JURNAL VARIAN
e-ISSN. 2581-2017

JURNAL VARIAN

Volume 5 | Issue 2 | April 2022

Susunan Dewan Redaksi Jurnal Varian

Vol. 5 No. 2 Tahun 2022

Penasehat/ Pembina
Dr. Ir. Anthony Anggrawan, M.Kom., Ph.D

Deputy Editor of Chief
Siti Soraya, M.Si

Reviewer
Prof. Dr. Rer.Pol. Heri Kuswanto, M.Si
Prof. Syafruddin Side, S.Si, M.Si, Ph.D
Santi Puteri Rahayu, S.Si., M.Si., Ph.D
Dr .Purhadi,M.Sc
Dr. Kusman Sadik, S.Si, M.Si
Dr. Irwansyah, S.Si., M.Si
Novri Suhermi, M.Sc
Dr. Agus M Soleh, S.Si., MT
M.Faris Fadillah Mardianto, S.Si., M.Si
Dr. Bambang W Otok, M.Si
M. Fathurahman, S.Si., M.Si
Bobby Poerwanto, S.Si., M.Si
Muh.Irwan, S.Si., M.Si
Harry Soepriyanto, S.Pd., M.Pd
Riska Yanu Fa,rifa, S.Si., M.Si
Amin Tohari, S.Si., M.Si
Dr. Walid, S.pd., M.Si

Editorial Board
Dr. Muhammad Ahsan, M.Si
Didiharyono, S.Si., M.Si
Habib Ratu Perwira Negara, S.Pd., M.Pd
Puspita Kartikasari, S.Si., M.Si
Vita Fibriyani, S.Si., M.Si
Maulida Nurhidayati, S.Si., M.Si
Gde Palguna Reganata, S.Si., M.Si
Andika Ellena Saufika Hakim Maharani S, S.Si., M.Si

Assistant Editorial Board
Abdul Muhaimi, S.Kom
Dinda Lestari, S.TP

Alamat Sekretariat/ Redaksi
LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
(LPPM)
UNIVERSITAS BUMIGORA
Jl. Ismail Marzuki - Mataram, Telp. (0370) 634498

JURNAL VARIAN

Vol.5 No. 2 April 2022

DAFTAR ISI

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 97–106
e-ISSN: 2581-2017

Convolutional Neural Network for Cataract Maturity Classification Based on LeNet

Radimas Putra Muhammad Davi Labib[1], Sirojul Hadi[2], Parama Diptya Widayaka[3], Irmalia Suryani Paradisa[4]

[1,4]Electrical Engineering, National Institute of Technology, Indonesia
[2]Information Technology, Universitas Bumigora, Indonesia
[3]Electrical Engineering, Surabaya State University, Indonesia

Article Info

Article history:

Received : 12-21-2021
Revised : 01-04-2022
Accepted : 04-11-2022

Keywords:

CNN;
Classification;
Cataract maturity;
LeNet.

ABSTRACT

The eyes are one of the vital organs owned by humans. One of the common eye diseases is cataracts. This disease is characterized by clouding of the lens of the eye and can interfere with vision. Worst case, sufferers can experience blindness. Cataract maturity can be divided into four categories, namely incipient, immature, mature, and hypermature. Cataracts can be removed through surgery when the cataract is in the mature or hypermature phase. Cataract examination is usually done using a slit lamp. The lack of hospitals that have this equipment can cause delays in the healing process for cataract sufferers. This study created an image processing algorithm for the maturity classification process of cataracts using the Convolutional Neural Network method with LeNet network architecture. The algorithm that has been built is capable of classifying the maturity of cataracts with an accuracy rate of 93.33%.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1629

Corresponding Author:

Radimas Putra Muhammad Davi Labib,
Electrical Engineering, Institut Teknologi Nasional
Email: radimas@lecturer.itn.ac.id

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The eyes are vital organs owned by humans that work to see the surrounding environment. The organ consists of a complex optical system that collects light from the surrounding environment (Atchison, 2018). Disorders of these organs can occur in anyone, especially in people with old age. One of the common eye disorders is cataracts. A cataract is an eye disease characterized by cloudiness in the eye's lens which can interfere with the process of entering light into the eye. This can result in blurred vision or even blindness. The disease can be caused by various factors such as age, diabetes, hypertension, and smoking habits (Harun et al., 2020). There are six types of cataracts, namely senile, congenital, traumatic, complicated, toxic, and secondary cataract. Cataract maturity is divided into four, namely incipient, immature, mature, and hyper mature. This disease can be cured by means of surgery, but the process is only when the cataract is in the mature and hyper mature phase, so it is necessary to classify the cataract maturity before surgery (Astari, 2018). Cataract examination is usually done using a slit lamp. This equipment has a very high price, so that not all health infrastructure has it. This results in delays in the healing process in cataract patients and also causes the number of cataract sufferers to increase.

The development of artificial intelligence system methods is very rapid at this time. One of the branches of artificial intelligence methods that have been widely developed is the Convolutional Neural Network (CNN). At the beginning of its appearance, this method was used as a handwriting recognition system based on images with a neural network architecture called LeNet. In its

development, the architecture can be used for the general object recognition process. In 2017, there was research on the use of LeNet architecture to perform the recognition process in handwritten Arabic numerals based on digital images (Sawy et al., 2017). There was a study in 2018 on an electronic nose gas identification system using the CNN method with LeNet as the architectural model (Wei et al., 2019). The research produced output in the form of a gas identification system with an accuracy rate of 98.67%. This research proves that LeNet is not only used for handwriting recognition but can also be used as a general classification algorithm. Research in 2019 on a sleep apnea detection system based on electrocardiogram signals using the CNN method with the LeNet architectural model (Wang et al., 2019). The study produced output in the form of a sleep apnea detection system with an accuracy rate of 97.1%.

There was a similar study in 2016, research conducted by Hariyanto and his team (Hariyanto et al., 2016) regarding the process of classifying cataracts based on pathological abnormalities using the Learning Vector Quantization (LVQ) algorithm. In this study, the classification process was carried out using one of the techniques in an artificial neural network system with a dataset in medical records from patients with cataracts. This study resulted in an accuracy of 99% cataract determination. This research focuses on the implementation of LVQ for cataract classification based on medical record datasets. In 2017, a study was conducted by Purba and his team (Purba et al., 2017) regarding the cataract diagnosis system using the concept of the retrograde method. The study used an expert system algorithm to diagnose based on the symptoms felt by the sufferer. This research produces an output in the form of a final diagnosis system regarding the severity of cataracts suffered in the form of a percentage. However, in this study, the accuracy of cataract eye detection was not explained. In 2019, Risma and her team (Risma et al., 2019) researched performance analysis of cataract detection systems based on digital images using Discrete Cosine Transform as a feature extraction technique and artificial neural networks as a classification technique. The research produced output in the form of a simulation that can detect and classify cataracts with an accuracy rate of 87.66%. In addition, a study conducted by Gifran and his team (Gifran et al., 2019) regarding the cataract classification process based on digital images using the Discrete Wavelet Transform method as a feature extraction technique and Support Vector Machine as a classification technique. The research produced output in a cataract classification system algorithm with an accuracy rate of 80%.

The purpose of this study is to be able to classify cataract maturity cheaply and have high accuracy. Based on previous research, the classification of cataract maturity was carried out using medical record data, the classification would require a longer time because it had to enter medical record data into the system to produce a cataract maturity classification. In this paper, the classification of cataract maturity is based on image data. The system was built using the CNN method with LeNet architecture, which can produce a high-accuracy classification process without having to use other methods to obtain certain features from digital images as in previous studies.

B. LITERATURE REVIEW

1. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is an artificial neural network that performs classification with high accuracy by inputting raw images (Suniantara et al., 2020). Before CNN, various feature extraction methods were used to describe the object in the picture. The resulting features were only suitable for describing specific objects, so they cannot be used to describe various objects universally. CNN provided a scalable approach by utilizing the principles of linear algebra to identify features in objects (Albawi et al., 2017). CNN can be used to describe various objects universally, but it requires high computational resources. There are three main types of layers, namely the convolutional layer, the pooling layer, and the fully connected layer (Bau et al., 2020).

The convolutional layer is the core of a CNN where most of the computation is done (Liu and Guo, 2019) (Suniantara et al., 2020). Generally, this layer is followed by an additional convolutional layer or it can also be followed by a pooling layer. The convolutional layer requires several components such as input data, kernel, and feature map. For example, the input data is an image with a Red Green Blue (RGB) color space. This means that the input is a three-dimensional matrix consisting of height, width, and depth. Height and width represent pixel data, while depth is a color channel. In addition to the input, there is also a feature detector (also known as the kernel) which will move through the receptive plane of the image to check for the presence of the required features. This process is known as convolution. A feature detector or kernel is a two-dimensional array that represents a part of the image. The size of the kernel can vary but is usually a 3×3 matrix. The kernel is then applied to the image area and the dot product calculation process is carried out between the input pixels and the kernel. The result of the dot product calculation is then entered into the output array. Then, the kernel is shifted and the process is repeated until it covers the entire image area. The final output result of a series of dot product calculations between the input and the kernel is known as a feature map, activation map, or convolved feature. Each output value in the feature map does not have to be related to the pixel

value of the input or image. The output value only needs to correspond to the receptive field to which the kernel is applied. In general, the convolution process can be seen in Figure 1.

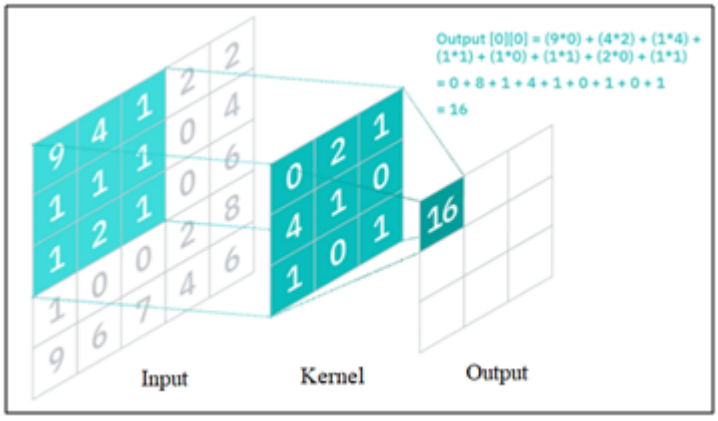


Figure 1. Convolution Process

After convolution operation, an activation function in the form of a Rectified Linear Unit (ReLu) transformation is applied to the feature map. It aims to introduce non-linearity to the model. Rectifier Linear Unit or commonly abbreviated as ReLU is a linear activation function. The ReLU activation function began to emerge in the context of visual feature extraction in neural networks in the late 1960s. ReLU has become the default activation function in many types of neural networks because it produces a model that is easy to train and has a strong biological motivation and mathematical justification and so often gives excellent performance (Agarap, 2018). ReLu activation function is mathematically described by Equation (1). The ReLU activation function is used to pass the input directly if it is positive. If the input is negative, this activation function will produce an output of zero.

$$y = \max(0, x) \tag{1}$$

The second layer in the Convolutional Neural Network is the pooling layer. This layer is also known as the down sampling process which has the main goal of reducing the dimensions of the input so that its parameters are reduced (Gholamalinezhad and Khosravi, 2020). The working principle of this layer is similar to the Convolutional Layer, which is to apply a filter to the input, but the kernel in this layer does not have a weight value. Instead, the kernel at this layer applies an aggregation function to the receptive field and the result is entered into an output array. There are two types of pooling, namely max pooling and average pooling. However, in this study only the max-pooling type was used, the way it works is shown in Figure 2.

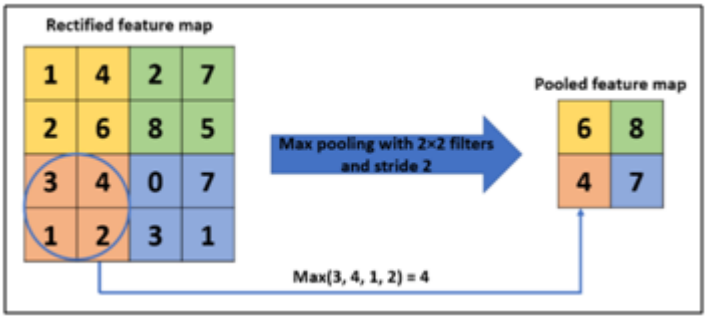


Figure 2. Max-Pooling

The third layer in the Convolutional Neural Network is the fully connected layer. After getting some features map from the convolutional layer and the pooling layer, the matrices are smoothed into a vector and put into a fully connected layer (Zhou et al., 2017). An example of a fully connected layer can be seen in Figure 3.

