Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 16, 2023, page 67 – 75

Modeling The Characteristics of Air Pollution Index Using Copula Approach

Chin Hui Shan, Shariffah Suhaila Syed Jamaludin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: suhailasj @utm.my

Abstract
The duration of air pollution and its severity are two main characteristics that used to describe the characteristics of unhealthy air pollution. Both are measured based on concentration of $PM_{2.5}$. Since these two variables are intercorrelated and have different marginal distributions, thus a copula approach can provide a good statistical tool to deal with these issues and obtained important information from air pollution data. Air pollution data for duration are determined from the Air Pollution Index (API) that exceed 100 while data for severity derived for the cumulative of the duration of air pollution events. The Kendall's tau correlation are applied to describe the characteristics of unhealthy air pollution. A case study is carried out in Klang, Malaysia to investigate the characteristics of unhealthy air pollution. By performing Cross-Validation Copula Information Criterion (cvCIC), Clayton copula is the best fitted copula to describe the correlation of duration and severity of unhealthy air pollution.

Keywords: $PM_{2.5}$; characteristics of unhealthy air pollution; dependence structure; correlation; copula

1. Introduction

Air pollution is defined as the combination of the air pollutants and form harmful gases that transforms around the Earth naturally. According to [1], the pollutants that included in the gases is particulate matter ($PM_{2.5}$ and PM_{10}), carbon monoxide (CO), ozone (O_3), sulphur dioxide (SO_2) and nitrogen dioxide (NO_2). The severity of the air pollution is measured and represented by Air Pollution Index (API). Department of Environment (DOE) Malaysia had carried out air quality monitory work in 1977 on the five pollutants based on research on [2]. From the study of [2], it demonstrated that PM_{10} is the main component that caused unhealthy air quality. According to [3], particle matter (PM) is the fine components that exist as liquid state in the air that are dangerous to human health. The fine particles are the $PM_{2.5}$ whereas the coast particle is the PM_{10}. According to [4], long interval exposure to the harmful pollutant will have high chance to get health diseases as it will accumulate inside the body. Based on Swiss air quality report technology company (IQAir) on 23 November 2022, Malaysia is at the moderate pollutant level compared to other countries and the most pollutant area of Malaysia is at Klang, Selangor.

The consequence of the relationship between emission and concentration of air pollutants need to be investigated by the government to minimize the pollution level in country. Based on [5], air pollution modelling is a tool to provide more deterministic description of air quality problem. In his study, Gaussian dispersion model act as indicator to air pollution since it has fastest response time and it can apply to only single wind situation by using American Meteorological Society/Environment Protection Agency Model (AERMOD). Another method to modelling air pollution is through Lagrangian model based on [6], Lagrangian method is used for a short period of simulation by calculating the trajectories of air pollution. Other than Lagrangian method, land use regression is also one of the ways. According to [7], by monitoring the globally satellite data and used Lasso variable selection, the amount of nitrogen dioxide and its residuals can be determined. Hence, the pollution level can be discovered. Since modelling air pollution is a complex case as it had strong dependence structure between variables and different marginal distribution exists, therefore copula method serves as a way to resolve it.

Based on research [8], copula had the ability to combine different marginal distributions and form multivariate distribution and measure the dependency structure between variables. Based on the research done, Pseudo Maximum Likelihood Estimation is used to estimate the parameters for the model based on the dataset. Copula is the platform of multivariate distribution that provide flexibility and can be used to solve higher dimensional dataset. By solving it, the analytical form of joint probability and statistical properties can be obtained. Besides its ability to combine different marginal distribution, [9] showed it can provide more function by considering the multivariate models.

Copula widely used in financial markets because it can figure out the non-parametric measure of dependence between variables according to [10]. From study of [11], for the prediction of the portfolio value at risk, the marginal distribution is modelled and measured dependence between margins using copula. In study of [12], by calculating the affected area and plotting the time series, the separation of drought will be prominent and obtain the duration and severity, 3D copula is ready to gain the joint probability and classes of variables. Copula used in the environmental issue according to the study of [13] showed copula applied in the forestry to analyse the tree growth and yield production. According to [14], C-vine copula had been used to describe the wind wave environment for sea crossing bridges by determine optimal marginal distribution and illustrate multiple correlation of the performance degradation in the accelerometer.

