ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

International Journal of Statistics and Probability

Vol. 11, No. 5 September 2022

CANADIAN CENTER OF SCIENCE AND EDUCATION

INTERNATIONAL JOURNAL OF STATISTICS AND PROBABILITY

An International Peer-reviewed and Open Access Journal for Statistics and Probability

International Journal of Statistics and Probability (ISSN: 1927-7032; E-ISSN: 1927-7040) is an open-access, international, double-blind peer-reviewed journal published by the Canadian Center of Science and Education. This journal, published **bimonthly** (January, March, May, July, September and November) in both **print and online versions**, keeps readers up-to-date with the latest developments in all areas of statistics and probability.

The scopes of the journal:

- Computational statistics
- Design of experiments
- Sample survey
- Statistical modelling
- Statistical theory
- Probability theory

The journal is included in:

- BASE
- Google Scholar
- JournalTOCs
- LOCKSS
- SHERPA/RoMEO
- Ulrich's

Submission Policy

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the authorities responsible where the work was carried out. However, we accept submissions that have previously appeared on preprint servers (for example: arXiv, bioRxiv, Nature Precedings, Philica, Social Science Research Network, and Vixra); have previously been presented at conferences; or have previously appeared in other "non-journal" venues (for example: blogs or posters). Authors are responsible for updating the archived preprint with the journal reference (including DOI) and a link to the published articles on the appropriate journal website upon publication.

iThenticate Professional Plagiarism Prevention

The publisher and journals have a zero-tolerance plagiarism policy. We check the issue using two methods: a plagiarism prevention tool (iThenticate) and a reviewer check. All submissions will be checked by iThenticate before being sent to reviewers.

creative commons

We insist a rigorous viewpoint on the self-plagiarism. The self-plagiarism is plagiarism, as it fails to contribute to the research and science.

IJSP accepts both Online and Email submission. The online system makes readers to submit and track the status of their manuscripts conveniently. For any questions, please contact ijsp@ccsenet.org.

International Journal of Statistics and Probability; Vol. 11, No. 5; September 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Editorial Team

International Journal of Statistics and Probability; Vol. 11, No. 5; September 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Contents

International Journal of Statistics and Probability; Vol. 11, No. 5; Setepmber 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Application of Logistic Regression on Heart Disease Data and a Review of Some Standardization Methods

Florence George[1], Sultana Mubarika Rahman Chowdhury[1] & Sneh Gulati[1]

[1] Mathematics and Statistics Department, Florida International University, Miami, Florida, USA

Correspondence: Florence George, Mathematics and Statistics Department, Florida International University, Miami, Florida, 33199, USA

Received: July 17, 2022 Accepted: August 25, 2022 Online Published: August 29, 2022

doi:10.5539/ijsp.v11n5p1 URL: https://doi.org/10.5539/ijsp.v11n5p1

Abstract

The purpose of this study is to do a review of logistic regression and its applications. In addition to the review, a comparison of four different methods of standardization of the β - coefficients was done using publicly available Heart Disease Data. The methods were compared using their performance in testing accuracy, training accuracy, and area under the curve (AUC). Based on the comparisons, it was evident that standardizing the coefficient did not affect the overall prediction accuracy of the model regardless of the method used. Although there was some difference found in the training and testing accuracies, the AUC's were similar to the unstandardized model for all methods. In essence, standardizing facilitates better interpretation and does not affect the predictive accuracy of the model.

Keywords: Logistic Regression, Logit model, Standardized Coefficients

1. Introduction

Logistic regression analysis is a specialized case of regression analysis, where the variable to be predicted is classified into two or more categories. In such cases, the traditional regression technique fails to explain the association between the independent variables and the response variable. Binary logistic regression model or logit model is the most common form of this method of analysis in which the response variable takes only two values (Menard, 2000).

The specific form of a binary logistic regression model generally used is

$$P(Y=1) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}} = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p)}}, \tag{1}$$

where Y is the dependent variable and $X_1, X_2, ..., X_p$ are the independent variables. The dependent variable Y takes on the values either 0 or 1; where 1 indicates the occurrence of a specific event and 0 indicates the absence. Therefore, $P(Y = 1)$ represents the probability of that event happening and $P(Y = 0)$ depicts the probability of the event is absent.

Logistic regression has a wide range of applications in various fields and its functionality has increased dramatically in the past several decades. While multiple linear regression falls short in analyzing data with response variable that is not continuous, logistic regression gives an essential tool is such cases. Application of this method is not limited to only binary cases as it can be easily modified for cases where response variables have more than two categories. Risk factor analysis and predictive modeling is one of the main implementations of logistic regression (Peterson, L. E. et al., 1995). Logistic regression can also be used in survival analysis by grouping event times into intervals and converting them to categories (Abbott, 1985). Hence, is broadly used in medical research fields to examine the association between risk factors and diseases (Kurt, I. et al., 2008; Hassanipour, S. et al., 2019).

The parameters, the standard error of the parameters, and the measures of the goodness of it are estimated using the methods of maximum likelihood estimation (Greene, 1993, Peng et al., 2002)

The logit transformation of $P(Y = 1)$ is defined as

$$logit(Y) = \ln\left(\frac{p(Y=1)}{p(Y=0)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_pX_p. \tag{2}$$

The model converts the nonlinear relationship between $P(Y = 1)$ and the independent variables to a linear equation that explains the effect they may have on the dependent variable. This linear form gives the opportunity to interpret the coefficients of the proposed model.

The interpretation of results is rendered using the widely used odds ratio technique for both categorical and continuous predictors (Peng et al., 2002). Even though the odds ratio can give an idea of the direction of the relationship between the response variable and explanatory variables, it is not enough to explain the overall extent of how they are related and also it falls short of comparing over models (Allison, 1999). It should also be noted that some alternate methods based on the effect measures are proposed in several papers to explain the effects of covariates on binary response variables in a logistic regression model (Agresti A., Kateri M.,2017; Agresti A., Tarantola C.,2018).

However, the primary focus of this paper is on the β-coefficients and does not investigate these alternate methods. S-tandardizing the β -coefficients is another approach found in various literature studies (Long J. S., 1997; Menard S.W. , 1995) , and different techniques to standardize the β-coefficients have also been proposed to allow for more meaningful interpretations. Standardized coefficients become invariant to the change in scale of measurement which enables one to compare the relative influence of different explanatory variables within logistic regression (Agresti A., 2018; Agresti A. , Finlay B., 1997). However, even though there are some proposed standardized, semi-standardized coefficients for logistic regression none of them can be universally defined. Robert L. Kaufman (Kaufman R.L. ,1996) in his study found that semi-standardized coefficients measuring the change in predictive probability of outcomes are preferable because they are intuitively appealing and as they are bounded in the interval [-1, +1], interpretation of their magnitude becomes easier. Some approaches of standardizing the coefficients were analyzed using a practical example by Scott Menard, which included both semi-standardized and completely standardized techniques (Menard S. , 2004).

In this paper, we will discuss the four methods discussed by Menard and in addition to that, we propose a modification of these four methods for standardization of logistic regression coefficient. These methods will also be compared based on the resulting testing accuracy, training accuracy, AUC (area under the curve).

The simplest method of partial standardization of logistic regression coefficients is to multiply the coefficients by their individual standard deviation. This method was mentioned by Menard (Menard S.W., 1995).

$$b_1 = b * S_x, \tag{3}$$

where, the standard deviation of the explanatory variable X (S_x) is multiplied with the unstandardized estimated coefficient of the corresponding variable b. This can be considered as the only predictor-based standardization technique. Another similar approach is to change the scale of both the dependent variable and the predictors using the standard deviation of the standard logistic distribution. That is,

$$b_2 = \frac{b * S_x}{\frac{\pi}{\sqrt{3}}}, \tag{4}$$

where, $(\pi\sqrt{3}) = 1.8$. This method has been adapted in SAS to standardize the coefficients in the PROC LOGISTIC procedure. Long suggested another approach for standardization which includes the standard deviation of the standard normal distribution (Long J. S., 1997).

The calculation of this method is similar to the previous one, the only difference is the standard deviation of the standard normal distribution is added with the standard deviation of the logistic distribution. Hence Equation (4) becomes,

$$b_3 = \frac{b * S_x}{\frac{\pi}{\sqrt{3}} + 1}. \tag{5}$$

All of these standardized coefficients only take into account the variation of the independent variable. Hence, they cannot be considered as fully standardized. To standardize the response variable standard deviation of *logit (y)* needs to be calculated, which is tricky. A way out of this is to use the standardization followed in OLS, which is defined as follows,

$$b^{**} = b * \frac{Sx}{Sy}.$$

Again, from the definition of Coefficient of Determination (R^2), we get

$$R^2 = \frac{S^2_{\hat{y}}}{S^2_y},$$

where, $\widehat{y}$ is the estimated value of y. Adjusting the equation for OLS we get,

$$S_y^2 = \frac{S_{\hat{y}}^2}{R^2},$$

Substituting *logit(y)* in case of y and *logit(ŷ)* in the place of $\hat{y}$ we get for logistic regression,

$$S_{logit(y)}^2 = \frac{S_{logit(\hat{y})}^2}{R^2}.$$

Hence, using the similar strategy used in OLS the estimated coefficients can be standardized as follows

$$b_4 = \frac{(b * S_x)(R)}{S_{logit(\hat{y})}}. \tag{6}$$

This coefficient can be considered as fully standardized as it also takes into account the variance of the response variable in contrast to the other coefficients discussed before where only the variation of the predictor was studied. For the purpose of comparing the above four standardization methods, they will be applied to z- scaled data using the mean and standard deviation. Since median and MAD may be better measures for scaling asymmetric data, we propose applying these standardization techniques to the median and the MAD scaled data.

In the next section, these standardized logistic regression coefficients for both z-scaled and median/MAD scaled data will be compared by applying the methods to Heart Disease Data.

2. Implementation of Standardization Methods

In order to illustrate the calculation of the standardization techniques and to review the outcomes, the Cleveland Heart disease dataset was used. It is a widely used dataset that is publicly available online (Detrano R., 1989). The aim was to apply logistic regression to develop a predictive model for heart diseases using the predictors. The four different coefficient standardization methods were applied to the coefficients of the customary model. After that, the resultant models were compared based on their prediction accuracy.

2.1 Dataset Details

Originally, the data set contained 76 attributes, but a subset of 14 variables are generally used by the researchers in all published experiments with a total of 313 observations. The 14 variables include a response variable ”target” which refers to the presence of heart disease in the patient. For the target variable, a value of 0 indicates no/ less chance of heart attack while a value of 1 indicates yes/ high chance of heart attack.

The 13 predictors considered in the dataset are as follows (Detrano R., 1989):

1. *AGE*: Continuous
2. *SEX*: Categorical (0 = Female, 1 = Male)
3. *Chest Pain Type(CP)*: Categorical (4 values) 0: typical angina 1: atypical angina 2: non - anginal pain 3: asymptomatic
4. *Trestbps*: Continuous, represents resting blood pressure on admission
5. *Chol*: Continuous, represents Serum cholesterol in mg/dl
6. *Fbs*: Categorical , represents fasting blood sugar level, (2 values) 1: True - fasting blood sugar is greater than 120 mg/dl 0: False - fasting blood sugar is less than 120 mg/dl
7. *Restecg*: Categorical,represents resting electrocardiographic outcomes (4 values) 0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes’ criteria)
8. *Thalach*: Continuous, represents maximum heart rate achieved
9. *Exang*: Categorical, represents the existence of exercise-induced angina (2 values Yes/No)
10. *Oldpeak*: Continuous, ST depression induced by exercise relative to rest
11. *Slope*: Categorical, represents the slope characteristics of the peak exercise ST segment

12. *Ca*: Discrete, represents the number of fluoroscopy colored major vessels (values 0-3);

13. *Thal*: Categorical, (3 values) 0: normal 1: fixed defect 2: reversible defect

2.2 Methodology and Results

Primarily, logistic regression was applied to the complete dataset. Four standardization techniques of the coefficients discussed in the previous section were applied to this result. Calculation of b_1 is done by simply multiplying the standard deviation of each explanatory variable with their corresponding coefficients. For instance, for *Age* b_1 = (-0.004908) * (9.0821010) = -0.04457922 and so on Table **??**.

Table 1. Modified coefficients using different standardization methods

	Customary model (Pvalue)	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Intercept	3.4505				
Age	- 0.0049(0.8323)	- 0.0446	- 0.0246	- 0.0158	- 0.0110
Sex	- 1.7582 (0.0002)	- 0.8193	- 0.4517	-0.2912	- 0.2019
Cp	0.8599 (0.000)	-0.8874	0.4893	0.3154	0.2189
Trestbps	-0.0195 (0.0596)	-0.3416	-0.1883	-0.1214	-0.0842
Chol	-0.0046 (0.2209)	-0.2400	-0.1323	-0.0853	-0.0591
Fbs	0.0349 (0.9475)	0.0124	0.0069	0.0044	0.0031
Restecg	0.4663 (0.1806)	0.2452	0.1352	0.0871	0.06043
Thalach	0.0232 (0.0265)	0.5317	0.2931	0.1889	0.1310
Exang	-0.9800 (0.0168)	-0.4604	-0.2538	-0.1636	-0.1135
Oldpeak	-0.5403 (0.0115)	-0.6273	-0.3458	-0.2229	-0.1546
Slope	0.5793 (0.0977)	0.3570	0.1968	0.1269	0.0880
Ca	-0.7733 (0.0000)	-0.7908	-0.4360	-0.2811	-0.1949
Thal	-0.9004 (0.0019)	-0.5513	-0.3040	-0.1959	-0.1359

Table 2. Logistic regression coefficients (Mean/SD scaled data)

	Customary model (Pvalue)	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Intercept	0.2319				
Age	-0.0419 (0.8323)	-0.0419	-0.0231	-0.0365	-0.0101
Sex	-0.8188 (0.0002)	-0.8172	-0.4505	-0.7106	-0.1966
Cp	1.0425 (0.0000)	1.0317	0.5688	0.8972	0.2483
Trestbps	-0.2409 (0.0596)	-0.2340	-0.1323	-0.2087	-0.0577
Chol	-0.2510 (0.2209)	-0.2297	-0.1266	-0.1997	-0.0553
Fbs	-0.0730 (0.9475)	-0.0755	-0.0416	-0.0657	-0.0182
Restecg	0.3668 (0.1806)	0.3711	0.2046	0.3228	0.0893
Thalach	0.3420 (0.0265)	0.3385	0.1866	0.2944	0.0815
Exang	-0.4276 (0.0168)	-0.4304	-0.2373	-0.3743	-0.1036
Oldpeak	-0.5950 (0.0115)	-0.6236	-0.3438	-0.5423	-0.1501
Slope	0.5568 (0.0977)	0.5641	0.3110	0.4905	0.1357
Ca	-0.7673 (0.0000)	-0.7983	-0.4402	-0.6943	-0.1921

Table 2 continued from previous page

	Customary model (Pvalue)	Method 1	Method 2	Method 3	Method 4
Thal	-0.5539 (0.0019)	-0.5676	-0.3129	-0.4936	-0.1366

To get b_2, (Equation 4) above result has to be divided by $\pi\sqrt{3}$, the numerical value of which is approximately 1.814. Hence, for *Age* the standardized coefficient becomes b_2 = (-0.004908) * (9.0821010)/ 1.814 = -0.02457. To obtain the standardized coefficient by the third method (Equation 5) discussed in the previous section, the calculation is similar but instead of dividing by $[\pi\sqrt{3}]$ the unstandardized coefficients are divided by $[\pi\sqrt{3}+1]$ which is equal to approximately 2.814. Therefore, for Age the calculation of the standardized coefficients is as follows: b_3 = (-0.004908) * (9.0821010)/ 2.814 = -0.01584. The fully standardized fourth approach utilizes the value of the coefficient of determination (R^2) to calculate the modified coefficients. This method, multiplies the first approach explained in equation 3 by $R/S_{logit(\hat{y})}$. In this example, the value of the square root of R^2 divided by the standard deviation of the $logit(\hat{y})$ was calculated to be 0.246434. So the modified coefficient for predictor Age changed in to b_4 = (-0.004908) * (9.0821010) * (0.246434) = -0.01098. Similar calculations have been done for all other variables and are presented in Table 1.

The column 'Customary model' in Table 1 refers to the calculated unstandardized coefficients from the logistic regression model. 'Method 1', 'Method 2', 'Method 3', and 'Method 4' represent the standardized coefficients computed using Equation 3, Equation 4, Equation 5, Equation 6 respectively. From the results in Table 1 it is evident that as the coefficients start from being partially standardized using method 1 to fully standardized in method 4, they seem to decrease in terms of magnitude. Techniques used in SAS have the closest values to the method suggested by Long. Predictor cp (chest pain) seems to have a comparatively higher relative effectiveness among the significant variables.

Table 3. Logistic regression coefficients (Median/MAD scaled data)

	Customary model (Pvalue)	Method 1	Method 2	Method 3	Method 4
Intercept	0.6920				
Age	0.0650 (0.9446)	0.0844	0.0465	0.0734	0.0177
Sex	-0.8415 (0.0007)	-0.8398	-0.4630	-0.7303	-0.1760
Cp	1.1343 (0.0000)	1.1226	0.6189	0.9762	0.2353
Trestbps	-0.1610 (0.5814)	-0.2322	-0.1280	-0.2019	-0.0487
Chol	0.1072 (0.9692)	0.1527	0.0842	0.1328	0.0320
Fbs	0.0673 (0.7820)	0.0696	0.0384	0.0606	0.0146
Restecg	0.1261 (0.7370)	0.1276	0.0703	0.1109	0.0267
Thalach	0.6672 (0.0004)	0.9032	0.4979	0.7854	0.1893
Exang	-0.7237 (0.0004)	-0.7284	-0.4016	-0.6335	-0.1527
Oldpeak	-0.8915 (0.0000)	-1.1391	-0.6280	-0.9906	-0.2387
Slope	0.9685 (0.0002)	0.9811	0.5409	0.8532	0.2056
Ca	-0.7982 (0.0000)	-0.8305	-0.4579	-0.7222	-0.1741
Thal	-0.6699 (0.0011)	-0.6864	-0.3784	-0.5969	-0.1439

In the next step, the target was to set up four different models using standardized coefficients calculated by these approaches and compare their performance based on prediction accuracy. To measure the prediction accuracy, the dataset was randomly divided into two sets; the testing set which contains 20% of the data and the training set which contains the rest of the data. The models were developed using the training set and the testing set was used to verify the overall accuracy. One of the major hurdles faced while setting up models to calculate their accuracies is that the predictors were measured using different scales. Hence, to make the comparison easier, the predictors were scaled before any kind of analysis was done. Firstly, all the variables were standardized using the mean and standard deviation of the corresponding independent variable. In addition, we computed standardized coefficients using the Median/ MAD standardized data.

Previously explained four methods of standardizing the coefficients were then applied to both of these scaled datasets. All

these calculations were done with the help of statistical software R. Outcomes of standardization of the coefficients are given in Table 2 and Table 3. Here in the Table 2 column 'Customary model' refers to the unstandardized coefficients of the dataset scaled by the mean and the standard deviation along with the four standardization methods for the coefficients in the following columns. Similarly, in the Table 3 column 'Customary model' refers to the unstandardized coefficients of the dataset scaled by the median and the mean absolute deviation along with the four standardization methods for the coefficients in the following columns.

2.3 Evaluation Criteria

To compare the performance of the models with different standardization techniques, we have used training accuracy and testing accuracy of the models. In predictive modeling for binary outcome variable the term accuracy refers to the fraction of correctly specified predictions made by the proposed model. The complete data is divided into two sets namely the training set and the testing set by a random split for instance in this analysis we have used 80% of the data for the training set and 20% for the test set. At first the prediction model is built on the training set and later applied on the test set to asses its prediction accuracy.

One predicament in this process is that, as the data are divided into training and testing sets randomly using R software, there is a chance of getting different results for different subsets which may result in bias. To solve this issue the complete process was repeated 1000 times and the average of these repetitions was taken for calculations of testing and training accuracy. Another criteria that is use in comparing the accuracy in binary predictive modeling is area under the Receiver operating characteristics (ROC) . The plot represents the proportion of correctly specification events versus the proportion incorrect specification of the non-events for different probability cutoff's. A high area under the ROC curve indicates a better predictive accuracy.

2.4 Results

Table 4 shows the testing accuracy and training accuracy of the models constructed by applying each of the four coefficient standardization methods along with the model of unstandardized coefficients, which is represented by the 'Customary model' column.

Results indicate that the testing accuracy of the customary model was slightly higher than all standardized models for median/MAD scaled data. However, for the mean/SD scaled data, the testing accuracy for the customary model and the models for the 4 methods were similar. Similarly, the training accuracies were somewhat similar for the unstandardized and standardized coefficients. Moreover, method 4 was seen to have the lowest prediction accuracy among all four methods. On the other hand, by comparing the testing and training accuracies for mean/SD scaled data and median/MAD scaled data it can be seen that median/MAD scaled data has approximately 4% to 5% higher accuracy overall.

Table 4. Table for Testing and Training Accuracy

Data	**Customary model**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Mean Standardized (Test set)	0.8193	0.8218	0.8193	0.8221	0.8126
Median Standardized (Train set)	0.8754	0.8767	0.8646	0.8766	0.7813
Mean Standardized (Test set)	0.8576	0.8574	0.8540	0.8569	0.8472
Median Standardized (Train set)	0.9005	0.9000	0.8859	0.8987	0.8083

However, by taking a look at the AUCs for these models in Table 5 it can be seen that even though the unstandardized model had slightly different AUCs, there was no difference in AUCs of the models constructed from different standardization techniques. This indicates that in terms of distinguishing between the two diagnostic groups, all of these models show similar performance.

In terms of improving the sensitivity or specificity of the models the standardization techniques seem to have no significant effect. As the overall accuracy for the standardized models were lower than the un-standardized one, evidently the

sensitivity ans specificity was also found to be less than the prior. Moreover, method 4 seems to have the higher sensitivity than all other models, which also means lower specificity than others.

Table 5. AUC's for Testing and training set

Data	Customary model	Method 1	Method 2	Method 3	Method 4
Mean Standardized (Test set)	0.8895	0.8899	0.8899	0.8899	0.8899
Median Standardized (Train set)	0.9230	0.9210	0.9210	0.9210	0.9210
Mean Standardized (Test set)	0.9262	0.9261	0.9261	0.9261	0.9261
Median Standardized (Train set)	0.9280	0.9279	0.9279	0.9279	0.9279

It is worth mentioning that the techniques used to scale the dataset seem to have some effect on improving the overall accuracy of the models. Test sets taken from the dataset for which the numerical variables were scaled using median/MAD standardization performed better than the one which was scaled using mean/standard deviation. For instance, for the customary model and the first three models, the testing accuracies were approximately 4% higher in the case of the dataset standardized by median/MAD Table 3. Additionally, from Table 5 it can be seen that the AUCs are slightly higher for the data which was standardized using median/MAD.

3. Discussion

The primary purpose of standardizing logistic regression coefficients is to set a ground on the basis of which the predictors can be ranked. The absolute value of the standardized coefficients enables one to order the independent variables in terms of importance. According to Menard (Menard S.W., 1995) standardized coefficients render a more precise idea than the un-standardized logistic regression coefficients. However, adapting such measures for the sake of interpretation may effect the overall performance of the model. In this study, the goal was to investigate how different standardization techniques effect the accuracy of the logistic regression model under study.