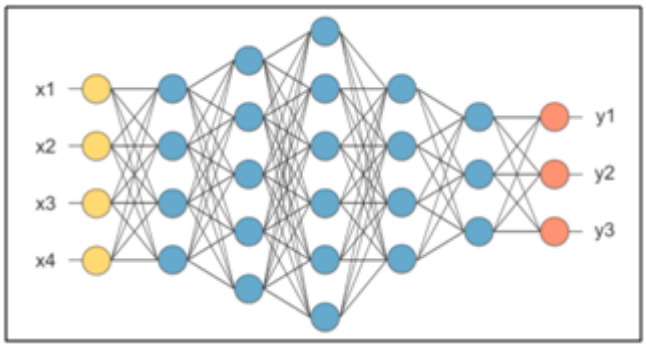


Figure 3. Fully Connected Layers

In Figure 3, the feature map matrix will be converted into a vector $(x1, x2, x3, x4)$. With a fully connected layer, the system can combine these features to produce a model. The output classification process is carried out by applying the softmax activation function. The softmax activation function or also known as the normalized exponential activation function is a generalization of the logistic function for many dimensions (Kanai et al., 2018). Softmax is used in multinominal logistic regression and is often used as the last activation function in neural networks to normalize the network output to a probability distribution over the predicted output class. Mathematically, the softmax activation function is described in Equation (2).

$$y_i = \frac{\exp(x_i)}{\Sigma_j \exp(x_j)} \tag{2}$$

2. LeNet Network Structure

LeNet is one of the earliest convolutional neural network structures (Rongshi and Yongming, 2019). In 1989, Yann LeCun and his team implemented a combined convolutional neural network that was trained to use backward propagation to read handwriting. The neural network structure succeeded in identifying the zip code numbers written by hand. This structure is a prototype of what is hereinafter referred to as LeNet. The classification system in this study uses LeNet as a CNN architecture with a network structure as shown in Figure 4.

In this study, the input used is an image of an eye with RGB color space and dimensions of 28×28 pixels. Next, a convolution process will be carried out with a 5×5 kernel to get a line of feature maps measuring 24×24 with a depth of 20 feature maps. After getting the first line of feature maps, then the max-pooling process is carried out with a 2×2 kernel to produce a line of feature maps measuring 12×12 with a depth of 20 feature maps. Then a second convolution process is carried out with a 5×5 kernel to get a line of feature maps measuring 8×8 with a depth of 50 feature maps. Then the max-pooling process is carried out with a 2×2 kernel to produce a row of 4×4 feature maps with a depth of 50 feature maps. The next process is flattened which serves to convert the last row of feature maps into vectors. The vector will be inserted into the fully connected layer to get the prediction results.

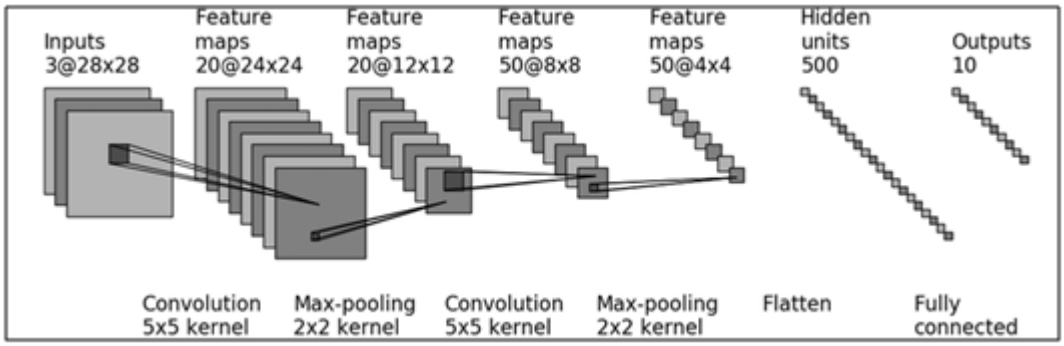


Figure 4. LeNet Network

DOI: https://doi.org/10.30812/varian.v5i2.1629

C. RESEARCH METHOD

1. Training Algorithm Design

Artificial neural networks are designed to imitate the work of the human brain through a combination of input data, weights, and biases. These elements work together to identify, classify, and accurately describe objects. The stages of machine learning can be seen in Figure 5.

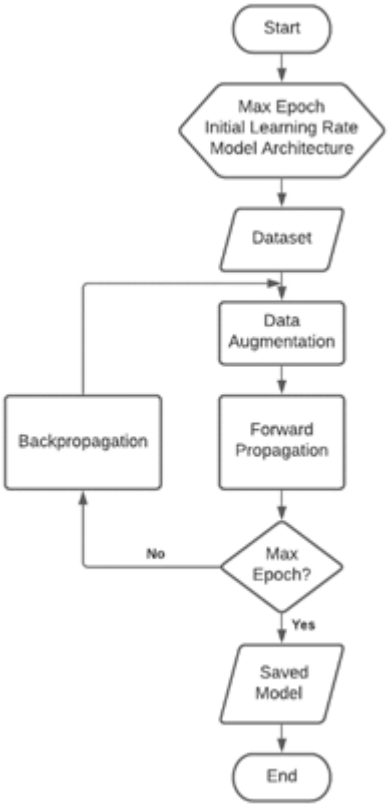


Figure 5. Training Algorithm Design

A deep neural network consists of several layers of interconnected nodes. The artificial neural network model requires a learning process based on datasets that have been prepared to get prediction results with a high level of accuracy. There are two main processes in a machine learning system on an artificial neural network, namely forward propagation and backward propagation. Forward propagation is a computational process through each layer of the network to get the prediction results from the classification system and get the loss-accuracy value. Backpropagation is an evaluation process to improve the elements in each layer of the neural network to improve the accuracy of the prediction process. Both processes are carried out internally by the Tensorflow module.

2. Prediction Algorithm Design

After doing the training process, an artificial neural network model will be obtained which is ready to be used for the prediction process. Before carrying out the prediction process, preprocessing is carried out on the input data to adjust the input standards for the neural network. After that, the prediction process will be carried out through the forward propagation process using an artificial neural network model generated by the learning process. Broadly speaking, the stages of the prediction system can be seen in Figure 6.

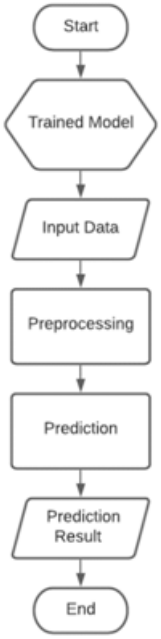


Figure 6. Algorithm of Prediction Process Design

3. Dataset

The dataset is used in the training process on an artificial neural network. In this study, a dataset of 37 images was used with details of 10 images of eyes with immature cataracts, 17 images of eyes with mature cataracts, and 10 images of normal eyes . The dataset used is obtained from Google Images. The dataset is shown in Figure 7.

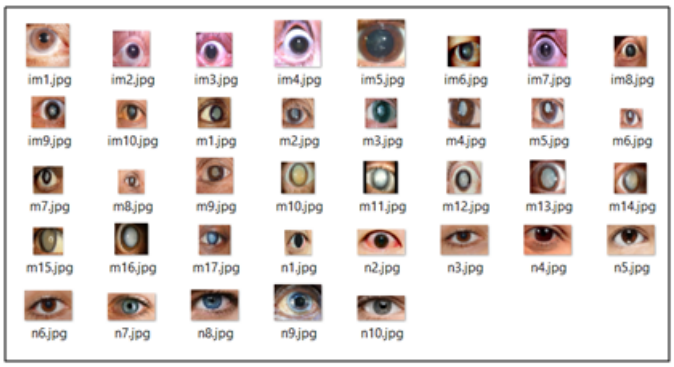


Figure 7. Dataset

D. RESULTS AND DISCUSSION

1. Machine Learning Result

The learning process was carried out three times with a different number of epochs in each experiment. The first experiment used 25 epochs with a training graph as shown in Figure 8 and a validation graph as shown in Figure 9. In the training process, there is a change in the value of accuracy and loss at each epoch. Although the change does not always increase, the final results show that the accuracy value is greater than in the previous epochs. This training process gives the final results in the form of an accuracy value of 81.82% and a loss value of 0.3957.

Figure 8. Training process on the first experiment

The validation process is carried out after getting the results of each epoch of the training process. The validation process of the model that has been trained in the last epoch produces an accuracy value of 73.33% and a loss value of 0.4085. Based on these results, the model generated can be used to classify cataract maturity because it has an accuracy rate above 50%.

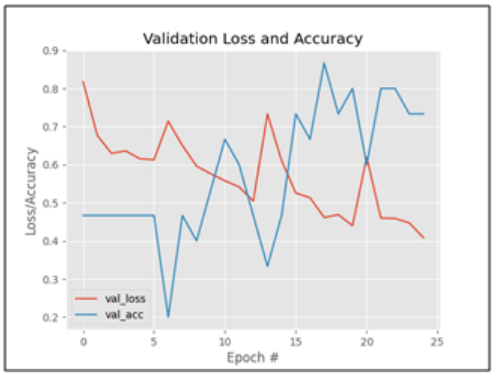


Figure 9. Validation process on the first experiment

The second experiment used 50 epochs with a training graph as shown in Figure 10 and a validation graph as shown in Figure 11. During the training process, the accuracy value in the last epoch was not the best. However, it shows that the accuracy value is still above 50%. This training process gives the final results in the form of an accuracy value of 86.36% and a loss value of 0.3092.

Figure 10. Training process on the second experiment

In the validation process, the final result is an accuracy value of 73.33% and a loss value of 0.4025. Based on these data, it can be shown that the loss value decreased slightly compared to the previous experiment. In addition, there is no difference in the accuracy obtained when compared with the previous experiment. However, the model generated in this experiment is also able to carry out the process of classifying cataract maturity.

Figure 11. Validation process on the second experiment

The third experiment used 100 epochs with a training graph as shown in Figure 12 and a validation graph as shown in Figure 13. The training process gives the final results in the form of an accuracy value of 95.45% and a loss value of 0.1424. It shows that the training results obtained are much better than the first and second experiments.

In the validation process, the resulting accuracy value is 93.33% and the resulting loss value is 0.2705. Based on these data, it can be shown that the validation results obtained are much better than the first and second experiments. In addition, the resulting accuracy rate is above 90% for the results of the training and validation, so that the developed model is capable of classifying the maturity of cataract s with very high accuracy. The value of accuracy and loss in all machine learning that has been carried out can be shown in Table 1.

Figure 12. Training process on the third experiment

Figure 13. Validation process on the third experiment

Table 1. Results of training and validation on cataract maturity

Epoch	Training		Validation	
	Accuracy	Loss	Accuracy	Loss
25	81.82%	0.3957	73.33%	0.4085
50	86.36%	0.3092	73.33%	0.4025
100	95.45%	0.1424	93.33%	0.2705

2. Prediction System Result

The prediction system uses an artificial neural network model generated by the learning process in the third experiment. The experiment has a higher accuracy value compared to the learning process in the first and second experiments. The results obtained in the prediction system can be shown in Figure 14.

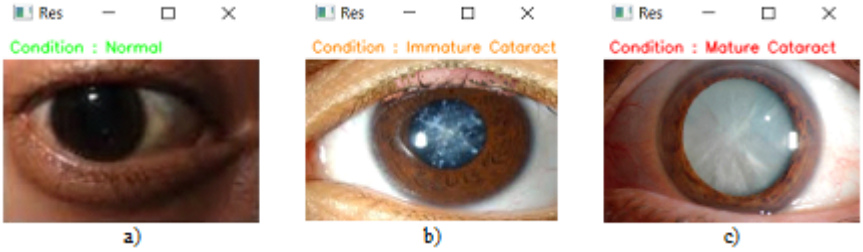


Figure 14. Prediction result: a) Normal eye; b) Immature Cataract; c) Maturation Cataract

In detail, the prediction process produces an output value for each label. From all these output values, the label with the highest value was chosen as a predictor of cataract maturity. The process can be shown in Table 2 . In the table, the output value written in bold is the highest value . So the label with the highest value will be used as a prediction result.

