Copula Modeling to Identify the Dependency Structure of Agricultural Production and Its Environment Indicators in Indonesia

Atina Ahdika[#*1], Dedi Rosadi[#2], Adhitya Ronnie E.[#3], Gunardi[#4]
[#]*Department of Mathematics, Universitas Gadjah Mada*
Yogyakarta, Indonesia
[*]*Department of Statistics, Universitas Islam Indonesia*
Yogyakarta, Indonesia
[1]atinaahdika@gmail.com, atina.a@uii.ac.id
[2]dedirosadi@gadjahmada.edu
[3]adhityaronnie@ugm.ac.id
[4]gunardiugm@yahoo.com

***Abstract*— Agriculture is a very potential field developed in agrarian countries such as Indonesia. The country has abundant natural wealth as a food source for plants. In addition, the natural condition also has an important role on the quality and quantity of agricultural products. This study aims to model dependency structure of rice production and its environment indicators, in this case, includes temperature change, CO_2 emission, and rainfall precipitation, in Indonesia using copula model. We identify the linearity of correlation between variables by comparing Pearson correlation with normality assumption and dependency structure modeled by copula function with any marginal distribution. We analyze and discuss how copula model shows the dependency between variables which cannot be identified by linear correlation.**

***Keywords*—** *agriculture, copula, dependency structure, linear correlation, Pearson correlation coefficient*

1. Introduction

Indonesia is one of the agrarian countries where the main livelihood of the population is farming. With a very strategic geographic position, Indonesia is lavished with natural resources and conditions that benefit the agricultural sector. There are two main factors affecting plant growth; internal and external factor. Internal or genetic factors include genes and plant hormone, whereas external factors include the acquisition of nutrients either from the provision of fertilizer or from natural conditions. In other words, the level of agricultural production largely is affected by natural and human capital [1]. Many studies have been conducted to identify the relationship between agricultural production and the factors that influence it.

International Journal of Supply Chain Management IJSCM, ISSN: 2050-7399 (Online), 2051-3771 (Print) Copyright © ExcelingTech Pub, UK (http://excelingtech.co.uk/)

[2] identified the association between soil variables and yield of paddy using a multiple linear regression model. They identified the relationship between crop yield and six soil variables; soil reaction, organic matter, total nitrogen, available phosphorus, potassium, and soil texture, by putting it into a Pearson correlation function. It was found that the most influence soil variables are total nitrogen, organic matter, and phosphorus. [3] predicted crop yield by analyzing the relationship between its environmental parameters such as area under cultivation, annual rainfall, and food price index using linear regression. [4] built weather analysis to predict rice cultivation time to escalate farmer's exchange rate. The weather variables used are average temperature, average humidity, rainfall, and solar radiation and were modeled using multiple linear regression.

The studies above conducted by using a linear regression model where its analysis was built based on the relationship between variables using Pearson correlation coefficient. The correlation coefficient was first introduced by [5], he found the idea of correlation in which two variables are said to be correlated if variations of one variable are followed (on average) by more or less variation of the other variable and in the same direction. This concept describes the linear relationship among variables. [6] stated some problems if the Pearson correlation coefficient is used as a dependency measure, some of which are; (1) the Pearson correlation coefficient is only a measure of scalar dependencies, it cannot give much information about the structure of non-linear dependencies between X and Y, (2) correlation value depends on the marginal distribution of variables, both must form a normal bivariate distribution, and (3) correlation is not invariant under transformations, for example $\log X$ and $\log Y$ do not have the same correlation

with X and Y. There are alternative correlations to measure dependency structure of some variables, which can accommodate both linear and non-linear correlation, i.e rank correlation and tail dependence. The measures can be expressed in a multivariate distribution function known as copula.

Copula was first introduced by Abe Sklar in 1959, it is a function that couple one-dimensional marginal distribution function forming a multivariate distribution function [7]. Copula is a function that invariant under strictly increasing transformation, therefore it is a robust statistical method. The small value of Pearson correlation coefficient does not indicate that the variables analyzed have no correlation. We can use copula model to identify the dependency structure between the variables. This study focuses on identifying the dependency structure of rice production and its environment indicators, and analyze how copula model can capture dependency structure which cannot be done by linear correlation.

