Pertanika J. Sci. & Technol. 29 (2): 999 - 1015 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Robust Multivariate Correlation Techniques: A Confirmation Analysis using Covid-19 Data Set

Friday Zinzendoff Okwonu[1,2], Nor Aishah Ahad[1]*, Joshua Sarduana Apanapudor[2] and Festus Irismisose Arunaye[2]

[1]School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[2]Department of Mathematics, Faculty of Science, Delta State University, P.M.B.1, Abraka, Nigeria

ABSTRACT

Robust multivariate correlation techniques are proposed to determine the strength of the association between two or more variables of interest since the existing multivariate correlation techniques are susceptible to outliers when the data set contains random outliers. The performances of the proposed techniques were compared with the conventional multivariate correlation techniques. All techniques under study are applied on COVID-19 data sets for Malaysia and Nigeria to determine the level of association between study variables which are confirmed, discharged, and death cases. These techniques' performances are evaluated based on the multivariate correlation (R), multivariate coefficient of determination (R^2), and Adjusted R^2. The proposed techniques showed R=0.99 and the conventional methods showed that R ranges from 0.44 to 0.73. The R^2 and the Adjusted R^2 for proposed methods are 0.98 and 0.97 while the conventional methods showed that R equals 0.53, 0.44, and 0.19 whereas Adjusted R^2 equals 0.52, 0.43, and 0.18, respectively. The proposed techniques strongly affirmed that for any patient to be discharged or die of the Covid-19, the patient must be confirmed Covid-19 positive, whereas the conventional method showed moderate to very weak affirmation. Based on the results, the proposed techniques are robust and show a very strong association between the variables of interest than the conventional techniques.

Keywords: Coefficient of determination, Covid-19, multivariate correlation techniques, robust

ARTICLE INFO

Article history:
Received: 13 October 2020
Accepted: 04 February 2021
Published: 30 April 2021

DOI: https://doi.org/10.47836/pjst.29.2.16

E-mail addresses:
o.friday.zinzendoff@uum.edu.my (Friday Zinzendoff Okwonu)
aishah@uum.edu.my (Nor Aishah Ahad)
japanpudor@yahoo.com (Joshua Sarduana Apanapudor)
ifarunaye@yahoo.com (Festus Irismisose Arunaye)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Karl Pearson established the product sum formula, widely known as the Pearson product-moment correlation coefficient denoted by "*r*" (Pearson, 1920). The multivariate correlation technique is the generalisation of the conventional Pearson product-moment correlation coefficient. The latter is pairwise involving one independent variable and one dependent variable while the former requires two or more independent variables and one dependent variable. It can be described as the squared correlation between the expected values and the observed values (Abdi, 2007). Historically, Bravais had discussed correlation involving two and three variables (Pearson, 1920). These three variables consideration is the generalisation of the two variable concepts, and by extension is regarded as the multivariate or multiple correlation techniques.

The bivariate correlation analysis has been extensively discussed with numerous applications to different fields of study (Armstrong, 2019; Bareinboim et al., 2014; Brown et al., 2012; Geiss et al., 1991; Mukaka, 2012; Nguyen et al., 2013; Wang et al., 2017). Different techniques describing multivariate correlation analysis and their performance analysis have been discussed (Wang et al., 2017). The emergence of multivariate correlation analysis is due to the ever-increasing data dimension with associated variables and the need to determine associations between the variables. Unlike the conventional correlation technique due to Karl Pearson, multivariate correlation techniques have no basic formulation fundamentals (Wang & Zheng, 2020). Techniques such as the partial correlation and coefficient of determination have been recruited to justify analytically to analyse the association of multivariate independent and dependent variables. Techniques such as unsigned correlation have been proposed to investigate multivariate variables association (Wang & Zheng, 2020).

The multivariate correlation coefficient determines the strength of the association between two or more variables of interest, unlike the univariate correlation coefficient (r) whose values range between ±1, the multivariate correlation coefficient values (r_{xyz}) range between 0 and 1. A strong association indicates that r_{xyz} tends to one. On the other hand, if the multivariate correlation coefficient value tends to zero, it implies little or no association between the variables. The uniqueness of r_{xyz} is that it is always positive. The amplifying nature of the multivariate correlation technique is the ability to handle all the independent and dependent variables provide simultaneously (Geiß & Einax, 1996). In most cases, multivariate correlation techniques often reveal a stronger association between variables of interest than the univariate correlation technique (Zhang et al., 2008).

Correlation is applied to determine the level of association between variables of interest (Asuero et al., 2006). It has been investigated that the classical correlation coefficient is not robust when the data set contains influential observations (Abdullah, 1990; Okwonu et al., 2020). Studies have shown that a nonparametric approach based on the Spearman

correlation coefficient is insensitive against the influential observation present in the paired variables (Abdullah, 1990). We observed that when the bivariate correlation method was transformed into a multivariate correlation technique, the issue of underperformance due to the presence of outliers in the data set persisted as such, the existing multivariate correlation techniques were susceptible to outliers when the data set contained random outliers. To solve this problem of outliers, we propose robust and high breakdown multivariate correlation techniques that utilizes all the variables simultaneously and does not require the distributional assumption of the data set to compute the multivariate correlation coefficient value. Therefore, the main contribution to this study is that the proposed methods are insensitive and not susceptible to outliers.