Table 2. Cataract maturity prediction system result

Input	Output Value			Prediction
	Normal	Immature	Maturation	
	0.6812226	0.3176026	0.0011747	Normal
	0.0039405	0.9110546	0.0850049	Immature
	0.0293672	0.2084033	0.7622294	Maturation

E. CONCLUSION AND SUGGESTION

Based on the experiments that have been carried out, it can be concluded that the image processing algorithm based on the Convolutional Neural Network designed can carry out the maturity classification process of cataracts. Unlike in previous studies, the algorithm that has been designed does not require additional methods to extract certain features. Based on the results of experiments that have been carried out, this algorithm can perform the classification process with an accuracy rate of 93.33%. This proves that the use of CNN with LeNet architecture for the classification process is more efficient and provides more accurate classification results

ACKNOWLEDGEMENT

We would like to thank Institut Teknologi Nasional Malang for funding our research.

REFERENCES

Agarap, A. F. (2018). Deep Learning using Rectified Linear Units (RELU). *http://arxiv.org/abs/1803.08375*.

Albawi, S., Mohammed, T. A. M., and Alzawi, S. (2017). Layers of A Convolutional Neural Network. *IEEE*, page 16.

Astari, P. (2018). Katarak: Klasifikasi, Tatalaksana, dan Komplikasi Operasi. *Cermin Dunia Kedokteran*, 45(10):748–753.

Atchison, D. A. (2018). Optics of The Human Eye. *Encyclopedia of Modern Optics*, 1-5:43–63.

Bau, D., Zhu, J. Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding The Role of Individual Units in A Deep Neural Network. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30071–30078.

Gholamalinezhad, H. and Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, A Review.

Gifran, N. A., Magdalena, R., and Fuadah, R. Y. N. (2019). Klasifikasi Katarak Menggunakan Metode Discrete Wavelet Transform dan Support Vector Machine Classification of Cataract Using Discrete Wavelet Transform and Support Vector Machine. *e-Proceeding of Engineering*, 6(2):4170–4177.

Hariyanto, R., Basuki, A., and Hasanah, R. N. (2016). Klasifikasi Penyakit Mata Katarak Berdasarkan Kelainan Patologis dengan Menggunakan Algoritma Learning Vector Quantization. *Jurnal Ilmiah NERO*, 2(3):177–182.

Harun, H. M., Abdullah, Z., and Salmah, U. (2020). Pengaruh Diabetes, Hipertensi, Merokok dengan Kejadian Katarak di Balai Kesehatan Mata Makassar. *Jurnal Kesehatan Vokasional*, 5(1):45.

Kanai, S., Yamanaka, Y., Fujiwara, Y., and Adachi, S. (2018). Sigsof Tmax: Reanalysis of The Softmax Bottleneck. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):286–296.

Liu, G. and Guo, J. (2019). Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification. *Neurocomputing*, 337:325–338.

Purba, W., Aisyah, S., and Tamba, S. P. (2017). Perancangan Sistem Pakar Diagnosa Penyakit Mata Katarak Menggunakan Konsep Metode Runut Mundur. *JUSIKOM PRIMA (Junal Sistem Informasi Ilmu Komputer Prima)*, 1(1).

Risma, H. A., Patmasari, R., and Magdalena, R. (2019). Analisis Performansi Sistem Pendeteksi Katarak Menggunakan DCT (Discrete Cosine Transform) dan Jaringan Saraf Tiruan Backpropagation (JST Backpropagation). *e-Proceeding of Engineering*, 6(1):364–371.

Rongshi, D. and Yongming, T. (2019). Accelerator Implementation of Lenet-5 Convolution Neural Network Based on FPGA with HLS. *3rd International Conference on Circuits, System and Simulation, ICCSS 2019*, pages 64–67.

Sawy, A. E., El-Bakry, H., and Loey, M. (2017). CNN for Handwritten Arabic Digits Recognition Based on LeNet-5. *Advances in Intelligent Systems and Computing*, 533:565–575.

Suniantara, I. K. P., Suwardika, G., and Soraya, S. (2020). Peningkatan Akurasi Klasifikasi Ketidaktepatan Waktu Kelulusan Mahasiswa Menggunakan Metode Boosting Neural Network. *Jurnal Varian*, 3(2):95–102.

Wang, T., Lu, C., Shen, G., and Hong, F. (2019). Sleep Apnea Detection From A Single-Lead ECG Signal with Automatic Feature-Extraction Through A Modified LeNet-5 Convolutional Neural Network. *PeerJ*, 2019(9):1–17.

Wei, G., Li, G., Zhao, J., and He, A. (2019). Development of A LeNet-5 Gas Identification CNN Structure for Electronic Noses. *Sensors (Switzerland)*, 19(1).

Zhou, Y., Song, S., and Cheung, N. M. (2017). On Classification of Distorted Images with Deep Convolutional Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (January):1213–1217.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 107–114
e-ISSN: 2581-2017

Workload and Performance of Nurses During The Covid-19 Pandemic: A Meta Analysis Study

Gde Palguna Reganata[1], I Gusti Ngurah Made Yudhi Saputra[2]
[1,2]Hospital Administration, Bali International University, Indonesia

Article Info

Article history:

Received : 01-11-2022
Revised : 03-07-2022
Accepted : 04-19-2022

Keywords:

Covid-19;
Meta-Analysis;
Nurses;
Workload;
Performance.

ABSTRACT

The surge in Covid-19 cases has caused hospitals and health workers to experience functional collapse. The high workload in handling Covid-19 cases by nurses is happening everywhere. Many studies have been conducted to look at the effect of workload on nurse performance during a pandemic. This research was conducted to determine the effect of workload on the performance of nurses with a meta-analysis approach. This type of research is observational with a retrospective approach. This research conducted through secondary data obtained from relevant sources related to the workload of nurses and nurse performance in various journals. The population and samples were taken from studies that met the criteria. Data analysis using meta-analysis. The result showed that there is a negative correlation between workload and performance of nurses, with $\rho = 0.334$ are in the reception area of the 95% (0.334 ± 0.219) confidence interval with p-value < 0.0001. Workload has a contradictory effect on performance, where when the workload of nurses is high, nurses tend to experience a decrease in performance. This needs to be a serious concern, because nurses are at the forefront of health services. If the nurse's performance has started to decline, then the patient's handling becomes not optimal and can increase the risk of death for the patient.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1657

Corresponding Author:

Gde Palguna Reganata,
Hospital Administration, Bali International University
Email: palgunareganata@iikmpbali.ac.id

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The Covid-19 pandemic is still a world health issue that gets high attention in handling. Starting from December 2019 until now the number of cases of Covid-19 worldwide is still relatively high. Even in July 2021, Indonesia had experienced the peak phase of the number of cases of Covid-19. Although in some cases the incidence has decreased because of effectiveness of the vaccine (Side et al., 2021). This certainly has an impact on increasing public access to health service providers.

As part of the front line in dealing with Covid-19 cases, not a few nurses experience fatigue both physically and mentally. The high workload in handling Covid-19 cases and the use of level 3 Personal Protective Equipment (PPE) greatly affect the decline in body immunity, so the risk of contracting the virus increase.

In addition, in this era of COVID-19, new factors can greatly affect the workload of nurses (Lucchini et al., 2020a). COVID-19 patients require prophylactic measures to prevent or contain the spread of the virus to other patients: wearing protective clothing, special decontamination procedures, specially isolated areas where certain supplies are stored. All these actions increase the workload of nursing (Giuliani et al., 2018), not only for the time required for its implementation but also for its organization and management.

Several early reports identified a very high nursing workload in COVID-19 patients (Lucchini et al., 2020a); (Lucchini et al., 2020b) (Kiba, 1969). In addition to the severity of the illness, the workload of nurses increases because of the need to provide humanistic care in the absence of a family. The introduction of cell phone calls (Negro et al., 2020) also helps patients to reduce their

sense of isolation and keeps them and their relatives updated, about what is happening outside and within the ”hospital walls”. When people with COVID-19 enter the hospital, they completely disappear from the lives of their relatives.

The heavy workload of hospital nurses is a major problem for the health care system because nurses are the cornerstone of any health care system regardless of the country. (Barton, 2009) for example have observed that globally, it appears that nurses are progressively experiencing an increase in workloads than before because of factors including inadequate supply of nurses, increased demand for nurses, reductions in patient length of stay, staff reductions. and increased overtime. Even more worrying is the consequences of a high nurse workload or perceived workload. For example, studies have shown that heavy nursing workloads have a negative impact on patient safety (Hegney et al., 2003). Another study also found that heavy nursing workloads adversely affect nurse job satisfaction, contributing to high nurse turnover and nurse shortages (Duffield and O'Brien-Pallas, 2003). Furthermore, it has been found that higher patient acuity, work system factors and expectations also contribute to nurses' workload: nurses are expected to perform schedules including delivery and collection of trays of food; transporting patients; and other household tasks, all of which are non-professional or service ancillary tasks (Aiken et al., 2002).

Research on the relationship between workload and nurse performance has been widely carried out around the world. From a quantitative point of view, the large number of studies conducted on this topic increases the possibility of variations in the results or research conclusions. In fact, it is not uncommon for studies on the same topic to show conflicting results. This situation, of course, creates problems, especially in constructing a comprehensive theory or making it the basis for decision making. Meta-analysis could find trends in the magnitude of the observed effect in a set of quantitative studies and all of them are included in the same research problem Meta-analysis appears to address research problems in various fields, especially health. Various study findings that initially seemed contradictory and difficult to accumulate eventually became more integrative and systematic with meta-analysis.

Based on the explanation above, it can be concluded that the large number of studies that have been conducted on similar topics open space to draw more precise conclusions by integrating various study findings into a solid foundation for theory development as well as decision making and policy determination, so that these results provide an overview for health policy makers to evaluate the workload of nurses in pandemic era.

B. LITERATURE REVIEW

1. Meta-Analysis

Meta-analysis is one type of systematic review. According to (Xiao and Watson, 2019) systematic literature review means identifying, evaluating, and interpreting all existing relevant studies for a specific research question, or a particular topic area or phenomenon of interest to the researcher. (Gough et al., 2012) revealed several reasons for the need for a systematic review including: individual research may be wrong, either by changes or because of how the research was designed and carried out; individual research is likely to be of limited relevance because of its scope and context; a review provides a more comprehensive and robust picture based on multiple studies and settings than a single study; and findings from a review provide context for interpreting the results of the new primary study. The goal of the meta-analysis is to provide estimates of results that represent results at all research levels. An important feature of meta-analysis is the ability to incorporate information about the quality and reliability of primary studies by giving greater weight to larger and more well-reported studies (Mikolajewicz and Komarova, 2019).

There are at least 11 artifacts that can be used as criteria to understand why there are differences in research results on the same topic that need to be corrected (Dennis et al., 2012). Statistical artifacts are evidenced as the main cause of variation across individual studies in meta-analysis. The artifacts are (1) Sampling error; (2) Measurement error on the dependent variable; (3) Measurement error on the independent variable; (4) Dichotomy on the dependent variable; (5) Dichotomy on independent variables; (6) Variation of the range in the independent variable; (7) Artifacts of attrition; (8) Imperfect construct validity on the dependent variable; (9) Imperfection of construct validity on independent variables; (10) Reporting or transcriptional errors; and (11) Variance caused by external factors.

The steps that must be taken in the meta-analysis include the following procedures:

1. Identify and formulate research problems.
2. Collecting data through the selection of articles or research results that are relevant to the research problem.
3. Explanation and evaluation of data
4. Analysis and interpretation of the results of the analysis itself.

2. Nurse Workload

According to Health Law No. 36 of 2009 states that workload is the product of the number of jobs with time and the amount of work that must be hit by a position or organizational unit. The nurse's workload is all activities in the nursing service unit carried out by a nurse (Marquis and Huston, 2017). The high workload can result in poor communication between nurses and patients, failure of collaboration between nurses and doctors, nurses leaving and job dissatisfaction.