Different methods of standardizing the coefficients assist in explaining the variation in the dependent variable and allow one to compare their contributions. It was also investigated if standardizing the coefficients would change the performance of the model. From the results, it can be seen that if the standardized values are only used in the case of relative comparison of the predictors, there is not much difference between the four methods. The overall magnitude of the influence is comparatively lower for the 4^{th} method but if the influences of the predictors were ranked, the ranking was found to be the same for all four methods.

By taking a closer look at the results it can be seen that standardizing the coefficients did not affect the overall prediction accuracy of the predictive logistic regression model. Similarly, no evidence was found that following a certain type of standardization technique would show better performance than the others; the unstandardized regression model, in general, had higher accuracy. However, method 4 would be a better approach compared to others, as method 2 suggested by Long and method 3 used in SAS, partially standardizes by only considering the predictors and does not include the outcome variable in calculation. Both of these methods make little difference to the outcomes thus are not recommended.

In essence, standardizing facilitates better interpretation and does not affect the predictive capacity of the model. This is evident from the AUC's computed for both unstandardized and standardized regression coefficients showen in Table 5. As the AUCs calculated from taking the average of multiple iterations, they turned out to be exactly equal for all standardization techniques, which was also similar to the un-standardized logistic regression model.

4. Conclusion

Logistic regression facilitates a wide range of techniques in conducting statistical analyses. In logistic regression like any other regression technique, the primary aim is to construct an equation based on the set of explanatory variables, which as a whole would explain the variation and predict the dependent variable better.

Therefore, it could be inferred that standardized coefficients can also be used for predictive modeling. Similarly, selecting

any specific method for standardizing the coefficients for interpretation is completely based on how one wants to interpret it. If the primary goal of conducting a logistic regression analysis is building up a predictive model which can also be used for comparing the predictor effects and does not affect the overall accuracy of the model, standardizing the regression coefficients may be advisable.

Acknowledgment

The authors are grateful to the associate editor and the two referees for their valuable suggestions which have helped improve this paper.

References

Menard, S. W. (1995). *Applied logistic regression analysis*.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24. https://doi.org/10.2307/2685605

Greene, W. H. (1993). *Econometric Analysis.* (2nd & 4th ed.) Prentice Hall, New Jersey.

Peng C. Y. J., Lee K. L., & Ingersoll G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research, 96*(1), 3-14. https://doi.org/10.1080/00220670209598786

Allison P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological methods & research, 28*(2), 186-208. https://doi.org/10.1177/0049124199028002003

Agresti A., & Kateri M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics, 73*(1), 214-219. https://doi.org/10.1111/biom.12565

Agresti A., & Tarantola C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica, 72*(3), 210-223. https://doi.org/10.1111/stan.12130

Abbott, R. D. (1985). Logistic regression in survival analysis. *American journal of epidemiology, 121*(3), 465-471. https://doi.org/10.1093/oxfordjournals.aje.a114019

Long J. S. (1997). *Regression models for categorical and limited dependent variables.* Sage: vol. 7.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Agresti, A., & Finlay, B. (1997). *Statistical Methods for the Social Sciences*.

Kaufman, R. L. (1996). *Comparing effects in dichotomous logistic regression: A variety of standardized coefficients.* Social Science Quarterly. (p. 90-109).

Menard S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58*(3), 218-223. https://doi.org/10.1198/000313004X946

Detrano R. (1989) *Cleveland heart disease database.*: VA Medical Center, Long Beach and Cleveland Clinic Foundation.

Peterson, L. E., Nachemson, A. L., Bradford, D. S., Burwell, R. G., Duhaime, M., Edgar, M. A., ... & Willner, S. V. (1995). Prediction of progression of the curve in girls who have adolescent idiopathic scoliosis of moderate severity. Logistic regression analysis based on data from The Brace Study of the Scoliosis Research Society. *Journal of Bone and Joint Surgery-Series A, 77*(6), 823-827.

Hassanipour, S., Ghaem, H., Arab-Zozani, M., Seif, M., Fararouei, M., Abdzadeh, E., ... & Paydar, S. (2019). Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis. *Injury, 50*(2), 244-250.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications, 34*(1), 366-374.

International Journal of Statistics and Probability; Vol. 11, No. 5; September 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

In Situ Flood Frequency Analysis Used for Water Resource Management in Kelantan River Basin

Nor Hidayah Hassim[1], Basri Badyalina [1*], Nurkhairany Amyra Mokhtar[1], Muhammad Zulqarnain Hakim Abd Jalal[1], Nur Diana Binti Zamani[1], Lee Chang Kerk[1], Amir Imran Zainoddin[2], Ahmad Syahmi Ahmad Fadzil[2] & Nur Fatihah Shaari[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, Cawangan Johor, Kampus Segamat, 85000 Segamat, Johor, Malaysia

[2]Universiti Teknologi MARA Cawangan Johor, Kampus Segamat, 8500 Segamat, Johor, Malaysia

*Correspondence: Basri Badyalina, Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, Cawangan Johor, Kampus Segamat, 85000 Segamat, Johor, Malaysia.

Received: July 24, 2022 Accepted: August 24, 2022 Online Published: August 29, 2022

doi:10.5539/ijsp.v11n5p9 URL: https://doi.org/10.5539/ijsp.v11n5p9

Abstract

One of the most prevalent and traditional uses of statistics in hydrology is flood frequency analysis. The flood can occur practically everywhere and is considered the leading cause of natural disaster death worldwide. This study aims to apply the flood frequency analysis of the Kelantan streamflow site to identify the optimal distribution that best fits the flood frequency data from the goodness-of-fit test (GOF). Five distributions were applied in this study; namely lognormal (LG), generalized extreme value (GEV), generalized Pareto (GP), log-Pearson three (L3) and generalized logistic (GL) distribution. to obtain the parameter estimates. The distribution performance evaluation is then performed utilizing the GOF and efficiency evaluations. The results indicate that the generalized GP distribution is the best possible function for determining the annual peak flow at the Kelantan streamflow site.

Keywords: generalized pareto, L-Moment, Malaysia

1. Introduction

In Malaysia, floods have been the most damaging natural event. This is because it is geographically subject to seasonal monsoon winds, which bring torrential rainfall to the country's north and east coasts (Mabahwi and Nakamura, 2020; Mokhtar et al., 2021; Mokhtar et al., 2021; Badyalina et al., 2022). In Kelantan, floods are considered an annual natural phenomenon and the worst record-setting flood of 2014 was recorded as a 'tsunami-like disaster' (Baharuddin et al., 2015). According to Weng et al. (2016), the flood that badly affected Kelantan in December 2014 was a very severe flood that resulted in flood losses in terms of lives lost, injuries, infrastructure destruction, property damage, crop loss, loss of livelihoods, interruption of routine services, and healthcare costs. Thus, estimating flood frequency is a critical issue, especially in water resource management, because it is used to design hydraulic structures and gives essential information (Badyalina et al., 2016; Jan et al., 2016; Kim and Lee, 2021). In flood frequency analysis, the parameters of distributions are often estimated using the L-moments approach. L-moments were first proposed by Hosking (1990) as a method for estimating distribution parameters using a linear combination of probability-weighted moments (Jan et al., 2016; Kang et al., 2019). According to David and Nagaraja (2004), L-moments are derived from the expectations of order statistics. Hence this may be a factor that is better than conventional moments for describing distribution form (Jan et al., 2018; Asquith, 2007). L-moments are widely used in applied research such as civil engineering, meteorology, and hydrology. L-moments are widely used in applied research such as civil engineering, meteorology, and hydrology. Among the benefits of L-moments include their capacity to work as a linear function of the data, being less prone to sample variability, being more robust to extreme values or outliers in the data and allowing for more confident inferences about the underlying probability distribution from small samples (Anas et al., 2021). This implies that L-moments are less affected by outliers, and the bias of their small sample estimates is kept to a minimum. In addition, because L-moments are linear combinations of order statistics, they have been demonstrated to be more useful in estimating statistical parameters than other methods, such as the method of moments and least squares (Maleki-Nezhad, 2014). This finding is supported by Šimková (2021) proposed that L-moments are the most accurate estimators for higher quantiles of the interest rate compared to moments and maximum likelihood methods (MLE). Furthermore, it is most effective when the tail of the distribution is heavier, and the sample size is small. The minimal amount of data available is well recognized to raise the level of uncertainty in both parameter and quantile estimates (Blain et al., 2021).

As a result, properly using the L-moments approach with desired constraints can help solve this problem. One of the interesting issues made by Shahzad et al. (2021) is that the L-moment can be represented for any random variable with a mean. Since the mean considers all data values, it is often applied to generate a better approximation of population parameters. Even though L-moment is more reliable than other traditional procedures, the underlying equations necessary to determine the L-moment parameters are difficult to solve and require a deep understanding of mathematics (Ilaboya and Otuaro, 2019). Thus, it can be summed up that L-moments are very easy to perceive as interval estimation, hypothesis testing, and estimation parameters and are very straightforward to understand as indicators of distributional form since they assist in summarising theoretical distributions and empirical samples. These advantages are especially relevant when data has heavy tails, severe skewness, or large variations. The statistical and probabilistic methods were applied to past events to forecast the exceedance likelihood of future events to minimize risk and maximize efficiency in design (Smithers and Schulze, 2001). However, this may be a concern when single station data are to be used. Reliable estimations require long station records and histories (Malekinezhad & Zare-Garizi, 2014). Therefore, flood frequency analysis (FFA) is utilized to forecast and justify extreme flood events to the more soluble issue of fitting distributions to the bulk of the data with the aid of refinement of techniques for incorporating historical and palaeoflood data (Kidson & Richards, 2005). In Malaysia, FFA is widely used for the river basin distributions such as LG, GEV, GP, L3 and GL distribution (Yue & Wang, 2004; Badyalina et al., 2014; Badyalina et al., 2015; Badyalina et al., 2021). According to Maposa & Cochran (2017), the Generalized Pareto distribution (GPD) models produced in this study were found to be statistically worthwhile for fitting flood heights in the lower Limpopo River basin of Mozambique and proven to be a better match compared to time-homogeneous GPD models based on the GOF. Similarly, it also applied to rainfall extremes for nine locations in the Lake Victoria basin (LVB) in Eastern Africa. Based on this finding, one of the best parameter estimation methods is L-moments, and moreover, normal-tailed GPD was found suitable to assess the observed and large number of global climate model rainfall time series (Onyutha & Willems, 2015). Apart from that, generalized extreme value distribution (GEV) has been known to be extremely useful, especially in regional flood frequency. In this approach, the shape parameter k of the GEV distribution and the ratio of size and location parameters are consistent throughout all basins in the region (Morrison & Smith, 2002). In the study conducted by Nimac et al. (2022), GEV was employed to estimate the return value curves for the Zagreb-Grič station from 1908 until 2020. The analysis showed that short-duration wet events (rainfall levels greater than the appropriate 10-year return values) became increasingly common after the 1970s. In Pakistan, three GOF tests, namely Kolmogorov–Smirnov, Anderson–Darling, and Chi-squared, were used to the fitted distributions at the 5% significant level. The analysis is performed using annual maximum discharge data from 1980 to 2016, and the results show that the generalized extreme value distribution (GEV) and the lognormal distribution are the top two distributions for all locations (Badyalina et al., 2013; Badyalina et al., 2016; Badyalina et al., 2021; Farooq et al., 2018). Meanwhile, log-Pearson distribution (P3) is widely employed in hydrologic fields. P3 distribution provides a reasonable model of the distribution of annual United States flood data. L-moment ratio relationships for the P3 distribution are then improved to be used to compare a region's summary statistics (Griffis & Stedinger, 2007). In similar cases in Tunisia, L - moments are used in identifying regional flood frequency distributions, which fully employed GOF from L-skewness and L-kurtosis. The most frequently used distributions are GEV, GL, GP, L3 and LG. The GNO distribution was shown to be the best-suited flood frequency distribution, while the GNO and GEV distributions provide the best fit in central and southern Tunisia (Abida& Ellouze, 2008).

2. Methodology

Hosking (1990) proposed L-moments as a linear combination of probability-weighted moments (PWMs). Assume $x_{1:n} \le x_{2:n} \le \ldots \le x_{n:n}$ are the data in a specific order with a sample size of n. Based on Badyalina et al. (2021), the procedure of unbiased sample estimator of the L Moments is as follows:

$$b_r = \frac{1}{n}\binom{n-1}{r}^{-1} \sum_{i=r+1}^{n} \binom{i-1}{r} x_{i:n} \tag{1}$$

From Eq.1, we can obtain the first four components of L-Moments.

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} x_{i:n} \tag{2}$$

$$b_1 = \frac{1}{n}\sum_{i=2}^{n}\frac{(i-1)}{(n-1)}x_{i:n} \tag{3}$$

$$b_2 = \frac{1}{n}\sum_{i=3}^{n}\frac{(i-1)(i-2)}{(n-1)(n-2)}x_{i:n} \tag{4}$$

$$b_3 = \frac{1}{n}\sum_{i=4}^{n}\frac{(i-1)(i-2)(i-3)}{(n-1)(n-2)(n-3)}x_{i:n} \tag{5}$$

The first four sample estimates for L-moments are referred to as:

$$l_1 = b_0 \tag{6}$$

$$l_2 = 2b_1 - b_0 \tag{7}$$

$$l_3 = 6b_2 - 6b_1 + b_0 \tag{8}$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{9}$$

The samples of the L-moments ratio are addressed as follows:

$$t_2 = \frac{l_2}{l_1} \tag{10}$$

$$t_3 = \frac{l_3}{l_2} \tag{11}$$

$$t_4 = \frac{l_4}{l_2} \tag{12}$$

Table 2 provides the potential probability distribution that has been employed in this study, namely GEV, GL, GP, L3 and LG. In Table 2, the estimation of the parameters for each potential distribution is described.

Table 2. Parameter estimation for potential distribution using L-Moment Method

Dist	Cumulative Density function	Parameter Estimation
GEV	$x(F) = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left\{1-(-\ln(F))^{\hat{k}}\right\}$	$\hat{k} = 7.85890c + 2.9554c^2$ where $c = \frac{2}{3+t_3} - \frac{\ln 2}{\ln 3}$ $\hat{\alpha} = \frac{l_2}{\Gamma(\hat{k})(1-2^{\hat{k}})}$; $\hat{\xi} = l_1 - \frac{\hat{\alpha}}{\hat{k}} + \hat{\alpha}\Gamma(\hat{k})$
GL	$x(F) = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left[1-\left\{\frac{(1-F)}{F}\right\}^{\hat{k}}\right]$	$\hat{k} = -t_3$; $\hat{\alpha} = \frac{l_2}{\Gamma(\hat{k})[\Gamma(1-\hat{k})-\Gamma(2-\hat{k})]}$

		$\hat{\varepsilon} = l_1 - \frac{\hat{\alpha}}{\hat{k}} + \hat{\alpha}\Gamma(\hat{k})\Gamma(1-\hat{k})$
GP	$x(F) = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left\{1 - \left[1-F\right]^{\hat{k}}\right\}$	$\hat{k} = \frac{1-3t_3}{1+t_3};\ \hat{\alpha} = l_2(\hat{k}+1)(\hat{k}+2)$ $\hat{\xi} = l_1 - \frac{\hat{\alpha}}{\hat{k}} + \frac{\hat{\alpha}}{\hat{k}(\hat{k}+1)}$
LG	$x(F) = \alpha + e^{\xi+uk};\quad u = \Phi^{-1}\left[1-F\right]$	$\hat{k} = -0.001005 + 0.997386z + 0.001027z^2 - 0.005853z^3 - 0.000154z^4 + 0.000141z^5$ where $z = \sqrt{\frac{8}{3}}\Phi^{-1}\left[\frac{1+t_3}{2}\right]$, Φ^{-1} is an inverse CDF of Normal Distribution. $\alpha = \ln(l_2) - \ln\left(2S_1(k) - e^{\frac{k^2}{2}}\right);\ \xi = l_1 - e^{\alpha+\frac{k^2}{2}}$
L3	$x(F) = \hat{\xi}\hat{k} + K_T\sqrt{\hat{\xi}^2\hat{k}}$ where $K_T = \frac{2}{C_s}\left[\left\{\frac{C_s}{6}\left(\Phi^{-1}\left[1-F\right] - \frac{C_s}{6}\right)+1\right\}^3 - 1\right]$ $C_s = \frac{2}{\sqrt{\hat{k}}}$	$\hat{k} = 0.0127331632 + \frac{1.0246130369}{3\pi(t_3)^2} - \frac{0.0024863669}{(3\pi(t_3)^2)^2} + \frac{0.0001169073}{(3\pi(t_3)^2)^3} - \frac{0.0000027751}{(3\pi(t_3)^2)^4} + \frac{0.0000000323}{(3\pi(t_3)^2)^5} - \frac{0.0000000001}{(3\pi(t_3)^2)^6}$ $\hat{\alpha} = \frac{l_2}{2S_1(\hat{k})-\hat{k}};\ \hat{\xi} = l_1 - \hat{k}\hat{\alpha}$ $S_r(k) = \int_0^\infty \left[\int_0^x \frac{1}{\Gamma(\hat{k})} t^{\hat{k}-1}e^{-t}dt\right]^r \frac{1}{\Gamma(\hat{k})} x^{\hat{k}}e^{-x}dx$

2.3 Evaluation Criteria

2.3.1 Accuracy Indicators

In accuracy indicators, three accuracy indicators are used in this study; namely, root mean square error (RMSE), mean absolute error (MAE) and mean absolute error (MAE). The MAE, MAPE and RMSE is define in Eq.13-Eq. 15, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |F(y_i) - F(\hat{y}_i)| \tag{13}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{F(y_i) - F(\hat{y}_i)}{F(y_i)}\right| \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(F(y_i) - F(\hat{y}_i))^2}{n}} \tag{15}$$

where $F(y_i)$ represent the actual data, n represent the total number of data, $\overline{F}(y_i)$ represent the average of the actual data and $F(\hat{y}_i)$ represent the estimated return period from the chosen distribution.

2.3.2 L-Moment Ratio Diagram

Hosking and Wallis (1993) proposed an L-Moment Ratio Diagram to identify the ideal distribution at the selected river basin. The L-Moment ratio diagram demonstrates the conceptual relation among both t_3 and t_4. The value of t_3 drawn from the peak flow data and plotted to L-Moment ratio diagram to identify which distribution lies closely.

2.3.3 GOF

Several applicable statistical procedures, such as the GOF tests, can be used to assess whether the probability distributions are appropriate for a particular study. The GOF tests can be used to justify choosing the best distribution in FFA (Badyalina et al.,2021). Two GOF tests, the Kolmogorov-Smirnov (KS) and Anderson Darling (AD) tests are employed in this study to determine how closely the observed data resembles the distributions.

3. Results and Discussion

Peak flow is generally related to implementing and developing flood management design. This information is essential for flood design at the targeted catchment. This study's targeted catchment is a Kelantan, Malaysia river. The annual peak flow is analyzed by fitting five normal distributions: GEV, GL, GP, L3 and LG. The estimated parameters for the potential distributions using L-Moments are shown in Table 3.

Table 3. Parameters for the candidate distributions

	Parameters		
Dist.	$\hat{\alpha}$	$\hat{\xi}$	$\hat{k}$
GEV	3379.46	2159.52	0.097
GL	4189.48	1352.17	-0.11
GP	841.06	5774.65	0.61
L3	4435.51	2478.35	0.67
LG	-6526.70	9.28	0.22

Table 3 shows the estimated parameters for the candidate distributions. The $\hat{\alpha}$, $\hat{\xi}$ and $\hat{k}$ are for GEV distributions are 3379.46, 2159.52 and 0.097, respectively. The $\hat{\alpha}$, $\hat{\xi}$ and $\hat{k}$ are for GL distributions are 4189.48, 1352.17 and -0.11, respectively. The $\hat{\alpha}$, $\hat{\xi}$ and $\hat{k}$ are for GP distributions are 841.06, 5774.65 and 0.61, respectively. The $\hat{\alpha}$, $\hat{\xi}$ and $\hat{k}$ are for L3 distributions are 4435.51, 2478.35 and 0.67, respectively. The $\hat{\alpha}$, $\hat{\xi}$ and $\hat{k}$ are for LG distributions are -6526.70, 9.28 and 0.22, respectively.

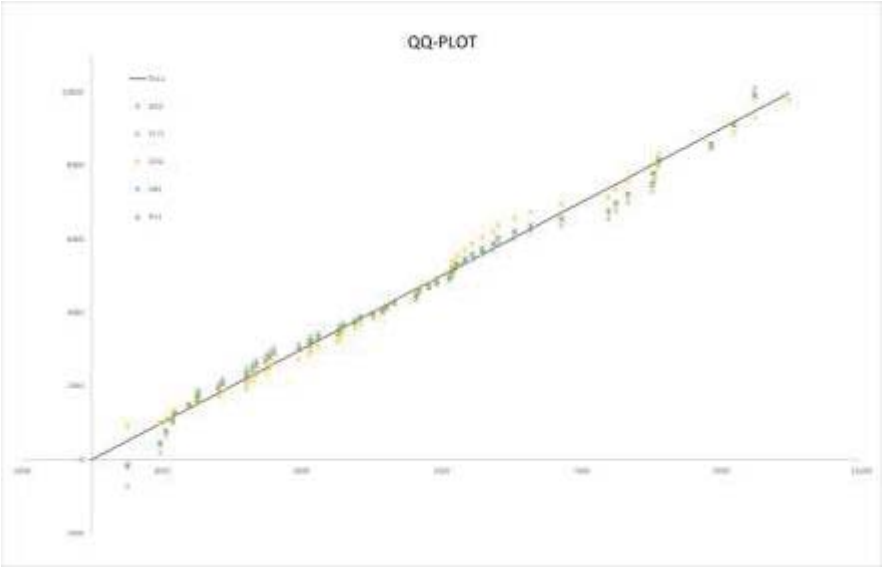


Figure 1. QQ-Plot for Sungai Kelantan river basin annual peak flow and potential distributions

Figure 1 illustrates QQ-plot for the annual peak flow of the Kelantan River basin and potential distribution. The QQ-plot is a visual tool for comparing potential distributions. Compared to other potential distributions, the GP

distribution closely resembles the black line based on observational observations. Accuracy indicators, the GOF test, and the L-Moment Diagram are added to the study to complement our visual method-based observation. The GOF test, a numerical accuracy measure, and an L-moment ratio diagram are the 3 measuring methods employed in this study to pick the most suitable model. This study uses the rank score technique to analyze the best distribution that fits the target basin. This method called for grading each distribution according to how well it resembled the real observation. A score of 5 is awarded for the best distribution that fits the data. A score of 1 is given to the distribution with the poorest fit. P-Value is employed to evaluate the GOF test distribution test. The larger P-value implies that the distribution fits the data well. In the L-moment diagram, estimations of the dimensionless ratios t_3 and t_4 are compared using sample data values. The distribution with the least RMSE, MAE, and MAPE is awarded the highest possible score of 5. The distribution with the largest MAPE, RMSE, and MAE is awarded the lowest possible score of 1.