The proposed methods and the conventional methods were applied to the Covid-19 data set. The level of association determined based on the data sets were used to evaluate the health risk of the Covid-19 pandemic. The analysis relying on the level of association could assist the researchers to advise people to abide by safety protocols established by healthcare providers and relevant organs of government.

The rest of this paper is organised as follows: In Section 2 we discussed the conventional methods: paired multivariate correlation; the median-based multivariate correlation technique; and the multivariate correlation by regression technique; followed by the proposed techniques: the multivariate cross-correlation technique, and the multivariate cross-correlation based on deviation method. The second section continued with the multivariate coefficient of determination, adjusted multivariate coefficient of determination, and ended with data collection. Results and discussions are presented in Section 3, and the conclusion follows in Section 4.

MATERIALS AND METHODS

Rodgers and Nicewander (1988) observed that the correlation coefficient (*r*) could be computed using different mathematical formulations. Different names have been ascribed to the mathematical complement, but the general objective is to determine the degree of associations between interest variables (Rodgers & Nicewander, 1988). This paper complies with the mathematical formulation concept elaborated and enunciated by Rodges and Nicewander (1988) in developing a new computational framework for the index. Before discussing the new computational framework, we looked at the different multivariate correlation index advanced by different researchers of different mindsets.

Paired Multivariate Correlation

Let $X, Y \in \hat{R}$ be independent random variables and $Z \in \hat{R}$ be the dependent variable. The idea is to investigate whether the independent and dependent variables correlate or not. From observation, the strength of correlation in Pearson's concept is stronger in dual sign

directions than the paired multivariate correlation. The conventional procedures involve pairing the random variables before the correlation values are plugged into the multivariate correlation formula. The following mathematical Equation 1 states the multivariate correlation coefficient method (Geiss et al., 1991; Geiß & Einax, 1996; Wang et al., 2018).

$$R = \frac{\varphi}{\theta}, \tag{1}$$

where $\left(R = r_{xyz}\right)$, $\varphi = \sqrt{R_{xz}^2 + R_{yz}^2 - 2R_{xz}R_{yz}R_{xy}}$ (Garnett, 1919) and $\theta = \sqrt{1 - R_{xy}^2}$. Equation 1 is defined by pairing the Pearson's correlation coefficient, where R_{xz}, R_{yz} and R_{xy} are well defined Pearson correlation coefficient. Equation 1 can be simplified as Equation 2-4:

$$D_{xy} = \frac{D_z}{D_{z1}} = \frac{D_{xi} \times D_{yi}}{\sqrt{D_{xi}^2 \times D_{yi}^2}}, \tag{2}$$

$$D_{xz} = \frac{D_k}{D_{z2}} = \frac{D_{xi} \times D_{zi}}{\sqrt{D_{xi}^2 \times D_{zi}^2}}, \tag{3}$$

$$D_{yz} = \frac{D_{kk}}{D_{zk}} = \frac{D_{yi} \times D_{zi}}{\sqrt{D_{yi}^2 \times D_{zi}^2}}, \tag{4}$$

where $D_{xi} = \sum_{i=1}^{n}(x_i - \bar{x})$, $D_{yi} = \sum_{i=1}^{n}(y_i - \bar{y})$ and $D_{zi} = \sum_{i=1}^{n}(z_i - \bar{z})$. Equation 1 can be written based on Equations 2-4 to form the paired multivariate correlation as shown in Equation 5

$$r\_xyz = \frac{d_{c(xyz)}}{t}, \tag{5}$$

where $d_{c(xyz)} = \sqrt{D_{xz}^2 + D_{yz}^2 - 2D_{xz}D_{yz}D_{xy}}$ and $t = \sqrt{1 - D_{xy}^2}$.

The Median-Based Multivariate Correlation Technique

This procedure involves substituting the mean by the median, that is $\widetilde{D_{xi}} = \sum_{i=1}^{n}(x_i - m(x))$, $\widetilde{D_{yi}} = \sum_{i=1}^{n}(y_i - m(y))$ and $\widetilde{D_{zi}} = \sum_{i=1}^{n}(z_i - m(z))$, where $m(x), m(y)$ and $m(z)$ are the median of the independent and dependent random variables. Substituting these into Equation 2-5, we have the median-based multivariate correlation as presented in Equation 6.

$$r_{\_dxyz} = \frac{d_{m(xyz)}}{m(t)}, \qquad 0 \le r_{\_dxyz} \le 1 \qquad (6)$$

where $d_{m(xyz)} = \sqrt{m(\widetilde{D_{xz}^2}) + m(\widetilde{D_{yz}^2}) - 2(m(\breve{D}_{xz})m(\breve{D}_{yz})m(\breve{D}_{xy})}$ and $m(t) = \sqrt{1 - m(\breve{D}_{xy}^2)}$.