C. RESEARCH METHOD

This study includes an observational study to analyse the relationship between variables with a meta-analytical approach, recursive (cause and effect), with approach retrospective that aims to determine the effect of workload on nurse performance during the Covid-19 pandemic. This study uses secondary data obtained from relevant sources related to the workload of nurses and nurses' performance in various journals. The data collection method used in this research is the documentation method which will be combined with meta-analysis. The criteria for an article to meet the requirements for a meta-analysis in this study are:

1. Primary studies that have primary data or data obtained directly from research subjects through questionnaires, interviews, or observations regarding the variables of a study.
2. The primary study contains results in the form of informative data or statistical values, namely the characteristics of the subject, the number of n (samples), r (correlation coefficient), d (Somers D test), t (t-student), and f (regression coefficient) which is the result of the statistical test of the study. Search for literature studies is done by accessing online media articles, namely Google Scholar with search keywords are workload and nurse performance.
3. The limitation of the research year starts from January 2020 to September 2021, where this range is chosen to show the impact of the Covid-19 pandemic in the hospital work environment, especially for nurses. From the search results that have been carried out, 14 studies have been published from 2020 to 2021.

D. RESULTS AND DISCUSSION

1. Inclusion Criterion

The research used in this study contain sufficient statistical information to be used. The inclusion criteria and sample descriptions are presented in Table 1.

Table 1. Cataract maturity prediction system result

Study	Researcher	Sample
1	(Chustianengseh, 2020)	72
2	(Kusuma et al., 2021)	62
3	(Kusumaningrum, 2020)	59
4	(Rianto et al., 2020)	78
5	(Maudul & Riskiyani, 2020)	67
6	(Enrico, 2020)	73
7	(Dasrin et al., 2020)	487
8	(Pourteimour et al., 2021)	139
9	(Yamin et al., 2020)	103
10	(Russeng et al., 2020)	96
11	Yosiana et al., 2020)	113
12	Hermawati & Yosiana, 2021)	30
13	(Nurulwiyah, 2020)	125
14	(Suryani et al., 2021)	135

2. Meta-Analysis Procedures

Based on the data that has been obtained, the stages of analysis according to the guidelines using the meta-analysis technique are described as follows (Dennis et al., 2012).

1. Convert the value of F to the values of t, d , and r.
   Based on the 14 primary studies that were used as research data, there were eleven studies that needed to convert the t-value into the r-value, namely studies number 1, 2, 4, 5, 6, 7, 10, 11, 12, 13, and 14. The equation formula for converting the value

of t to the value of r is:

$$d = \frac{2t}{\sqrt{N}} \text{ or } d = \frac{2r}{\sqrt{(1-r)^2}} \tag{1}$$

$$r = \frac{t}{\sqrt{t^2 + (n-2)}} \tag{2}$$

$$r = \frac{d/2}{\sqrt{1 + \left(\frac{d}{2}\right)^2}} \tag{3}$$

$$t = \sqrt{F} \tag{4}$$

$$D = \frac{\sum W_i d_i}{\sum N_i} \tag{5}$$

Here is the result of the conversion value t to the value r:

Table 2. Result of conversion study into t and r value

Study	t	r
1	2.982	0.336
2	3.240	0.386
4	0.125	0.014
5	1.794	0.217
6	17.344	0.899
7	-3.104	-0.140
11	4.773	0.442
11	-2.471	-0.228
12	-2.471	-0.423
13	14.722	0.799
14	1.981	0.169

2. Bare bones meta analysis of sampling error correction.
   The steps in bare bone meta-analysis of sampling error are as follows:
   (a) Calculating the mean population correlation (r_{xy} or $\check{r}$ or σ_{xy})
   The formula for calculating the mean population correlation:

$$\rho_{xy}(\check{r}) = \frac{\sum N_i r_i}{\sum N_i} \tag{6}$$

The result of calculating the mean population correlation after correction is 0.334.

DOI: https://doi.org/10.30812/varian.v5i2.1657

Table 3. Sample error correction

Study	N	r_{xy}	$N \times r_{xy}$
1	72	0.336	24.173
2	62	0.386	23.925
3	59	0.282	16.638
4	78	0.014	1.118
6	67	0.217	14.553
7	73	0.899	65.661
8	100	0.140	13.957
9	139	0.057	7.923
10	103	0.287	29.561
11	96	0.442	42.401
12	113	0.228	25.803
13	30	0.423	12.693
14	125	0.799	99.840
15	135	0.169	22.855
Total	1252	4.679341	401.1002
Mean	89.429		0.320

(b) Calculating the population correlation variance (σ_r^2)
The formula for calculating the population correlation variance:

$$\sigma_r^2 = \frac{\sum N_i(r_i - r)^2}{\sum N_i} \tag{7}$$

The result of calculating the population correlation variance after correction is 0.064.

Table 4. Population correlation variance

Study	N_i	r_{xy}	$r_i - \check{r}$	$(r_i - \check{r})^2$	$N_i * (r_i - \check{r})^2$
1	72	0.336	0.001	0.000	0.000
2	62	0.386	0.052	0.003	0.165
3	59	0.282	-0.052	0.003	0.161
4	78	0.014	-0.320	0.102	7.982
6	67	0.217	-0.117	0.014	0.918
7	73	0.899	0.565	0.319	23.322
8	100	0.140	-0.195	0.038	3.790
9	139	0.057	-0.277	0.077	10.684
10	103	0.287	-0.047	0.002	0.230
11	96	0.442	0.107	0.012	1.108
12	113	0.228	-0.106	0.011	1.267
13	30	0.423	0.089	0.008	0.237
14	125	0.799	0.464	0.216	26.968
15	135	0.169	-0.165	0.027	3.673
Total	1252	4.679	3.33E-16		80.505
Mean	89.429	0.334	2.38E-17		0.064

(c) Calculating the sampling error variance (σ_ϵ^2)
Here is the formula to calculate the sample variance error:

$$\sigma_\epsilon^2 = \frac{(1 - \check{r}^2)^2}{\check{N} - 1} \tag{8}$$

Sampling error variants obtained by knowing the average correlation $(\check{r})$ and the mean number of samples $(\check{N})$, then:

$$\sigma_\epsilon^2 = \frac{(1 - 0.334^2)^2}{89.429 - 1} = 0.009$$

DOI: https://doi.org/10.30812/varian.v5i2.1657

(d) Estimation of the actual population correlation variance or true score (σ^2_{xy})
The formula for calculating the actual population correlation variance:

$$\sigma^2_{\rho_{xy}} = \sigma^2_r - \sigma^2_\epsilon \tag{9}$$

So, the results are:

$$\sigma^2_{\rho_{xy}} = 0.064 - 0.009 = 0.055$$

Based on the results of the actual correlation variance $(\sigma^2_{\rho_{xy}})$, it can be calculated the size of the Standard Deviation $(SD) = \sqrt{\sigma^2_{\rho_{xy}}} = \sqrt{0.055} = 0.235$

(e) Confidence interval (M_ρ)
Here is the formula to calculate the confidence interval:

$$\begin{aligned} M_\rho &= \check{r} \pm 1.96(SD) \\ M_\rho &= \check{r} \pm 1.96(0.235) \\ M_\rho &= 0.334 \pm 0.444 \end{aligned} \tag{10}$$

Based on a meta-analysis on the above calculation, it can be concluded that there is a relationship between workload and performance of nurses, with $\check{r} = 0.334$ are in the reception area of the 95% confidence interval (0.334 ± 0.444). When the r value is less than 2 SD, the relationship that occurs is negative, then $0.334 < 0.888$, meaning that there is a negative relationship between the two variables.

(f) Impact of sampling error
The error in sampling is obtained by using the formula:

$$\begin{aligned} 1 - \text{Reliability} &= \frac{\sigma^2_{\rho_{xy}}}{\sigma^2_r} \\ 1 - \text{Reliability} &= 1 - \frac{0.05}{0.064} = 1 - 0.861 = 0.139 \end{aligned}$$

The correlation reliability in this study is 0.861, so the impact of sampling error is 0.139. These results indicate that the error in sampling is 13.9%.

3. Correction in measurement error
Another artifact that is corrected after sampling error is to correct the measurement error (Dennis et al., 2012). To calculate the measurement error requires statistical data or information, namely the reliability value of the measuring instrument of the two variables used. Based on the 14 studies used in this meta-analysis, only 5 studies contained data on the reliability of the independent variable (r_{xx}) and the dependent variable (r_{yy}). Systematically, calculating measurement error correction is carried out through the following steps:

(a) The formula for calculating the combined mean:

$$\begin{aligned} \hat{A} &= Ave(a) \times Ave(b) \\ \hat{A} &= 0.877 \times 0.882 = 0.773 \end{aligned} \tag{11}$$

Description:

$$\begin{aligned} (a) &= \sqrt{r_{xx}} \\ (b) &= \sqrt{r_{yy}} \\ Ave(a) &= Average(a) \\ Ave(b) &= Average(b) \end{aligned}$$

(b) Calculating the measurement error correction in x and y, ie the real correlation of population (ρ).
The formula to calculate an estimate of the population correlation:

$$\begin{aligned} \rho = Ave(\rho_i) &= \frac{\check{r}}{A} \\ \rho = Ave(\rho_i) &= \frac{0.320}{0.773} = 0.414 \end{aligned} \tag{12}$$

Estimation of the population correlation after correction of measurement error (ρ) that is equal to 0.414

DOI: https://doi.org/10.30812/varian.v5i2.1657

(c) The sum of the squared coefficients of variation (V)
The formula for calculating the sum of the squared coefficients of variation:

$$\begin{aligned} V &= \frac{SD(a)^2}{Ave(a)^2} + \frac{SD(b)^2}{Ave(b)^2} \\ V &= \frac{0.083}{0.877} + \frac{0.062}{0.882} = 0.014 \end{aligned} \tag{13}$$

(d) Variant referring to artifact variation (σ_2^2)
The formula for calculating variance referring to artifact variation:

$$\begin{aligned} \sigma_2^2 &= \rho^2 A^2 V \\ \sigma_2^2 &= 0.773^2 \cdot 0.414^2 \cdot 0.014 = 0.0014 \end{aligned} \tag{14}$$

(e) Varian real or true score correlations (ρ)
The formula for calculating the correlation variants real or true score:

$$\begin{aligned} Var(\rho) &= \frac{\sigma_{\rho_{xy}}^2 - \rho^2 A^2 V}{A^2} \\ Var(\rho) &= \frac{0.009 - 0.0014}{0.773^2} = 0.013 \end{aligned} \tag{15}$$

Based on the results of these calculations, standard deviation (SD) can be calculated using the formula below:

$$SD = \sqrt{Var(\rho)} = \sqrt{0.013} = 0.112$$

The estimated population correlation after correction of measurement error (ρ) is 0.334 with an SD of 0.112.

(f) Confidence interval (M_ρ)
The formula for calculating the confidence interval:

$$\begin{aligned} M_\rho &= \check{r} \pm 1.96(SD) \\ M_\rho &= 0.334 \pm 1.96(0.112) \\ M_\rho &= 0.334 \pm 0.219 \end{aligned} \tag{16}$$

Based on the meta-analysis calculation above, it can be concluded that there is a relationship between workload and nurse performance, with $\check{r} = 0.334$ being in the 95% confidence interval acceptance area (0.334 ± 0.219). Because the value of is smaller than 2 SD, then the relationship that occurs is negative, then $0.334 < 0.528$, meaning that there is a negative relationship between the two variables.

(g) Impact of variation in reliability or measurement error
The formula for calculating the variation in reliability or measurement error:

$$\begin{aligned} Var\ Rel &= \frac{\rho^2 A^2 V}{A^2} \\ Var\ Rel &= \frac{0.0014}{0.773^2} \\ Var\ Rel &= 0.00239 \end{aligned} \tag{17}$$

This result then used as a percentage $0.00239 \times 100\% = 0.24\%$. Based on the calculation of the variation in reliability, the results of the measurement error correction in this meta-analysis study are 0.24%.

E. CONCLUSION AND SUGGESTION

Based on the results of this meta-analysis study, it can be concluded that there is a negative relationship between workload and nurse performance. This study cannot be separated from weaknesses, thus, the researcher hopes that further research on workload and performance of nurses can use a more homogeneous primary study covering the study sample population area. In addition, further research is also expected to use research articles that thoroughly discuss these two variables, especially regarding the performance of nurses.