2. Literature Review

2.1. Linear Correlation

The idea about the linear relationship found by [5] is then developed by [8] become a dependency measure called Pearson correlation coefficient. The measure has a normality assumption that must be met by the variables to be measured.

Suppose that X and Y are r.vs having variance σ_X^2 and σ_Y^2 respectively, and $\sigma_{XY} = Cov(X,Y)$, then the Pearson correlation coefficient between X and Y is defined by

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{X,Y} \le 1 \tag{1}$$

If the random variables have bivariate normal distribution (elliptical distribution), then the correlation coefficient $\rho_{X,Y}$ can be used because it is distribution whose density is constant on ellipsoids. In two dimension, the contour lines of the density surface are ellipses [6]. It means that the increases of one variable will be followed by the other variables which describe that the relationship between variables is linear.

2.2. Copula Function

Copula is a function that combines one-dimensional marginal distribution function forms a multivariate distribution function.

Suppose that $H_{X,Y}(x,y)$ is a joint distribution function with marginal distribution functions $F_X(x)$ and $G_Y(y)$. Then there is a copula C such that $\forall x, y \in R$ [7]

$$H_{X,Y}(x,y) = C\big(F_X(x), G_Y(y)\big) \tag{2}$$

If $F_X(x)$ and $G_Y(y)$ continue, then C is unique. Otherwise if C is copula, $F_X(x)$ and $G_Y(y)$ are distribution functions, then $H_{X,Y}(x,y)$ is a joint distribution function with marginal distribution function $F_X(x)$ and $G_Y(y)$.

C is a two-dimensional copula with domain I^2 and having properties

1. Grounded, $\forall u, v \in I$

$$C(u,0) = 0 = C(0,v) \tag{3}$$

$$C(u,1) = u \text{ and } C(1,v) = v \tag{4}$$

2. 2-Increasing, $\forall u_1, u_2, v_1, v_2 \in I$ such that $u_1 \le u_2$ and $v_1 \le v_2$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0 \tag{5}$$

There are many families of Copula, one of which is Archimedean Copula. We used three types of Archimedean Copula; Clayton, Gumbel, and Frank. An n-dimensional Archimedean copula, defined by [7], can be expressed as follows:

$$C(u_1,u_2,\ldots,u_n) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + \ldots + \varphi(u_n)) \tag{6}$$

where $\varphi(\cdot)$ is generator function of Archimedean copula. The generator function for Clayton, Gumbel, and Frank copula is given in Table 1.

Table 1. Generator Function for Archimedean Copula

Copula	Generator Function	Range of Parameter
Clayton	$\frac{u^{-\theta}-1}{\theta}$	$[-1,+\infty)/\{0\}$
Gumbel	$(-\ln u)^{\theta}$	$[1,+\infty)$
Frank	$\ln(e^{-\theta}-1) - \ln(e^{-\theta u}-1)$	$(-\infty,0)\cup(0,+\infty)$

Based on the generator function given in Table 1 and Eq. (6), Clayton, Gumbel, and Frank copula is defined as [9], [10], [11]

$$C^{Clay}(u_1,u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \tag{7}$$

$$C^{Gumb}(u_1,u_2) = \exp\left\{-\left[(-\ln u_1)^{\theta}+(-\ln u_2)^{\theta}\right]^{\frac{1}{\theta}}\right\} \quad (8)$$

$$C^{Frank}(u_1,u_2) = -\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{(e^{-\theta}-1)}\right) \quad (9)$$

The relationship between copula parameter θ with Kendall's Tau τ for each Archimedean copula given by

$$\theta^{Clay} = \frac{2\tau^{Clay}}{1-\tau^{Clay}} \quad (10)$$

$$\theta^{Gumb} = \frac{1}{1-\tau^{Gumb}} \quad (11)$$

$$\tau^{Frank} = 1+\frac{4}{\theta^{Frank}}(D_1(\theta^{Frank})-1) \quad (12)$$

where $D_1(\cdot)$ is Debye function of the first kind.