Multivariate Correlation by Regression Technique (MCRT)

The concept of regression to compute the correlation coefficient was initiated by Mr. Yule around 1897 (Pearson, 1920; Weida, 1927). Standing on this note, this concept is considered suitable for this paper. The multivariate correlation technique based on regression was coined using the principle of multiple regression procedures (Huberty, 2003; Lewis-Beck et al., 2004). Recall the multiple regression formulation, that is Equation 7,

$$\hat{y} = b_0 + \sum_{k=1}^{p} b_k x_i, i = 1,2, \ldots\ldots, n \qquad (7)$$

where x_i are the independent variables (predictors), y denotes a single dependent variable, b_k, where $k = 1,2,..,p$ is the weights and $\hat{y}$ is the estimate of y called the predictor score. In this regard, the computation of b_k strictly relies on the information provided by the independent and dependent variables as such how large or small the correlation between y and $\hat{y}$ depends on the information provided by b_k. This method implies that the correlation between the dependent variable and the linear combination of the independent variables is called the multivariate correlation (Lewis-Beck et al., 2004). Based on Equation 7, the multivariate correlation denoted by $r_{\_y\hat{y}}$ (Asuero et al., 2006) can be started by applying the dependent variable and the predictor score, that is y and $\hat{y}$ as given in Equation 8.

$$r_{\_y\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}} \qquad (8)$$

where $\bar{\hat{y}} = \frac{\sum_{i=1}^{n}\hat{y}_i}{n}$. An alternative to Equation 8 is to apply the following procedure:

$$\hat{y} = Xb,$$

where $b = \left(X`X\right)^{-1}X`y$, $X_{n\times(p+1)}$ denote augmented matrix and $y_{n\times 1}$ vector (Abdi, 2007). Let γ denote the regression sum of square defined by Equation 9.

$$\gamma = b`X`y - \frac{(1`y)^2}{n} \qquad (9)$$

where $1^{`}$ is a row vector of 1 that align with the dependent variable *y* and γ as defined in Equation 9 and the total sum of square θ is defined as Equation 10.

$$\theta = y^{`}y - \frac{(1^{`}y)^2}{n} \tag{10}$$

$$\omega = y^{`}y - b^{`}X^{`}y \tag{11}$$

where ω is the sum of square of the residual (Equation 11). Applying Equation 9-11, we obtain the multivariate correlation as Equation 12.

$$\tau\hat{}2 = \frac{\gamma}{\varphi}, \varphi = \gamma + \omega \tag{12}$$

Equation 8 and 12 are based on the regression approach. The proposed methods are described as follows.

Multivariate Cross-Correlation Technique

Let $X, Y \in \hat{R}$ be independent random variables and $Z \in \hat{R}$ be a dependent random variable. These variables or vectors are paired such that each interacts to obtain the level of trend (Wang & Zheng, 2020). Recall the concept of vector, matrix products, and transpose, in this case, when a $(1 \times n)$ row vector is multiplied by an $(n \times 1)$ column vector, the product is a scalar or a real number. The independent and dependent variables may assume $(n \times 1)$ or $(1 \times n)$ row or column depending precisely on the variable transposed from row to column or column to row.

Let $S_{XX} = (XX^T)$ be the autocorrelation and $S_{XY} = (XY^T)$ be the cross-correlation as defined in Geiss et al. (1991) and Geiß and Einax (1996) and invoking the Cauchy-Binet concept we obtain the following Equation 13-15:

$$\Delta_1 = 1 - \left(\frac{X^TY}{X^TZ}\right), X^T = [x_i]^T, i = 1,2,..,n, X^TZ \neq 0 \tag{13}$$

$$\Delta_2 = 1 - \left(\frac{X^TZ}{Y^TZ}\right), Y^T = [y_i]^T, i = 1,2,..,n, Y^TZ \neq 0 \tag{14}$$

$$\Delta_3 = 1 - \left(\frac{Y^TZ}{X^TY}\right), Z^T = [z_i]^T, i = 1,2,..,n, X^TY \neq 0 \tag{15}$$

Where X^T, Y^T, Z^T are the transpose of the data set. Equation 13-15 can be written compactly as Equation 16-19.

$$\propto_1 = \left[\frac{X^T Y}{X^T Z}\right], \rightarrow \Delta_1 = 1 - \propto_1 \quad (16)$$

$$\propto_2 = \left[\frac{X^T Z}{Y^T Z}\right], \rightarrow \Delta_2 = 1 - \propto_2 \quad (17)$$

$$\propto_3 = \left[\frac{Y^T Z}{X^T Y}\right], \rightarrow \Delta_3 = 1 - \propto_3 \quad (18)$$

$$\tau = \sum_{j=1}^{k} \Delta_j^2, \vartheta = k \prod_{j=1}^{k+1} \Delta_j \text{ and } \emptyset = \sum_{j=1}^{k+1} \Delta_j^2 \quad (19)$$

The definition of Δ_j mimic Chatillon's correlation $r = \sqrt{1 - \left(\frac{c}{C}\right)^2}$ (Chatillon, 1984). Applying Equation 13-15, this technique produced Multivariate Cross-Correlation and can be stated as in Equation 20 $(sig\_del = \delta_\Delta)$: :

$$sig\_del = \delta_\Delta = \sqrt{\frac{\sum_{j=1}^{k} \Delta_j^2 - k\left\{\prod_{j=1}^{k+1} \Delta_j\right\}}{\sum_{j=1}^{k+1} \Delta_j^2}}, 0 \leq \delta_\Delta \leq 1. \quad (20)$$