Table 4. Rank score for potential distribution

Potential Distribution	GEV	GL	**GP**	L3	LG
MAPE	3	2	**5**	4	1
KS	4	2	**5**	1	3
AD	3	2	**5**	1	4
GLRMSE	3	1	**5**	4	2
MAE	3	1	**5**	2	4
LMR	4	2	**5**	1	3
Total Score	20	10	**30**	13	17

Table 4 demonstrates that the GP distribution is optimal for describing yearly peak flow data for the Kelantan River. After evaluating each distribution individually, Table 4 indicates that the GL distribution is unsuitable, as its rank score is the lowest. The GP distribution is ranked higher than the other distributions used in this investigation. Due to the unique characteristics of each river's data series, it is difficult to establish a particular probability distribution for all streamflow basins in Kelantan. The annual peak flow data fluctuates dramatically from year to year due to the unpredictability of the weather brought on by climate change. Estimating extreme streamflow with varying return times is the purpose of selecting the "optimal" model for each location. Consequently, various return periods of streamflow are predicted using the best suitable frequency model. Table 5 shows the estimated return period.

Table 5. Estimated peak flow for the Kelantan River basin

	Estimated peak flow (m^3/s)				
Return Period (Years)	GEV	GL	**GP**	L3	LG
2	4157.05	4189.48	**4108.82**	4162.17	4164.14
10	7745.44	7546.83	**8005.92**	7735.56	7717.18
25	9318.14	9325.80	**9010.31**	9286.64	9294.75
50	10394.90	10745.44	**9474.11**	10360.86	10405.59
100	11393.47	12256.66	**9778.72**	11376.43	11471.25
200	12323.38	13876.08	**9978.78**	12347.38	12505.20

4. Conclusion

FFA is suitable for predicting the long-term flow characteristics of a river. This study sought to determine the ideal potential distribution for FFA to represent the annual peak flow in the Kelantan River basin. GEV, GL, GP, L3 and LG are the distributions employed, representing the five most often used distributions in the analysis of extreme hydrologic variables. Regional flood frequency study has utilized L-moment because of the robustness of its GOF, which is superior to traditional single-basin GOF. The annual peak flow series data collected from the historical daily stream flow record of the Kelantan River basin in Malaysia is applied to the five potential distributions. The model's evaluation

uses the L-Moment Ratio Diagram and numerical performance criteria. Based on the performance evaluation, GP distribution is found to be optimal for describing yearly peak flow data for the Kelantan River. In future research, it would be interesting to implement the distribution to several rivers located near Kelantan River for regional FFA or in-situ FFA to provide guidance about the expected behaviour of future flooding.

References

Abida, H., & Ellouze, M. (2007). Probability distribution of flood flows in Tunisia. *Hydrology and earth system sciences discussions*, *4*(2), 957-981. https://doi.org/10.5194/hessd-4-957-2007

Anas, M. M., Ali, M., Shafqat, A., Shahzad, F., Abbass, K., & Alilah, D. A. (2021). L-Moments and Calibration-Based Estimators for Variance Parameter. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/9847714

Badyalina, B., & Shabri, A. (2013). Streamflow forecasting at ungauged sites using multiple linear regression. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 67-75.

Badyalina, B., & Shabri, A. (2015). Flood estimation at ungauged sites using group method of data handling in Peninsular Malaysia. *Jurnal Teknologi*, *76*(1). https://doi.org/10.11113/jt.v76.2640

Badyalina, B., & Shabri, A. (2015). Flood frequency analysis at ungauged site using group method of data handling and canonical correlation analysis. *Modern Applied Science*, *9*(6), 48. https://doi.org/10.5539/mas.v9n6p48

Badyalina, B., Mokhtar, N. A., Azimi, A. I. F., Majid, M., Ramli, M. F., & Yaa'coob, F. F. (2022). Data-driven Models for Wind Speed Forecasting in Malacca State. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 125-139.

Badyalina, B., Mokhtar, N. A., Jan, N. A. M., Hassim, N. H., & Yusop, H. (2021). Flood Frequency Analysis using L-Moment For Segamat River. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 47-62.

Badyalina, B., Mokhtar, N. A., Ramli, M. F., Majid, M., & Yusri, M. Y. (2021). Design of Simulation Studies for Flood Quantile Prediction Problems at Ungauged Site. *Applied Mathematical Sciences*, *15*(3), 137-140. https://doi.org/10.12988/ams.2021.914422

Badyalina, B., Shabri, A., & Jan, N. (2016). Prediction At Ungauged Site with Topological Kriging And Modified Group Method Of Data Handling. *Journal Of Environmental Hydrology*, *24*(6).

Badyalina, B., Shabri, A., & Marsani, M. F. (2021). Streamflow Estimation at Ungauged Basin using Modified Group Method of Data Handling. *Sains Malaysiana*, *50*(9), 2765-2779. https://doi.org/10.17576/jsm-2021-5009-22

Badyalina, B., Shabri, A., & Samsudin, R. (2014). Streamflow estimation at ungauged site using wavelet group method of data handling in Peninsular Malaysia. *International Journal of Mathematical Analysis, 8*(11), 513-524. https://doi.org/10.12988/ijma.2014.4251

Badyalina, B., Shabri, A., Mokhtar, N. A., Ramli, M. F., Majid, M., & Yusri, M. Y. (2021). Modified Group Method of Data Handling for Flood Quantile Prediction at Ungauged Site. *International Journal of Statistics and Probability*, *10*(6), 1-57. https://doi.org/10.5539/ijsp.v10n6p57

Baharuddin, K. A., Wahab, S. F. A., Ab Rahman, N. H. N., Mohamad, N. A. N., Kamauzaman, T. H. T., Noh, A. Y. M., & Majod, M. R. A. (2015). The record-setting flood of 2014 in Kelantan: challenges and recommendations from an emergency medicine perspective and why the medical campus stood dry. *The Malaysian journal of medical sciences: MJMS*, *22*(2), 1.

Blain, G. C., Sobierajski, G. D. R., Xavier, A. C. F., & De Carvalho, J. P. (2021). Regional Frequency Analysis applied to extreme rainfall events: Evaluating its conceptual assumptions and constructing null distributions. *Anais da Academia Brasileira de Ciências*, *93*. https://doi.org/10.1590/0001-3765202120190406

David, H. A., & Nagaraja, H. N. (2004). *Order statistics*. John Wiley & Sons. https://doi.org/10.1002/0471667196.ess6023

Farooq, M., Shafique, M., & Khattak, M. S. (2018). Flood frequency analysis of river swat using Log Pearson type 3, Generalized Extreme Value, Normal, and Gumbel Max distribution methods. *Arabian Journal of Geosciences*, *11*(9), 1-10. https://doi.org/10.1007/s12517-018-3553-z

Griffis, V., & Stedinger, J. (2007). Log-Pearson type 3 distribution and its application in flood frequency analysis. I: Distribution characteristics. *Journal of Hydrologic Engineering*, *12*(5), 482-491. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(482)

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 105-124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x

Hosking, J., & Wallis, J. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research*, *29*(2), 271-281. https://doi.org/10.1029/92WR01980

Ilaboya, I. R., & Otuaro, E. A. (2019). Simple to Use Microsoft Excel Template for Estimating the Parameters of Some Selected Probability Distribution Model by Method of L-Moment. *parameters*, *1*, 1.

Jan, N. A. M., Shabri, A., & Badyalina, B. (2016). Selecting probability distribution for regions of Peninsular Malaysia streamflow. AIP Conference Proceedings. https://doi.org/10.1063/1.4954619

Jan, N. A. M., Shabri, A., Hounkpè, J., & Badyalina, B. (2018). Modelling non-stationary extreme streamflow in Peninsular Malaysia. *International Journal of Water*, *12*(2), 116-140. https://doi.org/10.1504/IJW.2018.091380

Jan, N. A. M., Shabri, A., Ismail, S., Badyalina, B., Abadan, S. S., & Yusof, N. (2016). THREE-PARAMETER LOGNORMAL DISTRIBUTION: PARAMETRIC ESTIMATION USING L-MOMENT AND TL-MOMENT APPROACH. *Jurnal Teknologi*, *78*(6-11). https://doi.org/10.11113/jt.v78.9202

Kang, C., Park, K.-Y., & Cho, Y.-S. (2019). Numerical and Statistical Analyses of Tsunami Heights with the L-Moments Method. *Applied Sciences*, *9*(24), 5517. https://doi.org/10.3390/app9245517

Kidson, R., & Richards, K. (2005). Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography*, *29*(3), 392-410. https://doi.org/10.1191/0309133305pp454ra

Kim, S. U., & Lee, C.-E. (2021). Incorporation of cost-benefit analysis considering epistemic uncertainty for calculating the optimal design flood. *Water Resources Management*, *35*(2), 757-774. https://doi.org/10.1007/s11269-021-02764-z

Landwehr, J. M., Matalas, N., & Wallis, J. R. (1978). Some comparisons of flood statistics in real and log space. *Water Resources Research*, *14*(5), 902-920. https://doi.org/10.1029/WR014i005p00902

Mabahwi, N. A., & Nakamura, H. (2020). The Issues and Challenges of Flood-related Agencies in Malaysia. *Environment-Behaviour Proceedings Journal*, *5*(13), 285-290. https://doi.org/10.21834/e-bpj.v5i13.2069

Malekinezhad, H., & Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. *Atmósfera*, *27*(4), 411-427. https://doi.org/10.1016/S0187-6236(14)70039-6

Malekinezhad, H., & Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. *Atmósfera*, *27*(4), 411-427. https://doi.org/10.1016/S0187-6236(14)70039-6

Maposa, D., & Cochran, J. J. (2017). Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalized Pareto distribution. *Statistics and Its Interface*, *10*(1), 131-144. https://doi.org/10.4310/SII.2017.v10.n1.a12

Mokhtar, N. A., Badyalina, B., Chang, K. L., Yaa'cob, F., Ghazali, A., & Shamala, P. (2021). Error-in-Variables Model of Malacca Wind Direction Data with the von Mises Distribution in Southwest Monsoon. *Applied Mathematical Sciences*, *15*(9), 471-479. https://doi.org/10.12988/ams.2021.914521

Mokhtar, N. A., Zubairi, Y. Z., Hussin, A. G., Badyalina, B., Ghazali, A. F., Ya'acob, F. F., . . . Kerk, L. C. (2021). Modelling wind direction data of Langkawi Island during Southwest monsoon in 2019 to 2020 using bivariate linear functional relationship model with von Mises distribution. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1988/1/012097

Nimac, I., Cindrić Kalin, K., Renko, T., Vujnović, T., & Horvath, K. (2022). The analysis of summer 2020 urban flood in Zagreb (Croatia) from hydro-meteorological point of view. *Natural Hazards*, *112*(1), 873-897. https://doi.org/10.1007/s11069-022-05210-4

Onyutha, C., & Willems, P. (2015). Uncertainty in calibrating generalized Pareto distribution to rainfall extremes in Lake Victoria basin. *Hydrology Research*, *46*(3), 356-376. https://doi.org/10.2166/nh.2014.052

Shahzad, U., Ahmad, I., Almanjahie, I., & Al-Noor, N. H. (2021). Utilizing L-Moments and calibration method to estimate the variance based on COVID-19 data. *Fresenius Environmental Bulletin*, *30*(7A), 8988-8994.

Šimková, T. (2021). Confidence intervals based on L-moments for quantiles of the GP and GEV distributions with application to market-opening asset prices data. *Journal of Applied Statistics*, *48*(7), 1199-1226. https://doi.org/10.1080/02664763.2020.1757046

Smithers, J., & Schulze, R. (2001). A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments. *Journal of Hydrology*, *241*(1-2), 42-52. https://doi.org/10.1016/S0022-1694(00)00374-7

Yue, S., & Wang, C. (2004). Determination of regional probability distributions of Canadian flood flows using L-moments. *Journal of Hydrology (New Zealand)*, 59-73.

Nonlinear Mixed Models Applied to Ruminal Degradability Studies

Vanderly Janeiro[1], Robson Marcelo Rossi[1], Terezinha Aparecida Guedes[1],Ana Beatriz Tozzo Martins[1] & Lucimary Afonso dos Santos[2]

[1] Department of Statistics, State University of Maringá, Maringá, PR, Brazil.

[2] Department of Mathematics, Paraná State University, Paranavaí, PR, Brazil

Correspondence: Vanderly Janeiro, Department of Statistics, State University of Maring¢, Maringá, PR, Brazil

Received: July 26, 2022 Accepted: August 27, 2022 Online Published: September 7, 2022

doi:10.5539/ijsp.v11n5p18 URL: https://doi.org/10.5539/ijsp.v11n5p18

Abstract

This article presents an application of three classical models to studies of ruminal degradation kinetics, namely Ørskov and McDonald's model (1979); Van Milgen, Murphy and Berger's model (1991), and Richard's model proposed in France, Dijkstra, and Dhanoa (1996). Our approach is focused on accounting for animal effects given that measurements are repeated in the same animal. The models were studied under the perspective of nonlinear mixed-effects (NLME) models. In this way, we intended to accommodate the problems of response variance heterogeneity and correlations between repeated measures. To apply the proposed method, we used data from an experiment conducted in a Latin square design to assess the dry matter degradability of the following three silages: Elephant grass (*Pennisetum purpureum* Schumach.) silage treated with bacterial inoculant, Elephant grass silage treated with enzyme-bacterial inoculant, and corn (*Zea mays L.*) silage. Samples were incubation for 0, 2, 6, 12 , 24, 48, 72 and 96 h. For these experimental data, the Van Milgen, Murphy, and Berger's model showed better performance than the others. The proposed approach indicated that inclusion of animal effects is important for obtaining more accurate information and can be considered in NLME modeling. Furthermore, it was also possible to perform an easy-to-interpret analysis of contrasts between treatments by using Tukey's test.

Keywords: Degradation kinetics, Digestibility, *In situ* degradability, Intake, Random effects, Rumen degradation.

1. Introduction

Collections that generate correlated data are common in several areas of knowledge. In many situations, such collections are carried out longitudinally, implying that observations on the same individual are correlated and not independent, which limits the use of certain statistical methods and techniques. Correlations can be accounted for by using the mixed model theory, which relates a response variable to predictor variables as fixed and random effects factors under the assumption that the residual distribution is Gaussian (J. C. Pinheiro & Bates, 2000). Most of the time, however, linear mixed models are not appropriate for explaining relationships between variables. In such analyses, nonlinear mixed effects (NLME) models arise as an extension of linear mixed models for describing nonlinear parameters.

In the case of longitudinal data, the interest commonly lies in specific individual characteristics, given that the dependent variable or response is measured several times and the effect associated with the individual/subject is included in the model as a random effects factor. That is, mixed models are often used to deal with correlated or hierarchical data. Also, recognizing that there are random effects factors influencing the observed response can increase the accuracy and precision of fixed effects estimates, minimizing seriously inflated type I error rates (Wang, 2016).

The nonlinear mixed model is based on a mean curve that is fitted to the data, such that individual curves incorporating the random effects of each individual appear as deviations from this mean curve. In literature, there are several methods proposed to model continuous, unbalanced, and multilevel longitudinal data. One of the first models, proposed by Gregoire and Schabenberger (1996) incorporates subject random effects, whereas that developed by Littell (2006) directly models correlation structure. Gregoire and Schabenberger's (1996) approach employs nonlinear fixed effects models, inducing correlations in the marginal distribution of within-subject observations and using random effects that vary across subjects to reduce the impact of autocorrelation. The second procedure (Littell, 2006) uses a covariance structure and generalized least squares estimators, which are considered the best unbiased estimators (Tasissa & Burkhart, 1997).

In the field of cattle research, it is common to find articles that discuss longitudinal data without taking into account possible animal correlations (among observation in the same subject). An example is seen in rumen degradability studies using the model proposed by Mehrez and Ørskov (1977). This approach may lead to inappropriate conclusions because it ignores important effects and makes assumptions that are inconsistent with the reality of the data.

Rossi, Martins, Guedes, and Jobim (2010) noted that alternative and/or more innovative methods can provide a more parsimonious explanation for data of this nature. The authors emphasized the importance of Bayesian inference to make comparisons between parameters while considering different experimental treatments in a coherent manner, without having to resort, for example, to asymptotic procedures.

Frequentist methods can be applied to data sets with a longitudinal structure, as performed byMedeiros, Lima, Savian, Malheiros, and Werner (2020). The authors sought to address the problems of variance heterogeneity and correlations between repeated longitudinal measurements in in situ ruminal degradation kinetic studies by using NLME. From a statistical point of view, similar problems encountered in different contexts and areas can be properly addressed through the use of mixed models, such as seen in J. C. Pinheiro and Bates (2000); Sartrio (2013); Luwanda and Mwambi (2016); Wyzykowski, Custdio, Custdio, Gomes, and Morais (2015); Calama and Montero (2004); and Xu et al. (2014). The proposal to apply a mixed effects methodology involving fixed and random effects parameters and the construction of a data (co)variance matrix (Yang, Huang, Trincado, & Meng, 2009) seems adequate to capture between- and within-animal variabilities and allows modeling the degradability of each animal (subject-specific) as the average degradability of all animals (population specific) (Schabenberger & Pierce, 2002).

In this study we aimed to evaluate the ruminal degradation kinetics through the nonlinear models of Ørskov and McDonald (1979), Van Milgen, Murphy, and Berger (1991), and France, Dijkstra, and Dhanoa (1996) and compare them in order to determine the best berformance. For that, we will consider the three nonlinear models with the inclusion of mixed effects. So that the models include fixed effects and random effects, allowing the variability between animals to be evaluated, with different structures for the (co)variance matrix of errors and random effects. In addition, we will discuss the comparison of experimental fixed effects treatments.

In this study, we aimed to evaluate ruminal degradation kinetics using the nonlinear models proposed by Ørskov and McDonald (1979), Van Milgen, Murphy, and Berger (1991), and France, Dijkstra, and Dhanoa (1996) and compare the results in order to determine the model with the best performance. For this, we considered the three nonlinear models with the inclusion of mixed effects; that is, the models contain both fixed and random effects. Such an approach allowed us to assess between-animal variability using different structures for the (co)variance matrix of errors and random effects. We also provide a discussion of fixed effects treatments.

2. Materials and Methods

2.1 Material

For model comparison (Table 2), we used a set of observations from a ruminal degradability experiment carried out in the Dairy Cattle Sector of the Iguatemi Experimental Farm (FEI), State University of Maringá, Maringá, Paraná, Brazil. Ruminal degradation kinetics were assessed according to Rossi et al. (2010). Treatments consisted of Elephant grass (*Pennisetum purpureum* Schumach.) silage with bacterial inoculant (SCE-IBC) (Propiolact MS01), Elephant grass silage with enzyme-bacterial inoculant (SCE-IEZ) (Bacto Silo), and corn silage (SMI) (*Zea mays L.*), hereafter referred to as T_1, T_2 and T_3, respectively. Silages were stored in trench silos, without coating, with a capacity of approximately 20 t.

A 3×3 Latin square experimental design was used, with cows treated as a nuisance factor (three lactating Holstein cows, C_1, C_2 and C_3) and periods (P_1, P_2 and P_3) and treatments (T_1, T_2 and T_3) considered as factors of interest. For each animal/period/treatment combination, hereafter referred to as subject or individual (ind), ruminal degradation was evaluated at the following incubation times: 0, 2, 6, 12, 24, 48, 72, and 96 h (Table 1). The observed and analyzed response was dry matter (DM) disappearance. For more details on the experiment, see Rossi et al. (2010).

Table 1. Dataset structure.

			Time (h)							
ind	Combination	Treatment	0	2	6	12	24	48	72	96
1	C_1P_1	T_1	y_{11}				$\cdots$			y_{18}
2	C_1P_2	T_3	y_{21}				$\cdots$			y_{28}
3	C_1P_3	T_2	y_{31}				$\cdots$			y_{38}
4	C_2P_1	T_2	y_{41}				$\cdots$			y_{48}
5	C_2P_2	T_1	y_{51}				$\cdots$			y_{58}
6	C_2P_3	T_3	y_{61}				$\cdots$			y_{68}
7	C_3P_1	T_3	y_{71}				$\cdots$			y_{78}
8	C_3P_2	T_2	y_{81}				$\cdots$			y_{88}
9	C_3P_3	T_1	y_{91}				$\cdots$			y_{98}

ind: animal/period/treatment combination.

2.2 Methods

The sampling structure induces a correlation among observations of the same subject. When faced with this type of problem, several authors adopted a mixed effects modeling approach(Medeiros et al., 2020; Xu et al., 2014; Calama & Montero, 2004; J. C. Pinheiro & Bates, 2000). A mixed nonlinear model (1) considering the i-th subject in the j-th evaluation time, according to J. C. Pinheiro and Bates (2000), is such that:

$$\boldsymbol{y}_i = f(\boldsymbol{\phi}_i, \boldsymbol{v}_i) + \boldsymbol{\varepsilon}_i \quad i = 1, \ldots, N = 9 \tag{1}$$

where, $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i}]'$ denotes the vector of measurements from i-th subject (animal/period/treatment) in the j-th observation time, f is the differentiable function of parameter vector $\boldsymbol{\phi}_i$ $(k \times 1)$, k is the number of parameters in the model, $\boldsymbol{v}_i = [v_{i1}, \ldots, v_{ij}, \ldots, v_{in_i}]'$ is the predictor vector, and $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1}, \ldots, \varepsilon_{ij}, \ldots, \varepsilon_{in_i}]'$ is the vector residual terms. Still according to J. C. Pinheiro and Bates (2000) (2):

$$\boldsymbol{\phi}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \tag{2}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are, the incidence matrix (or design) for fixed and random effects, respectively, with the respective parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. As demonstrated by Calama and Montero (2004), the NLME model has as its basic assumptions:

$$\boldsymbol{b}_i \overset{iid}{\sim} \mathbf{N}_q(\mathbf{0}, \boldsymbol{D})$$
$$\boldsymbol{\varepsilon}_i \overset{iid}{\sim} \mathbf{N}_J(\mathbf{0}, \boldsymbol{R}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\rho}))$$

here $\mathbf{N}$ denotes a multivariate normal distribution with a null mean vector and $\boldsymbol{D}$ is the $q \times q$ positive-definite varianceCcovariance matrix for random effects, representing among subject variability. In this formulation $\boldsymbol{R}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\rho})$ is the $n_i \times n_i$ intraindividual varianceCcovariance matrix defining within-subject variability. $\boldsymbol{R}_i$ is allowed to depend on both random and fixed effects, and $\boldsymbol{\rho}$ represents a set of common but unknown parameters. The $\boldsymbol{R}_i$ matrix is able to describe within-subject heteroscedasticity and autocorrelation by including both correlation effects and weighting factors. It can be decomposed and written as (3):

$$\boldsymbol{R}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\rho}) = \sigma^2\boldsymbol{G}_i^{1/2}\boldsymbol{\Gamma}_i\boldsymbol{G}_i^{1/2} \tag{3}$$

where, for the i-th subject with n_i measurements, σ^2 is the scaling factor for the error dispersion, $\boldsymbol{G}_i$ is the $n_i \times n_i$ diagonal matrix that accommodates the variability of the error due to time, and $\boldsymbol{\Gamma}_i$ is the $n_i \times n_i$ of within-time error autocorrelation (Crecente-Campo, Tom, Soares, & Diguez-Aranda, 2010; Davidian & Giltinan, 2003).

Among the nonlinear f functions proposed in the literature, we focused on the exponential model (Ørskov & McDonald, 1979), Van Milgen's model (Van Milgen, Murphy, & Berger, 1991), and Richards model (France, Dijkstra, & Dhanoa, 1996), as depicted in Table 2. We used the parameterization presented in Teixeira et al. (2016).

