 ISSN 2362-0412

# Outlier Detection Method for Identifying Outliers that are not in Gaussian Distribution

K.K.L.B. Adikaram[a], M.A. Hussein[b], M. Effenberger[c] and T. Becker[d]

[a] *Faculty of Agriculture, University of Ruhuna*
[b,c] *Group Bio-Process Analysis Technology, Technische Universität München, Weihenstephaner Steig, Freising, Germany.*
[d] *Institut für Landtechnik und Tierhaltung, Vöttinger Straße Freising, Germany.*
[a] *lasantha@agricc.ruh.ac.lk*

## Abstract

The most statistical methods demand outlier (noise) in Gaussian distribution. When outliers are not in Gaussian distribution, these methods produce bias results. We introduce an outlier detection method that performs best when the outliers are in non-Gaussian distribution. The method is non-parametric and based on properties of arithmetic progression (AP). If the number of elements in AP is $n$, the maximumelement is $a_{max}$, the minimumelement is $a_{min}$, andthe sum of all elements is $S_n$. Then $R_{max} = \frac{a_{max}-a_{min}}{S_n-a_{min}*n}$ and $R_{min} = \frac{a_{max}-a_{min}}{a_{max}*n-S_n}$ is always equal to $2/n$. Usually, $R_{max} > 2/n$ implies that the maximumelementis an outlier and $R_{min} > 2/n$ implies that the minimum elementis an outlier. The value $2/n$ is nonparametric and always between 0 and 1. If $t$ is a threshold relevant to the considered domain, the value $2/n + t$ can be used to identify significant outliers where $0 \leq t \leq 1 - 2/n$. The method identifies one outlier at a time and continuous application of the method allows detection of multiple outliers. The algorithm was tested using several artificial and real data sets. The real data sets were the data which were automatically recorded with a frequency of twelve data points per day from a biogas plant, over a period of seven months. Among the different parameters, we selected the $H_2$ content, which we expected to maintain linear behavior during the stable operation. When the outliers are non-Gaussian, the Grubbs' test locates 0% - 17% as significant outliers at the significance level of 0.05. With the new method, there was $t$, which was capable of locating more outliers than Grubbs' test

***Keywords***: *Gaussian distribution, multiple outlier detection, non-parametric method, significant outliers*