From [15], Archimedean copula which is the family of copula plays an important role in constructing optimal function type by goodness-of-fit and it consists of only a single parameter. By [16], Clayton copula is the branch of the Archimedean copula and it relates to lower tail dependency only that focuses on higher correlations. Another Archimedean copula is Gumbel-Hougaard that used to model the joint probabilistic characteristics for multivariate hydrological events in designing hydraulic and civil infrastructures by [17].

Kendall's tau correlation function in figuring out the similarity of the degree between sets of the same object from study of [18]. Another study in [19] showed that usage of the Kendall's tau correlation and the coefficient of upper tail can construct the graphical method to describe the data.

The objectives of this study are to model the characteristics of air pollution by measuring the pollutants which is particulate matters ($PM_{2.5}$). The dependency structure between the severity and the duration of the air pollution by using Kendall τ correlation and Spearman's rho correlation are examined. One of the objectives is to determine most suitable copula model by using goodness of fit test. The data used in this research are Air Pollution Index from 2017 to 2020. The data were obtained from the Department of Environment. The method used in this project is the subpart of the copula method which is Clayton Copula and Gumbel-Hougaard (GH) Copula. Kendall's tau correlation was also used in testing the dependency of the time taken and the severity of air pollution.

2. Methodology

Determination of the duration and severity of the air pollution

Concentration of $PM_{2.5}$ is used to compute the API value by using the following formula:

$$\mathrm{API} = \frac{(AQI_{Hi}) - (AQI_{Lo})}{(Conc_{Hi}) - (Conc_{Lo})} \times ((Conc_i) - (\mathrm{Conc_{Lo}})) + (\mathrm{AQI_{Lo}}) \quad (1)$$

The indicator $I_i(API_j)$ is determined as follows:

$$I_i(API_j) = \begin{cases} 1, if\ I_i(API_j) > 100 \\ 0, if\ I_i(API_j)\ \leq 100 \end{cases} \quad (2)$$

The duration (D_i) of an air pollution is computed as:

$$D_i = \sum_{j=1}^{N} I_i(API_j) \tag{3}$$

The severity (S_i) of air pollution is determined by:

$$S_i = \sum_{j=1}^{D_i} (API_j) \tag{4}$$

Determination of marginal distribution

Several statistical models are used to determine the distribution of duration and severity of unhealthy air pollution in Klang, Malaysia. The selected distribution is common apply in hydrological analysis and it can be useful in modelling distribution of duration and severity of unhealthy air pollution events.

Exponential distribution

Exponential distribution consisted of only one parameter and widely used in many fields. The maximum likelihood estimator is:

$$\hat{\theta} = \bar{x} \tag{5}$$

The probability density function (PDF) and cumulative density function (CDF) is as follows:

$$f(x) = \frac{1}{\theta}\exp(\frac{-x}{\theta}) \tag{6}$$

$$F\,(x) = 1 - \exp\,(\frac{-x}{\theta}) \tag{7}$$

where, $\hat{\theta}$ is the estimate parameter and x is the sample data

Gamma distribution

Gamma distribution used in theoretical and applied research areas. The estimated of maximum likelihood estimator are by solving Equation (8) and Equation (9) simultaneously:

$$\hat{\beta} = \frac{\bar{x}}{\theta} \tag{8}$$

$$\ln(\hat{\alpha}) - \psi(\hat{\alpha}) = \ln\,(\tfrac{1}{n}\textstyle\sum_{i-1}^{n} x_i) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ln x_i) \tag{9}$$

The probability density function (PDF) and cumulative density function (CDF) is as follows:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\exp(\frac{-x}{\beta}) \tag{10}$$

$$F\,(x) = \frac{\gamma(\alpha, \frac{x}{\beta})}{\Gamma(\alpha)} \tag{11}$$

where, α is the shape parameter, β is the scale parameter and $\gamma(.)$ is lower incomplete Gamma function.