The values of Δ_j play a vital role in determining whether there is an association between the variables. Equation 20 detects an association if $\Delta_j < 1$ and $\propto_j \neq 0$. On the other hand, Equation 20 detects no association if $\Delta_j = 1, \propto_j = 0,$ and $\Delta_j = 0$ if $\propto_j = 1$. If the definitions of $\propto_j$ are achievable for the first instance, that is $\propto_j \neq 0,$ the association of the variables can be determined uniquely. On the other hand, if $\propto_j = 0, 1,$ this implies no association between the variables. It is pertinent to state that the value of $\propto_j$ is always positive. A similar computational framework has been expressed in Tan et al. (2011). Therefore, the following is true (Equation 21).

$$\delta_\Delta = \begin{cases} \leq 1, if \ \propto_j \neq 0 \\ 0, if \ \propto_j = 0,1 \end{cases} \quad (21)$$

Multivariate Cross-Correlation Based on Deviation Method

Let *X*, *Y* adn *Z* be as defined, then the deviation method is stated as Equation 22-24:

$$Z = (X - \bar{X}), \bar{X} = \frac{\sum_{i=1}^{l} x_i}{l} \quad (22)$$

$$B = (Y - \bar{Y}), \bar{Y} = \frac{\sum_{i=1}^{l} y_i}{l} \quad (23)$$

$$W = (Z - \bar{Z}), \bar{Z} = \frac{\sum_{i=1}^{l} z_i}{l} \quad (24)$$

Based on the deviations and invoking the concept defined in Geiss et al. (1991) and Urain and Peters (2019), the following argument applies and mimics Pearson's procedure. For each $C_j, j = 1,2,3,$ we have Equation 25-27.

$$C_1 = \frac{Z^T B}{ZW^T}, ZW^T \neq 0 \quad (25)$$

$$C_2 = \frac{Z^T W}{BW^T}, BW^T \neq 0 \quad (26)$$

$$C_3 = \frac{B^T W}{Z^T B}, Z^T B \neq 0 \quad (27)$$

The values of C_j, j strictly depends on the deviations from the variable of interest. Hence we have Equation 28.

$$\gamma = \sum_{j=1}^{k} c_j^2 - k\left(\prod_{j=1}^{k+1} c_j\right) \text{ and } \delta = \sum_{j=1}^{k+1} c_j^2. \quad (28)$$

Applying the definitions in Equation 28 we obtain Multivariate Cross-Correlation Based on Deviation Method ($\text{delta} = \nabla$) as stated in Equation 29

$$\text{delta} = \nabla = \sqrt{\frac{\sum_{j=1}^{k} c_j^2 - l\left(\prod_{j=1}^{k+1} c_j\right)}{\sum_{j=1}^{k+1} c_j^2}}, 0 \leq \nabla \leq 1 \quad (29)$$

The value of ∇ is always positive. Therefore Equation 29 is true if the following conditions hold as in Equation 30

$$\nabla = \begin{cases} \leq 1, if\, C_j \neq 0 \\ 0, if\, C_j = 1. \end{cases} \quad (30)$$

Therefore, the limits [0,1] for the proposed procedures are established based on the conditions explicitly stated in Equation 21 and 30. The proposed methods are insensitive against influential observations.

Multivariate Coefficient of Determination

In some situations, the correlation value is often questioned as lacking interpretations for its measurement unit as such; the correlation value is squared. This squared correlation

R^2

value is called the coefficient of determination. This can be described as the proportion of the variance in one variable explained by the differences in the other variables (Nagelkerke, 1991; Nakagawa et al., 2017; Pyrczak et al., 2018). The strength of the different multivariate correlation techniques discussed in this paper depends on the multivariate coefficient of determination R^2 (Equation 31).

$$R^2 = \left(r_{\_dxyz}\right)^2, 0 \leq R^2 \leq 1 \tag{31}$$

Equation 31 denotes the fraction of the sum of interaction explained by the methods discussed in this paper. In this case, R^2 close to one show superior interaction or association between the independent and dependent variables. On the other hand, if R^2 is close to zero it implies that the level of interaction between the variables is not suitably explained. It is vital to note that the univariate coefficient of determination and multivariate coefficient of determination play a similar explanatory role. Explaining Equation 31 further implies that $|R^2| < r_{_dxyz}$may signify a closer association between the independent and dependent variables than R^2 . However, R^2 allows for the comparison of the level of association (Asuero et al., 2006).

In this study, it is proper for us to investigate and compare the Pearson coefficient of determination from the regression approach. The formula in Equation 32 is applied to determine the coefficient of determination for Equation 8.

$$r_{y\hat{y}}^2 = 1 - \frac{\sum_{k=1}^{l}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{l}(y_k - \bar{y}_k)^2}, 0 \leq r_{y\hat{y}}^2 \leq 1 \tag{32}$$

Where $\hat{y}_k$ denotes the predicted values of y_k based on the information provided by the regression model, and $\bar{y}$ is the mean of the response variable. Equation 32 was compared with the usual technique of squaring *r* values. This argument was considered vital because Equation 8 yields the same results as Pearson's technique. Unfortunately, the coefficient of determination computed by applying Equation 32 is equal to the coefficient of determination obtained by squaring the Pearson's *r*.

Adjusted Multivariate Coefficient of Determination

The adjusted multivariate coefficient of determination is stated as Equation 33.

$$Adjusted\ R^2 = 1 - \frac{(l-1)}{l-(k+1)}(1 - R^2) \tag{33}$$

where *l* denotes the sample size, and *k* is the number of independent variables, and R^2 denotes the multivariate coefficient of determination. The basic objective of using Equation

33 is to minimise positive estimation bias (Huberty, 2003). Minimisation of bias is assumed large depending on the sample size divided by the independent variables' dimension, that is (n/p). In general, Equation 33 gives a true estimate and more reliable than the multivariate correlation coefficient of whatever derivation.

