On the development of the GE and the GGE interaction Biplot in the RCIM model and the evaluation of its’ robustness to the outlying observations

Alfian Futuhul Hadi1,a, H Sadiyah2 and M Hasan3

1Statistical Laboratory, Department of Mathematics, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia
2Biometrical Laboratory, Department of Agronomy, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia
3Department of Mathematics, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia
E-mail: aafhadi@unej.ac.id

Abstract. Our recent statistical research on modeling of the two-ways table data was focused especially on the robustness of Row Column Interaction Model (RCIM). It has been showed that the RCIM model provide better result in fitting the data with outliers better than others even for Normal distribution. In this paper we focus on the influenced of the outlying observations to the visualization of the interaction effects in the RCIM modeling. We proposed the Genotype by Environment (GE) and the Genotype and Genotype by Environment (GGE) graphics display by a Biplot on the RCIM model. We also evaluate the influence of the outlying observations to the two kinds of Biplot of GEI by adding the outlying observations to the data. According to the Mean Square Error (MSE) and the Procrustes analysis, the GGE Biplot of RCIM has better result than the GE. The GE Biplot fails to accommodate inflation variance by the presence of a single environmental outliers with large percentage. In addition, the GE Biplot has difficulties to hold total percentage of the variance explained.

1. Introduction
As part of the national strategy of food sufficiency, plant breeding provided some useful information determining the suitable genotype in the variety development. In wide archipelagic agricultural country area like Indonesia, not every region has a similar condition. There were some varieties of cultivar, that cannot grow well in any region. The variation of the environment may lead in observations having different characteristic to the other observations, known as outlier. Such outliers often excluded from the analytical data processing. in some cases of plant breeding research, the outliers have very useful information.

Various statistical methods have been developed to assess the Genotypes by Environments Interaction (GEI), the two of those are the Genotype by Environment (GE) and the Genotype and Genotype by Environment (GGE) Biplots. Biplot is an appropriate visualization tool for describing interactions in the data. The GE commonly used in the Additive Main Effects and Multiplicative Interactions (AMMI) model, it is considered an effective tool for determining the pattern of the GEI, graphically. While the GGE uses the environmentally centralized data, that is the G + GE, the GE
Biplot AMMI is known as with double-centered data. The AMMI analysis separates G from GE first and then puts them together again, whereas GGE biplot analysis deals with G+GE directly. Therefore, explicit separation of G from GE in AMMI analysis does not lead to the conclusion that it is superior to GGE biplot analysis [1].

The GE and the GGE Biplot were almost similar, they decompose the residual matrix by the Singular Value Decomposition (SVD) method then plotting the two main components into the two-dimensional superimposed graph. The difference is just that the GE Biplot has balanced weight (singular value) in each genotype and environmental score whereas the GGE gives an additional weight to the environmental score. But, this will result in different advantages in each Biplot. One unique merit of a GGE Biplot is that it can graphically show the which-won-where patterns of the data [2].

In more general overview of statistical modelling, there is an overlapping methodology of AMMI model, that is the RCIM of Yee and Hadi [3]. The RCIM model was identical to AMMI Model. The RCIM is also used for modelling count data that was potentially has unique robustness to outlier in Poisson data distribution [4]. The RCIM is an extension of the Reduced-Rank Vector Generalized Model (RR-VGLM) where the first linear predictor is modeled by the sum of row effect, column effect and the row by column intercations effect. The big issue here is that we knew that SVD is vulnerable to the outliers [5], so the construction of robust visualization of the interaction effects in the RCIM model is needed to be investigated.

This research was conducted to evaluate the robustness of RCIM model in the point of view of its graphical representative of the interactions by the GE and the GGE Biplot, especially in the case of there were outlying observations on the data. Our recent statistical research on the RCIM model’s robustness, it was shown that the RCIM model provide a better result in fitting the data with outliers rather than others, even for normally distribution data. For further reading please see the detail in [4]. We proposed here the GE and the GGE Biplot on the RCIM Model, and then evaluate the influence of the outlying observations to the two kinds of Biplot. We now will focus on the visualization of the interaction and the effect of the outlying observations in it.