For parameter estimation of copula function, we used maximum likelihood estimation method. The parameter can be estimated by maximizing the log-likelihood function of its copula density given by [12]

$$L = \sum_{i=1}^{T}\log c_\theta(F_1(x_{1i}),F_2(x_{2i}),\ldots,F_n(x_{ni})) \quad (13)$$

where

$$c_\theta(F_1(x_{1i}),F_2(x_{2i}),\ldots,F_n(x_{ni})) = \frac{\partial^n}{\partial_1\ldots\partial_n}C(F_1(x_1),F_2(x_2),\ldots,F_n(x_n)) \quad (14)$$

While for selecting the best copula fitted to the data, we used the distance measure between the empirical copula and the estimated copula. The distance measure is given by

$$d(C_e,\hat{C}_X) = \sqrt{\sum_{i=1}^{m}(C_e(i)-\hat{C}_X(i))^2} \quad (15)$$

where

$C_e(i)$: Empirical Copula

$\hat{C}_X$: Estimated Copula

See [13] for the steps of calculating empirical copula. We also test whether the copula is fitted or not by

$$H_0 : C = C_\theta$$
$$H_1 : C \neq C_\theta$$

H_0 is rejected if $p-value < \alpha$.

3. Result and Analysis

In this section, we analyze and identify the dependency structure of rice production and its environment indicators consist of temperature change, CO_2 emission, and rainfall precipitation in Indonesia from 1961-2016. The data used in this paper are obtained from the official website of Food and Agriculture Organization of United Nations (FAO-UN) which can be accessed through website address http://www.fao.org/faostat/en/\#data and described by Figure 1.

Figure 1 shows the data pattern of rice production, temperature change, CO_2 emission, and rainfall precipitation. From the figures we can see that rice production and CO_2 emission have ascending trend, temperature change also has an ascending trend but with considerable fluctuations at some points, while rainfall precipitation has a highly fluctuating data with no specific trend. By assuming that all the data follows Normal distribution, we identified the correlation between rice production and the other variables, and test whether the correlation is significant or not by setting null hypotheses as $\rho = 0$. The null hypotheses will be rejected if $p-value < \alpha$ with $\alpha = 0.05$. Values of the linear correlation are presented in Table 2.

Table 2. Linear Correlation

Model	Correlation	$p-value$	Decision
Rice Prod-Temperature Change	0.304	0.026	Reject H_0
Rice Prod-CO_2 Emission	0.856	0.000	Reject H_0
Rice Prod-Rainfall Precipitation	-0.179	0.196	Do not reject H_0