DOI: https://doi.org/10.30812/varian.v5i2.1657

This study illustrates that the workload of nurses has a worrying impact on the handling of Covid-19. This needs to be taken seriously, because nurses are at the forefront of health services. If the nurse's performance has started to decline, the patient's handling will not be optimal and can increase the risk of death for the patient. In addition, the slow handling also has an impact on other aspects, such as negative economic growth and domestic security instability.

Further researchers are also advised not to use Google Scholar to search for primary studies online because the scope is too broad and results in many articles that do not meet the research criteria. In addition, it is also recommended to use a statistical program in analyzing research data to be more accurate in carrying out data analysis calculations

ACKNOWLEDGEMENT

Acknowledgments are addressed to the Bali International University who have provided support so that this research can be realized properly.

REFERENCES

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J., and Silber, J. H. (2002). Hospital Nurse Staffing and Patient Mortality, Nurse Burnout, and Job Dissatisfaction. *Jama*, 288(16):1987–1993.

Barton, A. (2009). Patient Safety and Quality: An Evidence-Based Handbook for Nurses. *Aorn Journal*, 90(4):601–602.

Dennis, A., Wixom, B. H., and Roth, R. M. (2012). Systems Analysis and Design 5th Edition.

Duffield, C. and O'Brien-Pallas, L. (2003). The Causes and Consequences of Nursing Shortages: a Helicopter View of The Research. *Australian Health Review*, 26(1):186–193.

Giuliani, E., Lionte, G., Ferri, P., and Barbieri, A. (2018). The Burden of Not-Weighted Factors–Nursing Workload in a Medical Intensive Care Unit. *Intensive and Critical Care Nursing*, 47:98–101.

Gough, D., Oliver, S., and Thomas, J. (2012). *An Introduction to Systematic Reviews*. Sage.

Hegney, D., Plank, A., and Parker, V. (2003). Workplace Violence in Nursing in Queensland, Australia: A Self-Reported Study. *International journal of nursing practice*, 9(4):261–268.

Kiba, F. (1969). Nurse-Patient Relations. *[Kango gijutsu]:[Nursing technique]*, 34(4):326–337.

Lucchini, A., Giani, M., Elli, S., Villa, S., Rona, R., and Foti, G. (2020a). Nursing Activities Score is Increased in COVID-19 Patients. *Intensive & critical care nursing*, 59:102876.

Lucchini, A., Iozzo, P., and Bambi, S. (2020b). Nursing Workload in the COVID-19 Era. *Intensive & critical care nursing*, 61:102929.

Marquis, B. and Huston, C. (2017). Socializing and Educating Staff in a Learning Organization. In *Leadership roles and management functions in nursing*. Wolters Kluwer, Philadelphia.

Mikolajewicz, N. and Komarova, S. V. (2019). Meta-Analytic Methodology for Basic Research: a Practical Guide. *Frontiers in physiology*, 10:203.

Negro, A., Mucci, M., Beccaria, P., Borghi, G., Capocasa, T., Cardinali, M., Pasculli, N., Ranzani, R., Villa, G., and Zangrillo, A. (2020). Introducing the Video Call to Facilitate the Communication Between Health Care Providers and Families of Ppatients in the Intensive Care Unit During COVID-19 Pandemia. *Intensive & critical care nursing*, 60:102893.

Side, S., Hulinggi, P. K. M., Syam, H. K., Irfan, M., and Taufik, A. G. P. (2021). The Effectiveness of Vaccination Against The Spread of COVID-19 with SEIR Mathematical Modeling in Gowa District. *Jurnal Varian*, 5(1):17–28.

Xiao, Y. and Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, 39(1):93–112.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 115–124
e-ISSN: 2581-2017

Determinants of Leprosy Prevalence in Sulawesi Island Using Spatial Error Model

Geraldi Putra P Balebu[1], Siskarossa Ika Oktora[2]
[1,2]Departement of Applied Statistics, Politeknik Statistika STIS, Indonesia

Article Info

Article history:

Received : 12-25-2021
Revised : 04-09-2022
Accepted : 04-13-2022

Keywords:

Spatial Analysis;
Leprosy Prevalence;
Sulawesi Island;
Spatial Error Model;
Spatial Correlation.

ABSTRACT

Leprosy is one of the infectious diseases and has become a serious health problem in Indonesia. There are still many areas in Indonesia that have not met the leprosy elimination status. One of them is Sulawesi Island. Leprosy can spread across regions. The incidence of leprosy in an area can affect the condition of leprosy in other areas. Therefore, spatial regression is used to analyze the determinants of leprosy prevalence in Sulawesi Island. This study used data from Health Profile and Province in Figure publications with an analysis unit consisting of 81 districts in Sulawesi Island. The results show a spatial effect on leprosy prevalence exists in Sulawesi Island. Queen contiguity-based spatial weights are also considered while performing the spatial analysis. Using Spatial Error Models, the results show that population density, the number of multibacillary (MB) leprosy cases, and spatial effect significantly affect the leprosy prevalence. In contrast, a clean and healthy lifestyle, proper water access, and proper sanitation access do not significantly affect the leprosy prevalence. Because the spatial effect on the leprosy prevalence exists between districts in Sulawesi Island, so each local government should collaborate to reduce the prevalence of leprosy.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1632

Corresponding Author:

Siskarossa Ika Oktora,
Department of Applied Statistics, Politeknik Statistika STIS
Email: siskarossa@stis.ac.id

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

This Leprosy is an infectious disease and a serious health problem because of the disability it can cause. According to the *World Report on Disability* from World Health Organization (WHO), leprosy is one of the main causes of disability. It can be transmitted through respiration or contact with the patient. This causes leprosy to spread easily and generally occurs in developing countries, including Indonesia. The number of leprosy cases in Indonesia is also still fluctuating every year. Based on *Profil Kesehatan Indonesia*, the number of leprosy cases in Indonesia from 2013 to 2017 is still fluctuating, peaking at 17.202 cases in 2015. Meanwhile, in 2017 the number of leprosy cases in Indonesia has decreased significantly compared to previous years. However, based on data from WHO, Indonesia is the third country with the most leprosy case in the world, contributing roughly about 7.72% of all leprosy cases in 2017. In addition, based on the publication *Profil Kesehatan Indonesia* in 2017, cases of leprosy patients with level 2 disability is 4.26 per 1,000,000 population (Kemenkes RI, 2017). Some cases and relatively high disability shows that leprosy eradication in Indonesia has not been fully achieved. Therefore it is necessary to provide a counter-measures to help eliminate leprosy cases in Indonesia.

Indonesia has achieved the status of leprosy elimination. Elimination of leprosy in *Profil Kesehatan Indonesia* and WHO is defined as an area with a leprosy prevalence < 1 per 10,000 population. However, on a smaller scale, many areas have not reached leprosy elimination status, including Sulawesi Island. In 2017, the condition of leprosy prevalence on Sulawesi Island was still quite high. None of the regions on the Sulawesi Island have achieved the leprosy elimination target. The highest leprosy prevalence is

in North Sulawesi (leprosy prevalence: 2.04), and the lowest is in Central Sulawesi (leprosy prevalence: 1.1). In addition, the total cases of leprosy in Sulawesi Island, which reached 2,633 people, contributed about 16.5% of the total cases in Indonesia (Kemenkes RI, 2017). These findings show that the condition of leprosy on Sulawesi Island needs attention and studied further.

The incidence of leprosy in one area can spread to other surrounding areas. This is supported by the nature of leprosy, which is an infectious disease and can occur across regions. This situation also needs attention because there are indications that the incidence of leprosy in an area can affect the condition of leprosy in other areas. Therefore, spatial regression, particularly spatial error model was used to analyze the leprosy prevalence in areas of Sulawesi Island because it is considered capable of determining the factors that affect the leprosy prevalence, as well as taking into account the appropriate spatial effects that match the analysis units in this study.

Many studies research leprosy, but only a few focus on eastern Indonesia, such as Sulawesi. Using Spatial Durbin Model (SDM), (Ernawati et al., 2016) find that percentage of households that practice a clean and healthy lifestyle, population density, percentage of poor population, and percentage of public health centers per 100,000 population significantly affected leprosy prevalence in East Java. (Shovalina and Atok, 2016) use Geographically Weighted Regression (GWR) and found that every district have different significant variables, including the percentage of households that have non-brick walls, population density, poor population percentage, and the percentage of households that practice a clean and healthy lifestyle. (Emerson et al., 2020) also researched leprosy in 2020 by conducting a case-control study in North Gondar, Ethiopia. It is found that the Water, Sanitation, and Hygiene (WASH) factor that was significantly associated with leprosy is open sewage, absence of access to soap, access to water, hand washing practices, and water sources. Spatial analysis also has been conducted by (Pratiwi et al., 2020) and (Pratiwi et al., 2018). Using spatial econometric and spatial Durbin model, each research models economic growth, poverty, and unemployment.

The studies conducted by (Dzikrina and Purnami, 2013); (Ernawati et al., 2016); and (Prakoeswa et al., 2020) used a spatial analysis approach to study the effect of regional aspects on cases of leprosy. However, among studies on leprosy, the focus of research on the Sulawesi Island area has never been generated before. Based on the problem, it is necessary to study determinants of the leprosy prevalence on Sulawesi Island. Because leprosy is an infectious disease that can occur across regions, the analysis also needs to be studied spatially to determine if spatial aspects affect leprosy in Sulawesi Island in 2017. This study specifically uses the social condition variables that indicate as determinants of leprosy prevalence in Sulawesi Island, such as population density, the number of multibacillary (MB) leprosy cases, a clean and healthy lifestyle, proper water access, and proper sanitation access. Among similar studies on leprosy, the focus of research on the Sulawesi Island area has never been generated before.

Based on the problem, it is necessary to study determinants of the leprosy prevalence on Sulawesi Island in 2017. The main objective in this study is to find the factors that affecting the leprosy prevalence, while also finding the spatial effect on leprosy prevalence. Because leprosy is an infectious disease that can occur across regions, the analysis also needs to be studied spatially to determine if spatial aspects affect leprosy on Sulawesi Island in 2017.

B. LITERATURE REVIEW

This research uses cross-section spatial data which is used by (Bivand et al., 2021) and be written as follows:

$$\boldsymbol{y} = \rho \boldsymbol{W} y + \boldsymbol{X}\boldsymbol{\beta} + u \tag{1}$$

with:

$$\boldsymbol{u} = \lambda \boldsymbol{W} u + \boldsymbol{\varepsilon} \tag{2}$$

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I) \tag{3}$$

with:

$\boldsymbol{y}$ = vector of response variable
$\boldsymbol{X}$ = matrix of predictor variable
$\boldsymbol{\beta}$ = regression coefficient
ρ = parameter coefficient of Spatial Lag
λ = parameter coefficient of Spatial Error
$\boldsymbol{u}$ = error vector of standard model
$\boldsymbol{\varepsilon}$ = error vector of Spatial Error model
$\boldsymbol{W}$ = spatial weighting matrix

There are several possible models that can be formed from the General Spatial Model in Equations 1-3, including:

1. If $\rho = 0$ and $\lambda = 0$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4}$$

The model in Equation 4 is known as the classical regression model or commonly referred to as the Ordinary Least Square (OLS) regression model. OLS regression has no spatial effect in it.

2. If $\rho \neq 0$ and $\lambda = 0$

$$\boldsymbol{y} = \rho \boldsymbol{W} y + \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{5}$$

The equation is referred to Spatial Lag Model or Spatial Auto-Regressive Model (SAR). This model assumes a spatial lag effect on the dependent variable between regions, but no spatial effect on the error.

3. If $\rho = 0$ and $\lambda \neq 0$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \lambda \boldsymbol{W u} + \boldsymbol{\varepsilon} \tag{6}$$

This equation is the Spatial Error Model (SEM) regression. In this equation, the effect of spatial lag on the dependent variable does not exist, but there is a spatial effect in the error between one region and another.