Data Collection

The Covid-19 virus is a global problem, and some countries have experience in managing pandemics while others do not. In this study, we used the data set from two countries: Malaysia and Nigeria. These countries are both developing but they have similar weather or climate conditions, different continents, and geographical location but differ in population size. Unfortunately, Covid-19 does not recognise the lifestyle. We use Covid-19 data from these two countries because of easy access to data, similar lockdown measures, and the use of non-pharmaceutical interventions to combat the outbreak of the virus. The uses of the data set from these countries provide more information on the risk associated with the virus and non-compliance with safety protocols.

Case 1. Malaysian Covid-19 data set. The history of Covid-19 in Malaysia started on 25th January 2020 when the Covid-19 case was first reported in Malaysia. The data for this study was collected from 25th January to 31st July 2020 based on recorded data in the portal of the Ministry of Health Malaysia (KKM, 2020). This represents 189 days of the Covid-19 outbreak in Malaysia. As of 31st July 2020, the number of confirmed cases is 8,976, discharged cases are 8,645, representing 96.31% recovering, and 125 death cases representing 1.39% with 2.3% active cases.

Case 2. Nigerian Covid-19 data set: The data set are the daily recording of the Covid-19 outbreak in Nigeria which was first reported on the 27th of February 2020 to 4th August 2020 by the Nigeria centre for disease control (NCDC, 2020). The data consist of 163 days of recording with total confirmed cases (44,433), discharged cases (31,851) representing 71.68%, and death cases (910) representing 2.05%. This implies that as of 4th August 2020, Nigeria has approximately 11,672 (26.27%) active cases. Table 1 consists of variables of interest in this study.

Table 1
Study variables

Independent Variable	Confirmed cases
Dependent variables	Discharged and death cases

RESULTS AND DISCUSSIONS

In this section, we present the values of the multivariate correlation (R), coefficient of determination (R^2), and the adjusted coefficient of determination (Adjusted R^2). The Pearson correlation technique performance is evaluated based on the ±1, while the multivariate correlation technique' is evaluated between 0 and +1. The implication of this is that if R = *0*, it implies no association between the variables, and if R = 0.1,0.2,0.3,0.4 it implies weak association, and if R = 0.5,0.6,0.7,0.8,0.9,1.0 it implies moderate to strong association of the study variables. These performance evaluation procedures were adopted due to consistent and reliability to analyse situational data. However, as the value of R approaches one, the difference between R, R^2 and Adjusted R^2 becomes immaterial. In Figure 1, we used the Covid-19 data set from Malaysia for the period under review. The performance of the conventional multivariate correlation and the proposed techniques are compared.

In Figure1, the paired multivariate correlation (r_{xyz}), Equation 5 and the median-based paired multivariate correlation (r_{dxyz}), Equation 6 are defined, respectively. Equation 6 is a modification of Equation 5. At the same time, $r_{y\hat{y}}$ and τ^2 are the multivariate regression techniques, that is Equation 8 and Equation 12. The proposed methods are the multivariate cross-correlation technique (sig_del, Equation 20) and the multivariate cross-correlation based on the deviation method (delta, Equation 29). In Figure 1, the proposed methods (Equation 20 & 29) showed that there was a strong association between confirmed, discharged and death cases, the implication is that if a case is confirmed, two reasonable possibilities exist, that is the case may either be discharged or classified as dead. The modified approach (Equation 6) indicates a strong association between the study variables, and Equation 5 and Equation 8 show similar performance, and Equation 12 shows a weak association between the study variables. We observed that the regression techniques (Equations 8 & 12) and the conventional technique (Equation 5) are affected by influential observations (random outliers). This is justified because outliers often affect the mean, and the median is resistant against outliers. The implication is that the mean base techniques are affected when a high number of confirmed or discharge or death cases are reported. In this data set, influential observations or random outliers implies the day with the highest number of confirmed or discharge or death cases against the usual pattern of infection or discharge or death cases. From Figure 1, the comparative analysis indicates that the proposed techniques are robust against random outliers and hence showed a stronger variable association than the conventional techniques. Figure 2 consists of the Covid-19 data set from Nigeria. The performance analyses of the different methods based on the Nigeria Covid-19 data set showed similar comparative performance discussed for the Malaysian Covid-19 data in Figure 1.

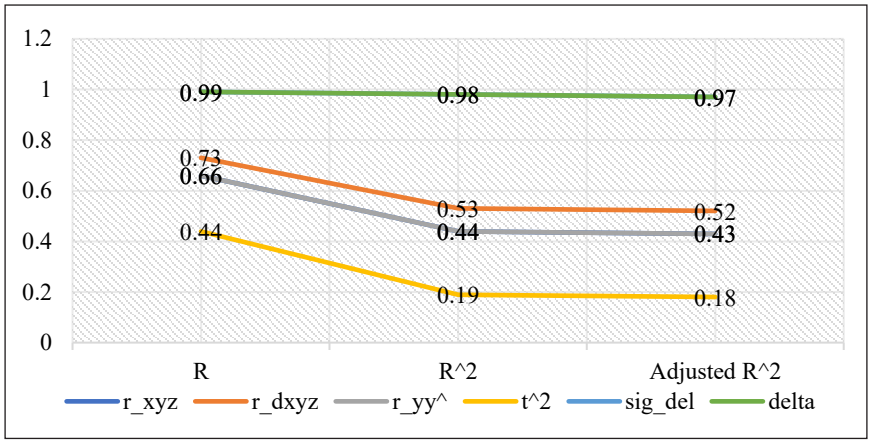

Figure 1. Comparative analysis of different Multivariate correlation techniques (Malaysia)

