Bhartiya Krishi Anusandhan Patrika, 36(4): 334-337 DOI: 10.18805/BKAP397

Outlier Removal in Sheep Farm Datasets Using Winsorization

Ambreen Hamadani, Nazir A. Ganai, Tariq Raja, Safeer Alam, Syed Mudasir Andrabi, Ishraq Hussain, Haider Ali Ahmad

Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar-190 006, Jammu and Kashmir, India.

Received: November 2021 Accepted: January 2022

ABSTRACT

Background: Sheep farm data is often biased by extreme values which are generally introduced due to errors in manual measurement. These values interfere with the accuracy of estimations especially in state-of-the-art techniques like Machine Learning.

Methods: Therefore, winsorization technique was attempted for the removal of outliers from sheep farm data for 11 years (2011-2021) for body weights at different ages. Some outliers were deliberately introduced into the data to check the efficiency of the technique. This study was conducted during the year 2021.

Result: Our results indicate that outlier values of 15.3, 42, 44, 60, 90 for birth weight, weaning weight, 6-month, 9 month and 12-month body weight which were far from the normal range were removed using this technique. The mean and standard deviation values were altered after winsorization. Winsorization technique works well for sheep farm data to remove the bias introduced by outliers and also removes, to a large extent, the need for manual outlier removal in data.

Key words: Body weights, Data correction, Outliers, Sheep data, Winsorization.

INTRODUCTION

During any statistical analysis of animal data, it is very important to clean it so as to ensure that all observations best represent the data. This is crucial for unbiased data analysis. However, sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data. These are called outliers and often machine learning modeling and model skill in general can be improved by understanding and even removing these outlier values.

An outlier is any observation that is different from all other observations. Such observations in animal sciences datasets differ significantly from other observations *e.g.* extreme body weights (Grubbs, 1969; Maddala 1992). An outlier may be due to variability in the measurement, or it may indicate experimental error; the latter are sometimes excluded from the data set (Grubbs, 1969). Outliers cause serious problems in statistical analyses and can occur by chance in any dataset especially related to the real-world dataset. Such values often indicate error in measurement.

Winsorization is a transformation technique of statistics. It limits extreme values present within the data and therefore reducing the effect of spurious outliers. The effect of winsorization is the same as clipping in signal processing. Considering all this, winsorization technique was attempted for the removal of outliers from sheep farm data for body weights at different ages.

MATERIALS AND METHODS

Data set

For the prediction of bodyweight, data for 11 years (2011-2021) for Corriedale breed was used. This data was obtained from Mountain Research Station on Sheep and Goat, SKUAST-K. Important traits that are regularly monitored at the farm like birth, weaning, 6-month, 9-month, 12-month were taken into consideration for the analysis. No manual cleaning of the data was done before the analysis. Some outliers were also deliberately introduced into the original dataset to check the efficiency of the technique. All data was digitized and stored in MySQL database in relational database tables (Widenius *et al.*, 2002).

Winsorization

Outlier detection was done using boxplots and histograms in Python (Hunter, 2007). Winsorization was done for limiting the effects of outliers (Hunter, 2007) in the dataset using the SciPy library (Gerard-Marchant, 2007). Winsorization was done at 1st and 99th percentile which implies that the values less than the value at 1st percentile were replaced by the value at 1st percentile and values greater than the value at 99th percentile were replaced by the value at 99th percentile.

RESULTS AND DISCUSSION

A comparison of graphs obtained (Fig 1-10) before and after winsorization indicate that the technique effectively removes

*Corresponding author's E-mail: escritor005@gmail.com

the outliers within the data. All the features in the dataset had distinct outliers which radically differed from real life values. All outliers were removed after the application of winsorization technique and the bar diagrams appeared normally distributed. The summary statistics before and after winsorization are given in Table 1 and 2.

Our results indicate that winsorization is an effective technique in outlier removal for the sheep breeding dataset. This may be seen both from the figures. The descriptive statistics also indicates a change in the mean and standard deviation values of the dataset. The mean number of observations was, however, not changed which is crucial in situations where the number of data points in a study are less and the removal of any data could lead to possible to information loss. This indicates that the technique successfully removes the effect of outliers on our dataset (Hargrave and Clarine, 2021).

Winsorized means obtained in our study are different the trimmed means because trimmed mean would remove data points which causes data loss which makes the winsorization technique better suited under a number of