Table 2. Candidate statistical models for describing ruminal degradability

Model	Statistical expression
Ørskov (OR)	$y_{ij} = \beta_1 + \beta_2(1 - e^{-\beta_3 t_{ij}/2}) + \varepsilon_{ij}$
Van Milgen (VM)	$y_{ij} = \beta_1 + \beta_2[1 - (1 + \beta_3 t_{ij})e^{-\beta_3 t_{ij}}] + \varepsilon_{ij}$
Richard's (RI)	$y_{ij} = \beta_1 + \beta_1\beta_2[\beta_1^{\beta_4} + (\beta_2^{\beta_4} - \beta_1^{\beta_4})e^{-\beta_3 t_{ij}/2}]^{-1/\beta_4} + \varepsilon_{ij}$

β_1: soluble fraction (%) ($\beta_1 \geq 0$);
β_2: potentially degraded insoluble fraction (%) ($\beta_2 \geq 0$);
β_3: joint fractional rate of latency and degradation ($\beta_3 \geq 0$);
β_4: parameter without biological meaning ($\beta_4 \geq -1$).

To adjust the models and analyze the data, we used the resources available in the `nlme` package (linear and nonlinear mixed effects models) (J. Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2021) of the R statistical environment (R Core Team, 2021). The following steps were taken:

- nonlinear models were fitted to individual curves considering only fixed effects, using the `nlsList` function to determine whether this approach would be sufficient to explain ruminal degradation kinetics. In the analyzed case, this model structure was not sufficient;
- random effects were added to all model parameters for selection of the $\boldsymbol{D}$ matrix;

- after the $\boldsymbol{D}$ matrix was chosen, models were adjusted by incorporating different random effects components into parameters;
- the best structure for the varianceCcovariance matrix of residuals ($\boldsymbol{R}_i$) was defined;
- the variance components of all three adjusted models were defined, and the final model was selected; and
- parameter estimates for the final model and pairwise contrasts were analyzed to compare the performance of treatments.

We considered `pdSymm` (positive definite matrix), `pdDiag` (diagonal matrix), and `pdIdent` (identity matrix) as structures for the (co)variance matrix ($\boldsymbol{D}$) of random effects of parameters (e.g. J. C. Pinheiro and Bates (2000), section 4.2.2).

Regarding the residual matrix ($\boldsymbol{R}$), three other correlation structures were considered for $\boldsymbol{\Gamma}$, namely `corAR1` (autoregressive of order 1 - AR1), `corCompSymm` (compound symmetry), and `corLin` (General Linear). The `nlm` package implements the `corARMA` function (autoregressive moving average - ARMA), which can be useful to decide between an AR1 and ARMA(p,q) correlation structures. A useful approach is to generate all possible combinations of ARMA models (for $p = 0$ to $p = 2$ and $q = 0$ to $q = 2$) and choose the one with the lowest Akaike information criterion (AIC) and/or Bayesian information criterion (BIC). The three matrices were considered with and without the `varIdent` class matrix ($\boldsymbol{G}$) to correct for possible heteroscedasticity within groups (treatments or times) (J. C. Pinheiro & Bates, 2000). To obtain the estimates presented in Tables 4, 5, 6, and 8, we used restricted maximum likelihood estimation (`method = "REML"` of the `nlme` function).

Several authors do not recommend the use of the determination of coefficient (R^2) to select an NLME model, according to Spiess and Neumeyer (2010). Therefore, of the models described in Table 2, that with the best performance in predicting DM was determined using the following criteria: intercept, slope, residual sum of squares (RSS), mean squared error (MSE), root-mean-square error (RMSE), and (R^2) of the simple linear model fit between DM values observed and DM values predicted by NLME model. Furthermore, model efficiency (ME), normalized model efficiency (NME), correlation between observed and fitted values (Corr), and concordance correlation (ConCorr) were determined after adjusting the models. These statistics were obtained by using the `IA_tab` function of the `nlraa` package (Miguez, 2021; Miguez, Archontoulis, & Dokoohaki, 2018).

3. Results

The observed response (DM) for each subject (animal/period/treatment) is displayed in Figure 1. The curves indicate that the proposed models are plausible and suggest differences between treatments.

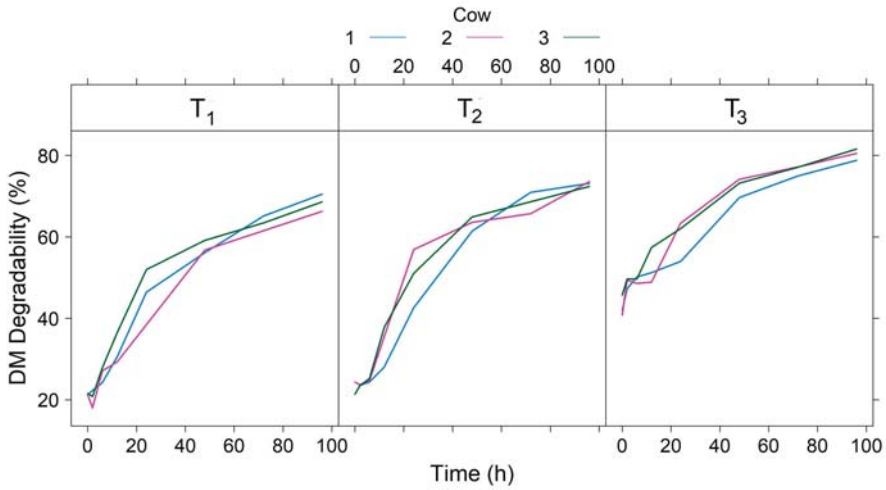


Figure 1. Rumen degradability of dry matter (DM) per animal and treatment

The models presented in Table 2 were adjusted for each individual (Table 1), totaling nine adjustments for each model. Adjustments were made using the `nlsList` function of the `nlme` package, which uses a nonlinear least squares procedure. Because of the difficulty in estimating Richards model parameters, we opted for a first fit considering $\beta_4 = 1$. With this adjustment, we obtained initial values for the simultaneous adjustment of the four parameters. A summary of parameter

estimates is presented in Table 3. For all models, the coefficient of variation (CV) was high (Table 3), indicating a large variation in parameter estimates. This suggests that an effect can be added through a probability distribution, that is, by including random effects components in the model. Such a procedure affords a mixed effects model.

Table 3. Summary of parameter estimates for models fitted to each curve (ind)

		Parameter			
Model	Statistic	β_1	β_2	β_3	β_4
	mean	27.6786	52.0225	0.0523	-
OR	SD	12.5968	8.8589	0.0177	-
	CV(%)	45.5108	17.0290	33.8497	-
	mean	30.0505	42.9881	0.0711	-
VM	SD	12.3312	7.2544	0.0190	-
	CV(%)	41.0349	16.8754	26.7380	-
	mean	14.6044	58.2513	0.1431	1
RI*	SD	5.9844	2.5824	0.0538	1
	CV(%)	40.9764	4.4332	37.6296	1
	mean	14.7507	60.2679	0.0990	0.4716
RI	SD	6.4449	3.2281	0.0342	1.5911
	CV(%)	43.6919	5.3562	34.5448	337.3897

RI*: Richard's model with $\beta_4 = 1$; SD: standard deviation of parameter estimates; and CV(%): coefficient of variation of parameter estimates.

In this next step, we made adjustments considering mixed effects models. First, we considered that all model parameters (Table 2) were associated with a random effect of subjects and a fixed effect of treatments.

To determine the best varianceCcovariance matrix for random effects ($\boldsymbol{D}$), we considered three matrix structures: multiple of the identity (`pdIdent`), diagonal (`pdDiag`), and general positive-definite (`pdSymm`).

According to the likelihood ratio test results, matrices were not considered to have significant differences, despite $p = 0.0485$ (see Table 4). AIC values of OR and VM models indicated better results for the identity matrix. For the RI model, the diagonal matrix was indicated as the best; however, the values obtained for identity and diagonal matrices were very similar. Considering the lowest BIC value, we concluded that the identity matrix afforded the best results, in addition to requiring the estimation of fewer parameters.

Table 4. Assessment of different $\boldsymbol{D}$ matrix structures for the three models

Model	$\boldsymbol{D}$ Matrix	df	AIC	BIC	logLik	LRT	$p-$value
	pdIdent	11	361.31	385.72	-169.66		
OR	pdDiag	13	365.31	394.16	-169.66	0.0000	1.0000
	pdSymm	16	365.81	401.32	-166.91	5.4994	0.1387
	pdIdent	11	348.99	373.40	-163.49		
VM	pdDiag	13	352.99	381.84	-163.49	0.0006	0.9997
	pdSymm	16	354.15	389.66	-161.07	4.8402	0.1839
	pdIdent	14	355.77	386.84	-163.88		
RI	pdDiag	17	353.88	391.62	-159.94	7.8838	0.0485
	pdSymm	23	365.01	416.06	-159.50	0.8710	0.9900

df: degrees of freedom; logLik: log-likelihood value; LRT: likelihood ratio statistic.

VarianceCcovariance identity matrices of the fitted models had the following estimates: $\widehat{\boldsymbol{D}}_{OR} = 2.12 \times 10^{-5}\mathbf{I}_{(3)}$; $\widehat{\boldsymbol{D}}_{VM} = 10.27 \times 10^{-5}\mathbf{I}_{(3)}$; $\widehat{\boldsymbol{D}}_{RI} = 9.48 \times 10^{-5}\mathbf{I}_{(4)}$ (where $\mathbf{I}$ is a matrix of ones on the main diagonal and zero otherwise).

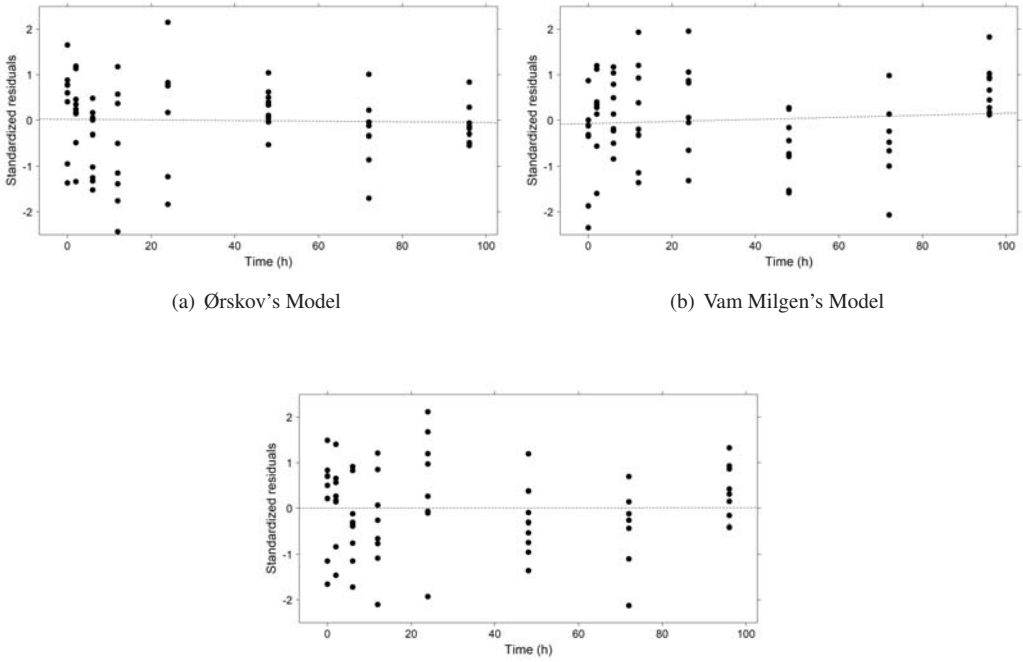
A random effect was considered for each model parameter with a varianceCcovariance identity matrix. In Table 5, we present the results of model adjustments. All possibilities for incorporating random effects were tested. Ørskovs and Van Milgens models had the lowest AIC, BIC, and logLik values when random effects were attributed to parameter β_3 only, being hereafter referred to as OR_3 and VM_3, respectively. For Richards model, this behavior was observed when random effects were attributed to β_4 only (RI_4). For these models, the estimated varianceCcovariance matrices were $\widehat{\boldsymbol{D}}_{OR_3} = 2.12 \times 10^{-5}\mathbf{I}_{(1)}$, $\widehat{\boldsymbol{D}}_{VM_3} = 10.27 \times 10^{-5}\mathbf{I}_{(1)}$, and $\widehat{\boldsymbol{D}}_{RI_4} = 0.1188\mathbf{I}_{(1)}$.

Table 5. Evaluation of the mixed effects of each model parameter using the identity varianceCcovariance matrix ($\boldsymbol{D}$)

Model	i	Mixed parameters	AIC	BIC	logLik
	1	β_1	362.9761$^{(5)}$	387.3906$^{(5)}$	-170.4880
	2	β_2	364.2511$^{(7)}$	388.6657$^{(7)}$	-171.1255
	3	β_3	361.3103$^{(1)}$	385.7249$^{(1)}$	-169.6552
OR_i	4	$\beta_1\beta_2$	363.5209$^{(6)}$	387.9354$^{(6)}$	-170.7604
	5	$\beta_1\beta_3$	361.3112$^{(4)}$	385.7258$^{(4)}$	-169.6556
	6	$\beta_2\beta_3$	361.3112$^{(3)}$	385.7258$^{(3)}$	-169.6556
	7	$\beta_1\beta_2\beta_3$	361.3104$^{(2)}$	385.7249$^{(2)}$	-169.6552
	1	β_1	359.7734$^{(5)}$	384.1880$^{(5)}$	-168.8867
	2	β_2	360.9057$^{(7)}$	385.3203$^{(7)}$	-169.4528
	3	β_3	348.9860$^{(1)}$	373.4005$^{(1)}$	-163.4930
VM_i	4	$\beta_1\beta_2$	360.4013$^{(6)}$	384.8159$^{(6)}$	-169.2007
	5	$\beta_1\beta_3$	348.9864$^{(3)}$	373.4010$^{(3)}$	-163.4932
	6	$\beta_2\beta_3$	348.9861$^{(2)}$	373.4007$^{(2)}$	-163.4931
	7	$\beta_1\beta_2\beta_3$	348.9866$^{(4)}$	373.4012$^{(4)}$	-163.4933
	1	β_1	359.6773$^{(13)}$	390.7504$^{(13)}$	-165.8386
	2	β_2	363.1060$^{(15)}$	394.1791$^{(15)}$	-167.5530
	3	β_3	355.7812$^{(10)}$	386.8543$^{(10)}$	-163.8906
	4	β_4	347.8875$^{(1)}$	378.9606$^{(1)}$	-159.9437
	5	$\beta_1\beta_2$	360.2088$^{(14)}$	391.2819$^{(14)}$	-166.1044
	6	$\beta_1\beta_3$	355.7872$^{(12)}$	386.8603$^{(12)}$	-163.8936
	7	$\beta_1\beta_4$	348.7045$^{(3)}$	379.7776$^{(3)}$	-160.3522
RI_i	8	$\beta_2\beta_3$	355.7867$^{(11)}$	386.8598$^{(11)}$	-163.8933
	9	$\beta_2\beta_4$	348.0757$^{(2)}$	379.1488$^{(2)}$	-160.0378
	10	$\beta_3\beta_4$	355.7630$^{(8)}$	386.8361$^{(8)}$	-163.8815
	11	$\beta_1\beta_2\beta_3$	355.7809$^{(9)}$	386.8540$^{(9)}$	-163.8905
	12	$\beta_1\beta_2\beta_4$	348.8564$^{(4)}$	379.9295$^{(4)}$	-160.4282
	13	$\beta_1\beta_3\beta_4$	355.7562$^{(6)}$	386.8293$^{(6)}$	-163.8781
	14	$\beta_2\beta_3\beta_4$	355.7557$^{(5)}$	386.8288$^{(5)}$	-163.8779
	15	$\beta_1\beta_2\beta_3\beta_4$	355.7579$^{(7)}$	386.8310$^{(7)}$	-163.8790

$^{(\cdot)}$:column values ranks by model

Having decided in which parameters to use random effects and their varianceCcovariance structure, we then applied the within-subject variance-covariance structure $\boldsymbol{R}_i$, in Eq. (3). The graphs depicted in Figure 2 show that the variability of residuals differs between models but not over time.



(a) Ørskov's Model (b) Vam Milgen's Model

(c) Richard's Model

Figure 2. Time versus Standardized residuals

In an attempt to remove such effect, we modeled residual variance as a function of time. For this, we considered different variances for each time period using the `varIdent` variance function class of the `nlme` package in $\boldsymbol{G}_i$ (3). For the autocorrelation $\boldsymbol{\Gamma}_i$ matrix, we used three standard structures from the `nlme` package, namely autoregressive of order 1 (AR), compound symmetry (CS), and general linear (GL). We fitted the three models using the proposed $\boldsymbol{\Gamma}_i$ matrices with and without `varIdent` (ID). The results demonstrated that CS associated with ID produced the best results for all models (Table 6).

Table 6. Comparison of model performance for different alternatives of the $\boldsymbol{R}$ matrix

Model	$\boldsymbol{R}$	df	AIC	BIC	logLik	Test	LRT	$p-$value
		11	361.3103	385.7249	-169.6552			
	AR	12	363.3103	389.9444	-169.6552	1 X 2	0.0000	0.9999
	AR+ID	19	349.3578	391.5285	-155.6789	2 X 3	27.9525	0.0002
OR_3	CS	12	362.9669	389.6010	-169.4835	3 X 4	27.6091	0.0003
	CS+ID	19	345.8534	388.0241	-153.9267	4 X 5	31.1135	0.0001
	GL	12	363.3103	389.9444	-169.6552	5 X 6	31.4569	0.0001
	GL+ID	19	349.3578	391.5285	-155.6789	6 X 7	27.9525	0.0002
		11	348.9860	373.4005	-163.4930			
	AR	12	350.9860	377.6200	-163.4930	1 X 2	0.0000	1.0000
	AR+ID	19	346.3521	388.5227	-154.1760	2 X 3	18.6339	0.0094
VM_3	CS	12	347.9360	374.5701	-161.9680	3 X 4	15.5839	0.0292
	CS+ID	19	346.0176	388.1883	-154.0088	4 X 5	15.9184	0.0259
	GL	12	350.9860	377.6200	-163.4930	5 X 6	18.9683	0.0083
	GL+ID	not convergence						
		14	347.8875	378.9606	-159.9437			
	AR	15	349.8874	383.1800	-159.9437	1 X 2	0.000075	0.9931
	AR+ID	22	340.7743	389.6035	-148.3872	2 X 3	23.113090	0.0016
RI_4	CS	15	353.9218	387.2145	-161.9609	3 X 4	27.147531	0.0003
	CS+ID	22	340.7083	389.5375	-148.3542	4 X 5	27.213515	0.0003
	GL	15	349.8836	383.1763	-159.9418	5 X 6	23.175322	0.0016
	GL+ID	22	340.7605	389.5897	-148.3803	6 X 7	23.123121	0.0016

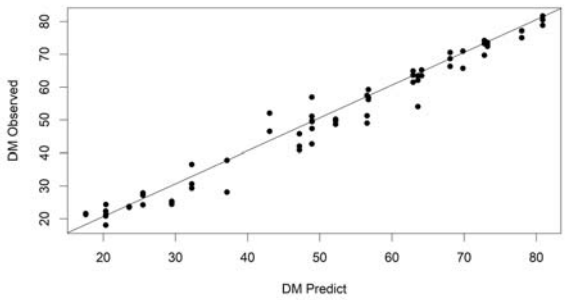
After analyzing the predictive capacity of models, as shown in Table 7 and Figures 3 and 4, we concluded that the

VM_3+CS+ID model had the best performance. The graph of observed versus predicted values of the model showed high linearity, with low RSS, MSE, and RMSE values. Furthermore, the high statistics for the NLME model corroborate this result.

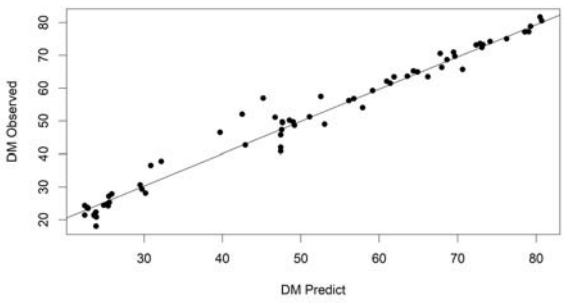
Table 7. Comparison of model performance

		Models		
	Statistics	OR_3+CS+ID	VM_3+CS+ID	RI_4+CS+ID
DM	intercept	0.8130	0.4242	1.4514
	slope	0.9716	0.9941	0.9785
	RSS	764.1455	638.3976	686.9378
	MSE	11.2374	9.3882	10.1020
	RMSE	3.3522	3.0640	3.1784
	R^2	0.9707	0.9756	0.9737
NLME	ME	0.9690	0.9755	0.9728
	NME	0.9699	0.9761	0.9735
	Corr	0.9853	0.9877	0.9868
	ConCorr	0.9847	0.9877	0.9865

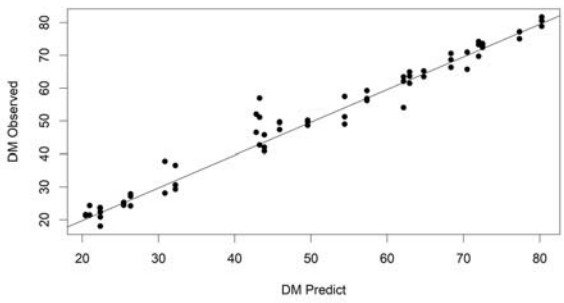
DM, curve of observed versus predicted dry matter degradability values; RSS, residual sum of squares; MSE, mean squared error; RMSE, root-mean-square error; ME, model efficiency; NME, normalized model efficiency.



(a) Ørskov's Model

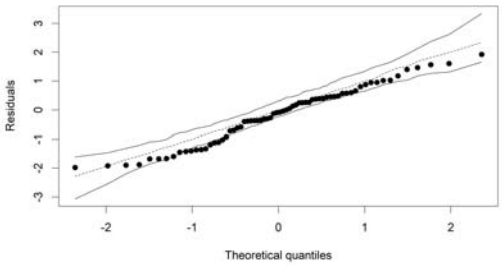


(b) Van Milgen's Model

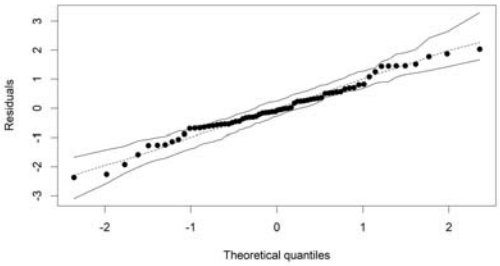


(c) Richard's Model

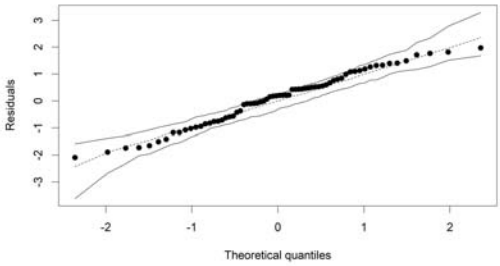
Figure 3. Observed dry matter (DM) degradability values versus DM values predicted using the with SC+ID



(a) Ørskov's Model



(b) Vam Milgen's Model



(c) Richard's Model

Figure 4. Simulated envelope for models fitted with the CS+ID

Table 8 shows the results of Tukeys test, obtained by the function `emmeans` of the package `emmeans` (Lenth, 2021). T_3 afforded a higher soluble fraction ($\hat{\beta}_1 = 47.4471$), differing significantly from treatments T_1 and T_2 at the 5% significance level. The largest degradable fraction was obtained with T_2 ($\hat{\beta}_2 = 51.2885$), and all treatments differed from each other in this parameter. Joint fractional rate of latency and degradation (β_3), however, did not differ between treatments. The results depicted in Figure 5 support the previous discussion, as T_3 starts at time zero with DM values clearly higher than that of other treatments. Total degradability was virtually equal for all treatments at 45 h; thus, the degradability rates of T_1 and T_2 were higher up to 45 h.

