Available online at www.isas.org.in/jisas
**Journal of the Indian Society of Agricultural Statistics 73(2) 2019 129–132**

# Application of Robust ANOVA Methods in Papaya having Outlier Data

**R. Venugopalan and B.L. Manjunath**
*ICAR-Indian Institute of Horticultural Research, Bengaluru*

*Received 02 June 2018; Revised 27 March 2019; Accepted 29 March 2019*

**SUMMARY**

Paul and Bhar (2011) advocated the use of M-estimation methods to address the issue of dealing with outliers in designed experimental data. An attempt has been made to elucidate the efficacy of this method over the regular ANOVA method using real time horticulture perennial crop experimental data. Results fortified the efficacy of robust methods while dealing with outliers as revealed by the three to five fold reduction in error sum of squares coupled with acceptable probability values for all the characters. Thus this study calls for adoption of robust ANOVA approach while dealing with outliers in perennial horticulture crop experiments in future research.

*Keywords:* Outlier, Papaya, Robust ANOVA.

## 1. INTRODUCTION

In a perennial crop field experiment, to compare the efficacy of several well defined treatments, the data generated from a designed experiment are analyzed under certain assumptions based on normality. The major hindrance to the requirement of normality is the presence of outlier among the replicated values for any of the observed trait. Outliers in any of the replications (of any treatment) lead to failure of normality assumption. One way out is to identify such an outlier(s) and delete them to have a meaningful comparison among treatments. However, deleting the outlying observation is not recommended because its deletion leads to violation of basic principles designs of experiment and from experimenter point of view every observation carries some information that should be exploited. This aspect is very much pertinent especially when we deal with perennial trees. Paul and Bhar (2011) advocated the use of M-estimation methods to address this issue. In the present communication, for real life perennial horticultural crop experiments (with an expected values of coefficient of variation beyond acceptable limits), an attempt has been made to adopt a suitable robust estimation method (without removing outliers), by employing the suggested approach to real life data on yield and associated traits of Papaya (Cv Red Lady). A thorough comparison of regular ANOVA and Robust ANOVA methods for various traits was also made for better understanding and further use in similar perennial crop research.

*Corresponding author:* R. Venugopalan
*E-mail address:* Venugopalan.R@icar.gov.in

## 2. MATERIALS AND METHODS

Primary data recorded on four important characters of Papaya (Cv Red Lady), *viz.,* Fruit number/plant, Fruit weight (kg)/plant, Mean Fruit weight (kg)/plant and Water productivity (kg/m$^3$) for twelve treatments (partial root zone drying irrigation treatments having alternate or fixed partial irrigation at different levels of evaporation replenishment. Normal irrigation meeting 80% ER using two emitters/plant served as control) with three replications evaluated in RCBD set up at the experimental plot of ICAR-Indian Institute of Horticultural Research, Bengaluru during 2016-17 season were used.

At the first step, as a measure of influence of the $i^{th}$ data point (suspected to be an outlier) on the estimation of treatment contrast, the value of Cook-statistic (Cook, 1977), when the first observation is an outlier, is computed as below:

$$D_1 = \frac{u'Su}{(p-1)\hat{\sigma}^2}\hat{\delta}^2$$

Substituting the value of S in $D_1$, we get

$$D_1 = \frac{S_{11}}{(p-1)\hat{\sigma}^2}\hat{\delta}^2$$

Where $S_{11} = u'B\Delta'C^{+}\Delta Bu$

Outliers were identified across treatments/traits based on the above measure and then the experimental data for all the traits were reanalyzed (simple two way analysis of variance) by deleting the identified outliers Further, since this process results in erroneous approach with regard to expressing the potential of a replicated value coupled with the rejection of randomization assumption, robust analysis of variance, which actually gives small weights to those outlying observations, thus extracting some information from that observation, was employed. Specifically, robust analysis of variance, using Huber's M-estimation (Huber,1973) and Andrew's M-estimation (Andrew's, 1974) were adopted.