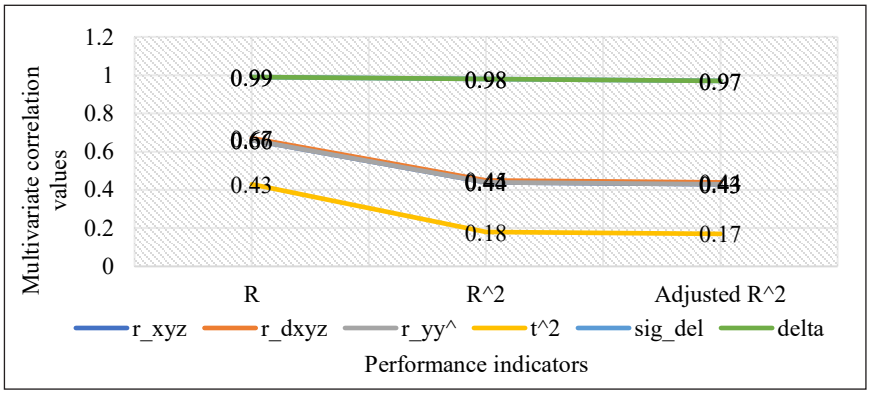

Figure 2. Comparative analysis of different Multivariate correlation techniques (Nigeria)

In Figure 1 and 2 above, τ^2 has the lowest coefficient of determination (R^2) and Adjusted R^2 values. The proposed methods based on R^2 and Adjusted R^2 showed strong association indicating that when somebody was infected with the Covid-19 virus, there was a strong association (relationship) that the person might be discharged or die of the virus. However, the other methods apart from the multivariate regression method τ^2 showed a moderate degree that an infected person might either be discharged or die. This indicates that the Covid-19 data set from the two different continents and countries have similar characteristics. The implication of R^2 = 0.98 for the proposed techniques showed that these two methods explained the data set very well.

The conventional methods imply that when someone is confirmed positive, there is a slightly moderate to strong association that the person may be discharged or die of the virus. However, the degree of association to be discharged or dead differ with the different

techniques. The proposed method showed a strong association for the confirmed person to be discharged or confirmed dead. Although, the other methods differ in association because R, $R\hat{}2$ and Adjusted $R\hat{}2$ differ strongly. We observed from Figure 1 and 2 that $R\hat{}2$ values are greater than the values of Adjusted $R\hat{}2$. From the study, we observed that *sig_del* and *delta* performed similarly for both data set. For the Nigeria data set, $r_{\_xyz}$ and $r_{\_dxyz}$ performed comparably.

The comparative analysis showed that there were possibilities that any confirmed case might either be discharged or die of the virus. The study had shown that there was a strong to a weak association for the study variables. The study also showed that the confirmed and discharged cases were strongly associated hence there was a minimum number of death cases for both countries. This affirmation concurred with the number of deaths recorded.

In Figure 3, we observed that the discharged cases were 8,770 (97.7%), including death cases for the period under consideration. This implies that about 206 (2.3%) active cases were still receiving treatment in the country's health facilities. While in Figure 4, about 32,761 (73.7%), including death cases had been discharged, and about 11,672 (26.3%) active cases remained in the country's health facilities. In Figure 3 and 4, we observed that the number of discharged cases was approaching the number of confirmed cases. The trend in Figures 3 and 4 continues for the two countries.

The total confirmed cases for the two countries for the period under review were 53,105 with 83.2% cases reported in Nigeria and 16.8% cases reported in Malaysia. Malaysia has recorded 96.3% recovering on country-wise analysis than Nigeria with 71.7% analysed