Weibull distribution

Weibull distribution applied in research area and its maximum likelihood estimator is:

$$\hat{\alpha} = [\left(\sum_{i=1}^{n} x_i{}^{\hat{\alpha}} \ln x_i\right)\left(\sum_{i=1}^{n} x_i{}^{\hat{\alpha}}\right)^{-1} - n^{-1}\sum_{i=1}^{n} \ln x_i]^{-1} \tag{12}$$

$$\hat{\beta} = [(\frac{1}{n}\sum_{i=1}^{n} x_i{}^{\hat{\alpha}}]^{\frac{1}{\hat{\alpha}}} \tag{13}$$

The probability density function (PDF) and cumulative density function (CDF) is as follow:

$$f(x) = \frac{\alpha}{\beta}\,(\frac{x}{\beta})^{\alpha-1}\exp\,[-\left(\frac{x}{\beta}\right)^{\alpha}] \tag{14}$$

$$F\,(x) = 1 - exp(-\frac{x}{\beta})^{\alpha} \tag{15}$$

3. Copula Models

Clayton copula

Clayton copula applied to get the positive dependency of bivariate variable in the lower tail with 0 ≤ θ ≤ ∞. The joint distribution function (CDF) and probability density function (PDF) are as follows:

$$C\,(u_1, u_2) = \,(u_1{}^{-\theta} + \,u_2{}^{-\theta} - 1)^{\,-\frac{1}{\theta}} \tag{16}$$

$$c\,(\,u_1, u_2) = \,(\theta + 1)\,(u_1{}^{-\theta} + \,u_2{}^{-\theta} - 1)^{\,-(\frac{1+2\theta}{\theta})}(u_1 u_2)^{\,-\theta-1} \tag{17}$$

The Kendall’s tau correlation specified for Clayton is:

$$\tau = \frac{\theta}{\theta + 2} \tag{18}$$

Gumbel-Hougaard

Gumbel-Hougaard is capable to portrait the positive correlation between variables in the upper tail. The joint distribution function (CDF) and probability density function (PDF) are as follows:

$$C(u_1, u_2) = \,e^{-[(-lnu_1)^{\theta}+(-lnu_2)^{\theta}]^{\frac{1}{\theta}}} \tag{19}$$

$$c(u_1, u_2) = \frac{C(u_1, u_2)[(-ln\,u_1) + (\,-ln\,u_2)]^{\theta-1}[\theta]^{\frac{2}{\theta}-2}\left\{[\theta - 1][\theta]^{\frac{-1}{\theta}} + 1\right\}}{u_1 u_2} \tag{20}$$

with parameter space 1 ≤ θ < ∞ and $\theta = (-ln\,u_1)^{\theta} + (-ln\,u_2)^{\theta}$ which is parameter. The Kendall tau correlation in Gumbel-Hougaard is as follows:

$$\tau = 1 - \frac{1}{\theta} \tag{21}$$

Frank

Frank copula can used to describe the positive or the negative dependency because it can provide versatile dependency structure of the range of dependencies from [-1,1]. The joint distribution function (CDF) and probability density function (PDF) is given as:

$$C\,(u_1, u_2) = -\frac{1}{\theta}\ln\,[1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}] \tag{22}$$

$$c\,(\,u_1, u_2) = \frac{-\theta e^{-\theta(u_1+u_2)}\,(e^{-\theta} - 1)}{[e^{-\theta(u_1+u_2)} - e^{-\theta u_1} - e^{-\theta u_2} + e^{-\theta}]^2} \tag{23}$$

The Kendall’s tau correlation for Frank copula is:

$$\tau = 1 + \frac{4[D_1(\theta) - 1]}{\theta} \tag{24}$$

where $D_1(\theta) = \frac{\int_0^\theta \frac{t^k}{e^{(t-1)}}dt}{\theta}$ is the Debye function.