Robust M-estimation approach instead of minimizing the sum of squared residuals in the classical ANOVA based approach, minimizes the sum of a less rapidly increasing function of the residuals ($\rho(e_i)$), as given below (Paul and Bhar, 2011).

$$Min\sum_{i=1}^{n}\rho\left(y_i - \sum x_{ij}\beta_j\right) = Min\sum_{i=1}^{n}\rho(e_i)$$

The solution is not scale equi-variant, and thus the residuals must be standardized by a robust estimate of their scale $\hat{\sigma}_e$ which is estimated simultaneously. As in the case of M-estimates of location, the median absolute deviation (MAD) is often used. Taking the derivative of above equation and solving, produces the score function

$$\sum_{i=1}^{n}\Psi\left(y_i - \sum x_{ij}\beta_j / \hat{\sigma}\right)x_{ik} = \sum_{i=1}^{n}\Psi\left(e_i / \hat{\sigma}_e\right)x_i = 0 \qquad (18)$$

Where $\hat{\sigma} = median\left|e_i - median(e_i)\right|/0.6745$

With $\Psi = \rho'$. There is now a system of k+1 equations, for which $\Psi$ is replaced by appropriate weights that decrease as the size of the residual increases

$$\sum_{i=1}^{n} w_i\left(e_i / \hat{\sigma}_e\right)x_i =$$

$$\sum_{i=1}^{n}\frac{x_{ij}\{\psi[(y_i - x_i'\beta)/s]/(y_i - x_i'\beta)/s\}/(y_i - x_i'\beta)}{s} = 0$$

j=0.1,…..,k

As

$$\sum_{i=1}^{n} x_{ij}w_{i0}\left(y_i - x_i'\beta\right) = 0 \qquad \text{j=0.1,…..,k}$$

$$w_{i0} = \begin{cases} \dfrac{\psi\left[\left(y_i - x_i'\hat{\beta}_0\right)/s\right]}{\left(y_i - x_i'\hat{\beta}_0\right)/s} & \text{where} \\ & \text{if } y_i \neq x_i'\hat{\beta}_0 \\ 1 & \text{if } y_i = x_i'\hat{\beta}_0 \end{cases}$$

Hence by matrix notation $X'W_0X\beta = X'W_0y$

where $W_0$ is n × n diagonal matrix of weights then one step estimator is -

$$\hat{\beta} = \left(X'W_0X\right)^{-1}X'W_0y$$

**Robust criterion functions**

| Criterion | $\rho(z)$ | $\psi(z)$ | $w(z)$ | Range |
|---|---|---|---|---|
| Least squares | ½ z2 | z | 1.0 | $\lvert z\rvert < \infty$ |
| Huber's function | ½ z2<br>$\lvert z\rvert t - 1/2t^2$ | z<br>t sign(z) | 1.0<br>$t/\lvert z\rvert$ | $\lvert z\rvert \le t$<br>$\lvert z\rvert > t$ |
| Andrew's function | a[1-cos (z/a)]<br>2a | Sin(z/a)<br>0 | Sin(z/a)/(z/a)<br>0 | $\lvert z\rvert \le a\pi$<br>$\lvert z\rvert > a\pi$ |

Here $\rho(z)$ is the function of residual, $\psi(z)$ is the derivative of $\rho(z)$ and w (z) is the weight function. SAS codes using SAS V 9.3 were generated for both the estimation procedures and used for analysis (SAS V 9.3,2012).

## 3. RESULTS AND DISCUSSION

The results of both classical and robust ANOVA methods for all the four characters studied are presented in Tables 1-10. Using Cook's measure, number of outliers (replications) identified were 2,1,2, and 2 across treatments for the respective characters. For the trait, fruit no/plant, Cook's measure was 3.45 and 7.22 for two replications pertaining to T2 and T7; for the trait, fruit weight/plant, it was 0.346 for a replication of T7; for the trait mean fruit weight, measures computed to be 8.22 and 11.276 for two replications corresponding to T3 and T2 and for the trait water productivity, Cook's measures were 6.33 and 11.07 two replications pertaining to T7 and T3, and over all the traits it was observed in the range of 0.346 to 11.276. Further, the efficacy of robust estimation methods based on the estimated average error variance is depicted in Table 9 and 10. Perusal of the results indicated the following salient conclusions.