Table 8. Estimates and contrasts of parameters β_1, β_2 and β_3, with standard errors (SE) and 95% confidence limits (CL) of the model VM_3+CS+ID

Parameter	T_i	Estimate	SE	df	Lower CL	Upper CL
β_1	1	23.6593	0.51416	8	22.4736	24.8449
	2	22.4619	0.52545	8	21.2502	23.6735
	3	47.4471	0.50861	8	46.2742	48.6200
β_2	1	46.1181	1.11124	52	43.8882	48.3479
	2	51.2885	1.01282	52	49.2561	53.3208
	3	33.9917	1.06929	52	31.8460	36.1374
β_3	1	0.0541	0.00306	52	0.0480	0.0603
	2	0.0621	0.00320	52	0.0557	0.0686
	3	0.0547	0.00341	52	0.0479	0.0616
	T_i contrast	Estimate	SE	df	$t-$ratio	$p-$value*
β_1	1 - 3	-23.7878	0.7232	8	-32.892	0.0000
	2 - 3	-24.9852	0.7312	52	-34.166	0.0000
	2 - 1	-1.1974	0.7351	8	-1.629	0.2890
β_2	1 - 3	12.1263	1.5421	52	7.863	0.0000
	2 - 3	17.2967	1.4728	52	11.744	0.0000
	2 - 1	5.1703	1.5035	52	3.439	0.0032
β_3	1 - 3	-0.0005	0.0045	52	-0.128	0.9909
	2 - 3	0.0074	0.0046	52	1.587	0.2600
	2 - 1	0.0080	0.0044	52	1.809	0.1765

*: Tukey test

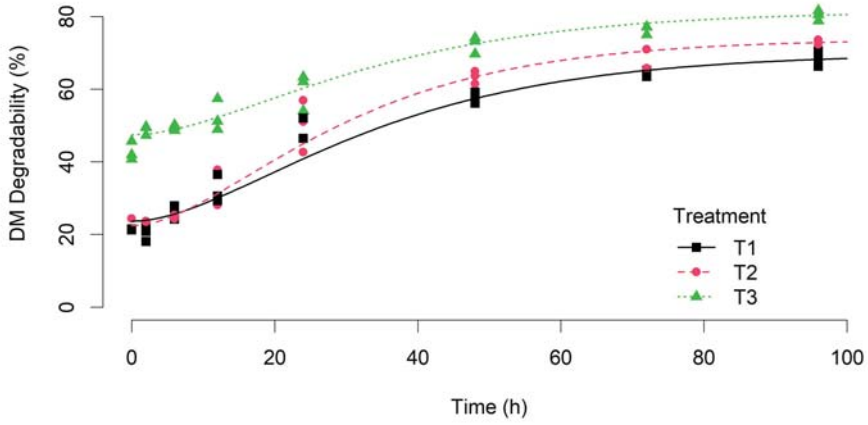


Figure 5. Prediction curves of the adjusted VM_3+CS+ID model

4. Conclusion

This study aimed to identify a nonlinear mathematical model for the study of ruminal degradability. The model proposed by Van Milgen provided better results than Ørskovs and Richards models.

Originally, these are fixed effects models that do not contemplate the addition of random effects to parameters nor the modeling of their variance and covariance structures.

However, assuming fixed effects for each animal/period/treatment combination, we observed high coefficient of variations of estimates, indicating that random effects components could improve the results. This observation was confirmed by addition of random effects of animals using NLME. This method made it possible to consider the compound symmetry autocorrelation matrix and identity covariance structure, resulting in improvements in model residuals and, consequently,

greater precision in parameter estimates.

In addition, it was possible and easy compare treatments by the parameters contrasts test using the EMM function.

References

Calama, R., & Montero, G. (2004). Interregional nonlinear heightdiameter modelwith random coefficients for stone pine in spain. *Canadian Journal of Forest Research*.

Crecente-Campo, F., Tom, M., Soares, P., & Diguez-Aranda, U. (2010). A generalized nonlinear mixed-effects height-diameter model for eucalyptus globulus l. in northwestern spain. *Forest Ecology and Management*, *259*(5), 943-952. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378112709008676 doi: https://doi.org/10.1016/j.foreco.2009.11.036

Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(4), 387.

France, J., Dijkstra, J., & Dhanoa, M. (1996, 01). Growth functions and their application in animal science. *Annales de Zootechnie*, *45*, 165-174.

Gregoire, T. G., & Schabenberger, O. (1996). A non-linear mixed-effects model to predict cumulative bole volume of standing trees. *Journal of Applied Statistics*, *23*(2-3), 257-272. Retrieved from https://doi.org/10.1080/02664769624233

Lenth, R. V. (2021). emmeans: Estimated marginal means, aka least-squares means [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=emmeans (R package version 1.6.2-1)

Littell, R. (2006). *Sas for mixed models*. SAS Institute. Retrieved from https://books.google.com.br/books?id=QstNmAEACAAJ

Luwanda, A. G., & Mwambi, H. G. (2016). A nonlinear mixed-effects model for multivariate longitudinal data with dropout with application to hiv disease dynamics. *Journal of Agricultural, Biological, and Environmental Statistics*, *21*(2), 277–294. Retrieved from https://doi.org/10.1007/s13253-015-0242-1

Medeiros, S. D. S., Lima, C. G., Savian, T. V., Malheiros, E. B., & Werner, S. S. (2020). Mixed nonlinear models in ruminal in situ degradability trials. *Ciłncia Animal Brasileira*, *21*. Retrieved from https://doi.org/10.1590/1809-6891v21e-57596

Mehrez, A. Z., & Ørskov, E. R. (1977). A study of artificial fibre bag technique for determining the dig estibility of feeds in the rumen. *The Journal of Agricultural Science*, *88*(3), 645650.

Miguez, F. (2021). nlraa: Nonlinear regression for agricultural applications [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=nlraa (R package version 0.89)

Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Nonlinear regression models and applications. In *Applied statistics in agricultural, biological, and environmental sciences* (p. 401-447). John Wiley AND Sons, Ltd. Retrieved from https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/appliedstatistics.2016.0003.c15

Ørskov, E. R., & McDonald, I. (1979). The estimation of protein degradability in the rumen from incubation measurements weighted according to rate of passage. *The Journal of Agricultural Science*, *92*(2), 499503.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=nlme (R package version 3.1-152)

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus*. New York, NY [u.a.]: Springer.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rossi, R. M., Martins, E. N., Guedes, T. A., & Jobim, C. C. (2010). Bayesian analysis for comparison of nonlinear regression model parameters: an application to ruminal degradability data. *Revista Brasileira de Zootecnia [online]*, *39*(2), 419-424.

Sartrio, S. D. (2013). *Modelos n£o lineares mistos em estudo de degradabilidade ruminal in situ*. phdthesis, Escola Superior de Agricultura Luiz de Queiroz, Universidade de S£o Paulo, Piracicaba.

Schabenberger, O. v., & Pierce, F. J. v. (2002). *Contemporary statistical models for the plant and soil sciences*. Boca Raton (Fla.) : CRC press. Retrieved from http://lib.ugent.be/catalog/rug01:002054341

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of r2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach. *BMC Pharmacology*, *10*(1), 6. Retrieved from https://doi.org/10.1186/1471-2210-10-6

Tasissa, G., & Burkhart, H. (1997, 02). Modeling thinning effects on ring width distribution in loblolly pine (pinus taeda). *Canadian Journal of Forest Research*, *27*, 1291-1301.

Teixeira, U. H. G., Simionir, T. A., Bezerra, R. P., Soares, K. A. R. S. C., Romualdo, T. G., Torres, R. N., Santos, ... Pina, D. S. (2016). Mathematical models for estimating the parameters of ruminal degradation kinetics of protein

concentrates. *Revista Brasileira de Sade e Produ£o Animal [online]*, *17*(1), 73-85.

Van Milgen, J., Murphy, M., & Berger, L. (1991). A compartmental model to analyze ruminal digestion. *Journal of Dairy Science*, *74*(8), 2515-2529. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022030291784294 doi: https://doi.org/10.3168/jds.S0022-0302(91)78429-4

Wang, K. (2016). Linear and non-linear mixed models in longitudinal studies and complex survey data. *J Biom Biostat*, *7*(290), 2.

Wyzykowski, J., Custdio, A. A. d. P., Custdio, A. A. d. P., Gomes, N. M., & Morais, A. R. d. (2015). Analysis of the diameter coffee canopy after pruning through nonlinear mixed model. *RevistaBrasileira de Biometria*, *33*(3), 243-256. (S£o Paulo)

Xu, H., Sun, Y., Wang, X., Fu, Y., Dong, Y., & Li, Y. (2014). Nonlinear mixed-effects (nlme) diameter growth models for individual china-fir (cunninghamia lanceolata) trees in southeast china. *PloS one*, *9*, e104012.

Yang, Y., Huang, S., Trincado, G., & Meng, S. X. (2009). Nonlinear mixed-effects modeling of variable-exponent taper equations for lodgepole pine in alberta, canada. *European Journal of Forest Research*, *128*(4), 415–429. Retrieved from https://doi.org/10.1007/s10342-009-0286-2

International Journal of Statistics and Probability; Vol. 11, No. 5; September 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Forecasting Hydropower Generation in Ghana Using ARIMA Models

Smart Asomaning Sarpong[1], Akwasi Agyei[1,]

[1] Institute of Research, Innovation and Development-IRID, Kumasi Technical University, Ghana

Received: April 18, 2022 Accepted: May 26, 2022 Online Published: September 7, 2022

doi:10.5539/ijsp.v11n5p30 URL: https://doi.org/10.5539/ijsp.v11n5p30

Abstract

In this study, an Autoregressive Integrated Moving Average (ARIMA) model was used to forecast Ghana's Akosombo dam level and hydropower generation by the end of year 2022. Data used for this study span from January 2010 to December 2019. Base on the final ARIMA model, power generation is forecasted to decrease from 398 Megawatts/hr in December 2019 to approximately 374 Megawatts/hr by December 2022. On the other hand, water level of the Akosombo dam is predicted to decrease marginally from 264.8 ft in December 2019 to approximately 255.19 ft by December 2022. The Volta River Authority (VRA) and managers of the electricity production in Ghana are encouraged to be proactive in expanding energy production by turning more to renewable energy sources. In the coming years, as they seek to provide sustainable electricity for their cherished customers, investment decisions should be directed towards protecting the volta river from drying up due to human and climatic activities as well as expanding energy mix.

Keywords: Hydropower, Akosombo Dam, ARIMA, Ghana

1. Introduction

Energy is a key component to sustaining economic growth and general well-being. Many of our daily activities is dependent on stable electricity. In our global quest to achieving the sustainable development goals, stable electricity has been deemed crucial (Owusu et al., 2016)**.** According to the International Energy Agency (IEA, 2015), about 10 % of the population of the world are without electricity, with 22% out the population are without electricity living in sub-Saharan Africa and some parts of Asia. More scientific studies must be focusing on helping to improve electricity supply in developing countries within sub-Saharan Africa and Asia.

According to the World Energy Outlook report by the International Energy Agency (2015), fossil fuels account for about 67% of the worldwide power generation. Despite its availability worldwide, the negative effects on the environment as a results of generating electricity from fossil fuel, climate change, depletion of fossil fuel reserves, and price volatility need a worldwide increase in the use of renewable resources for power generation (IEA, 2015; NREL, 2015; Panwar et al., 2011**)**.

Globally, hydropower remains the largest renewable energy resource due to its cost-effectiveness and reliability (Zhou et al., 2020). Ghana has two major hydropower system made up of two plants. The hydropower plants are located at Akosombo and Kpong and were commissioned in 1965 and 1982 respectively. Akosombo has six turbine-generator units which operates between 84.15m (276ft) maximum and 75.59m (248ft) minimum of headwater elevation. The Kpong plant has four turbines which operates between 17.7m maximum and 14.5m maximum of normal water level. Ghana's electricity generation was once powered by diesel generators prior to the construction of Akosombo dam (Eshun & Amoako Tuffour, 2016). It is estimated that, about 65% of Ghana's electricity is sourced from Akosombo dam and the remaining 35% generated from other sources (Gyamfi et al., 2015). However, the sustainability of hydropower is directly influence by the water level at any given point in time (Harrison, & Whittington, 2001; Miescher, 2021).

Ghana is among the West African countries blessed with so many renewable resources (IRENA, 2015). With the need to increase energy access and also protect the environment, renewable energy development has become key element in the sustainable future energy agenda. The energy sector has been searching for sustainable solution to the periodic power outages and load shedding. The many years of reliance on hydropower generation, the slow pace of innovative energy mix, and growth in economic activities and population are some of the causes of the unstable electricity supply to many parts of Ghana (Asumadu-Sarkodie & Owusu, 2016). Recent statistics indicate that the electricity generating mix of the country is mainly made up of thermal and hydro sources, with steps being made to add non-renewable sources like solar. Currently, about 59.9% of the total electricity generated in Ghana is derived from thermal sources (natural gas, light crude oil and diesel fuels), 39.9% is derived from hydro and the 0.2% left is derived from solar technology (Ankrah & Lin, 2020).

However, despite the 39.9% of electricity derived from hydro, there has being a significant decline in hydropower since 2014. Ashong (2016) indicated that inadequate water inflow into the hydro dams as a results of low rainfall have been the main reason for the decline in hydropower generation, hence the major cause of unstable renewable energy state in Ghana.

There are several models used for forecasting in time series. This study mainly applies the Autoregressive Integrated Moving Average (ARIMA) model for predicting power generation and dam level. ARIMA models depend on past values to predict the future. The ARIMA model consist of three components (p, d, q), where p is the order of AR process, d is the difference order, and q is the order of MA process. The ARIMA model is one of the most used techniques by many researchers due to its reliability (Debnath & Mourshed, 2018; Ediger & Akar, 2007). Sarpong (2013) found out that the use of ARIMA model is adequate for forecasting. In addition, El Desouky, & Elkateb, 2000) revealed that the use of ARIMA model for forecasting provides smaller errors.

In previous studies, the Owusu et al. (2018) found out that, electricity generation will decline if alternative power sources are not urgently considered by Government. Boadi and Owusu, (2019) in their study on climate change and its effect on hydropower in Ghana using monthly data from 1970 to 2010 concluded that, 21% of Ghana's unstable electricity supply was due to shortfall in water levels of Akosombo hydroelectric power station. Michieka et al., (2021) found out that, long-run positive shock in temperature increases electric power production. According to Asian Development Bank (2007), drought causes more shortages resulting in outages and insufficient cooling water which ultimately decrease hydropower production.

Moreover, Ediger and Akar (2007) forecasted primary energy demand by fuel using ARIMA model. The forecasted result shows that primary energy demand will decrease between 2005 and 2020. Mite-León and Barzola-Monteses (2018) used ARIMA models in forecasting hydropower generation in Ecuador. The outcome of the study showed monthly increased in hydropower generation in Ecuador. Kabo-bah et al. (2016) found out that, regular low flow of water into the Akosombo dam affects power generation.

However, most of the previous studies developed on energy focused more on renewable resources, predicting energy consumption and in particular, overall energy production (Dind etal., 2018; Katani, 2019; Kaur & Ahuja, 2017; Sarkodie, 2017; Wu et al., 2017). Some too focused on factors affecting hydropower generation (Kabo-bah et al, 2016; Michieka et al., 2021). Notwithstanding the above, hydropower generation forecast in developing countries, like Ghana, has attained very little attention. Energy production forecast is of great importance to the operators of electrical system and decision makers to defined better policies and manage risks.

Also, from previous studies, different models have been used by different researchers in forecasting hydropower generation. Owusu et al. (2018) used Polynomial regression. Zolfaghari and Golabi (2021) used adaptive wavelet transform (AWT), long short-term memory (LSTM) and random forest (RF) algorithm (AWT-LSTM-RF) to predict the electricity production in hydroelectric power plant. Dmitrieva (2015) combined Neural Networks, SVM and ARIMA models in forecasting hydropower plant production. Mite-León and Barzola-Monteses (2018) **u**sed ARIMA with seasonal component in predicting hydropower generation in Ecuador.

From the year 2003, the Energy Commission of Ghana decided on an annual increase in power supply of 0.9 to 1.8% due to increasing population and economic activities (EC, 2013)**.** It is clear from the above that, trends for future hydropower generation and water level of the longest serving source of electricity in Ghana (the Akosombo dam) is crucial to overcoming power supply challenges.

This study therefore attempts to forecast Ghana's hydropower generation as well as water level of the Akosombo dam using the Autoregressive Integrated Moving Average (ARIMA) technique. Many of the existing literature on ARIMA forecasting models usually ignore analysis of forecasting errors (Koutroumanidis et al., 2009; Ömer Faruk, 2010; Khashei & Bijari, 2011). In this study however, error analysis is performed using Root Mean Square Error (RMSE) and Mean Absolute Percentage error (MAPE) to evaluate the forecasting accuracy of the selected model. The Volta River Authority (VRA) and managers of the electricity company of Ghana may find this study very useful in the planning for the coming years as they seek to provide sustainable electricity for their cherished customers. This study may promote the need for intervention programs to protect the Volta river from drying up due to human and climatic activities.

2. Materials and Methods

The study used two secondary univariate time series data; power generation and dam level, which span from January 2010 to December 2019. Since the data was measured over time, and uniformly spaced, we considered utilizing the Box-Jenkins strategy (Shumway et al., 2000). The time series forecasting by using ARIMA models can be performed in four basic steps namely, Identification, Estimation, Diagnosis and Forecasting (Box et al., 2015), to end up with a specific formula that satisfies all the underlying conditions as much as possible to produce good and accurate forecast.

2.1 Autoregressive Integrated Moving Average Process (ARIMA)

ARIMA model is a type of Box-Jenkins series analysis which depends on past values to predict the future (Devi et al., 2013). The modelling is done using the integrated autoregressive and moving average processes. The ARIMA (p, d, q) model is divided into three main parts: The Autoregressive (AR) part of order p, which explains the present value of a series by the function of p past values, the Moving Average (MA) part of order q, which indicates that the output variable depends linearly on current and various past values, and the differenced (d) part which indicates that the data values have been replaced with the difference between the values and the previous values.

The Box-Jenkins methodology apply the maximum likelihood principles in parameter estimation. Using a modified form of Mite-León and Barzola-Monteses (2018) model approch, the ARIMA (p, d, q) model is expressed as:

$$X_t^d = c + \phi_1 X_{t-1}^d + \cdots + \phi_p X_{t-p}^d + \theta_1 X_{t-1}^d + \cdots + \theta_q X_{t-q}^d + \varepsilon_t \quad (1)$$

where,

X_t^d is the series with a difference order (d)

$\phi_1,\ldots,\phi_p$ and $\theta_1,\ldots,\theta_q$ are the model parameters

ε_t represent white noise with i.i.d.

The ARIMA (p, d, q) model is to make nonstationary time series stationary by d order difference.

2.2 Unit Root Evaluation

Test for a unit root is one of the basic assumptions underlying any time series data. This study made use of time-plot of the data and the ADF statistical test in evaluating the stationarity of the two series. The ADF was based on the assumption that the series can approximate an autoregressive of order 1 (Mite-León and Barzola-Monteses, 2018). The ADF test is performed under the null hypothesis, the series has a unit root. The regression equation of the ADF test is given by:

$$\Delta y_t = \alpha + \beta_t + \gamma_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_p \Delta y_{t-p} + \varepsilon_t \quad (2)$$

Where,

y_t is the observed time series

α is constant

β is the coefficient of the time trend

p is the order of AR process.

If $\gamma = 0$, the series is random walk and if $-1 < 1+\gamma < 1$, the series is stationary.

2.4 Model Identification

Before applying the ARIMA model, the Autoregressive (AR) component p, and the Moving Average (MA) component q, was identified using ACF and PACF plots respectively. According to Box and Pierce (1970), the ACF and the PACF are correlogram functions that help to decide the degree of association between two successive values of the series and give an idea of the possible parameters of the ARIMA model. By following Polprasert et al., (2021) method of identification, the ACF and PACF were drawn to stationary time series and the p and q values were evaluated based on truncation and trailing nature of the function. Truncation refers to the nature in which the ACF or PACF time sequence is zero (0) after some time, and trailing refers to the nature in which the ACF or PACF slowly shrinks to zero (0). If the PACF is truncating and ACF is trailing, then p equals the truncation order, q equals 0, and it can be concluded that the sequence fits AR model. If PACF of the stationary series is trailing and ACF is truncating, then q equals the truncating order, p equals 0 and it can be concluded that the sequence fits MA model; if both the PACF and ACF are trailing, then p equals the PACF truncation order, q equals the ACF truncation order, and the model fits the ARMA model.

2.5 Model Estimation

After the parameters (p, d, q) of the ARIMA model have been identified, the model is then estimated to obtain the coefficients. The maximum likelihood estimation is used in this study to get the estimates of the coefficients of the suggested models at the identification stage. We fit all the suggested models at the identification stage to the series to obtain estimates of the coefficients.

2.6 Model Selection

After a successful estimation of the model, the Akaike Information Criteria (AIC) and the significance of the models will be accessed to determine the best model for our series. It is expected that the ARMA components would be

significance at 5% level of significance after estimation and return the minimum AIC value. AIC is mathematically expressed as:

$$AIC = 2k - 2\log(L) \tag{3}$$

where,

L is the maximum likelihood value

k is the number of parameters to be estimated.

When these conditions are satisfied, that model is then selected as the best model for the series.

2.7 Model Diagnostics

The Box-Jenkins methodology also provides an avenue to access the goodness of fit of the selected model. It is expected that after a complete estimation, the residuals of the selected model would exhibit the following characteristics: the residuals should be white noise, the ARMA process should be covariance stationary, thus all the AR roots must lie inside the unit circle, the ARMA process should be invertible, thus all the MA roots must lie inside the unit circle. The study employs the Ljung-Box test to test whether the residuals are white noise or not. The test is express as:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \tag{4}$$

Where Q is asymptotic distribution which a Chi-square distribution with degrees of freedom $h = m - p - q$, p and q are the orders of AR and MA, n is the sample size, r_k is the estimated autocorrelation of the time series at lag k, and m is the number of lags to be tested.

2.8 Model Forecasting and Evaluation

Once the selected model has been verified, the model will then be used to predict power generation and dam level in the next 36 months. After the forecasting, this study employs the Root Mean Square Error (RMSE) and Mean Absolute Percentage error (MAPE) to evaluate the forecasting accuracy of the selected model. The RMSE and MAPE are given by:

$$RMSE = \sqrt{\frac{\sum(P_i - O_j)^2}{n}} \tag{5}$$

$$MAPE = \frac{100}{n}\left|\frac{(P_i - O_j)}{O_j}\right| \tag{6}$$

where P_i is the predicted value for the ith observation, O_i is the observed value for the jth observation, n is the number of non-missing residuals.

3. Results

3.1 Graphs of the Series

It can be observed from figure 1 and 2 that, both series depicts a change in mean over time which suggest that both series are non-stationary.