2. Methodology
This research used the data of Yan [1] with a simple scheme of simulation about outliers. This data was originally available on GGEBiplotGUI package, which has 18 variants of genotypes determined from the 9 different locations or environments. We then conducted a simple scheme of simulation for adding outliers to the data. The outliers were added to the data, placed randomly as we conducted before on [5], for 2%, 5% and 10% of outlier of the whole cells in the data matrix. In other words, we put the number of outliers 4, 7, and 17 respectively, from the total number of 162 observations. The outliers were generated randomly following the normal distribution [8]:

\[ N(\mu_j + k\sigma_j,\sigma_j^2) \]

where
\( \mu_j \) is the average value of the data for the j-th column.
\( \sigma_j^2 \) is the variance of the data.
\( k \) is a constant value of the magnitude of the outliers, \( k= 1, 2, 3 \).

2.1 Outlier
Outlier is defined as the part of observation which has different characteristics from most corresponding observation data set. An observation is considered as an outlier when its value of the \( k \)-multiplied of standard deviation is greater than its original mean, where the \( k \) is greater than 3. Mathematically, it can be expressed by the following [9]:

\[ y_i \geq \mu_j + k \times \text{stdev}(y_i) \]
The value of $k$ is the magnitude of the outlier, it shows how far the outlier value from its original mean. In this case, there are three values of $k$, i.e. $k=1$, 2, 3, that were used to evaluate the magnitude of the outlier. Furthermore, we simulated three scenarios of the number of outlier as proportion to the total cell of data matrix: there are 2%, 5% and 10%. Subsequently, the outlier then be placed on the data, follow [10] that proposed some placement methods, some of them which are our interest here:

1. **Scattered Outliers**
   The (scattered) outlier will be allocated randomly in some representing positions, by choosing genotype (row) and environment (column) randomly. The outlier then was being placed on the certain row and column. The next outliers were placed at the same fashion.

2. **Single Environment Outliers**
   Firstly, the outlier was placed randomly by choosing column (environment) randomly then the others following outliers were placed on the chosen row (genotype), which had been chosen randomly, until they filled in all elements of its column.

2.2 **RCIM for GE and GGE Biplot in R**

The GE and GGE Biplot were built from SVD of the residuals of the RCIM2 Model. However, there is a different model between the GE Biplot and the GGE Biplot of RCIM. We used the `rcim` command with rank=0 --that is similar as the `vglm` command in VGAM packages-- to estimate the model [3]. The RCIM with normal distribution for the GE model will run the `rcim` with `gaussianff` family function:

$$\text{rcim}(\text{data}, \text{family} = \text{gaussianff}(), \text{Rank} = 0)$$

while the RCIM with normal distribution for the GGE model will use the `vglm` with `gaussianff` family function:

$$\text{vglm}(\text{yield}\sim\text{G}, \text{family} = \text{gaussianff}(), \text{data})$$

after that we use decomposed the residual using the `svd` for both model respectively to get the first 2 principal components (PC1 and PC2) than we plotted it by the `ggplot2` packages.

2.3 **Procrustes**

Two ordinations could be similar. Two ordinations are said to be similar if they are only distinguished by a certain transformation we can find. Unfortunately, it is hard to find such a congruence, since the usually axes have slightly different orientation and scaling. The best way to compare ordinations is to use the procrustes rotation [11]. The procrustes aims to compare two configurations. In principle, to see the similarity of the shape and size of the two configurations, one of the configurations is fixed, while the other is transformed to fit the first one [12]. This process is interpretation by figure 1:

![Figure 1](image)

**Figure 1.** The procrustes superimposition of two shapes (a) begins by translating them to superimpose their centroids (b) before scaling (dialating) (c) and rotating them (d) to maximize their coincidence [12].

The procrustes rotates a configuration to maximum similarity with another configuration and also tests the non-randomness between two configurations. Procrustes rotation rotates a matrix to maximum similarity with a target matrix minimizing sum of squared differences. Procrustes rotation is typically used in comparison of ordination results. It is particularly useful in comparing alternative solutions in multidimensional scaling. The Procrustes function returns an object of class `Procrustes` with items. Function `protest` inherits from Procrustes then provides some items interesting: (1) The Sum of
Squared (SS) measure the differences between X and (rotated) Y. The smaller SS the more similar between the two matrices. (2) The “signif” is the p-value that the smaller p-value, the more similar the two matrices were [13].