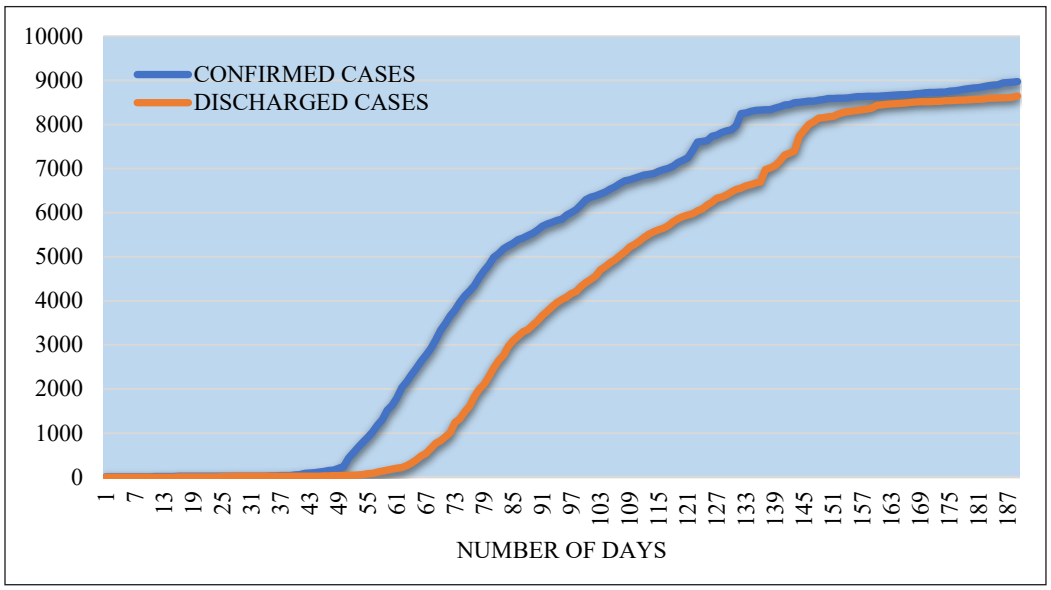

Figure 3. Comparative analysis of confirmed and discharged cases of Covid-19 in Malaysia

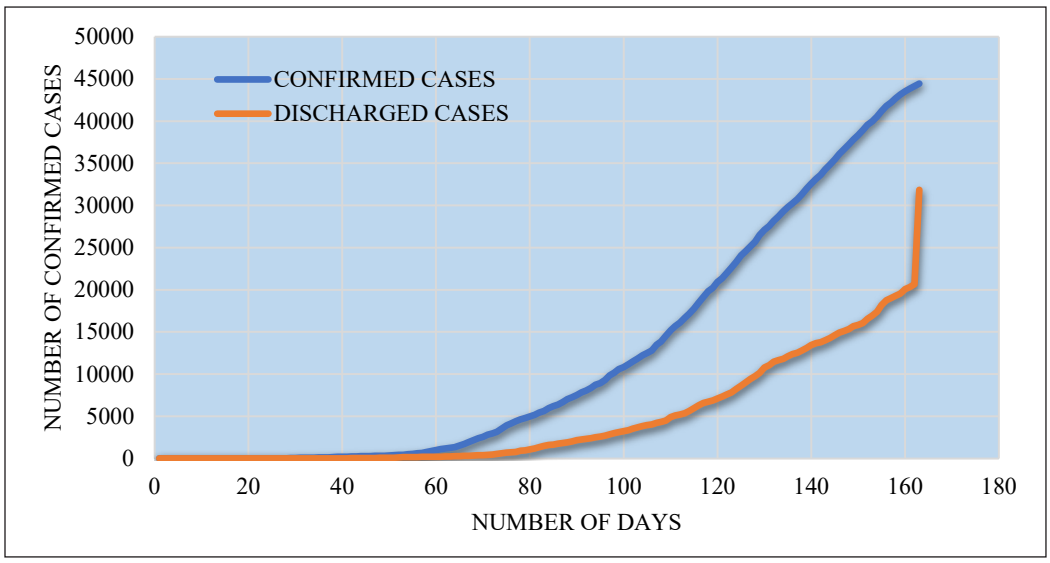

Figure 4. Comparative analysis of confirmed and discharged cases of Covid-19 in Nigeria

using similar conditions for the period under review. As at 31st July 2020, the active cases based on the data source (KKM, 2020), Malaysia had 206 (2.3%) active cases while as at 4th August 2020, Nigeria (NCDC, 2020) had 11,672 (26.3%) active cases (number of deaths inclusive). Based on the results and analysis above, this study confirmed a strong association between confirmed and discharged cases. This implies a fragile association between the confirmed and death cases.

In general, without vaccines and recommended drugs to treat Covid-19 patients, the two countries studied have done remarkably well to fight the pandemic. Based on the current circumstances whereby on availability of drugs and vaccines, it is better to examine the level of association between the confirmed, discharged, and death cases. Therefore, due to the nature of global data on daily confirmed, discharged and death cases, Covid-19 data may not be normally distributed. Using conventional methods to determine the association level may indicate a lower level of association due to the outlier effects on conventional methods. Apart from the proposed method, every other method, especially the multivariate regression technique is heavily affected by the presence of influential observations. Therefore, we may conclude that due to the present scenario surrounding Covid-19 management, robust and precise procedures be applied to determine the level of association to enable a caregiver to infer that there is a strong association that Covid-19 infected person may recover or die. Therefore, this study confirmed that a strong association between the confirmed and recovered cases implied that the number of casualties might be small depending on the infected cases.

CONCLUSIONS

The comparative analysis of the proposed methods and the conventional methods showed that the proposed methods were insensitive to influential observations or random outliers than the conventional methods. The study based on these methods showed moderate to a strong association between confirmed, discharged, and death cases. It implies that someone infected with the virus can only be discharged or die. Although the different methods revealed different degrees of association, the proposed methods showed a robust association indicating that the infected persons could either recover or die. Other techniques showed a different association level based on the multivariate correlation, multivariate coefficient of determination, and adjusted coefficient of determination. Therefore, there is a very weak association between confirmed and death cases and a very strong association between confirmed and discharged cases. The study affirmed that the strong association between confirmed and discharged cases implies a minimal number of deaths recorded in both countries, thereby confirming a high survival rate and minimal death rate. Therefore, it is recommended that the method with strong association be accepted due to the lack of vaccine and globally accepted drugs to treat Covid-19 for effective case management. These strengths of the association will encourage people to adopt non-pharmaceutical interventions such as the "Movement control order" in Malaysia, social distancing, and lockdown in Nigeria. However, these two countries still maintain the same trend of infection, discharge, and low death rate.