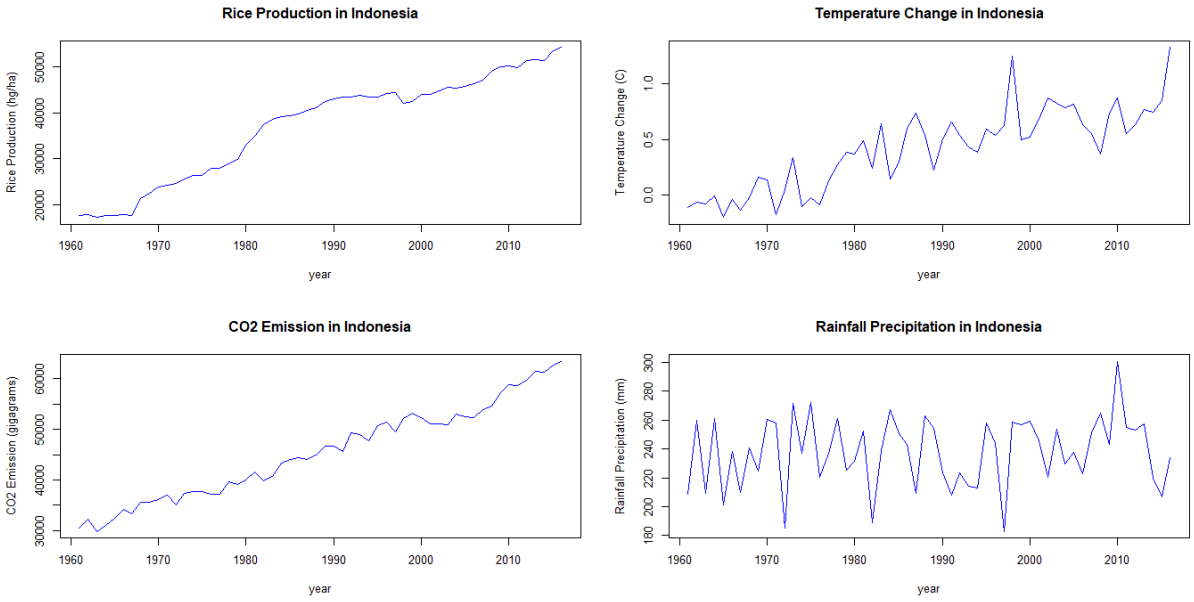
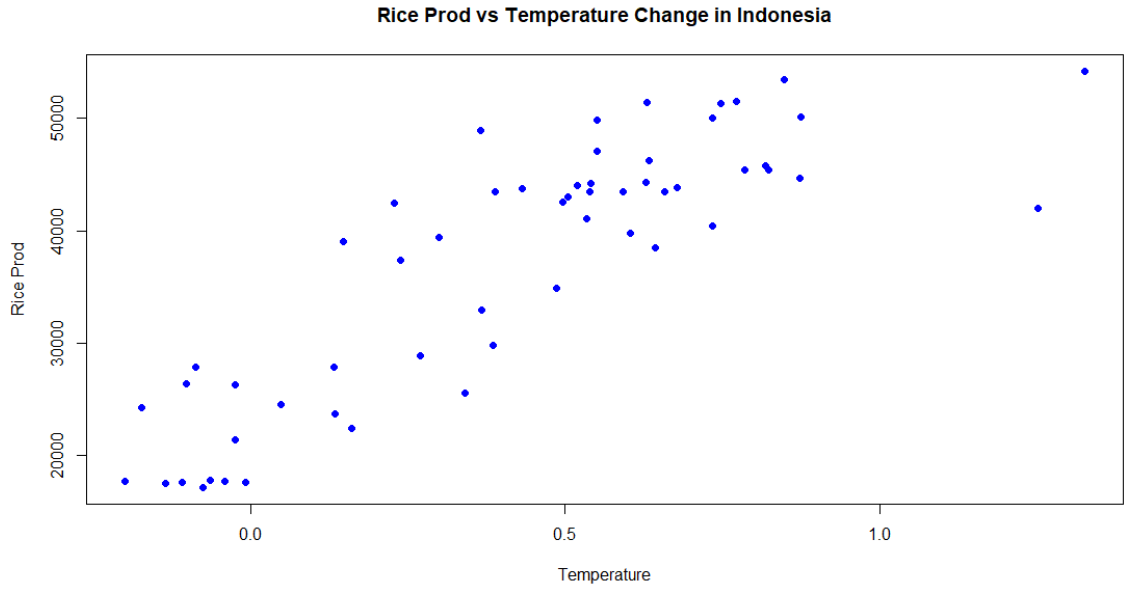


Figure 1. Description of Variables

The result above shows that rice production has a weak positive correlation with temperature change, strong positive correlation with CO_2 emission, and weak negative correlation with rainfall precipitation. The decisions also show that temperature change and CO_2 emission has correlation with rice production, while rainfall precipitation has no correlation. The weak correlations do not indicate that the variables have a weak relationship, there is a possibility that the variables have non-linear relationship. The relationship between rice production and the indicators, graphically displayed by Figure 2.