3. Results

3.1 The influence of the outlier on the MSE

For this dataset without any outlier, Tables 1a & 1b show us that the GGE model was provided smaller MSE than the GE model. This means that by default, the GGE model was fitting the data better than the GE one. So does when we took a look to the result of simulated data with scattered or single environment outliers, we can see here that generally, the GGE model was fit the data properly. This indicate that potentially, the GGE model had robustness to the outlier better than the GE one.

The MSE was inflated to become higher than before in the original data without outlier. There was different influence corresponding to the GE or the GGE model; the outlier was seen to affect the inflation of the MSE of the GE model higher than that of the GGE model. Generally, either for the GE or the GGE model, there was a tendency that the higher value of outlier the higher MSE we got. Specifically, for the GE model, the more number of outliers in the data the larger value of MSE, either of scattered outliers or single environment outliers. We can say here that both the GE and the GGE model still facing the vulnerability of scattered outliers. But not so for the GGE model by the single environment outliers, there was something interesting here. For all $k$ means, lower, middle or higher value of outliers, there was decreasing MSE with more percentage of 10% of single environment outliers. One may pay attention to the right below corner cell of table 1b at, bold printed. This indicated that the GGE model potentially hold the robustness to the SE outliers.

3.2 The Procrustes: The influence of the outliers to the Biplot configuration

Procrustes is applied to test how far two configurations differ each other’s. In this case, to test the influence of the outlier on the RCIM model, we verify the difference between the residual matrix of RCIM2 from the original data versus it’s from the simulated data with outlier (scattered and single environment outliers). For small value scattered outlier, all of the Procrustes’s $p$-value are slightly significant, it means that the small value outliers, both scattered and single environment outliers, was less influential on the RCIM Biplot either for GE or GGE model. It is indicating the potential robustness of RCIM Biplot to the small value of outlier. Table 2 show the similarity result as the data without outliers when the value of $k = 3$ with any percentage of scattered outliers, and so were the single environment outliers. But it was no longer parallel result with the higher value and percentage of outliers, the RCIM2 of the GGE model provide different result. But for higher value on scattered outliers $k=10$, the more number of outliers the more tendencies to get higher $p$-value in Procrustes analysis, that means the more influence impacted in the Biplot of the GGE model.

| Table 1a. The MSE of the GE model. |
|-----------------------------------|
| Outlier                      | Percentage of outliers | The magnitude of the outlier |
| Without Outlier              | 0%                     | 0.830  | 0.830  | 0.830  |
| Scattered Outliers           | 2%                     | 1.130  | 2.279  | 3.103  |
|                                | 5%                     | 1.083  | 3.292  | 6.838  |
|                                | 10%                    | 1.281  | 5.780  | 12.367 |
| Single Environment (SE) Outliers | 2%                  | 0.972  | 2.016  | 3.051  |
|                                | 5%                     | 1.131  | 3.190  | 5.431  |
|                                | 10%                    | 2.325  | 7.302  | 9.310  |

| Table 1b. The MSE of the GGE model. |
|-----------------------------------|
| Outlier                      | Percentage of outliers | The magnitude of the outlier |
| Without Outlier              | 0%                     | 0.123  | 0.123  | 0.123  |
| Scattered Outliers           | 2%                     | 0.282  | 1.336  | 2.460  |
|                                | 5%                     | 0.313  | 2.420  | 5.528  |
|                                | 10%                    | 0.514  | 4.444  | 9.586  |
| Single Environment (SE) Outliers | 2%                  | 0.214  | 0.990  | 2.301  |
|                                | 5%                     | 0.235  | 1.215  | 3.042  |
|                                | 10%                    | 0.180  | 0.470  | 0.826  |
In the case of focusing on single environment outliers, Biplot of the GGE RCIM2 model with outliers also provides the similar result as RCIM2 model without outlier when the value of $k = 3$. But, there is some interesting information here, for higher value of single environment outlier, there was still possibility of robustness of the GGE Biplot of RCIM2 model. The highest value outlier $k = 10$ and $k = 15$ did not affect the Biplot of GGE when there were 10% outliers in the simulated data.

The procrustes sum of square between two residuals from GE model with outlier (scatter and single environment) and GE model without outlier show that the results are almost similar where the model with small value of outlier ($k = 3$) provides the similar result as the data without outliers. When we focus on 10% single environment outlier from GGE model and GE model, the residuals model with this single environment outlier and without outlier are almost similar. This indicates that RCIM is robustness with single environment outlier.