ACKNOWLEDGEMENT

We would like to express our gratitude to Universiti Utara Malaysia and Delta State University for the support to carry out this research work. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Abdi, H. (2007). Multiple correlation coefficient. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 648-651). Sage Publication.

Abdullah, M. B. (1990). On a robust correlation coefficient. *Journal of the Royal Statistical Society: Series D (The Statistician), 39*(4), 455-460. https://doi.org/10.2307/2349088

Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics, 39*(5), 316-327. https://doi.org/10.1111/opo.12636

Asuero, A. G., Sayago, A., & Gonzalez, A. G. (2006). The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry, 36*(1), 41-59. https://doi.org/10.1080/10408340500526766

Bareinboim, E., Tian, J., & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 28, No. 1). PKP Publishing Services Network.

Brown, G., Pocock, A., Zhao, M. J., & Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research, 13*(1), 27-66.

Châtillon, G. (1984). The balloon rules for a rough estimate of the correlation coefficient. *American Statistician*, *38*(1), 58-60. https://doi.org/10.1080/00031305.1984.10482875

Garnett, J. C. (1919). General ability, cleverness and purpose. *British Journal of Psychology, 9*(3), 345-366.

Geiß, S., & Einax, J. (1996). Multivariate correlation analysis - A method for the analysis of multidimensional time series in environmental studies. *Chemometrics and Intelligent Laboratory Systems*, *32*(1), 57-65. https://doi.org/10.1016/0169-7439(95)00067-4

Geiss, S., Einax, J., & Danzer, K. (1991). Multivariate correlation analysis and its application in environmental analysis. *Analytica Chimica Acta*, *242*, 5-9. https://doi.org/10.1016/0003-2670(91)87040-E

Huberty, C. J. (2003). Multiple correlations versus multiple regression. *Educational and Psychological Measurement*, *63*(2), 271-278. https://doi.org/10.1177/0013164402250990

Lewis-Beck, M. S., Bryman, A., & Futing Liao, T. (2004). *The SAGE Encyclopedia of Social Science Research Methods* (Vols. 1-0). Sage Publications, Inc. https://doi.org/10.4135/9781412950589

KKM. (2020). *Distribution of covid-19 cases according to date of confirmation*. Retrieved October 01, 2020, from http://covid-19.moh.gov.my/

Mukaka M. M. (2012). Statistics corner: A guide to the appropriate use of the Correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69-71.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. https://doi.org/10.1093/biomet/78.3.691

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), Article 20170213. https://doi.org/10.1098/rsif.2017.0213

NCDC. (2020). *The official Twitter account of the Nigeria Centre for Disease Control.* Retrieved June 19, 2020, from https://twitter.com/ncdcgov

Nguyen, H. V., Müller, E., Vreeken, J., Keller, F., & Böhm, K. (2013). CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 198-206). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972832.22

Okwonu, F. Z., Asaju, B. L., & Arunaye, F. I. (2020, September). Breakdown analysis of pearson correlation coefficient and robust correlation methods. In *IOP Conference Series: Materials Science and Engineering* (Vol. 917, No. 1, p. 012065). IOP Publishing. https://doi.org/10.1088/1757-899X/917/1/012065

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25-45. https://doi.org/10.1093/biomet/13.1.25

Pyrczak, F., & Oh, D. M. (2018). *Making sense of statistics: A conceptual overview* (7th ed.). Routledge.

Rodgers, L. J., & Nicewander, W. L. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66. https://doi.org/10.1080/00031305.1988.10475524

Tan, Z., Jamdagni, A., He, X., Nanda, P., & Liu, R. P. (2011). Denial-of-service attack detection based on multivariate correlation analysis. In *International Conference on Neural Information Processing* (pp. 756-765). Springer. https://doi.org/10.1007/978-3-642-24965-5_85

Urain, J., & Peters, J. (2019). Generalized multiple correlation coefficient as a similarity measurement between trajectories. In *IEEE International Conference on Intelligent Robots and Systems* (pp. 1-7). IEEE Conference Publication. https://doi.org/10.1109/IROS40897.2019.8967884

Wang, J., & Zheng, N. (2020). Correlation with applications (1): Measures of correlation for multiple variables. In *IEEE International Conference on Intelligent Robots and Systems* (pp. 1-18). Cornell University Press.

Wang, L., Tang, X., Zhang, J., & Guan, D. (2018). Correlation Analysis for Exploring Multivariate Data Sets. *IEEE Access, 6*, 44235-44243. https://doi.org/10.1109/ACCESS.2018.2864685

Wang, Y., Romano, S., Nguyen, V., Bailey, J., Ma, X., & Xia, S. T. (2017). Unbiased multivariate correlation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1). PKP Publishing Services Network.

Weida, F. M. (1927). On various conceptions of correlation. *The Annals of Mathematics*, *29*(1/4), 276-312. https://doi.org/10.2307/1968000

Zhang, X., Pan, F., Wang, W., & Nobel, A. (2008). Mining non-redundant high order correlations in binary data. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* (Vol. 1, No. 1, p. 1178). NIH Public Access. https://doi.org/10.14778/1453856.1453981