1. Moran's Error Test

Moran's I statistic (error) was developed to capture the spatial dependencies between observations. Moran's I statistics (error) is used as an index to identify the distribution of observations in each location, whether clustered, random, or uniform (dispersion) pattern. The null hypothesis is there is no spatial dependency in error $(I_e = 0)$, and the statistic test:

$$Z_{\text{score}} = \frac{I_e - E(I_e)}{\sqrt{var(I_e)}} \sim N(0,1) \tag{7}$$

$$I_e = \frac{\boldsymbol{\varepsilon' W \varepsilon}}{\boldsymbol{\varepsilon' \varepsilon}} \tag{8}$$

$$E(I_e) = \frac{tr(\boldsymbol{MW})}{n-k} \tag{9}$$

$$Var(I_e) = \frac{tr(\boldsymbol{MWMW'}) + tr(\boldsymbol{MWMW}) + [tr(\boldsymbol{MW})]^2}{(n-k)(n-K+2)} - [E(I_e)]^2 \tag{10}$$

$$\boldsymbol{M} = \boldsymbol{I - X(X'X)^{-1}X'} \tag{11}$$

where:

$\boldsymbol{M}$ = projection matrix
n = number of observations
k = number of parameters
$\boldsymbol{W}$ = spatial weighting matrix

The null hypothesis is rejected when $|Z_{\text{score}}| > Z_{\frac{\alpha}{2}}$ or $p - value < \alpha$.

2. Lagrange Multiplier Test (LM Test)

Lagrange Multiplier (LM) test is used to test spatial dependencies. LM Test performs tests on the lag coefficient and spatial error, which aims to determine the right model to be used in spatial regression analysis. The null hypothesis of LM test for the lag coefficient is no spatial lag effect on the dependent variable, and the statistic test:

$$LM_\rho = \frac{\left[\frac{\boldsymbol{\varepsilon' WY}}{\sigma^2}\right]^2}{\frac{\boldsymbol{B}}{\sigma^2}} \sim \chi^2_{(1)} \tag{12}$$

$$B = \left[(\boldsymbol{W} \times \boldsymbol{\beta})' \boldsymbol{M} (\boldsymbol{W} \times \boldsymbol{\beta}) + \boldsymbol{T}\sigma^2\right] \tag{13}$$

$$\boldsymbol{T} = tr[(\boldsymbol{W'} + \boldsymbol{W})\boldsymbol{W}] \tag{14}$$

DOI: https://doi.org/10.30812/varian.v5i2.1632

where:

$\boldsymbol{B}$ = projection matrix
$\boldsymbol{T}$ = trace matrix
λ = parameter coefficient of$Spatial\ Error$
σ^2 = variance of regression model

While the null hypothesis of LM test for error is no spatial in error, and the statistic test:

$$LM_\lambda = \frac{\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{W}\boldsymbol{\varepsilon}}{\sigma^2}\right]^2}{\boldsymbol{T}} \sim \chi^2_{(1)} \tag{15}$$

Both tests follow the Chi-Square distribution with a degree of freedom of one. If the LM statistic is greater than the critical value of Chi-Square, then H_0 is rejected.

3. Spatial Autoregressive Model (SAR)

According to (Ver Hoef et al., 2018), to test the significance of the spatial lag coefficient (ρ), the Likelihood Ratio Test (LRT) is used. The null hypothesis is spatial lag coefficient is not significant, with with a statistical test:

$$LR = \left\{-2\ln|\boldsymbol{I}-\rho\boldsymbol{W}| + \frac{1}{\sigma^2}\left[(\boldsymbol{I}-\rho\boldsymbol{W})\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right]^T\left[(\boldsymbol{I}-\rho\boldsymbol{W})\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right] - \frac{1}{\sigma^2}\left[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right]^T\left[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right]\right\} \tag{16}$$

The null hypothesis is rejected if the LR value is greater than $\chi^2_{1-\alpha(1)}$.

4. Spatial Error Model (SEM)

The Likelihood Ratio Test is used to check for spatial dependencies in error. The null hypothesis is spatial Error coefficient is not significant, with a statistical test:

$$\begin{aligned} LR = -2\Bigg\{&-\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\left|(\boldsymbol{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{I}-\boldsymbol{\beta})^T\right| - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\left[(\boldsymbol{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{I}-\boldsymbol{\beta})^T\right]^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &+\frac{n}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\Bigg\} \end{aligned} \tag{17}$$

The null hypothesis is rejected if the LR value is greater than $\chi^2_{1-\alpha(1)}$.

5. Wald test

Wald test is used to test the significant effect of each independent variable in the spatial model. The null hypothesis is k-th independent variable has no significant effect on the dependent variable, with statistic test:

$$Wald = \left(\frac{\widehat{\beta}_k}{se(\widehat{\beta}_k)}\right)^2 \sim \chi^2_1 \tag{18}$$

Null Hypoheses is rejected if Wald $> \chi^2_{\alpha,1}$

C. RESEARCH METHOD

This study examines the leprosy prevalence on Sulawesi Island in 2017. The analysis units of this study are all districts in Sulawesi Island, which consists of 81 districts. The study uses secondary data from the publication of Health Profiles of each province on Sulawesi Island in 2017, as well as the publication of Provinces in Figures of each province on Sulawesi Island in 2018. The dependent variable is the leprosy prevalence in each district. The independent variables are population density (”Dense”), the percentage of household who practice clean and healthy lifestyle (”PHBS”), the percentage of household who has good sanitation (”Sanitation)” , the percentage of household who access safe water (”Water)” , and the number of MB leprosy (”LepraMB”). These variables were formulated based on previous studies about leprosy.

This study uses spatial analysis with the analytical steps carried out are as follows:

DOI: https://doi.org/10.30812/varian.v5i2.1632

1. *Generating multiple linear regression model*
   After estimating the parameters, classical assumptions are tested. The classical assumption tests that must be fulfilled:
   (a) Normality error
   This test is carried out using the Kolmogorov-Smirnov statistic test. Errors are normally distributed when the Kolmogorov-Smirnov statistics is greater than the value of the Kolmogorov Smirnov table or when the $p-value > 0.05$.
   (b) Non-multicollinearity of independent variables
   Multicollinearity means a high relationship (correlation) between independent variables in the regression. One way to find out the existence of multicollinearity is to check the value of Variance Inflation Factor (VIF_k). If there is an independent variable that has a value of $VIF_k > 10$, then it indicates the existence of multicollinearity.
   (c) Homoscedasticity
   The assumption of homoscedasticity can be tested using the Breusch-Pagan statistical test. When the Breusch-Pagan value is greater than the Chi-Square table value (χ^2) or when value $p-value < 0.05$, then the assumption is violated.
2. Identification of spatial autocorrelation
   This is carried out to identify whether the data has spatial autocorrelation. This test uses Moran's I to test globally and Moran's Scatterplot to test locally as written in Equation 7.
3. Spatial dependency diagnosis
   This is carried out to determine the spatial regression model. The test statistic used is the Lagrange Multiplier for both LM-error and LM-lag as written in Equations 12 and 15.
4. Likelihood Ratio (LR) test
   This is performed to decide if the spatial regression model formed more suitable than the multiple linear regression modelthe decision to reject H_0 when $p-value < 0.05$.
5. Constructing the spatial regression model. Formed as follows:

$$y_i = \rho \sum_{j=1,i\neq j}^{n} w_{ij}y_j + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \lambda \sum_{j=1,i\neq j}^{n} w_{ij}u_j + \varepsilon_i \quad (19)$$

6. Conducting a Wald test
   Carried out to determine which independent variables significantly affect the dependent variable. When the $p-value < 0.05$, H_0 is rejected, which shows that the independent variable j significantly affects leprosy prevalence.

D. RESULTS AND DISCUSSION

1. Overview of The Leprosy Prevalence on the Sulawesi Island

The leprosy prevalence on Sulawesi Island spans from 0 to 8.49 per 10,000 population, with an average of 1.34. The area with the highest leprosy prevalence is Siau Tagulandang Biaro Islands (8.48) in North Sulawesi, while areas with the lowest prevalence consist of 3 regions, namely Palopo City (0), Mamasa District (0), and Mamuju District (0). Figure 1 shows that most of the areas have a higher prevalence compared to both Indonesia prevalence (0.7) and leprosy elimination criteria (1.0). This condition is also in line with the publication from Kemenkes RI, where almost all provinces are high endemic for leprosy concentrated in the eastern part of Indonesia, including the island of Sulawesi (Kemenkes RI, 2018). In a study conducted by (Kansil, 2014), it was found that the people in Tagulandang Biaro District, which is the district with the highest leprosy prevalence on the island Sulawesi, have limited access to health facilities (Kansil, 2014). This results in many leprosy patients that cannot get adequate treatment and are at great risk of transmitting the disease to other people. In addition, based on the report of Dinas Kesehatan Sulawesi Selatan, it is also known that in Makassar city, high cases of leprosy occur because it is influenced by the amount of negative stigma attached to people who contracted leprosy, thus making the patients more reluctant to seek for some treatment (DinkesSulsesl, 2018). This impacts the increase of leprosy and the leprosy prevalence in these areas.

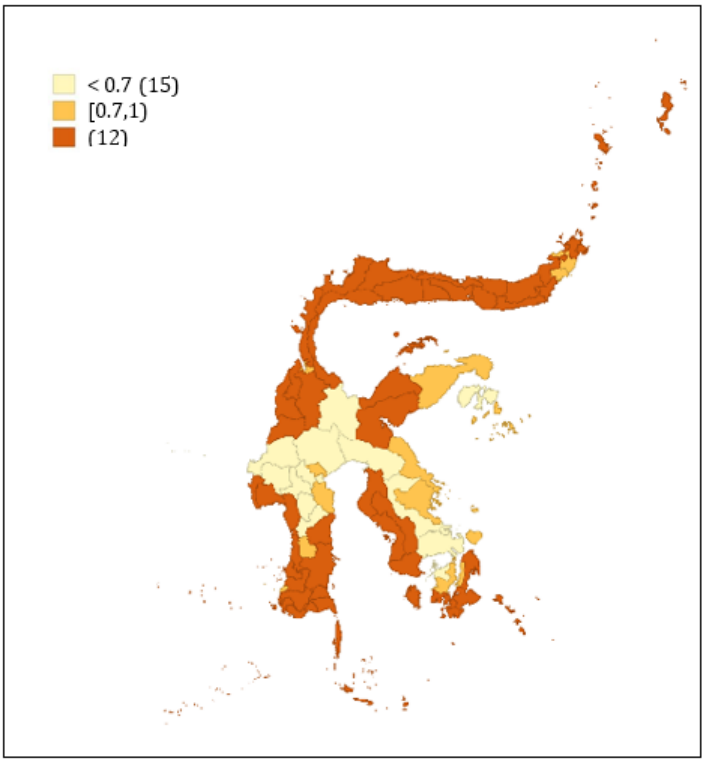


Figure 1. The leprosy prevalence per district in Sulawesi Island in 2017

2. Leprosy Prevalence Modeling with Ordinary Least Square Method

Table 2 shows the results of the partial parameter test on the linear regression model. Out of the five independent variables used, two significant variables are affecting the leprosy prevalence on Sulawesi Island in 2017, namely population density and the number of MB leprosy cases. Based on Kolmogorov-Smirnov test, the p-value is less than 0.05. This means that the error in the OLS model is not normally distributed. Abnormalities in error can be caused by several things, one of which is extreme values in a set of data that will produce a skewness distribution. This condition is also known as outliers. On the other hand, according to (Bivand et al., 2013), on a dataset that has observations in the form of islands which is an isolated area, the countermeasures that can be taken care of by creating a new subset of the data that excludes these isolated observations. Based on this basis, a test is carried out to see which data is an outlier in datasets. The test is carried out in the form of a univariate graphical test using boxplots graphs. Based on the boxplots, it is found that the outlier is observation 81, namely of the Siau Tagulandang Biaro Islands in North Sulawesi. This district has the highest leprosy prevalence in the Island of Sulawesi, which is 8.49. Also, this district is an observation with no neighboring areas. Therefore, this district will be excluded from the dataset. After outliers are removed, the data is re-tested for normality assumptions, with the results showing the normality assumption is fulfilled. The next test is the homoscedasticity test using the Breusch-Pagan test. The p-value of the Breusch-Pagan statistic is 0.1492 and greater than 0.05. It means the error variance is constant. Based on the non-multicollinearity test, it is found that there is no indication of multicollinearity in the model.

3. Spatial Autocorrelation and Spatial Dependencies

The Global Moran's Index is 0.325. A positive value indicates a positive spatial autocorrelation. The p-value is less than 0.05, so it means that there is a significant spatial autocorrelation of the prevalence of leprosy.