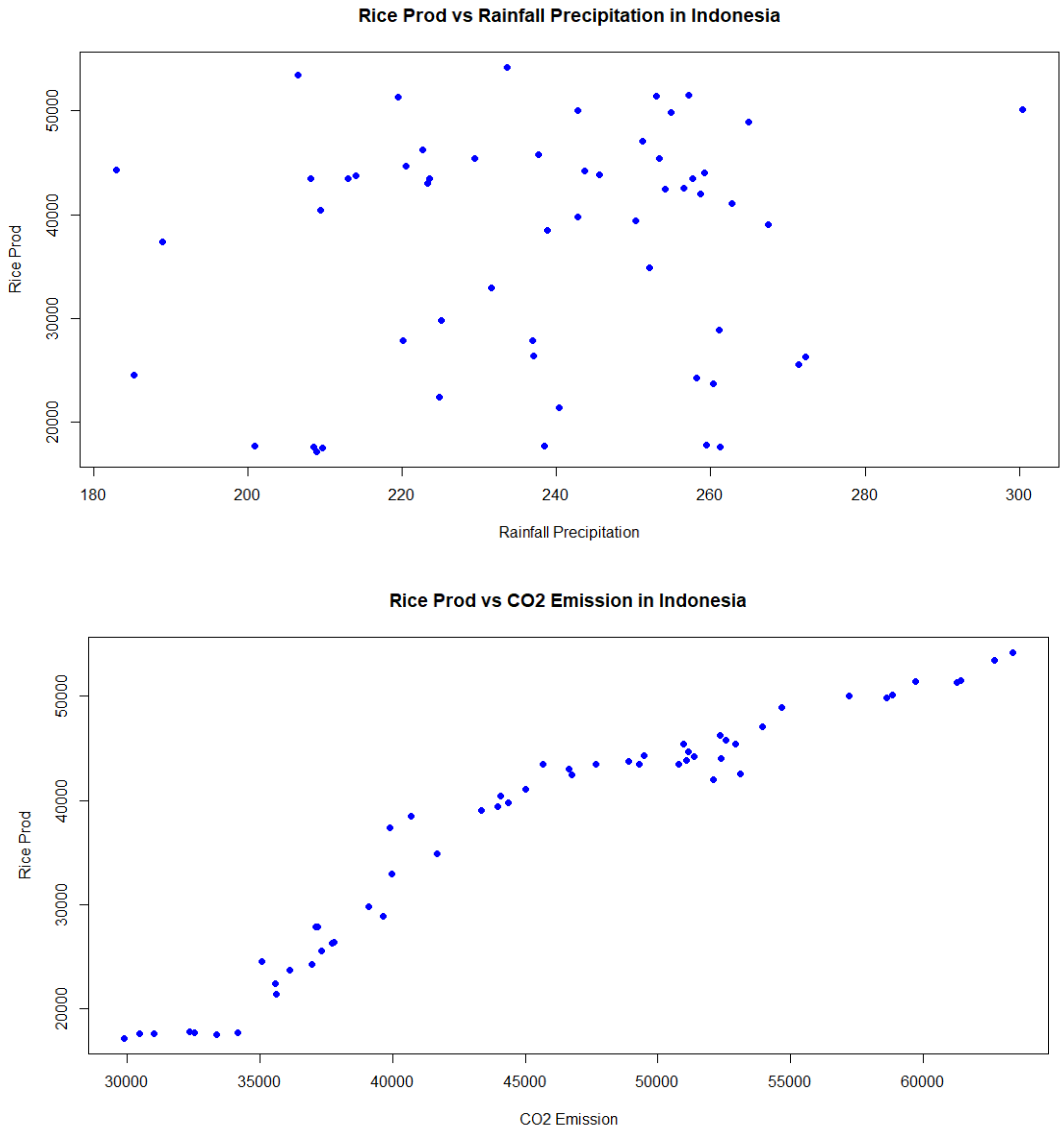


Figure 2. Scatterplot of The Relationship Between Rice Production and Its Environment Indicators

Figure 2 indicates that there is a relationship between rice production and the indicators, although the structure of the relationship is unknown. Therefore, we analyzed the relationship using copula function which can accommodate the possibility that there is non-linear relationship among variables. First of all, we identified the distribution of each variable, whether it has Normal distribution or not. We tested the hypotheses using Kolmogorov-Smirnov test by setting null hypotheses as Normal distribution, with a significant level of 5% . The result shows in Table 3 and Figure 3.