Table 2. The $p$-value of Procrustes of the GGE Biplot and the GE Biplot.

| Outliers         | The GGE Biplot | The GE Biplot |
|------------------|---------------|---------------|
|                  | The magnitude of the outlier | The magnitude of the outlier |
|                  | $k = 3$ | $k = 10$ | $k = 15$ | $k = 3$ | $k = 10$ | $k = 15$ |
| Scattered Outliers | 2%    | 0.002 | 0.001 | **0.568** | 0.001 | 0.001 | **0.603** |
|                  | 5%    | 0.003 | 0.294 | 0.610 | 0.001 | 0.303 | 0.626 |
|                  | 10%   | **0.070** | 0.159 | 0.035 | **0.081** | 0.174 | 0.043 |
| Single Environment (SE) Outliers | 2% | 0.001 | 0.012 | **0.179** | 0.002 | 0.018 | **0.194** |
|                  | 5%    | 0.001 | 0.067 | 0.214 | 0.001 | 0.066 | 0.250 |
|                  | **10%** | 0.001 | 0.031 | 0.049 | 0.001 | 0.048 | 0.054 |

Table 3a. The percentages of variance explained by the Biplot of the GE model.

| Outlier | The magnitude of the outlier |
|---------|----------------------------|
| Without Outlier | 0% | **66.15** |
| Scattered Outliers | 2% | 63.89 | 60.38 | 47.80 |
|                  | 5% | 58.25 | 48.34 | 40.93 |
|                  | 10% | 51.01 | 42.39 | 48.03 |
| Single Environment (SE) Outliers | 2% | 67.53 | 74.86 | 77.72 |
|                  | 5% | 66.51 | 76.43 | 81.26 |
|                  | **10%** | 63.89 | 60.38 | 47.80 |

Table 3b. The percentages of variance explained by the Biplot of the GGE model.

| Outlier | The magnitude of the outlier |
|---------|----------------------------|
| Without Outlier | 0% | **48.75** |
| Scattered Outliers | 2% | **55.32** | 64.29 | 48.52 |
|                  | 5% | 48.07 | 47.6 | 40.96 |
|                  | 10% | 39.65 | 41.75 | 46.22 |
| Single Environment (SE) Outliers | 2% | **56.05** | 71.98 | 78.38 |
|                  | 5% | 56.18 | 71.1 | 80.15 |
|                  | **10%** | **51.62** | 65.09 | 69.06 |

Now we turn to see whether outliers affect the percentage of total variance explained by the Biplot. Tables 3a and 3b show the percentage of total variance explained after the scattered and the single environment outliers by the Biplot of GE and GGE respectively. The GE Biplot seems to be affected by the scattered outlier since there was a noticeable drop in the percentage of variance explained for all number and value of the scattered outliers (table 3a). But for the GGE one, based on table 4b there is still higher percentage of variance explained at simulated data with small value and small number of scattered outliers.

On the other hand, in single environment outliers, again, there are strong indications that Biplot GE and GGE have a good robustness. Table 3 shows that in the presence of single environments it is followed by an increase in the percentage of total variance explained by the GGE Biplot, for both small and large outliers as well as for large and small percentage outliers. The exceptions only occur.
for the GE Biplot with the data containing many outliers with any magnitude \( k = 3, 10 \), and also \( k=15 \).

This shows us that Biplot GE fails to accommodate inflation variance by the presence of a single environmental outlier in large percentage.

4. Discussions: the GGE Biplot and its robustness

The RCIM Biplot of GGE model seems to hold a good robustness to the single environment outliers according to the MSE and also the Procrustes analysis. This robustness came out from the model of decomposing the interaction, i.e. the GGE model was accommodates the interaction terms in the G plus the pure interaction of G by E. When there is an increasing observation value by outliers in a single environment, the G effects from that “single environment” will increase as well as the GE effects, then the GGE model will immediately model the increase in the G + GE terms and decompose the GGE interactions by the Biplot, properly.