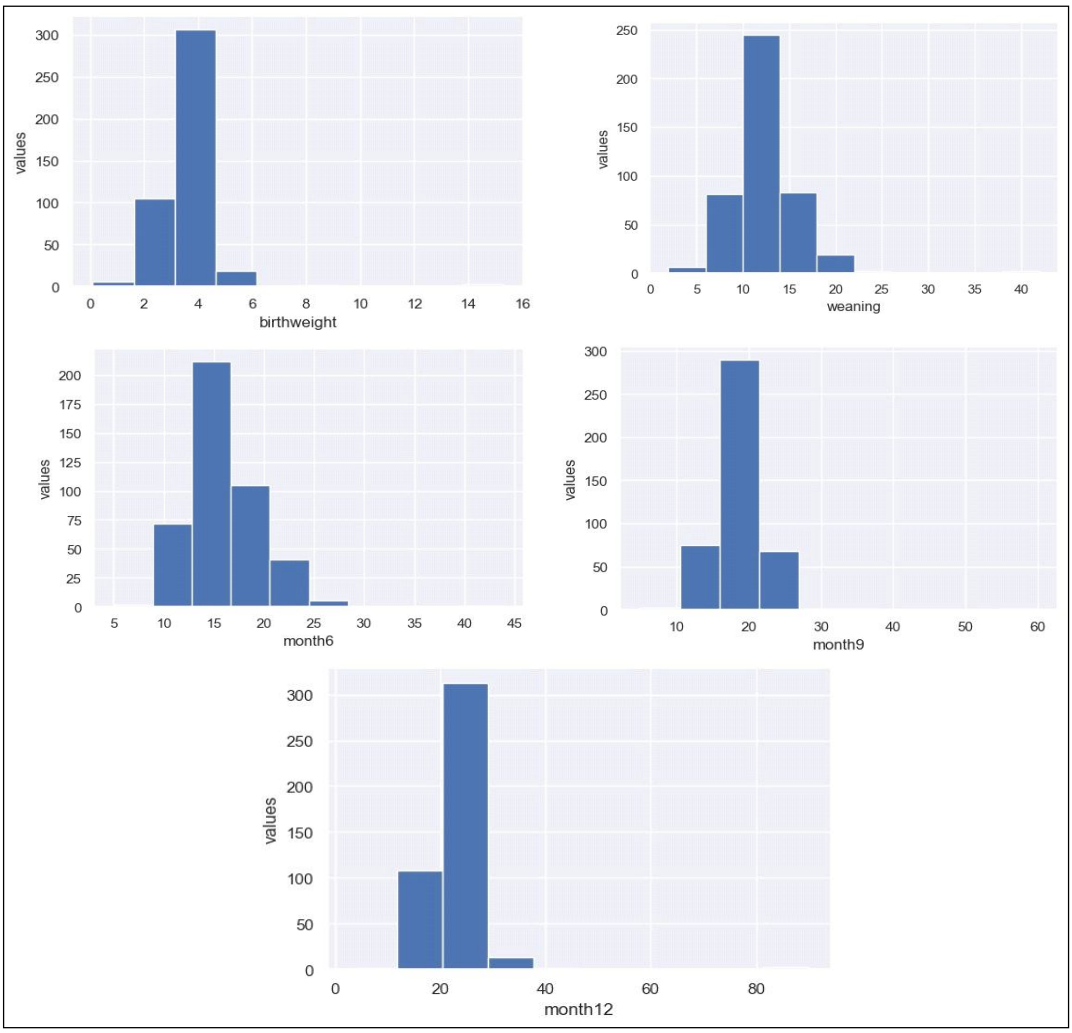

Fig 1-5: Bar-diagrams depicting body weights of sheep at birthweight, weaning, 6-, 9- and 12-month weights before winsorization.

Table 1: Descriptive statistics of data before winsorization.

	Birth weight	Weaning	Month 6	Month 9	Month 12
Count	438	438	438	438	438
Mean	3.618437	12.20766	16.2341386	18.59524	22.78154
Std	1.08470785	3.523652	3.60782333	3.523294	5.744751
Min	0.1	2	5	5	3
25%	3.125	10	13.625	16.85121	20.44991
50%	3.6	12	16.126162	18.57452	22.60238
75%	4	13.4	18	20	24.34025
Max	15.3	42	44	60	90

Table 2: Descriptive statistics of data after winsorization.

	Birth weight	Weaning	Month 6	Month 9	Month 12
Count	438	438	438	438	438
Mean	3.5727749	12.11827	16.1985023	18.51535	22.52045
Std	0.70245493	2.901577	3.32446508	2.770303	3.319951
Min	1.3	4.9	9.6	12	14.58702
25%	3.125	10	13.625	16.85121	20.44991
50%	3.6	12	16.126162	18.57452	22.60238
75%	4	13.4	18	20	24.34025
Max	5.4	20	25.5598899	25.45197	31.2531