Table 3. Kolmogorov-Smirnov Test for Normal Distribution

Variable	$p-value$	Decision
Rice Production	0.000	Reject H_0
Temperature Change	0.200	Do not reject H_0
CO_2 Emission	0.200	Do not reject H_0
Rainfall Precipitation	0.161	Do not reject H_0

From the Kolmogorov-Smirnov test, we can identify that temperature change, CO_2 emission, and rainfall precipitation follow Normal distribution, while rice production does not. Figure 3 also shows that, graphically, the distribution of rice production has different pattern with Normal distribution. By using the result, the assumption of normality to be met by each variable when calculating linear correlation is not met. Therefore, the dependence structure of rice production and its indicators are then analyzed by performing copula modeling. We used Archimedean copula consists of

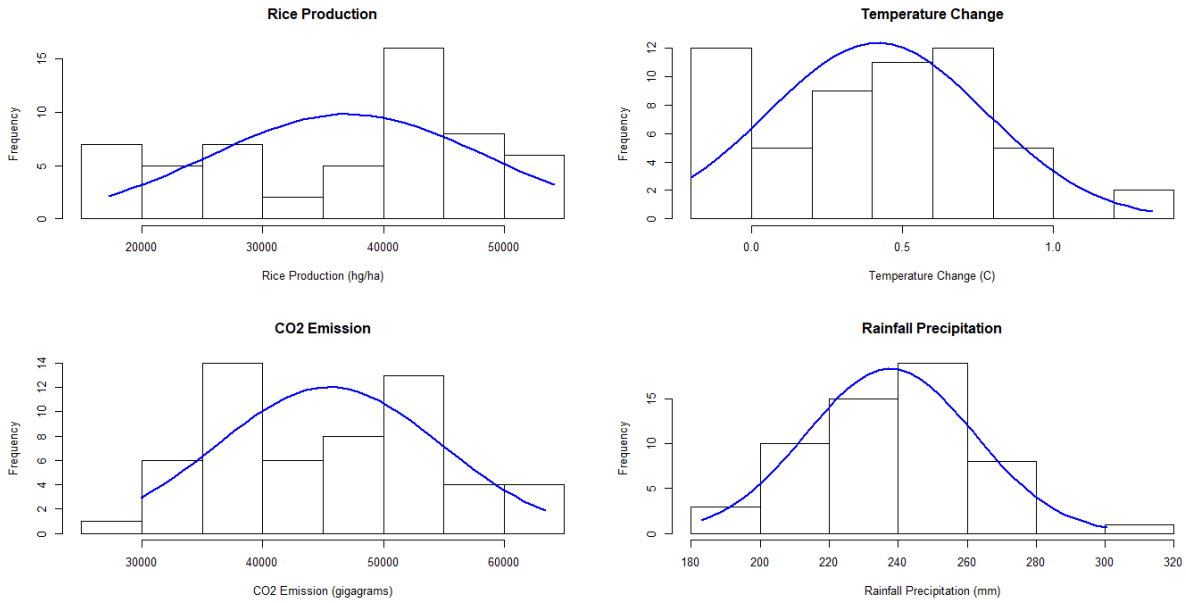


Figure 3. Histogram and Normal Curve of Variables

Clayton, Gumbel, and Frank copula. The result of parameter estimation, include θ and rank correlation τ, distance measure, statistics value, and $p-value$ are presented by Table 4

Table 4. Parameter Estimation of Archimedean Copula

Rice Production and Temperature Change						
Copula	θ	τ	DM	Statistics	$p-value$	Decision
Clayton	0.24	0.55	1.512891	0.072594	0.02941	Reject H_0
Gumbel	2.5683	0.6160637	1.432172	0.049097	0.009840	Reject H_0
Frank	9.65	0.66	0.846071	0.020807	0.402	Do not reject H_0
Rice Production and CO_2 Emission						
Copula	θ	τ	DM	Statistics	$p-value$	Decision
Clayton	9.62	0.83	1.083509	0.04779	0.02941	Reject H_0
Gumbel	9.1772	0.891034	0.749114	0.017044	0.2451	Do not reject H_0
Frank	34.25	0.89	0.747864	0.01819	0.1471	Do not reject H_0
Rice Production and Rainfall Precipitation						
Copula	θ	τ	DM	Statistics	$p-value$	Decision
Clayton	0.2	0.09	0.866617	-	-	-
Gumbel	1.0289	0.028088	0.8707	0.026024	0.5588	Do not reject H_0
Frank	0.34	0.04	0.846935	0.023865	0.598	Do not reject H_0