![Figure 2](image)

**Figure 2.** The GGE Biplot with (a) no outlier (original data), (b) with 10% SE outliers of \( k = 3 \), (c) 2% SE outliers of \( k=10 \), (d) 10% SE outliers of \( k=15 \).
As shown in Figure 2b, for the small value outliers but many, some environments increase in small magnitudes; therefore Biplot GGE produces a little change in configuration even it is very similar to the Biplot of original data (Figure 2a.). The total variance explained increase slightly in accordance with the addition of variance in some environments with small magnitudes. Conversely, if an increase only occurs in a particular environment with a large magnitude (2%, k = 10) then the increase only occurs in certain genotypes in particular environments resulting in a spike in the variance of the data. The GGE Biplot then modelled it accurately as Figure 2c. With very high value of % total variance explained, 71.98%, it shows the capability to accommodate the variance inflation.

Meanwhile, when the single environment outliers with large magnitude occur in many environments it is not necessarily resulting in a variance inflation, as followed by a rising shift mean in several environments simultaneously, so that the variance of the data with these outliers does not uphill dramatically. Thus, the Biplot will model a moderate variance rather than a variance with large inflation.

This is evident in low percentage of total variance explained by the GGE Biplot and even more by the GE Biplot. However, this decrease in total variance explained by the GGE was not lower than the original data, not as worse as the GE Biplot was.

5. Concluding Remark
The GGE Biplot in RCIM model has good robustness to the single environment outliers according to the MSE and also the Procrustes analysis. This robustness came out from the model of decomposing the interaction, i.e. the GGE Biplot was accommodates the interaction terms in the G plus the pure interaction of G by E. The GE Biplot fails to accommodate inflation variance by the presence of a single environmental outlier with large percentage. When the single environment outliers with large magnitude occur in many environments, it is not necessarily resulting in a variance inflation. In this situation, the Biplot will be expected to model a moderate variance rather than a variance with large inflation, but the GE Biplot faces difficulties to hold the total percentage of the variance explained.

Acknowledgments
This research was supported by Ministry of Research, Technology & Higher Education of Indonesia, Grant No. 1301/UN25.3.1/LT/2018. We thanks to D. Anggraeni, Dimas Bagus CW, Mia Purnama and Upik Susilowati (Statistical Laboratory, Department of Mathematics of UNEJ) for the preparation of this paper.

References
[1] Yan W, Kang M S, Ma B, Woods S and Cornelius P L 2007 GGE biplot vs AMMI analysis of genotype-by-environment data Crop Sci 47(2) 643-653
[2] Yan W, Hunt L A, Sheng Q and Szlavnics Z 2000 Cultivar evaluation and mega-environment investigation based on the GGE biplot Crop Sci 40 597–605
[3] Yee T W and Hadi A F 2014 Row-column interaction models, with an R implementation. Computational Statistics 29(6) 1427-1445
[4] Hadi A F, Sadiyah H and Hasan M 2018 Handling outlier in two-ways table data: the robustness of row-column interaction model J. Phys.: Conf. Ser. 1028 012222
[5] Hadi A F 2011 Handling outlier in two-ways table by robust alternating regression of FANOVA models: towards robust AMMI models Journal ILMU DASAR 12(2) 123–131
[6] Gauch Junior H G and Zobel R W 1996 AMMI analysis of yield trials Genotype by Environment Interaction (Kang M and Gauch Junior HG, eds., CRC Press, Boca Raton)
[7] Hadi A F, Sadiyah H and Iswanto R 2017 On generalization of additive main effect and multiplicative interaction (AMMI) models: an approach of row column interaction models for counting data Malaysian Journal of Mathematical Sciences 11 115-141
[8] Rocke D M and Woodruff D L 1996 Identification of outliers in multivariate data Journal of the American Statistical Association 91(435) 1047-1061
[9] Kriegel H P, Kröger P P and Zimek A 2010 Outlier detection techniques Proc. 16th ACM Int. Conf. Knowl. Discovery Data Mining (SIGKDD) (Washington, DC, USA)

[10] Rodrigues P C, Monteiro A and Lourenco V M 2015 A robust AMMI model for the analysis of genotipe by environment data Bioinformatics 32 58–66

[11] Oksanen J 2015 Multivariate analysis of ecological communities in R: vegan tutorial Trends Ecol. Evol. 3 121

[12] Buttigieg P L and Ramette A 2014 A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses FEMS Microbiol Ecol. 90 543–550 https://mb3is.megx.net/gustame/other-methods/procrustes-analysis

[13] Peres-Neto P R and Jackson D A 2001 How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test Oecologia 129 169-178