The results of Huber's M-estimation for the character fruit no/plant showed that treatments are significant at P<0.0143 and the error mean square (EMSS)was 0.265. The results of Andrew's M-estimation showed that treatments are significant at P<0.0201 and EMSS was 0.568. The results of Huber's M-estimation for the character fruit weight (kg/plant) showed that treatments are significant at 0.0005 level and EMSS was 0.092. The results of Andrew's M-estimation showed that treatments are significant at P<0.0018 with EMS as 0.134. The results of Huber's M-estimation for the character mean fruit weight (kg/plant) showed that treatments are significant at p<0.0014 with EMSS as 2.471. The results of Andrew's M-estimation showed that treatments are significant at P<0.0005 with EMSS as 2.322. The results of Huber's M-estimation for the character water productivity (kg/m3) showed that treatments are not significant. The results of Andrew's M-estimation showed that treatments are significant at P<0.0116 with EMSS as 8.789.

Comparison of ANOVA and Robust ANOVA methods (Table 10) indicated that the probability values for treatment differences were within the critical region when the aberrant replications are removed, which were non-significant in the regular ANOVA approach. However, as discussed earlier this procedure is scientifically incorrect. Robust ANOVA approach when adopted increased the level precision coupled with reduction in error mean square for all the traits studied, which calls for adoption of this approach while dealing with outliers in perennial crop data. Further, this approach helps the researchers to overcome the problem of ending with on par results for all the treatments imposed, due to the presence of one or more outlier replicated values, , especially in tree crop studies due to the use of regular ANOVA approach.

**Table 1.** Results of robust ANOVA using M-estimation (Huber's function: Fruit number/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 6.415 | 1.605 | 6.190 | 0.9603 |
| Treatment | 11 | 66.238 | 6.210 | 29.392 | **0.0143** |
| Error | 44 | 11.565 | 0.265 | | |
| Total | 59 | 78.308 | | | |

**Table 2.** Results of robust ANOVA using M-estimation (Andrew's function: Fruit number/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 8.462 | 2.112 | 3.712 | 0.5201 |
| Treatment | 11 | 48.420 | 4.405 | 7.705 | 0.0201 |
| Error | 44 | 25.003 | 0.568 | | |
| Total | 59 | 81.885 | | | |

**Table 3.** Results of robust ANOVA using M-estimation (Huber's function: Fruit weight (kg)/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 11.221 | 2.814 | 30.530 | 0.334 |
| Treatment | 11 | 37.214 | 3.384 | 37.800 | **0.0005** |
| Error | 44 | 4.006 | 0.092 | | |
| Total | 59 | 52.491 | | | |

**Table 4.** Results of robust ANOVA using M-estimation (Andrew's function: Fruit weight (kg)/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 11.845 | 2.964 | 22.360 | 0.194 |
| Treatment | 11 | 33.434 | 3.083 | 26.620 | **0.001** |
| Error | 44 | 5.761 | 0.134 | | |
| Total | 59 | 51.040 | | | |

**Table 5.** Results of robust ANOVA using M-estimation (Huber's function: Mean Fruit weight (kg)/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 10.322 | 2.581 | 1.074 | 0.4820 |
| Treatment | 11 | 904.125 | 82.544 | 40.145 | **0.001** |
| Error | 44 | 109.145 | 2.471 | | |
| Total | 59 | 1023.582 | | | |

**Table 6.** Results of robust ANOVA using M-estimation (Andrew's function: Mean Fruit weight (kg)/plant)

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 9.350 | 4.675 | 2.010 | 0.041 |
| Treatment | 11 | 845.250 | 76.811 | 37.449 | **0.0005** |
| Error | 44 | 102.25 | 2.322 | | |
| Total | 59 | 956.85 | | | |

**Table 7.** Results of robust ANOVA using M-estimation (Huber's function: Water Productivity (kg/m3))

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 3.196 | 0.811 | 1.560 | 0.0761 |
| Treatment | 11 | 17.171 | 1.520 | 0.138 | **0.051** |
| Error | 44 | 4.187 | 11.061 | | |
| Total | 59 | 24.544 | | | |