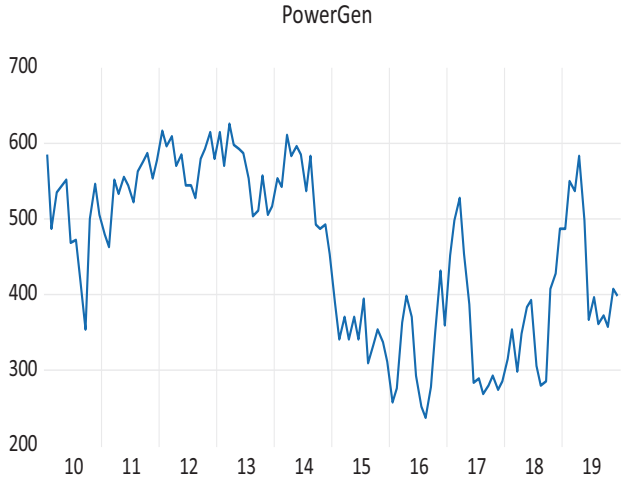


Figure 1. Power generation

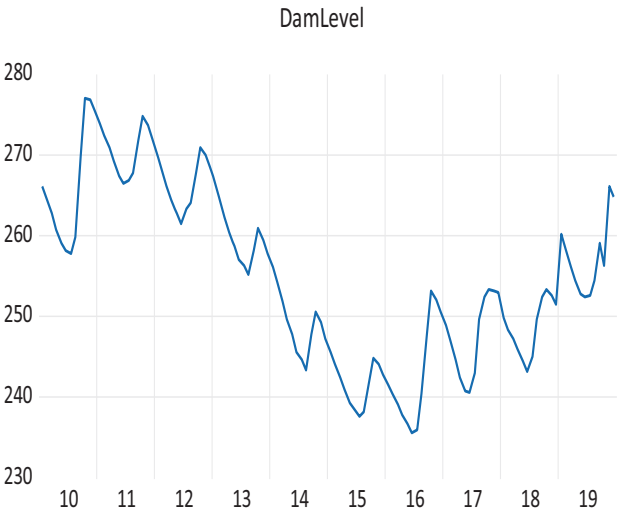


Figure 2. Dam level

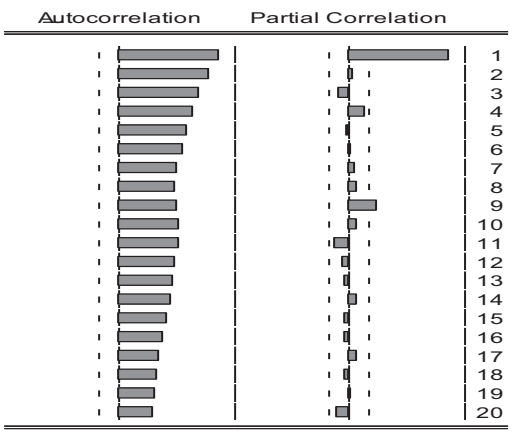


Figure 3. Correlogram for power generation

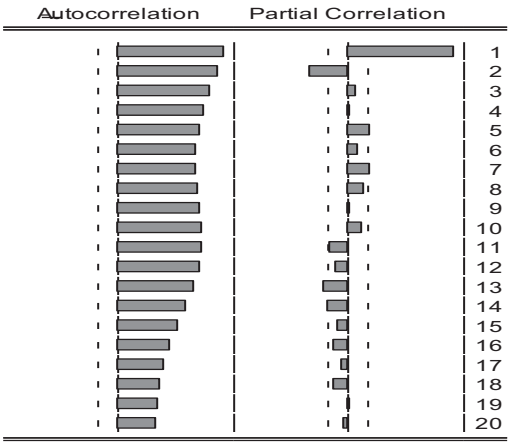


Figure 4. Correlogram for dam level

It can be observed from both correlogram plots (Figure 3 and 4) that, the ACF decline very slowly, which also suggest that both series are not stationary. This can be confirmed using the ADF test. Table 1 and 2 below display the results of ADF statistical test for stationarity 4 both power generation and dam level respectively.

Table 1. ADF test results for Power generation

Without differencing

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-2.512378	0.1151

After first differencing

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-7.532736	0.0000

Table 2. ADF test results for dam level

Without differencing

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-0.909780	0.7820

After first differencing

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-6.823674	0.0000

It can be confirmed from the ADF test that both series were not stationary at their levels and became stationary after first differencing. This confirmed the use of ARIMA (p,d,q) model to estimate our models and make predictions.

3.2 Model Estimation

It is therefore appropriate to determine the various ARMA components and the suggested models for both series.

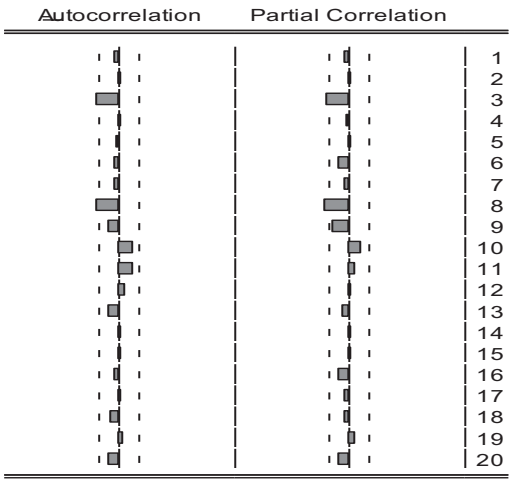


Figure 5. ACF and PACF for power gen

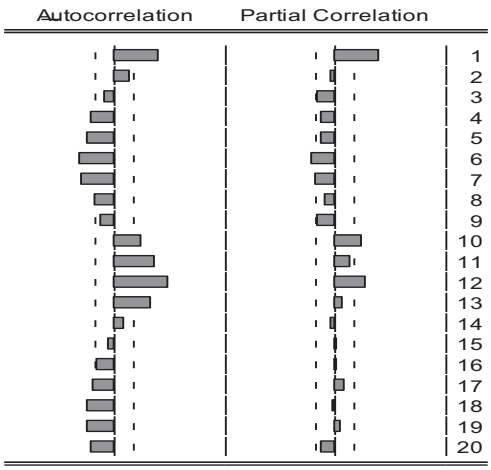


Figure 6. ACF and PACF for dam level

Possible models for power generation are ARIMA (3,1,3), (3,1,8), (8,1,8) and (8,1,3) and the possible models for dam level are ARIMA (1,1,1), (1,1,4), (1,1,5), (1,1,6), (6,1,1), (6,1,4), (10,1,1) etc. After a successful estimation of the possible models, ARIMA (8,1,3) returned significance ARMA coefficients with least AIC and SIC value. Therefore, it is selected as the best model for power generation.

Table 3. Parameters estimates of the selected model for power generation

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.040110	2.863109	-0.363280	0.7171
AR(8)	-0.204761	0.091469	-2.238583	0.0271
MA(3)	-0.234018	0.098148	-2.384329	0.0187
SIGMASQ	2145.712	302.7057	7.088443	0.0000

Akaike info criterion	10.58063
Schwarz criterion	10.67405

Also, ARIMA (10,1,1) returned significance ARMA and least AIC and SIC value. Therefore, it is selected as the best model for dam level.

Table 4. Parameters estimate of the selected model for dam level

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.054919	0.948797	0.057882	0.9539
AR(10)	0.409277	0.077391	5.288431	0.0000
MA(1)	0.427821	0.092820	4.609148	0.0000
SIGMASQ	5.545471	0.789410	7.024829	0.0000

Akaike info criterion	4.635189
Schwarz criterion	4.728605

3.3 Model Diagnostics

It is very advisable to check the goodness of fit of the selected model to see whether it adequately fit the data before forecasting is perform.

3.3.1 Power Generation Model Diagnostics

Table 5. Ljung-Box results

Q-statistic probabilities adjusted for 2 ARMA terms

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.075	-0.075	0.6927	
		2	0.006	-0.000	0.6964	
		3	0.018	0.018	0.7345	0.391
		4	-0.004	-0.001	0.7366	0.692
		5	-0.074	-0.075	1.4301	0.698
		6	-0.059	-0.071	1.8718	0.759

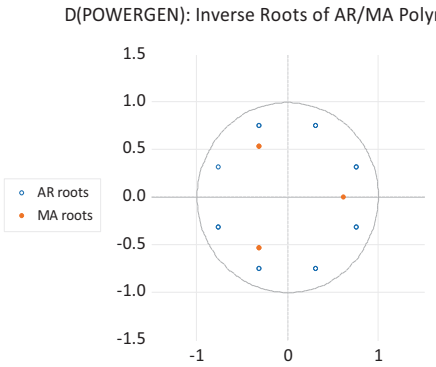


Figure 7. AR/MA roots results

All the AR and MA roots lie inside the unit circles, which shows that the ARMA process is covariance stationary and invertible. Also, all the p-values of Ljung-Box Q-statistics are greater than 5% level of significance, there we fail to reject the null hypothesis and conclude that the residuals are white noise. This confirm that ARIMA (8,1,3) model adequately fits the data.

3.3.2 Dam Level Model Diagnostics

Table 6. Ljung-Box results

Q-statistic probabilities adjusted for 2 ARMA terms

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.034	0.034	0.1370	
		2	0.032	0.031	0.2643	
		3	-0.115	-0.117	1.8984	0.168
		4	-0.105	-0.100	3.2775	0.194
		5	-0.109	-0.097	4.7738	0.189
		6	-0.127	-0.133	6.8149	0.146

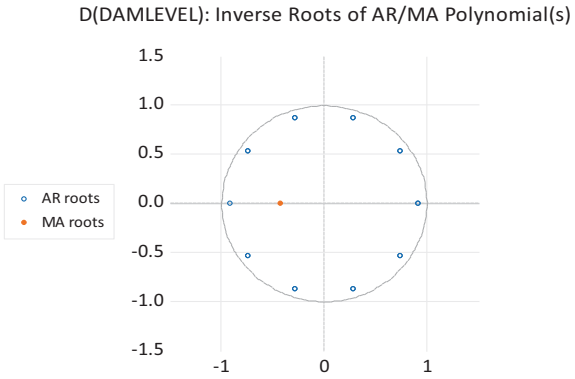


Figure 8. AR/MA root results

The AR and MA roots lie inside the unit circles, which shows that the ARMA process is covariance stationary and invertible. Also, all the p-values of Ljung-Box Q-statistics are greater than 5% level of significance, there we fail to reject the null hypothesis and conclude that the residuals are white noise. This confirm that ARIMA (10,1,1) model adequately fits the data.

3.4 Model Forecasting in the Next 36 Months

Since the selected model has successfully passed the diagnostic stage, power generation and dam level are predicted as below. From figure 9, the vertical axis represents the quantity of power generation in megawatts while the horizontal axis denotes the time in months in which the power was generated. It can be observed that there were fluctuations in power generation from October 2019 to June 2020. Right from June 2020, power generation is forecasted to decrease from 395 Megawatts in June 2019 to approximately 355 Megawatts by December 2022. From figure 10, the vertical axis denotes the volume of the water level in meters whiles the horizontal axis denotes the time in the value which is recorded in months. Based on the predicted values, water level of the Akosombo dam is predicted to increase marginally from 269.6 meters in December 2019 to approximately 275.5 meters by December 2022, despite the fluctuations that may occur.

Table 7. Forecasted values of power generation

Month/Year	Actual values	Predicted values
Oct-19	357	349.2051
Nov-19	407	361.0639
Dec-19	398	355.4225
Jan-20		371.5740
Feb-20		397.3494
*		*
*		*
*		*
Oct-21		370.0246
Nov-21		368.8737
Dec-21		367.8731
Jan-22		366.6854
Feb-22		365.4151
Mar-22		364.4296
Apr-22		363.3315
May-22		362.3126
Jun-22		361.2342
Jul-22		360.2168
Aug-22		359.1689
Sep-22		358.1587
Oct-22		357.1657
Nov-22		356.1144
Dec-22		355.0862

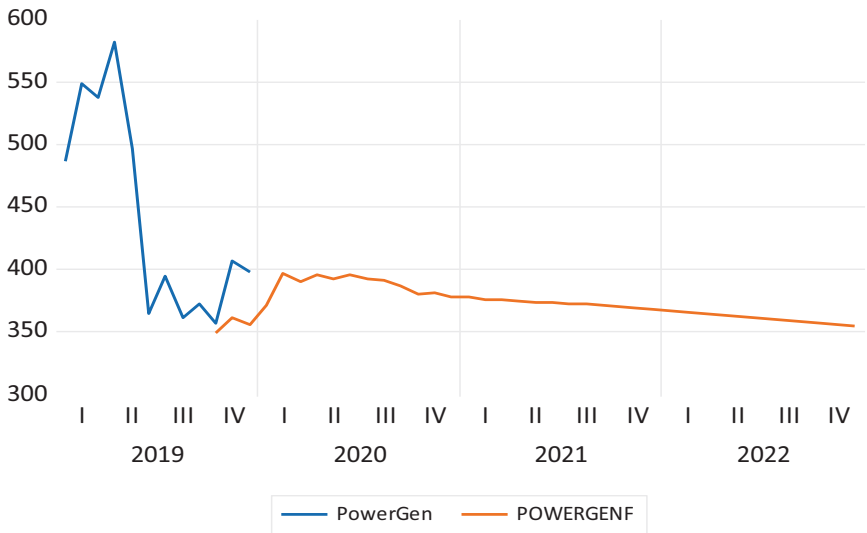


Figure 9. Actual (brown) and Forecasted (blue) graph for power generation

Table 8. Forecasted values of Dam level

Month/Year	Actual values	Predicted values
Oct-19	256.2	268.8725
Nov-19	266	272.1425
Dec-19	264.8	269.6202
Jan-20		268.7932
Feb-20		268.0890
*		*
*		*
*		*
Oct-21		276.8238
Nov-21		276.7653
Dec-21		276.7890
Jan-22		276.8539
Feb-22		277.0353
Mar-22		277.4018
Apr-22		277.2610
May-22		277.9840
Jun-22		278.2834
Jul-22		278.2778
Aug-22		278.2807
Sep-22		278.2892
Oct-22		278.3313
Nov-22		278.3903
Dec-22		278.4970

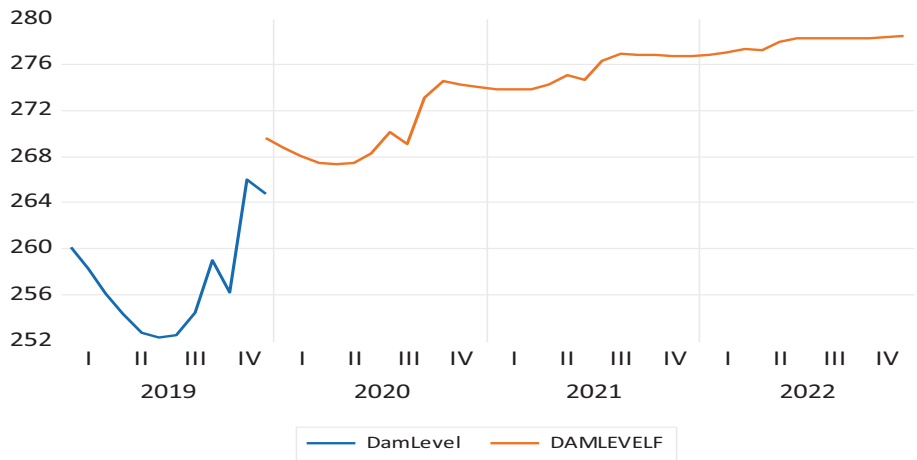


Figure 10. Actual (brown) and Forecasted (blue) graph for dam level

3.5 Forecasting Evaluation

The table below display the evaluation results of our forecasted models for hydropower generation and dam level respectively.

Table 9. Power generation forecasted evaluation results

Model fit statistics	Forecasted value
RMSE	36.4405
MAPE	8.0559

Table 10. Dam level forecasted evaluation results

Model fit statistics	value
RMSE	4.8203
MAPE	1.8203

It can be observed that, both tables presented values which are smaller. This means that the selected model for hydropower generation and that of dam level provide best forecasting accuracy results.

4. Discussion

The objective of the study was to obtain an appropriate ARIMA model that will help forecast hydropower generation and dam level which has a significant impact on hydropower. The Box-Jenkins method was employed to obtain the suitable model for our series. The study made use of two different univariate time series which was obtained monthly. The two datasets were monthly recorded data for power generation and dam level. The study first looked at hydropower and its significance to the development of a country, irregularities in dam levels because of climate and environmental conditions and lastly some of the models that have been used for forecasting hydropower generation. The graph of both series was obtained, and it was found that, there is a change in mean of the two series which shows nonstationary nature of the series.

There were rise and fall in dam level across the sample period, but in a downward pattern from 2011 to 2015 and upward pattern from 2016 to 2019. There is also sharp decline in power generation from 2014 to 2015. ACF plots for both series decline slowly which suggest nonstationary for both series. The ADF statistical test also proved that power generation and dam level are nonstationary. Transformation of the series was done by taking the first difference of both series to obtained stationarity. The correlogram of both series suggested different orders of p and q for AR and MA process. After different estimations, ARIMA (8,1,3) and ARIMA (10,1,1) fulfilled all the model selection criteria for power generation and dam level respectively.

For the validation of the model, the residuals hypotheses were tested. First, the Ljung-Box test was used to determine if the residuals of the selected models are white noise. The test returned p-values greater than 5% significance level up to six lags for both power generation and dam level. Also, from the root statistics, all the AR and MA roots fell inside the unit circle which shows that, the selected models were stationary and invertible respectively. With this validation, a forecast was made.

The results of Ghana's hydropower generation and Akosombo dam level are depicted in Figure 9 and Figure 10 respectively. The fitted model was used to forecast the observed series (2010-2019) and based on that the future series were forecasted as presented in Table 7 and Table 8 for hydropower generation and dam level respectively. Table 7 shows that there were fluctuations in Ghana's hydropower generation from October, 2019 to June 2020. But there is a decrease from 395mWh in June, 2019 to 355.0862mWh in 2022 in a low decrease scenario. Though there may be fluctuations but it won't be much as those observed from October, 2019 to June, 2020. But in all, there is going to be a slight decrease in power generation from December 2019 to December 2022 (355.4225meters to 355.0862meters). Also, Table 8 shows that there was a decrease in the Akosombo dam level from 269.8ft in December, 2019 to 267.4496ft in June 2020. Despite the fluctuations that may occur, there will be an increase in dam level from 269.8m in December, 2019 to 278.4970m by December, 2022 in slow increase scenario. This forecast supports earlier works of (Owusu et al., 2018 and Boadi & Owusu 2019).

Hydropower production phantom the sustainability of a country and Ghana has suffered from power outages since 2014. Jude et al., (2011) found out that, decrease in hydropower generation also decrease energy consumption, which in turn decrease economic growth. According to Ashong (2016), the decline in hydropower generation is mainly due to the inadequate inflow of water into the hydro dams as a results of low rainfall. This signifies that climate variabilities and environmental conditions are main variables that affect water levels and generation of hydropower (Michieka et al., 2021; Miescher, 2021).

The forecasting evaluation results for hydropower generation and dam level are displayed in Table 9 and 10 respectively.

From Table 9, RMSE of 1.16 means that the average distance between the observed series and the predicted values is 1.16. MAPE of 0.4% means that the ARIMA (8,1,3) model-predicted level varies by 0.4% from the observed series. Also, from Table 10, RMSE of 3.85 means that the average distance between the observed series and the predicted values is 3.85. MAPE of 1.45% means that the ARIMA (10,1,1) model-predicted level varies by 0.4% from the observed series. This means that the selected models are statistically sound to make future forecast. In a similar study, Sarkodie (2017) used RMSE and MAPE to analyze forecasting errors in predicting electricity consumption in Ghana.

5. Conclusions

The ARIMA models has revealed that, following the trend of past values of hydropower and water level from the Akosombo dam, both variables will trend downwards in future. Based on the forecasting results obtained, Ghana will experience a slight decrease in hydropower generation despite the increase in water level that will occur, as the results showed. This implies that there should an introduction of new hydro plants that will utilized the excess water to produce more of electricity for the country. The Volta River Authority (VRA) and managers of the electricity production in Ghana are also encouraged to be proactive in expanding energy production by turning more to renewable energy sources. In the coming years, as they seek to provide sustainable electricity for their cherished customers, investment decisions should be directed towards protecting the volta river from drying up due to human and climatic activities as well as expanding energy mix. Government of Ghana should devote funding to support scientific research in renewable energy and energy-harvesting technologies to ensure that enough energy is available for citizens and industries. Future research is encouraged to be done using new variables, to study patterns of power outages, construct economic models and make predictions. This will increase effective decision making process in the energy sector and also help sustain the growth of the economy.

Author Contributions: Conceptualization, S.A.S. and A.A.; methodology, S.A.S., and A.A,; software, S.A.S.; validation, S.A.S., and A.A.; formal analysis, S.A.S. and A.A; investigation, S.A.S; resources, S.A.S., and A. A.; data curation, S.A.S.; writing—original draft preparation, S.A.S., A.A, and A,A; writing—review and editing, S.A.S., and A.A.; supervision, S.A.S.; project administration, S.A.S.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reasons.

Acknowledgments: Staff of the Institute of Research, Innovation and Development - IRID, Kumasi Technical University—KsTU, Ghana.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ankrah, I., & Lin, B. (2020). Renewable energy development in Ghana: Beyond potentials and commitment. *Energy*, *198*, 117356. https://doi.org/10.1016/j.energy.2020.117356

Ashong, J. N. T. (2016). How Effectively Has Ghana Implemented Its Policy for Large-Scale Renewable Electricity Deployment: A Qualitative Assessment. *Renewable Energy Law and Policy Review*, *7*(2), 133-144.

Asian Development Bank. (2007). Asian water development outlook 2007. Retrieved from http://www.adb.org/sites/default/files/publication/27971/awdo.pdf

Asumadu-Sarkodie, S., & Owusu, P. A. (2016). The potential and economic viability of solar photovoltaic power in Ghana. *Energy Sour., Part. A: Recovery, Utilization, Environ. Eff., 38,* 709–716. https://doi.org/:10.1080/15567036.2015.1122682

Boadi, S. A., & Owusu, K. (2019). Impact of climate change and variability on hydropower in Ghana. *African Geographical Review*, *38*(1), 19–31. https://doi.org/10.1080/19376812.2017.1284598

Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, *65*(332), 1509-1526. https://doi.org/10.1080/01621459.1970.10481180

Debnath, K. B., & Mourshed, M. (2018). Forecasting methods in energy planning models. *Renewable and Sustainable Energy Reviews*, *88*(September 2017), 297–325. https://doi.org/10.1016/j.rser.2018.02.002

Devi, B. U., Sundar, D., & Alli, P. (2013). An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. *International Journal of Data Mining & Knowledge Management Process*, *3*(1), 65. https://doi.org/10.5121/ijdkp.2013.3106

Ding, N., Liu, J., Yang, J., & Lu, B. (2018). Water footprints of energy sources in China: Exploring options to improve water efficiency. *Journal of Cleaner Production*, *174*(2018), 1021–1031. https://doi.org/10.1016/j.jclepro.2017.10.273

Dmitrieva, K. (2015). Forecasting of a hydropower plant energy production. 192.

Ediger, V. Ş., & Akar, S. (2007). ARIMA forecasting of primary energy demand by fuel in Turkey. *Energy Policy*, *35*(3), 1701–1708. https://doi.org/10.1016/j.enpol.2006.05.009

El Desouky, A., & Elkateb, M. (2000). Hybrid adaptive techniques for electric-load forecast using ANN and ARIMA. *IEE Proceedings-Generation, Transmission and Distribution*, *147*(4), 213–217. https://doi.org/10.1049/ip-gtd:20000521

Energy Commission Ghana. (2013). Energy Commission Supply and Demand 2013. http://www.energycom.gov.gh/.Accessed October 29, 2021.