Pseudo Maximum Likelihood Estimation

Pseudo Maximum Likelihood Estimation is a method used to find the parameter of the given model. Itl estimates a reduced system of likelihood function by replacing the original parameter in the model. Pseudo MLE is obtained by maximizing the log-likelihood function and it will give large sample properties. Pseudo maximum likelihood estimation is the technique to replace the original with the new parameters to fix the copula method. It functions to be marginal distribution for variable by using a nonparametric distribution. Thus, the nonparametric distribution is given as follows:

$$\hat{F}_i(u) = \frac{\sum_{j=1}^{n} 1(U_{ij} \leq u)}{n+1} \quad , i = 1{,}2 \tag{25}$$

Next, the estimation of copula parameter is by maximizing the pseudo log likelihood function:

$$\log L(\theta) = \sum_{j=1}^{n} ln[c(\hat{F}_1(u_{1j}), (\hat{F}_2(u_{2j}); \theta)] \tag{26}$$

Marginal distribution selection

Akaike Information Criterion only require a short time compared others test such as multiplier good-of-fit test. The most suitable model can be determined by choosing the smaller AIC value. Akaike Information Criterion is defined as:

$$AIC = -2\iota(\theta_n) \ + 2q \tag{27}$$

where, (θ_n) is the maximized value of log pseudo likelihood function and *q* is number of free parameters.

Copula model selection

One of the approaches to choose best fitted copula is Cross-Validation Copula Information Criterion (cvCIC). The best fitted model is selected based on the maximum value of it. The fitted copula model is measured as follows:

$$xv_n = \frac{\sum_{i=1}^{n} lpg[c\theta_{n,-i}\left(\mathbf{F}_{n,-i}(\mathbf{U}_i)\right)]}{n} \tag{28}$$

where, $\theta_{n,-i}$ is the maximum pseudo-likelihood estimate and $\mathrm{F}_{n,-i}(\mathbf{U}_i) = [\mathrm{F}_{n,1,-i}(\mathrm{u}_1), \mathrm{F}_{n,2,-i}(\mathrm{u}_2)]$ computed by the following equation:

$$\mathrm{F}_{n,j,-i}(u) \begin{cases} \frac{\sum_{k=1}^{n} 1(U_{kj} \leq u)}{n}, if\ u \geq \min\limits_{k\in[1,2,\dots,n][i]} U_{kj} \\ \frac{1}{n}, otherwise \end{cases} \tag{29}$$

4. Results

From Figure 1, it seems that severity and the duration are correlated. By performing the correlation analysis, the correlation value for Kendall's tau is 0.981 with *p*-Value of $2.2e^{-16}$ which is less than 0.05 which is reject the null hypothesis. This also showed that the relationship between duration and severity is significant. Since the variables are significant, hence copula models are used to measure the dependency of the variables.