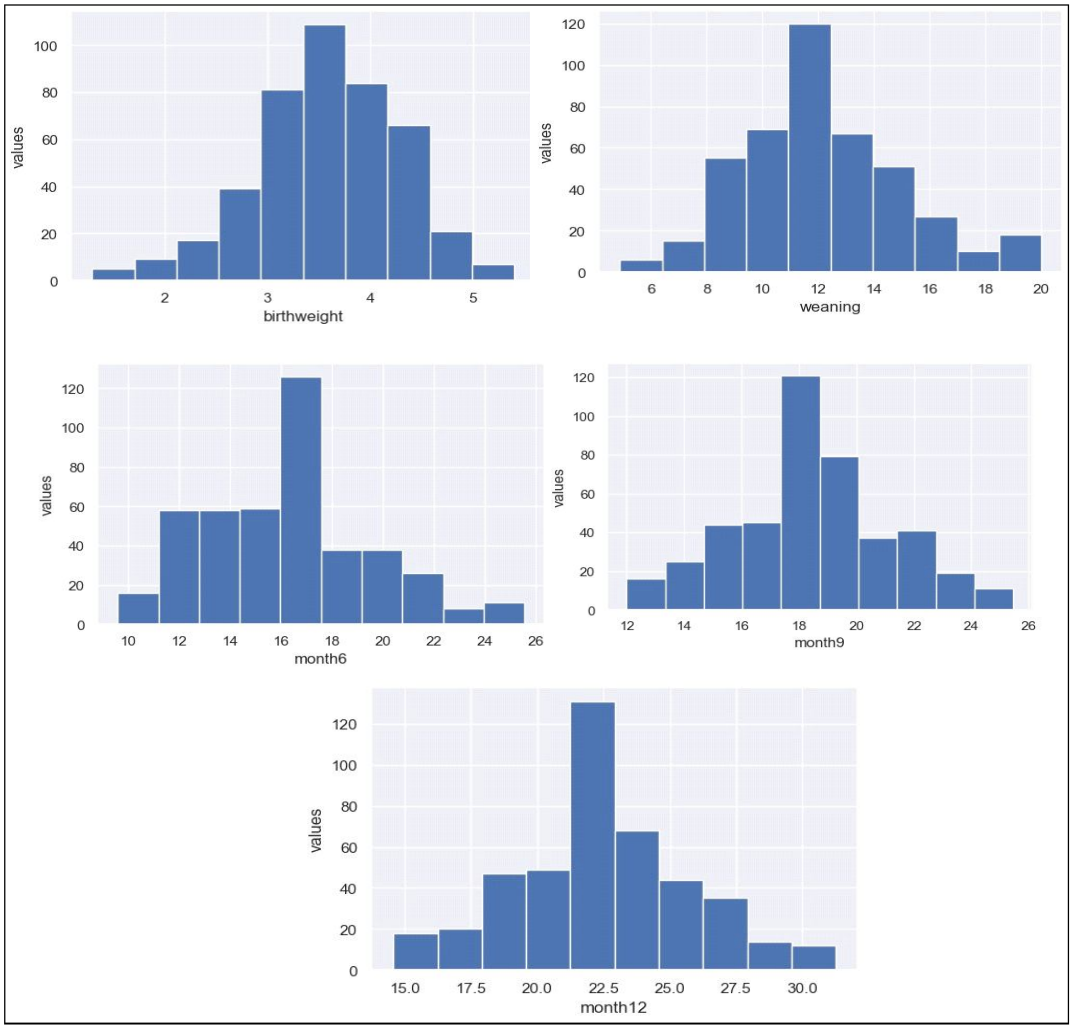

Fig 6-10: Bar-diagrams depicting body weights of sheep at birthweight, weaning, 6-, 9- and 12-month weights after winsorization.

circumstances. Winsorized mean is also less sensitive to outliers because it replaces them with less extreme values and therefore would not introduce more bias in the dataset. Winsorization would also not introduce more bias in the farm data because the anomalies are data points that do not correspond to normal behavior.

We used a two-sided winsoriztion approach in our study which was also reported to be better than the one-sided approach by Chambers *et al.* (2000). Winsorization, thus, is a crucial approach for detecting significant occurrences in the farm data. It is the process of identifying observations that deviate from the norm (Moso *et al.*, 2021).

CONCLUSION

Our results indicate that the winsorization technique works well for sheep farm data to remove the bias introduced by outliers and also removes, to a large extent, the need for manual outlier removal in data.

REFERENCES

Chambers, R., Kokic, P., Smith, P., Cruddas, M. (2000). Winsorization for identifying and treating outliers in business surveys. Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association Alexandria, Virginia. pp. 717-726.

Gerard-Marchant, P.G. (2007). scipy.stats.mstats.winsorize. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.winsorize.html.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. Technometrics. 11(1): 1-21. doi: 10.1080/00401706.1969.10490657.

Hargrave, M., Clarine, S. (2021). Winsorized Mean. https://www.investopedia.com/terms/w/winsorized_mean.asp.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering. 9(3): 90-95. 10.1109/MCSE.2007.55.

Maddala, G.S. (1992). Outliers. Introduction to Econometrics (2nd ed.). New York: MacMillan. pp. 89. ISBN 978-0-02-374545-4.

Moso, J., Cormier, S., Fouchal, F., de Runz, C., Wandeto, J. (2021). Anomaly Detection on Data Streams for Smart Agriculture. Agriculture. 11: 1083. https://doi.org/ 10.3390/agriculture 11111083.

Widenius, M., Axmark, D., DuBois, P. (2002). Mysql Reference Manual (1st. ed.). O'Reilly and Associates, Inc., USA. ISBN:978-0-596-00265-7. pp 712.