Eshun, M. E., & Amoako-Tuffour, J. (2016). A review of the trends in Ghana's power sector. *Energy, Sustainability and Society*, *6*(1), 1-9. https://doi.org/10.1186/s13705-016-0075-y

Gyamfi, S., Modjinou, M., & Djordjevic, S. (2015). *Improving electricity supply security in Ghana — The potential of renewable energy*, *43*, 1035–1045. https://doi.org/10.1016/j.rser.2014.11.102

Harrison, G., & Whittington, B. (2001). Climate change: A storm brewing for hydro. *International Water Power and Dam Construction*, *53*(9), 26–29.

I. R. E. N. A. (2015). Ghana renewables readiness assessment. *The International Renewable Energy Agency*, Available at https://www.irena.org//media/Files/IRENA/Agency/Publication/2015/IRENA_RRA_Ghana

I.E.A. (2015). World Energy Outlook. *International Energy Agency*, OECD/IEA, Paris.

Jude, C., Bangaké, C., & Rault, C. (2011). *Energy consumption and Economic growth; Revisited in African countries*

Kabo-bah, A. T., Diji, C. J., Nokoe, K., Mulugetta, Y., Obeng-ofori, D., & Akpoti, K. (2016). *Multiyear Rainfall and Temperature Trends in the Volta River Basin and their Potential Impact on Hydropower Generation in Ghana*. https://doi.org/10.3390/cli4040049

Katani, E. K. (2019). Forecasting the total energy consumption in Ghana using grey models. *Grey Systems: Theory and Application*, *9*(4), 488–501. https://doi.org/10.1108/gs-05-2019-0012

Kaur, H., & Ahuja, S. (2017). Time series analysis and prediction of electricity consumption of health care institution using ARIMA model. *Advances in Intelligent Systems and Computing*, *547*, 347–358. https://doi.org/10.1007/978-981-10-3325-4_35

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft. Comput., 11,* 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015

Koutroumanidis, T., Ioannou, K., & Arabatzis, G. (2009). Predicting fuelwood prices in Greece with the use of ARIMA models, artificial neural networks and a hybrid ARIMA–ANN model. *Energy Policy, 37,* 3627–3634. https://doi.org/10.1016/j.enpol.2009.04.024

Michieka, N. M., Razek, N. A., & Gearhart, R. S. (2021). The Relationship Between Climate Factors, Hydropower Production and Economic Growth in Ethiopia: A Nonlinear and Asymmetric Approach. *The Journal of Developing Areas*, *55*(3), 291–315. https://doi.org/10.1353/jda.2021.0066

Miescher, S. F. (2021). Ghana's Akosombo Dam, Volta Lake Fisheries & Climate Change. *Daedalus*, *150*(4), 124–142. https://doi.org/10.1162/daed_a_01876

Mite-León, M., & Barzola-Monteses, J. (2018). Statistical model for the forecast of hydropower production in Ecuador. *International Journal of Renewable Energy Research*, *8*(2), 1130–1137

Ömer Faruk, D. (2010). A hybrid neural network and ARIMA model for water quality time series prediction. *Eng. Appl. Artif. Intell., 23,* 586–594.

Owusu, P., & Asumadu-Sarkodie, S. (2016). A review of renewable energy sources, sustainability issues and climate change mitigation. Cogent. Eng. 3:1167990. doi:10.1080/23311916.2016.1167990

Owusu, P. A., Asumadu-Sarkodie, S., & Ameyo, P. (2016). A review of Ghana's water resource management and the

future prospect. Cogent. Eng. https://doi.org/10.1080/23311916.2016.1164275

Owusu, M., Parn, E., Kutin-Mensah, E., & Edwards, D. (2018). Power Infrastructure Sector Reforms, Power Generation, and Private Investments: Case Study from Ghana's Power Sector. *Journal of Food System Research*. https://doi.org/10.5874/jfsr.15.2_44

Panwar, N. L., Kaushik, S. C., & Kothari, S. (2011). Role of renewable energy sources in environmental protection: A review. *Renewable and sustainable energy reviews*, *15*(3), 1513-1524. https://doi.org/10.1016/j.rser.2010.11.037

Polprasert, J., Nguyên, V. A. H., & Charoensook, S. N. (2021, March). Forecasting Models for Hydropower Production Using ARIMA Method. In *2021 9th International Electrical Engineering Congress (iEECON)* (pp. 197-200). IEEE. https://doi.org/10.1109/iEECON51072.2021.9440293

Sarkodie, S. A. (2017). Estimating Ghana's electricity consumption by 2030: An ARIMA forecast. *Energy Sources, Part B: Economics, Planning and Policy*, *12*(10), 936–944. https://doi.org/10.1080/15567249.2017.1327993

Sarpong, S. A. (2013). Modeling and forecasting maternal mortality; an application of ARIMA models. *International Journal of Applied*, *3*(1), 19-28

Wu, H., Hao, Y., & Weng, J. H. (2019). How does energy consumption affect China's urbanization? New evidence from dynamic threshold panel models. *Energy Policy*, *127*(August 2018), 24–38. https://doi.org/10.1016/j.enpol.2018.11.057

Zhou, F., Li, L., Zhang, K., Trajcevski, G., Yao, F., Huang, Y., Zhong, T., … Liu, Q. (2020). Forecasting the Evolution of Hydropower Generation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2861–2870. https://doi.org/10.1145/3394486.3403337

Zolfaghari, M., & Golabi, M. R. (2021). Modeling and predicting the electricity production in hydropower using conjunction of wavelet transform, long short-term memory and random forest models. *Renewable Energy*, *170*, 1367–1381. https://doi.org/10.1016/j.renene.2021.02.017

International Journal of Statistics and Probability; Vol. 11, No. 5; Setepmber 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Parsimonious Bivariate T-distribution Type Symmetry Models for Square Contingency Tables

Kiyotaka Iki[1] & Sadao Tomizawa[2]

[1] Faculty of Economics, Nihon University, Chiyoda Ward, Tokyo, Japan

[2] Department of Information Science, Meisei University, Hino City, Tokyo, Japan

Correspondence: Kiyotaka Iki, Faculty of Economics, Nihon University, Chiyoda Ward, Tokyo, Japan.
E-mail: iki.kiyotaka@nihon-u.ac.jp

Received: July 23, 2022 Accepted: August 24, 2022 Online Published: September 12, 2022

doi:10.5539/ijsp.v11n5p44 URL: https://doi.org/10.5539/ijsp.v11n5p44

Abstract

For square contingency tables with ordered categories, Iki, Ishihara and Tomizawa (2013) considered the t-distribution type symmetry model and Iki, Okada and Tomizawa (2018) extended this model. These models are appropriate for a square contingency table if it is reasonable to assume an underlying bivariate t-distribution having any degrees of freedom. This study proposes three kinds of parsimonious models for these models. Additionally, this paper provides the decompositions of the parsimonious symmetry model using the proposed model. Some simulation studies based on bivariate t-distribution show the performances of the proposed models.

Keywords: bivariate t-distribution, square contingency table, symmetry, underlying distribution

1. Introduction

For analysis of contingency tables, we are interested in whether the two classificatory variables are independent of each another. When the independence does not hold, we may use Pearson's correlation coefficient to estimate the correlation between the two variables. Additionally, it is important to interpret the data, and propose models that fit the data well. Goodman (1979) considered the uniform association model, and Agresti (1983a) considered the linear-by-linear association model.

In particular, we consider tables with the same row and column classifications, which are known as square contingency tables. For square contingency tables, the independence between the row and column is unlikely to hold because many observations fall in the main diagonal cells, which indicates that the value of the row category is the same as the value of the column category. Therefore, for the analysis of square contingency tables, instead of independence, we are interested in whether or not the row variable is symmetric with the column variable. The symmetry (S) model (Bowker, 1948), the marginal homogeneity model (Stuart, 1955) and the quasi-symmetry model (Caussinus, 1965) have been proposed as models of symmetry. Moreover, for the research of the symmetry model, see Yoshimoto et al. (2019), Ando et al. (2021) and Shinoda et al. (2021).

We consider an $r \times r$ square contingency table with the same row and column ordinal classifications. Let p_{ij} denote the probability that an observation will fall in the ith row and jth column of the table ($i = 1, \ldots, r$; $j = 1, \ldots, r$). The S model is defined by

$$p_{ij} = p_{ji} \quad (i < j);$$

see Bishop et al. (1975, p.282). This model indicates a structure of symmetry of the probabilities with respect to the main diagonal of the table. Agresti (1983b) considered the linear diagonals-parameter symmetry (LDPS) model defined by

$$p_{ij} = \theta^{j-i} p_{ji} \quad (i < j).$$

This indicates that the probability of an observation falling in the (i, j)th cell, $i < j$, is θ^{j-i} times higher than the probability of it falling in the (j, i)th cell. A special case of the LDPS model obtained by putting $\theta = 1$ is the S model. Tomizawa (1991) proposed an extended linear diagonals-parameter symmetry (ELDPS) model defined by

$$p_{ij} = \theta_1^{j-i} \theta_2^{j^2-i^2} p_{ji} \quad (i < j).$$

This indicates that the probability of an observation falling in the (i, j)th cell, $i < j$, is $\theta_1^{j-i} \theta_2^{j^2-i^2}$ times higher than the probability of it falling in the (j, i)th cell. Agresti (1983; 1984, p.216) described the relationship between the LDPS model

and the joint bivariate normal distribution as follows: the LDPS model may be appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate normal distribution with equal marginal variances. Moreover, Tomizawa (1991) pointed out that the ELDPS model may be appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate normal distribution with different marginal variances.

For any fixed constant m ($m > 2$), Iki et al. (2013) proposed the t-distribution type symmetry (TS(m)) model defined by

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \eta_m(j-i) \quad (i<j).$$

A special case of this model can be obtained by putting $\eta_m = 0$ in the S model. The TS(m) model indicates that the difference between the two symmetric probabilities raised to the power $[= -2/(m+2)]$ is proportional to the distance from the main diagonal of the $r \times r$ table. The TS(m) model may be appropriate if it is reasonable to assume an underlying bivariate t-distribution with equal marginal variances having m degrees of freedom (see Iki et al., 2013). For any fixed constant m ($m > 2$), Iki et al. (2018) proposed the extended t-distribution type symmetry (ETS(m)) model defined by

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \gamma_m(j^2-i^2) + \eta_m(j-i) \quad (i<j).$$

A special case of this model can be obtained by putting $\gamma_m = 0$ in the TS(m) model. The ETS(m) model may be appropriate if it is reasonable to assume an underlying bivariate t-distribution with different marginal variances having m degrees of freedom (see Iki et al., 2018).

Now, we are interested in considering more parsimonious t-distribution type symmetry models, which can be described in terms of fewer parameters than the TS(m) (ETS(m)) models.

The purpose of this paper is to propose new models which may appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate t-distribution. The new models are different from the S, TS(m) and ETS(m) models. Section 2 proposes models and describes the properties of the new models. Section 3 includes the decompositions using the proposed models. Section 4 shows the maximum likelihood estimates of expected frequencies under the proposed models. Section 5 describes the relationships between the proposed models and t-distribution by the simulation study. Section 6 provides some concluding remarks.

2. Models

We consider random variables U and V having a joint bivariate t-distribution with m ($m > 2$) degrees of freedom, meaning $\mathrm{E}(U) = \mu_1$, $\mathrm{E}(V) = \mu_2$, variances $\mathrm{Var}(U) = m\sigma_1^2/(m-2)$, $\mathrm{Var}(V) = m\sigma_2^2/(m-2)$, and correlation coefficient $\mathrm{Corr}(U,V) = \rho$. The probability density function $f(u,v)$ is

$$f(u,v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\left(1+\frac{Q(u,v)}{m}\right)^{-\frac{m+2}{2}},$$

where,

$$Q(u,v) = \frac{1}{1-\rho^2}\left[\left(\frac{u-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho}{\sigma_1\sigma_2}(u-\mu_1)(v-\mu_2) + \left(\frac{v-\mu_2}{\sigma_2}\right)^2\right];$$

see Muirhead (2005, p.48). The probability density function is also expressed as

$$f(u,v) = c\left[1+\frac{1}{m}(a_1u+b_1v+a_2u^2+b_2v^2+d(u,v))\right]^{-\frac{m+2}{2}}, \tag{1}$$

where

$$\begin{aligned}
c &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},\\
a_1 &= \frac{2}{\sigma_1(1-\rho^2)}\left(\frac{\rho\mu_2}{\sigma_2}-\frac{\mu_1}{\sigma_1}\right), \quad b_1 = \frac{2}{\sigma_2(1-\rho^2)}\left(\frac{\rho\mu_1}{\sigma_1}-\frac{\mu_2}{\sigma_2}\right),\\
a_2 &= \frac{1}{\sigma_1^2(1-\rho^2)}, \quad b_2 = \frac{1}{\sigma_2^2(1-\rho^2)},\\
d(u,v) &= \frac{1}{1-\rho^2}\left(-\frac{2\rho}{\sigma_1\sigma_2}uv + \frac{\mu_1^2}{\sigma_1^2} + \frac{\mu_2^2}{\sigma_2^2} - \frac{2\rho\mu_1\mu_2}{\sigma_1\sigma_2}\right),
\end{aligned}$$

and $d(u,v) = d(v,u)$. When $\mathrm{Var}(U) = \mathrm{Var}(V)$, that is, $\sigma_1^2 = \sigma_2^2 \ (= \sigma^2)$, $f(u,v)$ is expressed as

$$f(u,v) = c\left[1 + \frac{1}{m}\left(a_1 u + b_1 v + t(u^2+v^2) + d(u,v)\right)\right]^{-\frac{m+2}{2}}, \tag{2}$$

where

$$\begin{aligned}
c &= \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}},\\
a_1 &= \frac{2}{\sigma^2(1-\rho^2)}\left(\rho\mu_2 - \mu_1\right), \quad b_1 = \frac{2}{\sigma^2(1-\rho^2)}\left(\rho\mu_1 - \mu_2\right),\\
t &= \frac{1}{\sigma^2(1-\rho^2)}\\
d(u,v) &= \frac{1}{\sigma^2(1-\rho^2)}\left(-2\rho uv + \mu_1^2 + \mu_2^2 - 2\rho\mu_1\mu_2\right),
\end{aligned}$$

and $d(u,v) = d(v,u)$. Moreover, when $\mathrm{E}(U) = \mathrm{E}(V)$ and $\mathrm{Var}(U) = \mathrm{Var}(V)$, that is, $\mu_1 = \mu_2 \ (= \mu)$ and $\sigma_1^2 = \sigma_2^2 \ (= \sigma^2)$, $f(u,v)$ is expressed as

$$f(u,v) = c\left[1 + \frac{1}{m}\left(k(u+v) + t(u^2+v^2) + d(u,v)\right)\right]^{-\frac{m+2}{2}}, \tag{3}$$

where

$$\begin{aligned}
c &= \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}},\\
k &= -\frac{2\mu}{\sigma^2(1+\rho)}\\
t &= \frac{1}{\sigma^2(1-\rho^2)}\\
d(u,v) &= \frac{2}{\sigma^2(1-\rho^2)}\left(-\rho uv + \mu^2 - \rho\mu^2\right),
\end{aligned}$$

and $d(u,v) = d(v,u)$.

We consider the $r \times r$ square contingency table with ordered categories. For any fixed constant m ($m > 2$), we propose a model defined by

$$p_{ij} = \left[1 + \frac{1}{m}\left(\mu + \kappa(i+j) + \tau(i^2+j^2) + \phi ij\right)\right]^{-\frac{m+2}{2}} \quad (i = 1,\ldots,r; j = 1,\ldots,r).$$

We shall refer to this model as a parsimonious symmetry (PaS(m)) model. From the form of equation (3), the PaS(m) model may be appropriate if it is reasonable to assume an underlying bivariate t-distribution with same marginal means and variances having m degrees of freedom. Under the PaS(m) model, we see that

$$p_{ij} = p_{ji} \quad (i < j).$$

Namely, the PaS(m) model implies the S model.

Next, for any fixed constant m ($m > 2$), we propose a model defined by

$$p_{ij} = \left[1 + \frac{1}{m}\left(\mu + \alpha_1 i + \beta_1 j + \tau(i^2+j^2) + \phi ij\right)\right]^{-\frac{m+2}{2}} \quad (i = 1,\ldots,r; j = 1,\ldots,r).$$

We shall refer to this model as a parsimonious t-distribution type symmetry (PaTS(m)) model. From the form of equation (2), the PaTS(m) model may be appropriate if it is reasonable to assume an underlying bivariate t-distribution with same marginal variances (and different marginal means) having m degrees of freedom. A special case of the PaTS(m) can be obtained by putting $\alpha_1 = \beta_1$ in the PaS(m) model. Under the PaTS(m) model,

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \frac{\beta_1 - \alpha_1}{m}(j-i) \quad (i < j).$$

Namely, the PaTS(m) model implies the TS(m) model. Additionally, under the PaTS(m) model, setting $\omega_{ij} = \mu + \alpha_1 i + \beta_1 j + \tau(i^2 + j^2) + \phi ij$, we see that

$$\begin{aligned}
\lim_{m\to\infty} \frac{p_{ij}}{p_{ji}} &= \lim_{m\to\infty} \frac{(1+\frac{\omega_{ij}}{m})^{-\frac{m+2}{2}}}{(1+\frac{\omega_{ji}}{m})^{-\frac{m+2}{2}}} \\
&= \lim_{m\to\infty} \frac{\{(1+\frac{\omega_{ij}}{m})^{\frac{m}{\omega_{ij}}}\}^{-\frac{\omega_{ij}}{2}(1+\frac{2}{m})}}{\{(1+\frac{\omega_{ji}}{m})^{\frac{m}{\omega_{ji}}}\}^{-\frac{\omega_{ji}}{2}(1+\frac{2}{m})}} \\
&= \frac{\exp[-\frac{\omega_{ij}}{2}]}{\exp[-\frac{\omega_{ji}}{2}]} \\
&= \exp\left[\frac{1}{2}(\alpha_1 - \beta_1)(j-i)\right] \\
&= \theta^{j-i} \quad (i<j),
\end{aligned}$$

where

$$\theta = \exp\left[\frac{\alpha_1 - \beta_1}{2}\right].$$

Namely, the PaTS(m) model approaches the LDPS model as m becomes larger.

Moreover, for any fixed constant m ($m > 2$), we propose a model defined by

$$p_{ij} = \left[1 + \frac{1}{m}\left(\mu + \alpha_1 i + \beta_1 j + \alpha_2 i^2 + \beta_2 j^2 + \phi ij\right)\right]^{-\frac{m+2}{2}} \qquad (i = 1, \ldots, r;\ j = 1, \ldots, r).$$

We shall refer to this model as a parsimonious t-distribution type symmetry (PaETS(m)) model. From the form of equation (1), the PaTS(m) model may be appropriate if it is reasonable to assume an underlying bivariate t-distribution with different marginal means and variances having m degrees of freedom. A special case of the PaETS(m) can be obtained by putting $\alpha_2 = \beta_2$ in the PaTS(m) model. Under the PaETS(m) model,

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \frac{\beta_1 - \alpha_1}{m}(j-i) + \frac{\beta_2 - \alpha_2}{m}(j^2 - i^2) \quad (i<j).$$

Namely, the PaETS(m) model implies the ETS(m) model. Further, under the PaETS(m) model, we see that

$$\begin{aligned}
\lim_{m\to\infty} \frac{p_{ij}}{p_{ji}} &= \exp\left[\frac{1}{2}(\alpha_1 - \beta_1)(j-i) + \frac{1}{2}(\alpha_2 - \beta_2)(j^2 - i^2)\right] \\
&= \theta_1^{j-i}\theta_2^{j^2-i^2} \quad (i<j),
\end{aligned}$$

where

$$\theta_1 = \exp\left[\frac{\alpha_1 - \beta_1}{2}\right],\ \theta_2 = \exp\left[\frac{\alpha_2 - \beta_2}{2}\right].$$

Namely, the PaETS(m) model approaches the ELDPS model as m becomes larger.

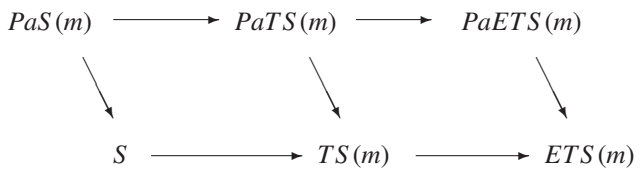


Figure 1. Relationships among models

In Figure 1, we show the relationships among models. In Figure, $A \to B$ indicates that model A implies model B.

3. Decompositions of Models

Consider the $r \times r$ square contingency table. Let X and Y denote the row and column variables, respectively. We refer to the model of equality of marginal means, that is, $\mathrm{E}(X) = \mathrm{E}(Y)$, as the ME model. Additionally, we refer to model of equality of marginal means and variances, that is, $\mathrm{E}(X) = \mathrm{E}(Y)$ and $\mathrm{Var}(X) = \mathrm{Var}(Y)$, as the MVE model. Then, we obtain the following theorems.

Theorem 1 *The PaS(m) model holds, if and only if both the PaETS(m) and MVE models hold.*

Proof. If the PaS(m) model holds, then the PaETS(m) and MVE models hold. Assuming that the PaETS(m) and MVE models hold, then we shall show that the PaS(m) model holds. From the PaETS(m) model, we see

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \frac{1}{m}\left[(\alpha_1 - \beta_1)(i-j) + (\alpha_2 - \beta_2)\left(i^2 - j^2\right)\right] \quad (i < j).$$

Then, because the MVE model is given by to $\mathrm{E}(X) = \mathrm{E}(Y)$ and $\mathrm{E}(X^2) = \mathrm{E}(Y^2)$,

$$\begin{aligned}
&\sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij}\left(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}\right) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{p_{ij}}{m}\left[(\alpha_1 - \beta_1)(i-j) + (\alpha_2 - \beta_2)\left(i^2 - j^2\right)\right] \\
&= \frac{\alpha_1 - \beta_1}{m}\sum_{i=1}^{r}\sum_{j=1}^{r}(i-j)\,p_{ij} + \frac{\alpha_2 - \beta_2}{m}\sum_{i=1}^{r}\sum_{j=1}^{r}\left(i^2 - j^2\right)p_{ij} \\
&= \frac{\alpha_1 - \beta_1}{m}\left(\mathrm{E}(X) - \mathrm{E}(Y)\right) + \frac{\alpha_2 - \beta_2}{m}\left(\mathrm{E}(X^2) - \mathrm{E}(Y^2)\right) \\
&= 0.
\end{aligned}$$

Additionally, we have

$$\begin{aligned}
&\sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij}\left(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}\right) \\
&= \sum_{i<j}\sum p_{ij}\left(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}\right) + \sum_{i>j}\sum p_{ij}\left(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}\right) \\
&= \sum_{i<j}\sum \left(p_{ij} - p_{ji}\right)\left(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}\right).
\end{aligned}$$

For any $i < j$, if $p_{ij} \neq p_{ji}$, then $(p_{ij} - p_{ji})(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}) < 0$, if $p_{ij} = p_{ji}$, then $(p_{ij} - p_{ji})(p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}}) = 0$. Thus, when we assume that the PaETS(m) and MVE models hold, we can obtain $p_{ij} = p_{ji}$ for all $i < j$. Moreover, $p_{ij} - p_{ji} = 0$ for all $i < j$, that is,

$$(\alpha_1 - \beta_1)(i-j) + (\alpha_2 - \beta_2)(i^2 - j^2) = 0 \quad \text{for all } i < j.$$

Therefore we obtain $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$. Namely, the PaS(m) model holds. The proof is completed.