**Table 8.** Results of robust ANOVA using M-estimation (Andrew's function: Water Productivity (kg/m3))

| Source | DF | Sum of Square | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Replication | 4 | 5.626 | 1.402 | 0.155 | 0.064 |
| Treatment | 11 | 31.665 | 2.854 | 0.355 | **0.012** |
| Error | 44 | 5.002 | 8.789 | | |
| Total | 59 | 42.293 | | | |

**Table 9.** Efficiency of M-estimation methods for Yield and associated traits

| Character | Average Error Variance (Huber's method) | Average Error Variance (Andrew's method) |
|---|---|---|
| Fruit Number/plant | 0.265 | 0.568 |
| Fruit weight (kg/plant) | 0.092 | 0.134 |
| Mean Frit Weight (kg) | 2.471 | 2.332 |
| Water Productivity (kg/m3) | 11.061 | 8.789 |

## 4. CONCLUSIONS

An attempt has been made to compare the efficacy of normally adopted ANOVA approach with that of robust ANOVA methods while analyzing real life perennial crop experimental data having outlier/aberrant observations. More specifically, primary experimental data on four different yield and associated characters of Papaya (Cv Red Lady), for twelve treatments with three replications evaluated in RCBD set up in the experimental plot of ICAR-Indian Institute of Horticultural Research, Bengaluru during 2016-17 season were utilised. Results fortified the efficacy of robust methods while dealing with outliers as revealed by the three to five fold reduction in error sum of squares coupled with acceptable probability values for all the characters. Thus this study calls for adoption of robust ANOVA approach while dealing with outliers in perennial horticulture crop experiments in future research and suggested to develop similar robust approach for carrying out pooled (over years/seasons/locations) ANOVA based real life experimental data, especially in long term perennial crop experiments, not only based on RCBD set up but also based on other design set up.

## ACKNOWLEDGEMENTS

Authors wish to thank the Director, ICAR-IIHR for providing all facilities, SSCNARS for providing the SAS software to its nodal centre ICAR-IIHR and the Anonymous referee for critical comments, which has led to the improvement of the quality of the paper.

## REFERENCES

Andrews, D.F (1974). A robust method for multiple linear regression. *Technometrics*. **16**, 523-531.

Bhar, L.M. and V.K. Gupta (2001). *A useful statistic for studying outliers in experimental designs*. *Sankhya*, **B63**, pp. 338-350.

Cook, R.D (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15-18.

Huber, P.J (1973). Robust regression: Asymptotic, conjectures, and Monte carlo. *Ann. Stat*., **1**, 799-821.

Paul, R.K and L.M. Bhar (2011). *M-estimation in block design*. *J. Ind. Soc. Agril. Statis.*, **65(3)**, pp. 323-330.

SAS V 9.3 (2012). *Statistical Analysis System Version 9.3 SAS Institute, Cary NC.*

**Table 10.** Comparison of regular ANOVA and Robust ANOVA methods for Yield and associated traits

| Character | Regular ANOVA | | ANOVA after removing outliers* | | Huber's Robust ANOVA | | Andrew's robust ANOVA | |
|---|---|---|---|---|---|---|---|---|
| | EMS | P-value | EMS | P-value | EMS | P-value | EMS | P-value |
| Fruit Number/plant | 11.336 | 0.072 | 4.352 | 0.046 | 0.265 | 0.024 | 0.568 | 0.010 |
| Fruit weight (kg/plant) | 21.566 | 0.084 | 4.254 | 0.042 | 0.092 | 0.008 | 0.134 | 0.012 |
| Mean Frit Weight (kg) | 12.175 | 0.091 | 4.52 | 0.084 | 2.471 | 0.042 | 2.332 | 0.046 |
| Water Productivity (kg/m3) | 19.554 | 0.042 | 12.252 | 0.040 | 11.061 | 0.00001 | 8.789 | 0.0001 |

*Number of outliers identified were 2,1,2, and 2 across treatments for the respective characters.