Table 4 gives some results as follow. For variable rice production and temperature change, the most appropriate copula can describe its dependency structure is Frank copula because it has the smallest distance measure. After testing the hypotheses, copula Frank is the most fitted copula because $p-value \geq \alpha$. The value of τ indicates that there is a moderate dependence between rice production and temperature change. For rice production and CO_2 emission, the most appropriate copula with the smallest distance measure is Frank copula. But after testing the hypotheses, Gumbel and Frank copula can be considered as the copula which can describe the dependence structure because both have $p-value \geq \alpha$. The value of τ indicates that there is a strong dependence between rice production and CO_2 emission. For rice production and rainfall precipitation, the copula having the smallest distance measure is Frank copula. Same with CO_2 emission variable, after testing the hypotheses, Gumbel and Frank copula can be considered as the most appropriate copula to describe the dependency

structure because both have $p-value \geq \alpha$. While Clayton copula did not give a result because there is infinite value in optimization while calculating the goodness of fit for the parameter. The value of τ indicates that there is a weak dependence between rice production and rainfall precipitation.

The distance measures between the empirical and estimated copula of each variable are presented in Figure 4.

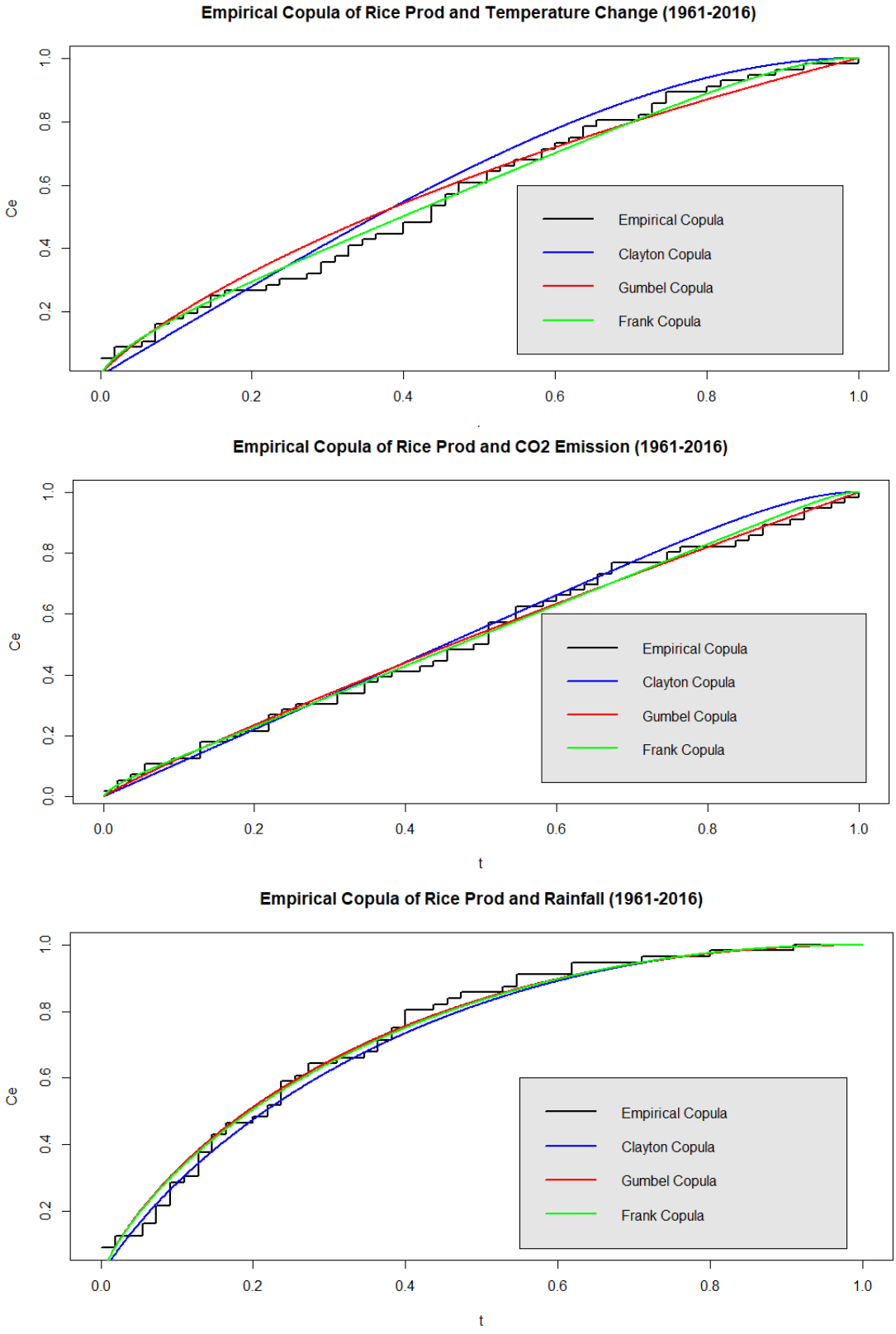


Figure 4. Distance Measures of Empirical and Estimated Copula

Figure 4 shows how close the estimated copula to the empirical copula. It can be used to determine the most appropriate copula to the data. From copula (black line) compared to Clayton copula. This result strengthens the analysis discussed from the Table 4.

Overall, we can say that the copula model can identify dependency structure better than linear correlation because it can accommodate both linear and non-linear correlation between variables.

4. Conclusion and Remarks