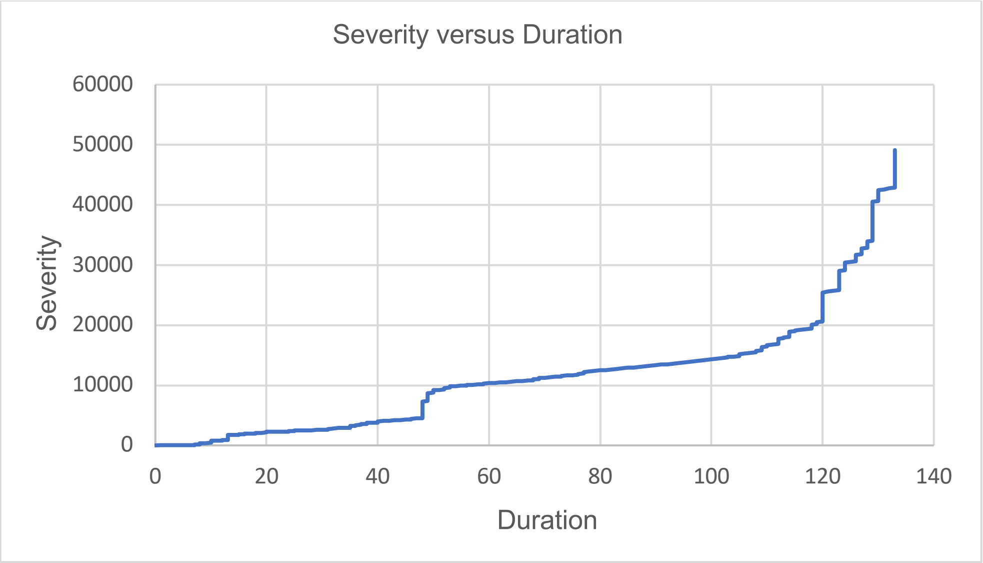

Figure 1: Correlation Analysis of Duration and Severity

Table 1: Correlation Analysis of duration and severity

Correlation Test	Correlation Value	p-Value
Kendall’s tau test	0.981	$2.2e^{-16}$

Table 2 showed the summary statistics for the duration of an unhealthy air pollution that happen in Klang, Malaysia. From the Table 2, the measure of the spread can be measured by the range of minimum value and maximum value. The larger difference of duration and severity showed that the data consist of variation. Other than that, the larger of the standard deviation showed that both of the variables are significant. The duration and severity displayed a positive skewed data since the mean is larger than the median. Since the skewness and kurtosis are not 0 hence, the data is asymmetric.

Table 2: Descriptive statistics of unhealthy air pollution events

Variables	Mean	Median	Min Value	Max Value	Standard Deviation	Skewness	Kurtosis
Duration	67.83	49.5	0	133	48.77	0.11	-1.65
Severity	14681.04	8714	0	49109	14916.82	0.83	-0.66

After visualized that for duration and severity, the marginal distribution had to be determined. Most possible distributions according to the plot of duration and severity is Exponential function, Gamma function and Weibull function. According to the Akaike Information Criterion (AIC), the best fit marginal distribution for duration is the Exponential distribution whereas for severity is the Gamma distribution. Both of the distributions have the lower distributions which are 7556.23 and 15322.46 respectively. In Kendall’s tau test, the relationship of the duration and severity is higher which is 0.981. Since the value is approximately to 1 then, it indicated that there exists a positive relationship between duration and severity. By using Spearman’s rho test, the value is 0.999 and showed strong positive relationship between duration and severity.

Table 3: Akaike Information Criterion for Duration and Severity

Duration	
Marginal Distribution	**Akaike Information Criterion (AIC)**
Exponential	7556.23
Gamma	7567.40
Weibull	9199.19
Severity	
Exponential	15342.56
Gamma	15322.46
Weibull	15344.56

By obtaining the suitable marginal distribution for the duration and severity, it can estimate the parameter of copula by pseudo maximum likelihood estimation. The estimated parameter for each copula model is as below:

Table 4: Estimation of the parameter for copula model
\

Copula Model	**Parameter (θ)**
Clayton	87.09
Gumbel-Hougaard	15.42
Frank	131.94

Since the marginal distributions and the estimated parameters had been known, the copula can be formed. To find the most fitted copula, Cross-Validation Copula Information Criterion (cvCIC) are employed. Based on the cvCIC, the best fitted copula is Clayton copula to describe the relation between the duration and severity of unhealthy air pollution event. This is because Clayton had the highest value which is 2478.33 compared to Gumbel-Hougaard and Frank copula. Based on Figure 2, the Clayton is the best fitted to the data of the duration and severity.

Table 5: Cross-Validation Copula Information Criterion (cvCIC) for copula models

Copula Models	**Parameter (θ)**	**cvCIC**
Clayton	87.09	2478.33
Gumbel-Hougaard	15.42	1695.97
Frank	131.94	2060.09