Two models can be formed in spatial regression, namely the spatial lag model and the spatial error model. Model selection is made by testing the Lagrange Multiplier (LM).

Table 1. Summary of Lagrange Multiplier test

Test	Queen Contiguity	
	Value	$p-value$
Moran's I (error)	3.5368	0.0004
LM error	10.1980	0.0014
LM lag	1.9161	0.1663
Robust LM error	11.0350	0.0009
Robust LM lag	2.7533	0.0971
SARMA	12.9510	0.0015

Table 1 shows that the appropriate spatial modeling in this study is the spatial error model. It is because the LM-error test results are significant, while the LM-lag test results are not significant. It means that the value of the leprosy prevalence variable in a district is influenced by the error value of its neighboring districts. Thus, it is necessary to establish a spatial error model. Queen contiguity is used in this model while considering the geographical conditions of regencies/cities on Sulawesi Island, which have many intersections with the surrounding area. It is also chosen based on a preliminary test comparing Morans Index of Queen and Rook contiguity. The Queen contiguity shows a more significant result.

4. Modeling the Leprosy Prevalence using Spatial Error Model

Based on the regression results, two variables are significant in the model, namely population density and cases of MB leprosy. In addition, the value of lambda of 0.5129 also has a p-value of less than 5% alpha. The results of this spatial error model indicate a spatial dependence on the coefficient lambda or u_i , because it is significant and has a positive sign. This means that there is a relationship between leprosy prevalence in a district with other districts, where the leprosy prevalence in a region is affected by errors from other neighboring regions

Table 2. Summary of Model Estimation

Independent Variables	Estimate	$z-value$	$p-value$
Intercept	0.851	21.314	0.03306
Density	-0.000212	-2.45239	0.01419*
PHBS	-0.000523	-0.103035	0.91793
Sanitation	0.00689	1.32023	0.18676
Water	0.000497	0.130819	0.89592
LepraMB	0.0193	54.601	0.0000*
λ	0.5129	46.905	0.0000*

$$\begin{aligned}\text{LeproPrev}_i =& 0.8510 - 0.0002\text{Dense}_i^* - 0.0005\text{PHBS}_i + 0.0069\text{Sanitation}_i + 0.0005\text{Water}_i \\ &+ 0.0193\text{LepraMB}_i^* + u_i^*\end{aligned} \tag{20}$$

with

$$u_i = 0.5129 \sum_{i=1, i\neq j}^{n} w_i u_j$$

*significant at $\alpha = 5\%$

Based on equation 20 of the spatial error model formed, it can be interpreted that the leprosy prevalence in i-th district will decrease by 0.0002 percent when population density increases by 1, and other variables are held constant. The population density variable is significant, as in the study conducted by (Ernawati et al., 2016), where population density is also influenced by regional proximity, significantly affecting leprosy prevalence in East Java Province. This result is contrary to the research results by (Kurniawan et al., 2018), where the incidence of leprosy in Blora district is not significantly affected by population density. Meanwhile, in the research by (Shovalina and Atok, 2016), population density also significantly affects the leprosy prevalence in the Province of East Java. These findings are also similar to research by (Franco-Paredes and Rodriguez-Morales, 2016), where congested and unstructured housing can intensify the risk of leprosy transmission. A similar thing was also stated by (Setiani and Patmawati, 2015), where a large residential density will facilitate the transmission of leprosy to other people. High population

DOI: https://doi.org/10.30812/varian.v5i2.1632

density is generally found in urban areas. A denser area means that the space for a population to move will be getting smaller. This small space of movement makes the possibility of spreading infectious diseases such as leprosy higher.

However, there is an anomaly where normally the population density should be proportionally linear with the leprosy prevalence but was inversely proportional. This case can be explained when we look at the condition between population density and leprosy prevalence in each district on the island of Sulawesi. It is known that the district with high population density is centered on big cities only, while other districts tend to have a low population density. When compared to the condition leprosy prevalence, it was found that there was a tendency for a large leprosy prevalence in areas with a lower population density than in urban a areas with a high population density. Prevalence conditions that tend to be smaller in urban areas align with the adequate health facilities network in big cities, such as in Central Sulawesi. The most widely available health services and facilities are in Palu city, which includes the health office or clinics, as well as a much larger number of individual physician practices if compared to other surrounding areas. More and adequate health facilities enable patients to undergo treatment, as well as increase the chances of the activity of finding people with leprosy to be carried out by leprosy officers at the health center, which supports these activities (DinkesSulsesl, 2017).

The leprosy prevalence in the i-th district will increase by 0.0193 when the total number of cases of leprosy type MB increases by 1 unit, and other variables are held constant. This finding is also similar to (Dzikrina and Purnami, 2013) research which reveals that the cases of leprosy multibacillary type affect the leprosy prevalence in East Java Province. (Zuhdan et al., 2017) found that living at home with non-lepromatous leprosy patients will increase the risk of developing leprosy by 9.5 times. This is also similar to the research of (Barreto et al., 2014) where the closer the relationship and family interaction with people who have leprosy, the higher the risk for leprosy to be transmitted. This also applies when the distance from where someone lives is getting smaller or is next to leprosy patients and increases the risk of contracting leprosy. In addition, if we look at the descriptive discussion, it is known that there is a tendency for people not to treat leprosy caused by several things, such as the absence of facilities and the amount of negative stigma from the community towards people with leprosy. These things make patients reluctant or ashamed to seek treatment, thus making leprosy suffered more severe, even lead to disability. The same condition is also proposed by (Sari et al., 2018), where people with leprosy tend to be embarrassed to take treatment because of the stigmatization of people with leprosy. This condition makes it easier for other people around the patients to be infected by leprosy.

The leprosy prevalence in the i-th district will increase by 0.5129 times spatial weighting if there is an increase in the average error in neighboring districts by 1 unit and other variables are considered constant. The average of other indicators that are not known or not included in the model in all regions that are considered neighbors will increase the prevalence of leprosy by 0.5129 units. A significant error coefficient indicates that there are other variables other than those used in this study which also affects the leprosy prevalence on the island of Sulawesi.

In related studies, it was found that the variable percentage of the population that performs a clean and healthy lifestyle has a significant effect on leprosy prevalence. In research conducted by (Dzikrina and Purnami, 2013), the percentage of households that perform a clean and healthy lifestyle significantly affects leprosy prevalence in East Java. In addition, based on research conducted by (Pramesti et al., 2020) in East Java, the clean and healthy lifestyle variable also affects leprosy prevalence. Meanwhile, in the research of (Prakoeswa et al., 2020), it was found that there is a relationship between a house's physical environment, waste disposal facilities, and personal hygiene, which are components of a clean and healthy lifestyle in women with leprosy in Gresik district. Research with similar results was also conducted by (Aprizal et al., 2017), where the cases of leprosy in Pekalongan District are influenced by the condition of the physical environment of the house, where bad environmental conditions make a person's risk of contracting leprosy become greater. On the other hand, based on the modeling results in this study, it is known that a clean and healthy lifestyle has no significant effect on leprosy prevalence. This condition is similar to (Pertiwi et al., 2020) research, where the percentage of households with clean and healthy lifestyles does not significantly affect the number of cases of leprosy. This is thought to be related to the clean and healthy lifestyle condition, which, although not evenly distributed, is, in general, have had a good percentage. This condition is supported by the Indonesian health performance promotion report, where the achievement of district indicators that already have clean and healthy lifestyle policies in 2017 reached 60.89%. The negative sign on the clean and healthy lifestyle coefficient also indicates that when the percentage of clean and healthy lifestyle increases, the leprosy prevalence decreases, but is not significant due to the uneven distribution of clean and healthy lifestyle achievements.

Sanitation is one of the variables that does not significantly affect leprosy prevalence. On the contrary, in a study conducted by (Sabil et al., 2018), it was found that that the percentage of the population with access to proper sanitation has a significant effect on the leprosy prevalence in South Sulawesi Province. In addition, the research conducted by (Rismawati, 2013), finds that

house sanitation has a significant effect on the incidence of multibacillary leprosy at the leprosy polyclinic of Tugurejo Hospital Semarang. This condition is also similar to the research of Emerson et al., where low sanitation is closely related to leprosy infection in the Ethiopian region (Emerson et al., 2020). This variable is not significant, allegedly influenced by the distribution of access to proper sanitation on the Sulawesi Island. The highest percentage of access to proper sanitation in Sulawesi Island is in South Sulawesi (62.84%), which is still far from the target of Rencana Pembangunan Jangka Menengah Nasional (RPJMN). Other provinces are also still below this value.

Access to proper water is a variable that has no significant effect on the leprosy prevalence on the island of Sulawesi. This result is contrary to the research done by (Prakoeswa et al., 2020), where clean water facilities correlate with the incidence of leprosy in women in Gresik District. The same thing also occurred in the study by Emerson et al., where the absence of access to clean water correlated with the incidence of leprosy in the Ethiopian region (Emerson et al., 2020). The insignificance of this variable is allegedly caused by the conditions of the percentage of access to proper water, which is not evenly distributed in all areas on the island of Sulawesi. The achievement of households with access to proper water in each province on the island of Sulawesi is still too far from the target in the RPJMN, which is 100%.

E. CONCLUSION AND SUGGESTION

The spatial effect on the leprosy prevalence exists between districts in Sulawesi Island in 2017. This indicates that the leprosy prevalence in an area will impact the leprosy prevalence in other surrounding areas, especially by the error of the leprosy prevalence from the surrounding area. In addition, based on modeling using SEM, it is found that population density and the number of MB leprosy significantly affect the leprosy prevalence.

In dealing with leprosy, the government is expected to consider that leprosy is an infectious disease that can spread within and between regions, particularly adjacent or neighboring areas. This way, each local government should collaborate to reduce the prevalence of leprosy, so the policies. will be more sustainable and tackle leprosy. For future research, it is highly recommended to account more variation of available variables, while also considering a different method of spatial approach.

ACKNOWLEDGEMENT

The authors thank Politeknik Statistika STIS for the support

REFERENCES

Aprizal, A., Lazuardi, L., and Soebono, H. (2017). Faktor Risiko Kejadian Kusta di Kabupaten Lamongan. *Berita Kedokteran Masyarakat*, 33(9):427–432.

Barreto, J. G., Bisanzio, D., Guimaraes, L. d. S., Spencer, J. S., Vazquez-Prokopec, G. M., Kitron, U., and Salgado, C. G. (2014). Spatial Analysis Spotlighting Early Childhood Leprosy Transmission in a Hyperendemic Municipality of the Brazilian Amazon Region. *PLoS neglected tropical diseases*, 8(2):e2665.

Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the J acobian in G aussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, 45(2):150–179.

Bivand, R., Millo, G., and Piras, G. (2021). A Review of Software for Spatial Econometrics in R. *Mathematics*, 9(11):1276.

DinkesSulsesl (2017). *Profil Kesehatan Provinsi Sulawesi Tengah Tahun 2017*. https://dinkes.sultengprov.go.id/.

DinkesSulsesl (2018). *Rencana Kerja Tahun 2018 Dinaas Kesehatan*.

Dzikrina, A. M. and Purnami, S. W. (2013). Pemodelan Angka Prevalensi Kusta dan Faktor-Faktor yang Mempengaruhi di Jawa Timur dengan Pendekatan Geographically Weighted Regression (GWR). *Jurnal Sains dan Seni ITS*, 2(2):275–281.

Emerson, L. E., Anantharam, P., Yehuala, F. M., Bilcha, K. D., Tesfaye, A. B., and Fairley, J. K. (2020). Poor WASH (Water, Sanitation, and Hygiene) Conditions are Associated with Leprosy in North Gondar, Ethiopia. *International journal of environmental research and public health*, 17(17):6061.

Ernawati, E., Latra, I. N., and Purhadi, P. (2016). Analisis Faktor-Faktor yang Memengaruhi Angka Prevalensi Penyakit Kusta di Jawa Timur dengan Pendekatan Spatial Durbin Model. *Jurnal Sains dan Seni ITS*, 5(2).

Franco-Paredes, C. and Rodriguez-Morales, A. J. (2016). Unsolved matters in leprosy: a Descriptive Review and Call for Further Research. *Annals of clinical microbiology and antimicrobials*, 15(1):1–10.

