D 1997 Colorado extreme storm precipitation data study URL https://ccc.atmos.colostate.edu/pdfs/Climo_97-1_Extreme_ppt.pdf
[27] Castillo E 1988 Extreme Value Theory in Engineering (London: Academic Press)