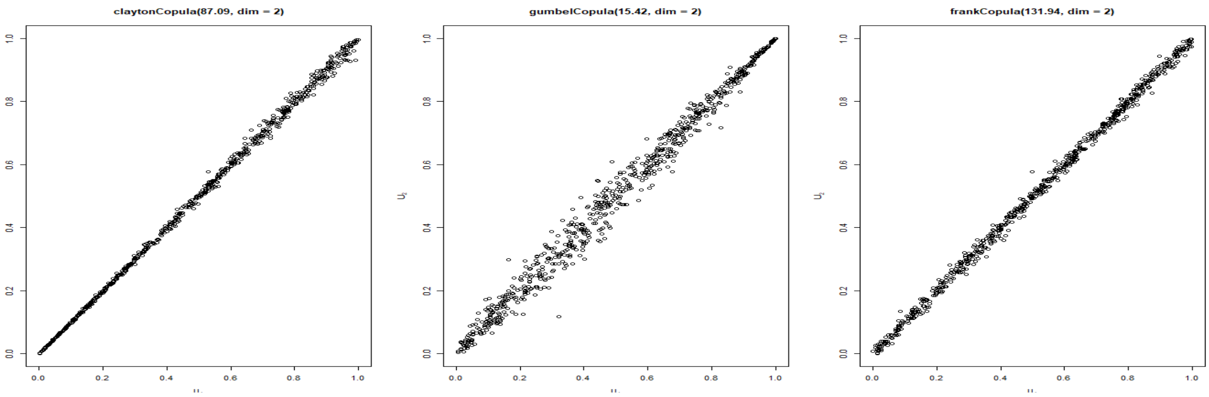

Figure 2: Plots of copula models

Conclusion

This study is focus on the characteristic of unhealthy air pollution event in Klang, Malaysia. The main focus is to evaluate the duration of the unhealthy air pollution events happen and the severity level of an unhealthy air pollution event. The duration of the air pollution is determined by the cumulative of subindices of Air Pollution Index (API) higher than 100. The API value of above 100 is considered that particular region had higher level of pollution. The severity of the air pollution is obtained by the cumulative value of the duration. Hence, the concentration of the $\boldsymbol{PM_{2.5}}$ is the indicator to measure the severity of the air pollution. Higher concentration of $\boldsymbol{PM_{2.5}}$ contribute to longer exposure to the highest level of pollution. Ultimately, reduce the activities that cause release of $\boldsymbol{PM_{2.5}}$ in the air is one of the ways to minimize the air pollution occur in the surrounding.

The correlation of the duration and severity is crucial to be known. By performing Kendall's tau test, the correlation value is 0.981 with *p*-value less than 0.05 confirmed that the relationship of the duration and severity is significant. Since the relationship is significant, the dependency between variables had to be determined by copula.

By performing the Cross-Validation Copula Information Criterion (cvCIC), the best fitted copula is Clayton copula. Clayton copula obtained the higher cvCIC value of 2478.32 compared to Gumbel-Hougaard and Frank copula. Gumbel-Hougaard is the least fitted copula as it obtained the lowest cvCIC value of 1695.97. From the Figure 4.11, the Clayton copula fit to the data well especially at the lower tail. Frank copula fit the data better than Gumbel-Hougaard copula as the outliers are less than in Gumbel-Hougaard copula.

References

[1] WHO. (2022). Air pollution. In: Compendium of WHO and other UN guidance on health and 42 environment, 2022 update. Geneva: World Health Organization; (WHO/HEP/ECH/EHD/22.01). WHO Fact Sheet, 2019(December), 5. https://www.who.int/en/news-room/fact-sheets/detail/arseni

[2] Mohd Zizi, N. A., Mohamed Noor, N., Izzah Mohamad Hashim, N., & Yusuf, S. Y. (2018). Spatial and Temporal Characteristics of Air Pollutants Concentrations in Industrial Area in Malaysia. IOP Conference Series: Materials Science and Engineering, 374(1). https://doi.org/10.1088/1757-899X/374/1/012094

[3] Salvador, S., & Salvador, E. (2012). Overview of Particle Air Pollution Air Quality Communication Workshop.

[4] Yen, C. C., & Chen, P. L. (2022). Regional air pollution severity affects the incidence of acute myocardial infarction triggered by short-term pollutant exposure: a time-stratified casecrossover analysis. Environmental Science and Pollution Research, 29(6), 8473– 8478. https://doi.org/10.1007/s11356-021-16273-4

[5] Daly, A., & Zannetti, P. (2007). Air Pollution Modeling – An Overview. Ambient Air Pollution, I(2003),15 – 28.