Theorem 2 *The PaS(m) model holds, if and only if both the PaTS(m) and ME models hold.*

The proof of Theorem 2 is omitted because that is obtained in a way similar to Theorem 1.

4. Goodness-of-fit Test

For an $r \times r$ contingency table, let n_{ij} denote the observed frequency in the ith row and jth column of the table, where $n = \sum\sum n_{ij}$ and let m_{ij} denote the corresponding expected frequency ($i = 1, \ldots, r$; $j = 1, \ldots, r$). Assume that the observed frequencies have a multinomial distribution. Let $G^2(M)$ denote the likelihood ratio chi-squared statistic, defined by

$$G^2(M) = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

where $\hat{m}_{ij}$ is the maximum likelihood estimate of expected frequency m_{ij} under model M. Under model M, these statistics have a asymptotically central chi-squared distribution with the corresponding degrees of freedom. For the PaS(m) model, $\{p_{ij}\}$ are determined by μ, κ, τ and ϕ. Therefore, the numbers of degrees of freedom for the PaS(m) model are $r^2 - 4$.

Similarly, the numbers of degrees of freedom for the PaTS(m) and PaETS(m) models are r^2-5 and r^2-6, respectively. We consider the maximum likelihood estimates of expected frequencies $\{m_{ij}\}$ under the PaS(m), PaTS(m) and PaETS(m) models in the log-likelihood equation. For the PaS(m) model, we must maximize the Lagrangian

$$L=\sum_{i=1}^{r}\sum_{j=1}^{r}n_{ij}\log p_{ij}-\lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r}p_{ij}-1\right)-\sum_{i<j}\sum\psi_{ij}(p_{ij}-p_{ji})$$
$$-\sum_{(i,j)\in D}\sum\lambda_{ij}\left\{p_{ij}^{-\frac{2}{m+2}}-\left(\mu+\kappa(i+j)+\tau(i^2+j^2)+\phi ij\right)\right\},$$

where

$$\mu=\frac{1}{2}\left(11p_{11}^{-\frac{2}{m+2}}-13p_{12}^{-\frac{2}{m+2}}+3p_{13}^{-\frac{2}{m+2}}+p_{23}^{-\frac{2}{m+2}}\right),$$
$$\kappa=\frac{1}{2}\left(-6p_{11}^{-\frac{2}{m+2}}+9p_{12}^{-\frac{2}{m+2}}-2p_{13}^{-\frac{2}{m+2}}-p_{23}^{-\frac{2}{m+2}}\right),$$
$$\tau=\frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}}-2p_{12}^{-\frac{2}{m+2}}+p_{13}^{-\frac{2}{m+2}}\right),$$
$$\phi=\frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}}-p_{12}^{-\frac{2}{m+2}}-p_{13}^{-\frac{2}{m+2}}+p_{23}^{-\frac{2}{m+2}}\right),$$
$$D=\{(i,j)|i<j,(i,j)\neq(1,1),(1,2),(1,3),(2,3)\},$$

with respect to $\{p_{ij}\}$, λ, $\{\psi_{ij}\}$ and $\{\lambda_{ij}\}$. For the PaTS(m) model, we must maximize the Lagrangian

$$L=\sum_{i=1}^{r}\sum_{j=1}^{r}n_{ij}\log p_{ij}-\lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r}p_{ij}-1\right)$$
$$-\sum_{(i,j)\in E_1}\sum\psi_{ij}\left(p_{ij}^{-\frac{2}{m+2}}-p_{ji}^{-\frac{2}{m+2}}-(j-i)p_{12}^{-\frac{2}{m+2}}+(j-i)p_{21}^{-\frac{2}{m+2}}\right)$$
$$-\sum_{(i,j)\in E_2}\sum\lambda_{ij}\left\{p_{ij}^{-\frac{2}{m+2}}-\left(\mu+\alpha i+\beta j+\tau(i^2+j^2)+\phi ij\right)\right\},$$

where

$$\mu=\frac{1}{2}\left(11p_{11}^{-\frac{2}{m+2}}-10p_{12}^{-\frac{2}{m+2}}+3p_{13}^{-\frac{2}{m+2}}-3p_{21}^{-\frac{2}{m+2}}+p_{23}^{-\frac{2}{m+2}}\right),$$
$$\alpha=\frac{1}{2}\left(-6p_{11}^{-\frac{2}{m+2}}+6p_{12}^{-\frac{2}{m+2}}-2p_{13}^{-\frac{2}{m+2}}+3p_{21}^{-\frac{2}{m+2}}-p_{23}^{-\frac{2}{m+2}}\right),$$
$$\beta=\frac{1}{2}\left(-6p_{11}^{-\frac{2}{m+2}}+8p_{12}^{-\frac{2}{m+2}}-2p_{13}^{-\frac{2}{m+2}}+p_{21}^{-\frac{2}{m+2}}-p_{23}^{-\frac{2}{m+2}}\right),$$
$$\tau=\frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}}-2p_{12}^{-\frac{2}{m+2}}+p_{13}^{-\frac{2}{m+2}}\right),$$
$$\phi=\frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}}-p_{13}^{-\frac{2}{m+2}}-p_{21}^{-\frac{2}{m+2}}+p_{23}^{-\frac{2}{m+2}}\right),$$
$$E_1=\{(i,j)|i<j,(i,j)\neq(1,2)\},$$
$$E_2=\{(i,j)|i<j,(i,j)\neq(1,1),(1,2),(1,3),(2,3)\},$$

with respect to $\{p_{ij}\}$, λ, $\{\psi_{ij}\}$ and $\{\lambda_{ij}\}$. For the PaETS(m) model, we must maximize the Lagrangian

$$L=\sum_{i=1}^{r}\sum_{j=1}^{r}n_{ij}\log p_{ij}-\lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r}p_{ij}-1\right)$$
$$-\sum_{(i,j)\in F_1}\sum\psi_{ij}\left[p_{ij}^{-\frac{2}{m+2}}-p_{ji}^{-\frac{2}{m+2}}+\frac{(j-i)}{2}\left\{(2i+2j-8)(p_{12}^{-\frac{2}{m+2}}-p_{21}^{-\frac{2}{m+2}})\right.\right.$$
$$\left.\left.-(i+j-3)(p_{13}^{-\frac{2}{m+2}}-p_{31}^{-\frac{2}{m+2}})\right\}\right]$$
$$-\sum_{(i,j)\in F_2}\sum\lambda_{ij}\left\{p_{ij}^{-\frac{2}{m+2}}-\left(\mu+\alpha_1 i+\beta_1 j+\alpha_2 i^2+\beta_2 j^2+\phi ij\right)\right\},$$

where

$$\mu = \frac{1}{2}\left(11p_{11}^{-\frac{2}{m+2}} - 6p_{12}^{-\frac{2}{m+2}} + p_{13}^{-\frac{2}{m+2}} - 7p_{21}^{-\frac{2}{m+2}} + p_{23}^{-\frac{2}{m+2}} + 2p_{31}^{-\frac{2}{m+2}}\right),$$
$$\alpha_1 = \frac{1}{2}\left(-6p_{11}^{-\frac{2}{m+2}} + p_{13}^{-\frac{2}{m+2}} + 9p_{21}^{-\frac{2}{m+2}} - p_{23}^{-\frac{2}{m+2}} - 3p_{31}^{-\frac{2}{m+2}}\right),$$
$$\beta_1 = \frac{1}{2}\left(-6p_{11}^{-\frac{2}{m+2}} + 8p_{12}^{-\frac{2}{m+2}} - 2p_{13}^{-\frac{2}{m+2}} + p_{21}^{-\frac{2}{m+2}} - p_{23}^{-\frac{2}{m+2}}\right),$$
$$\alpha_2 = \frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}} - 2p_{21}^{-\frac{2}{m+2}} + p_{31}^{-\frac{2}{m+2}}\right),$$
$$\beta_2 = \frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}} - 2p_{12}^{-\frac{2}{m+2}} + p_{13}^{-\frac{2}{m+2}}\right),$$
$$\phi = \frac{1}{2}\left(p_{11}^{-\frac{2}{m+2}} - p_{13}^{-\frac{2}{m+2}} - p_{21}^{-\frac{2}{m+2}} + p_{23}^{-\frac{2}{m+2}}\right),$$
$$F_1 = \{(i,j)|i < j, (i,j) \neq (1,2), (1,3)\},$$
$$F_2 = \{(i,j)|i < j, (i,j) \neq (1,1), (1,2), (1,3), (2,3)\},$$

with respect to $\{p_{ij}\}$, λ, $\{\psi_{ij}\}$ and $\{\lambda_{ij}\}$. Setting the partial derivations of L equal to zero using the Newton-Raphson method, we can obtain the maximum likelihood estimates of $\{m_{ij}\}$ under the PaS(m), PaTS(m) and PaETS(m) models.

5. Simulation Study

As described in Section 2, the PaS(m), PaTS(m) and PaETS(m) models may be appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate t-distribution having m degrees of freedom. We shall consider the relationships between the proposed models and bivariate t-distribution in terms of simulation studies, and the comparison between the proposed models and S, TS(m) and ETS(m) models.

Consider random variables U and V having a bivariate t-distribution with m degrees of freedom, meaning $\mathrm{E}(U) = 0$, $\mathrm{E}(V) = \mu_2$, variances $\mathrm{Var}(U) = m/(m-2)$, $\mathrm{Var}(V) = m\sigma_2^2/(m-2)$, and correlation coefficient $\mathrm{Corr}(U,V) = \rho$. Suppose that there are some conditions; $m = 30, 100$, $\mu_2 = 0, 0.2$, $\sigma_2^2 = 1, 1.2$, $\rho = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$, a 4×4 table of sample size 5000 is formed using cut points for each variable at $-0.7, 0, 0.7$.

We count the frequencies of acceptance (at the 0.05 significance level) based on the likelihood ratio chi-squared statistic for testing the hypothesis that the models with the corresponding m degrees of freedom hold per 10000 times for 4×4 tables on each condition.

From Tables 1 and 2, we see that the ETS(m) model is a good fit for all conditions. Further the TS(m) model is a good fit when $\sigma_2^2 = 1$, and the S model gives good fit on when $\mu_2 = 0$ and $\sigma_2^2 = 1$. In contrast, the PaS(m), PaTS(m) and PaETS(m) models show a similar trend when ρ is close to 0. Thus, from the result of this simulation, we obtain that if it is reasonable to assume an underlying bivariate t-distribution with a low correlation coefficient, the parsimonious models would fit the data well.

6. Concluding Remarks

Each of the S, TS(m) and ETS(m) models is saturated on the main diagonal cells of the table, but the PaS(m), PaTS(m) and PaETS(m) models are unsaturated on them. Thus, under the PaS(m), PaTS(m) and PaETS(m) models, the estimated expected frequencies on the main diagonal are always not equal to the observed frequencies on the main diagonal. The PaS(m), PaTS(m) and PaETS(m) models may be useful when we want to utilize the information on the main diagonal.

From Section 5, when observations are not so concentrated in the main diagonal cells, that is, a correlation coefficient between row and column variables is close to 0, the proposed models (PaS(m), PaTS(m) and PaETS(m)) may be better for application to a square table than the S, TS(m) and ETS(m) models.

Acknowledgements

The authors would like to thank the referee and the editorial team for many comments and suggestions.

Table 1. The frequencies of acceptance (at the 0.05 significance level) per 10000 times for 4×4 tables based on the likelihood ratio chi-squared statistic for testing the hypothesis that the S, TS(30), ETS(30), PaS(30), PaTS(30) or PaETS(30) model hold

μ_2	σ_2^2	ρ	S	TS(30)	ETS(30)	PaS(30)	PaTS(30)	PaETS(30)
0	1	0.1	9501	9520	9494	9156	9150	9116
0.2	1	0.1	0	9273	9241	0	8888	8837
0	1.2	0.1	1535	1308	9490	2403	2234	9082
0.2	1.2	0.1	0	1107	9289	0	1806	8927
0	1	0.2	9474	9460	9483	8589	8549	8489
0.2	1	0.2	0	9265	9246	0	8336	8304
0	1.2	0.2	1575	1317	9508	1936	1779	8462
0.2	1.2	0.2	0	1055	9294	0	1418	8258
0	1	0.3	9490	9493	9472	7417	7316	7171
0.2	1	0.3	0	9168	9137	0	7064	6935
0	1.2	0.3	1437	1212	9489	1237	1118	7124
0.2	1.2	0.3	0	983	9117	0	922	6765
0	1	0.4	9459	9505	9497	5159	4981	4751
0.2	1	0.4	0	9133	9058	0	4720	4565
0	1.2	0.4	1338	1111	9478	567	486	4599
0.2	1.2	0.4	0	901	9086	0	400	4390
0	1	0.5	9504	9491	9504	2349	2211	2025
0.2	1	0.5	0	9068	9003	0	2081	1866
0	1.2	0.5	1214	1017	9480	145	126	1888
0.2	1.2	0.5	0	765	8990	0	92	1755
0	1	0.6	9499	9487	9518	533	449	384
0.2	1	0.6	0	9028	8993	0	424	350
0	1.2	0.6	897	737	9455	7	7	315
0.2	1.2	0.6	0	636	8921	0	8	263

Table 2. The frequencies of acceptance (at the 0.05 significance level) per 10000 times for 4×4 tables based on the likelihood ratio chi-squared statistic for testing the hypothesis that the S, TS(100), ETS(100), PaS(100), PaTS(100) or PaETS(100) model hold

μ_2	σ_2^2	ρ	S	TS(100)	ETS(100)	PaS(100)	PaTS(100)	PaETS(100)
0	1	0.1	9486	9485	9495	9350	9350	9352
0.2	1	0.1	0	9271	9259	0	9117	9089
0	1.2	0.1	1525	1283	9502	2583	2412	9371
0.2	1.2	0.1	0	1062	9225	0	1840	9108
0	1	0.2	9480	9490	9486	8970	8951	8880
0.2	1	0.2	0	9242	9175	0	8647	8610
0	1.2	0.2	1469	1222	9489	2060	1897	8854
0.2	1.2	0.2	0	981	9247	0	1516	8585
0	1	0.3	9497	9483	9496	7942	7842	7739
0.2	1	0.3	0	9174	9154	0	7560	7445
0	1.2	0.3	1407	1168	9480	1397	1272	7700
0.2	1.2	0.3	0	921	9102	0	973	7252
0	1	0.4	9501	9521	9492	6038	5857	5672
0.2	1	0.4	0	9087	9049	0	5324	5151
0	1.2	0.4	1358	1145	9485	694	597	5407
0.2	1.2	0.4	0	785	8994	0	419	4833
0	1	0.5	9468	9470	9481	2880	2728	2508
0.2	1	0.5	0	8942	8847	0	2531	2337
0	1.2	0.5	1122	917	9465	165	141	2269
0.2	1.2	0.5	0	651	8881	0	110	1959
0	1	0.6	9529	9517	9512	669	609	541
0.2	1	0.6	0	8907	8844	0	525	467
0	1.2	0.6	838	651	9451	14	11	396
0.2	1.2	0.6	0	534	8764	0	6	326

References

Agresti, A. (1983a). A survey of strategies for modeling cross-classifications having ordinal variables. *Journal of the American Statistical Association, 78*, 184-198. https://doi.org/10.1080/01621459.1983.10477950

Agresti, A. (1983b). A simple diagonals-parameter symmetry and quasi-symmetry model. *Statistics and Probability Letters, 1*, 313-316. https://doi.org/10.1016/0167-7152(83)90051-2

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons.

Ando, S., Tahata, K., & Tomizawa, S. (2021). Extended diagonal uniform association symmetry model for square contingency tables with ordinal categories. *Statistics and Probability Letters, 169*, 1-4. https://doi.org/10.1016/j.spl.2020.108973

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: The MIT Press.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43*, 572-574. https://doi.org/10.1080/01621459.1948.10483284

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse, 29*, 77-182. https://doi.org/10.5802/afst.519

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74*, 537-552. https://doi.org/10.1080/01621459.1979.10481650

Iki, K., Ishihara, T., & Tomizawa, S. (2013). Bivariate t-distribution type symmetry model for square contingency tables with ordered categories. *Model Assisted Statistics and Applications: An International Journal, 8*, 315-319. https://doi.org/10.3233/MAS-130269

Iki, K., Okada, M., & Tomizawa, S. (2018). An extended bivariate t-distribution type symmetry model for square contingency tables. *Open Journal of Statistics, 8*, 249-257. https://doi.org/10.4236/ojs.2018.82015

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. New Jersey: John Wiley & Sons.

Shinoda, S., Tahata, K., Yamamoto, K., & Tomizawa, S. (2021). Marginal continuation odds ratio model and decomposition of marginal homogeneity model for multi-way contingency tables. *Sankhya: The Indian Journal of Statistics, Ser. B, 83*, 304-324. https://doi.org/10.1007/s13571-020-00228-9

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*, 412-416. http://dx.doi.org/10.1093/biomet/42.3-4.412

Tomizawa, S. (1991). An extended linear diagonals-parameter symmetry model for square contingency tables with ordered categories. *Metron: International Journal of Statistics, 49*, 401-409.

Yoshimoto, T., Tahata, K., Iki, K., & Tomizawa, S. (2019). Moment symmetry models and decompositions of symmetry for multi-way contingency tables. *Calcutta Statistical Association Bulletin, 71*, 83-98. https://doi.org/10.1177/0008068319880443

Copyrights

International Journal of Statistics and Probability; Vol. 11, No. 5; September 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reviewer Acknowledgements

International Journal of Statistics and Probability wishes to acknowledge the following individuals for their assistance with peer review of manuscripts for this issue. Their help and contributions in maintaining the quality of the journal is greatly appreciated.

Many authors, regardless of whether *International Journal of Statistics and Probability* publishes their work, appreciate the helpful feedback provided by the reviewers.

Reviewers for Volume 11, Number 5

Adekola Lanrewaju Olumide, Bells University of Technology, Nigeria

Adeyeye Awogbemi, National Mathematical Centre, Nigeria

Afsin Sahin, Gazi University, Turkey

Chin-Shang Li, School of Nursing, USA

Emmanuel Akpan, Federal School of Medical Laboratory Technology, Nigeria

Faisal Khamis, Al Ain University of Science and Technology, Canada

Mohammed Elseidi, Mansoura University, Egypt

Philip Westgate, University of Kentucky, USA

Poulami Maitra, NORC at the University of Chicago, India

Pourab Roy, US Food and Drug Administration, USA

Vilda Purutcuoglu, Middle East Technical University (METU), Turkey

Vyacheslav Abramov, Swinburne University of Technology, Australia

Wendy Smith

On behalf of,

The Editorial Board of *International Journal of Statistics and Probability*

Canadian Center of Science and Education

ACADEMIC AND OPEN ACCESS

Canadian Center of Science and Education

CCSENET.ORG

➢ CALL FOR MANUSCRIPTS

International Journal of Statistics and Probability is a peer-reviewed journal, published by Canadian Center of Science and Education. The journal publishes research papers in all aspects of statistics and probability. The journal is available in electronic form in conjunction with its print edition. All articles and issues are available for free download online.

We are seeking submissions for forthcoming issues. All manuscripts should be written in English. Manuscripts from 3000–8000 words in length are preferred. All manuscripts should be prepared in LaTeX or MS-Word format, and submitted online, or sent to: ijsp@ccsenet.org

Paper Selection and Publishing Process

a) Submission acknowledgement. If you submit manuscript online, you will receive a submission acknowledgement letter sent by the online system automatically. For email submission, the editor or editorial assistant sends an e-mail of confirmation to the submission's author within one to three working days. If you fail to receive this confirmation, please check your bulk email box or contact the editorial assistant.

b) Basic review. The editor or editorial assistant determines whether the manuscript fits the journal's focus and scope. And then check the similarity rate (CrossCheck, powered by iThenticate). Any manuscripts out of the journal's scope or containing plagiarism, including self-plagiarism are rejected.

c) Peer Review. We use a double-blind system for peer review; both reviewers' and authors' identities remain anonymous. The submitted manuscript will be reviewed by at least two experts: one editorial staff member as well as one to three external reviewers. The review process may take four to ten weeks.

d) Make the decision. The decision to accept or reject an article is based on the suggestions of reviewers. If differences of opinion occur between reviewers, the editor-in-chief will weigh all comments and arrive at a balanced decision based on all comments, or a second round of peer review may be initiated.

e) Notification of the result of review. The result of review will be sent to the corresponding author and forwarded to other authors and reviewers.

f) Pay the article processing charge. If the submission is accepted, the authors revise paper and pay the article processing charge (formatting and hosting).

g) E-journal is available. E-journal in PDF is available on the journal's webpage, free of charge for download. If you need the printed journals by post, please order at http://www.ccsenet.org/journal/index.php/ijsp/store/hardCopies.

h) Publication notice. The authors and readers will be notified and invited to visit our website for the newly published articles.

More Information

E-mail: ijsp@ccsenet.org

Website: http://ijsp.ccsenet.org

Paper Submission Guide: http://ijsp-author.ccsenet.org

Recruitment for Reviewers: http://www.ccsenet.org/journal/index.php/ijsp/editor/recruitment

➢ JOURNAL STORE

To order back issues, please contact the journal editor and ask about the availability of journals. You may pay by credit card, PayPal, and bank transfer. If you have any questions regarding payment, please do not hesitate to contact the journal editor or editorial assistant.

Price: $40.00 USD/copy Shipping fee: $20.00 USD/copy

ABOUT CCSE

The Canadian Center of Science and Education (CCSE) is a private for-profit organization delivering support and services to educators and researchers in Canada and around the world.

The Canadian Center of Science and Education was established in 2006. In partnership with research institutions, community organizations, enterprises, and foundations, CCSE provides a variety of programs to support and promote education and research development, including educational programs for students, financial support for researchers, international education projects, and scientific publications.

CCSE publishes scholarly journals in a wide range of academic fields, including the social sciences, the humanities, the natural sciences, the biological and medical sciences, education, economics, and management. These journals deliver original, peer-reviewed research from international scholars to a worldwide audience. All our journals are available in electronic form in conjunction with their print editions. All journals are available for free download online.

Mission

To work for future generations

Values

Scientific integrity and excellence

Respect and equity in the workplace

CONTACT US

1595 Sixteenth Ave, Suite 301,
Richmond Hill, Ontario, L4B 3N9,
Canada
Tel: 1-416-642-2606
E-mail: info@ccsenet.org
Website: www.ccsenet.org

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

Aerospace Database
BASE (Bielefeld Academic Search Engine)
EZB (Elektronische Zeitschriftenbibliothek)
Google Scholar
JournalTOCs
Library and Archives Canada
LOCKSS
MIAR
PKP Open Archives Harvester
SHERPA/RoMEO
Standard Periodical Directory
Ulrich's

International Journal of Statistics and Probability

Bimonthly

Publisher	Canadian Center of Science and Education
Address	1595 Sixteenth Ave, Suite 301, Richmond Hill, Ontario, L4B 3N9, Canada
Telephone	1-416-642-2606
E-mail	ijsp@ccsenet.org
Website	http://ijsp.ccsenet.org

ISSN 1927-7032