We have identified the dependency structure between rice production and its indicators using linear correlation and copula model. Archimedean copula, particularly Gumbel and Frank copula, has proven their capability and flexibility in capturing dependence structure between variables affecting rice production. From the analysis, we concluded that rice production has moderate dependency with temperature change, strong dependency with CO_2 emission, and weak dependency with rainfall precipitation. For the future studies, we can develop this research to the copula-based multiple regression model to predict the influence of the environment indicators to rice production by basing its model on the dependency structure analysis using copula.

References

[1] H. M. G. Van Der Werf and J. Petit, "Evaluation of the environmental impact of agriculture at the farm level: A comparison and analysis of 12 indicator-based methods," *Agric. Ecosyst. Environ.*, vol. 93, pp. 131–145, 2002.

[2] H. Dahal and J. K. Routray, "Identifying associations between soil and production variables using linear multiple regression models," *J. Agric. Environ.*, vol. 12, pp. 27–37, 2011.

[3] V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016.

[4] Luminto and Harlili, "Weather Analysis to Predict Rice Cultivation Time Using Multiple Linear Regression to Escalate Farmer ' s Exchange Rate," in *International Converence on Advanced Informatics: Concepts, Theory, and Applications*, 2017, pp. 1–4.

[5] F. Galton, "Co-relations and their measurement," *Proc. R. Soc. London*, vol. 45(273-279, pp. 135–145, 1889.

[6] P. Embrechts, A. McNeil, and D. Straumann, "Correlation Pitfalls and Alternatives," *Risk Mag.*, pp. 69–71, 1999.

[7] R. B. Nelsen, *An introduction to copulas, second edition*, Second. 2006.

[8] K. Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.

[9] D. G. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, vol. 65, no. 1, pp. 141–151, 1978.

[10] E. J. Gumbel, "Distributions des valeurs extremes en plusieurs dimensions," *Publ. Inst. Stat. Univ. Paris*, vol. 9, pp. 171–173, 1960.

[11] M. J. Frank, "On the simultaneous associativity of F(x,y) and x+y-F(x,y)," *Aequationes Math.*, vol. 19, no. 1, pp. 194–226, 1979.

[12] A. Charpentier, J. Fermanian, and O. Scaillet, "The estimation of copulas: theory and practice," in *Copulas: from theory to application in finance*, 2007, pp. 35–64.

[13] P. Kumar and M. M. Shoukri, "Copula based prediction models : an application to an aortic regurgitation study," vol. 7, no. 21, pp. 1–9, 2007.