[6] Leelőssy, Á., Molnár, F., Izsák, F., Havasi, Á., Lagzi, I., & Mészáros, R. (2014). Dispersion modeling of air pollutants in the atmosphere: a review. Central European Journal of Geosciences, 6(3), 257–278. https://doi.org/10.2478/s13533-012-0188-

[7] Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., & Hystad, P. (2017). Global Land Use Regression Model for Nitrogen Dioxide Air Pollution. Environmental Science and Technology, 51(12), 6957–6964. https://doi.org/10.1021/acs.est.7b01148

[8] Masseran, N. (2021). Modeling the Characteristics of Unhealthy Air Pollution Events : A Copula Approach.

[9] Smith, M., & Smith, M. S. (2011). Modeling Multivariate Distributions Using Copulas : Applications in Marketing Modeling Multivariate Distributions Using Copulas : Applications in Marketing. January.

[10] Zhang, L., Morisaki, H., Wei, Y., Li, Z., Yang, L., Zhou, Q., Zhang, X., Xing, W., Hu, M., Shima, M., Toriba, A., Hayakawa, K., & Tang, N. (2019). Characteristics of air pollutants inside and outside a primary school classroom in Beijing and respiratory health impact on children *. Environmental Pollution, 255, 113147. https://doi.org/10.1016/j.envpol.2019.113147

[11] Karmakar, M., & Paul, S. (2019). Intraday portfolio risk management using VaR and CVaR:A CGARCH-EVT-Copula approach. *International Journal of Forecasting*, *35*(2), 699–709. https://doi.org/10.1016/j.ijforecast.2018.01.010

[12] Hou, W., Yan, P., Feng, G., & Zuo, D. (2021). A 3D Copula Method for the Impact and Risk Assessment of Drought Disaster and an Example Application. Frontiers in Physics, 9(April), 1–14. https://doi.org/10.3389/fphy.2021.656253

[13] Bhatti, M. I., & Do, H. Q. (2019). ScienceDirect Recent development in copula and its applications to the energy, forestry and environmental sciences. International Journal of Hydrogen Energy, 44(36), 19453–19473. https://doi.org/10.1016/j.ijhydene.2019.06.015

[14] Fang, C., Xu, Y., & Li, Y. (2022). Journal of Wind Engineering & Industrial Aerodynamics Optimized C-vine copula and environmental contour of joint wind-wave environment for sea-crossing bridges. Journal of Wind Engineering & Industrial Aerodynamics, 225(April), 104989. https://doi.org/10.1016/j.jweia.2022.104989

[15] Zhonghui, J., & Xueqin, L. (2019). Comparative analysis of PM2.5 pollution risk in China using three-dimensional Archimedean copula method. Geomatics, Natural Hazards and Risk, 10(1), 2368–2386. https://doi.org/10.1080/19475705.2019.1697761

[16] Boateng, M. A., Omari-Sasu, A. Y., Avuglah, R. K., & Frempong, N. K. (2022). A Mixture of Clayton, Gumbel, and Frank Copulas: A Complete Dependence Model. Journal of Probability and Statistics, 2022, 1–7. https://doi.org/10.1155/2022/1422394

[17] Qian, L., Zhao, Y., Yang, J., Li, H., Wang, H., & Bai, C. Z. (2022). A New Estimation Method for Copula Parameters for Multivariate Hydrological Frequency Analysis With Small Sample Sizes. Water Resources Management, 36(4), 1141–1157. https://doi.org/10.1007/s11269-021-03016-w

[18] Abdi, H. (2008). Kendall Rank Correlation Coefficient. The Concise Encyclopedia of Statistics, 278–281. https://doi.org/10.1007/978-0-387-32833-1_211

[19] Serinaldi, F. (2008). Analysis of inter-gauge dependence by Kendall's τ K, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. Stochastic Environmental Research and Risk Assessment, 22(6), 671–688. https://doi.org/10.1007/s00477-007-0176-4