Kansil, O. M. (2014). Implementasi Kebijakan Pelayanan Kesehatan Bagi Penduduk Miskin di Puskesmas Ondong Kecamatan Siau Barat Kabupaten Siau Tagulandang Biaro. *JURNAL ADMINISTRASI PUBLIK*, 2(001).

Kemenkes RI (2017). *Profil Kesehatan Indonesia 2017*.

Kemenkes RI (2018). *Laporan Kinerja*.

Kurniawan, J., Radiono, S., and Kusnanto, H. (2018). Analisis spasial kejadian kusta di kabupaten Blora. *Berita Kedokteran Masyarakat*, 34(1):6–10.

Pertiwi, N. M. S., Sukarsa, I. K. G., and Susilawati, M. (2020). Pemodelan Jumlah Kasus Penyakit Kusta di Provinsi Jawa Timur. *E-Jurnal Matematika*, 9(1):42–50.

Prakoeswa, F. R. S., Ilhami, A. Z., Luthfia, R., Putri, A. S., Soebono, H., Husada, D., Notobroto, H. B., Listiawan, M. Y., Endaryanto, A., and Prakoeswa, C. R. S. (2020). Correlation Analysis Between Household Hygiene and Sanitation and Nutritional Status and Female Leprosy in Gresik Regency. *Dermatology Research and Practice*, 2020.

Pramesti, R. G., Ratna, M., and Budiantara, I. N. (2020). Pemodelan Faktor-Faktor yang Mempengaruhi Angka Prevalensi Kusta di Jawa Timur dengan Menggunakan Regresi Nonparametrik Spline Truncated. *Jurnal Sains dan Seni ITS*, 8(2):357–364.

Pratiwi, L. P. S., Hanief, S., and Suniantara, I. K. P. (2018). Pemodelan Menggunakan Metode Spasial Durbin Model untuk Data Angka Putus Sekolah Usia Pendidikan Dasar. *Jurnal Varian*, 2(1):8–18.

Pratiwi, L. P. S., Hendayanti, N. P. N., and Suniantara, I. K. P. (2020). Perbandingan pembobotan seemingly unrelated regression–spatial durbin model untuk faktor kemiskinan dan pengangguran. *Jurnal Varian*, 3(2):51–64.

Rismawati, D. (2013). Hubungan antara Sanitasi Rumah dan Personal Hygiene dengan Kejadian Kusta Multibasiler. *Unnes Journal of Public Health*, 2(1).

Sabil, R. M., Sastri, R., and Si, M. (2018). Analisis Spasial Determinan Prevalensi Kusta di Provinsi Sulawesi Selatan Tahun 2016. *Unnes Journal of Public Health*, pages 1–15.

Sari, D. A. K. W., Soewondo, S., and Supriati, L. (2018). Stigma Sosial Sebagai Indikator Penilaian Harga Diri pada Pasien Kusta di RS. Kusta Kediri. *Jurnal Penelitian Keperawatan*, 4(1).

Setiani, N. O. and Patmawati, P. (2015). Faktor Risiko Lingkungan dan Perilaku Penderita Kusta di Kabupaten Polewali Mandar. *Indonesian Bulletin of Health Research*, 43(3):20132.

Shovalina, M. R. and Atok, R. M. (2016). Pemodelan dan Pemetaan Prevalensi Kusta di Kabupaten/Kota Jawa Timur dengan Pendekatan Mixed Geographically Weighted Regression. *Jurnal Sains dan Seni ITS*, 5(2).

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018). Spatial Autoregressive Models for Statistical Inference from Ecological Data. *Ecological Monographs*, 88(1):36–59.

Zuhdan, E., Kabulrachman, K., and Hadisaputro, S. (2017). Faktor-Faktor yang Mempengaruhi Kejadian Kusta Pasca Kemoprofilaksis (Studi pada Kontak Penderita Kusta di Kabupaten Sampang). *Jurnal Epidemiologi Kesehatan Komunitas*, 2(2):89–98.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 125–136
e-ISSN: 2581-2017

Forecasting Stock Price PT. Indonesian Telecommunication with ARCH-GARCH Model

Wahidah Alwi[1], Aprilia Pratiwi S.[2], Ilham Syata[3]
[1,2,3]Departement of Mathematics, Universitas Islam Negeri Alauddin Makassar, Indonesia

Article Info

Article history:

Received : 11-11-2021
Revised : 04-21-2022
Accepted : 04-26-2022

Keywords:

Forecasting;
ARIMA;
ARCH-LM Test;
ARCH-GARCH;
Time Series.

ABSTRACT

This research discusses the modeling of time series using R software, focusing on forecasting the stock price of PT. Indonesian telecommunications with ARCH-GARCH model. The data used daily closing data on stock prices from January 6, 2020, to January 6, 2021 was obtained from the website www.finance.yahoo.com. The goal is to find out the best model arch-garch on PT. Indonesian telecommunications to find out the results of stock price forecasting the next day using the ARCH-GARCH model. The best model was ARIMA (2,1,3). The results of the ARCH-LM test showed the data contained heteroskedasticity effects or ARCH elements. The research models proposed in this study are ARCH (1) and ARCH-GARCH (1,1). The smallest AIC and BIC values of these two models are ARCH-GARCH (1,1) which is the best model for forecasting the stock price of PT. Indonesian telecommunications for the next 10 days. The study attempts to conduct stock price forecasting with the ARCH-GARCH model. The result of the forecasting of the share price of PT. Indonesian telecommunications from January 07, 2021 to January 20, 2021 respectively except for holidays is IDR 3374.884, IDR 3379.617,IDR 3378.305, IDR 3376.610, IDR 3380.050, IDR 3376.372, IDR 3379.071, IDR 3377.964, IDR 3377.515, IDR 3379.002. Forecasting results are close to factual data for forecasting the next 10 days so that they can be taken into consideration in investing by investors.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1543

Corresponding Author:

Aprilia Pratiwi S.,
Department of Mathematics, Universitas Islam Negeri Alauddin Makassar.
Email: apts.tiwy@gmail.com

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The position of supply and demand for shares in the Indonesian capital market makes shares have a selling price. The higher the demand, the stock price will increase, and the higher the supply the stock price will decrease. The movement of stock prices in Indonesia is volatile, meaning it can go up or down, therefore it is necessary to do modeling and predictions to determine conditions and prepare strategies to deal with stock price declines or spikes (Prasetya et al., 2020).

Information in finance such as stock indices is usually very random and has large volatility or variable error that is not constant (heteroscedasticity). The model that can be used to test the efficiency of the weak form of the capital market with heteroscedasticity is the ARCH-GARCH model. The ARCH (Autoregressive Conditional Heteroskedasticity) model was originally introduced by Engle which was developed to respond to the case of volatility in financial data. The basic concept of the ARCH model is the residual variance depending on the past square. Then this method was developed into GARCH (Generalized Autoregressive Conditional Heteroskedasticity) by Bollerslev, the use of the GARCH model on time series data experiencing heteroscedasticity is very useful in increasing efficiency because the dependence of most of the past volatility can be reduced. In this modeling, the residual variance does not only depend on the square of the past residual but also the variance of the past residual (Juliana et al., 2019).

Several studies that apply the ARCH-GARCH model include: (Bakhtiar et al., 2021) research uses the GJR-ARCH model in analyzing price data and measuring share price losses using AVAR as a measuring tool to assess the worst losses experienced by

investors. (Larasati et al., 2016) research using the GARCH model in predicting the price of nine staple foods to increase in 2015. In this study, the best GARCH model was used for analysis and research, and predicted the volatility of 9 staple foods or other studies that had a heteroscedasticity effect. (Arumningtyas et al., 2021) research analyzed using the ARIMAX-GARCH approach and using VaR to measure risk. The results showed that the major loss in investing using the ARIMAX-GARCH method was estimating VaR. (Angraeny, 2019) research uses the ARCH-GARCH model to analyze and forecast the value of Indonesia's exports. Using the help of the Eviews software. From the forecast results, it was found that the value of Indonesian exports increased and decreased from one month to the next. The research of (Bilondatu et al., 2019) with ARCH (1) and GARCH (1,1) models gives an indication that the forecasting results are close to factual data.

PT. Indonesian Telecommunications is the largest telecommunications company in Indonesia and is one of the largest state-owned enterprises owned by the government where all levels of society know it through its telecommunications products. PT. Indonesian Telecommunications is listed as a grade A stock on the Indonesian stock exchange, it is also included in the LQ45 list which contains the 45 most liquid on the stock. as a material consideration in investing by investors (Heriyanto, 2022). Increasing need for planning in business and economic activities, so that accurate predictions of future conditions continue to be a necessity. Modern time series analysis processes are based on statistical modeling and weight calculations that are possible with today's computers and software, such as the R program (Krispin, 2019). The growth of computing technology supports the development of various procedures and forecasting methods to predict future conditions that will answer these needs (Firdaus, 2020). Until the author is interested in doing Forecasting Stock Price PT. Indonesian Telecommunications With the ARCH-GARCH Model. Based on the above background, the purpose of this research is to get the best model of ARCH-GARCH at PT. Indonesian Telecommunications and Estimating the share price of PT. Indonesian Telecommunications the next day using the ARCH-GARCH Model.

B. LITERATURE REVIEW

1. Time Series Analysis

Time series analysis is an analysis based on time-oriented or chronological data or observations on the observed variables. This analysis is very useful in data whose changes are influenced by time or previous observations. In its development time series analysis is widely used in several fields such as economics, finance, transportation, and many more. To create a model suitable for data forecasting, there are several stages of the process in time series analysis, namely: data stationarity, parameter estimation, model specification, model checking, unit root testing, and forecasting (Prasetya et al., 2020).

2. ARIMA (Autoregressive Integrated Moving Average)

The ARIMA model or commonly referred to as the Box-Jenkins method is a model formed from two estimation models, namely the Autoregressive (AR) and Moving Average (MA) models. The Autoregressive (AR) model is a model that illustrates that the dependent variable that produces accurate short-term forecasts is influenced by the dependent variable itself in the previous period and the current error (error) or residual, in general, the AR (p) model can be written as:

$$r_t = \phi_0 + \sum_{1-i}^{p} \phi_1 r_{t-1} + a_t \tag{1}$$

Information:

r_t	=	dependent variable
p	=	Positive integer
ϕ_0	=	constant
$\phi_1, \ldots, \phi_p$	=	coefficient of the first autoregressive parameter to p
$r_{t-1}, \ldots, r_{t-p}$	=	Returns the past time, the independent variable which is the lag value of the dependent variable
a_t	=	Residual (white noise)

The moving average model is a model that states the dependence of observations (r_t) with a continuous error value from period t to $t-q$. Where $MA(q)$ can be written as follows:

$$r_t = C_0 + \sum_{1-i}^{p} \theta_i a_{t-1} + a_t \tag{2}$$

Information:

DOI: https://doi.org/10.30812/varian.v5i2.1543

q = Positive integer
C_0 = constant
$\theta_1, \ldots, \theta_q$ = Moving Average parameter coefficient to 1 to q
a_t = Residual at time t

Meanwhile, for non-stationary data, reduction or reduction is carried out until the data is stable and stationary. The ARIMA model (p, d, q) has the following formula:

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B)a_t \tag{3}$$

Where,

$$\phi_p(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$$
$$\phi_q(B) = (1 - \phi_1 B - \cdots - \phi_q B^q)$$

Information:
$(1-B)^d$ = differencing non-seasonal on order d

To estimate the time series data model used, the ARIMA Box Jenkins procedure is needed. The ADF (Augmented Dickey-Fuller) test was used to test the stationarity of the data. By hypothesis

$H_0 : \delta > 0.05$ = or not stationary
$H_0 : \delta < 0.05$ = or stationary

$$\Delta Z = \delta Z_{t-1} + u_t \tag{4}$$

$$\tau^* = \frac{\widehat{\delta}}{s.e(\widehat{\delta})} \tag{5}$$

If the value of $|\tau^*|$ is greater than the critical value of ADF with n degrees of freedom and significance level then H_0 is rejected, which can indicate that it has been corrected or the data is stationary. If the data is not stationary, then differencing is carried out until the data is stationary (Soeksin and Fatanah, 2020).

The ACF and PACF correlograms are used to estimate the ARIMA model, after which this model can be tested to estimate and test the significance of the parameters. The autocorrelation function (ACF) proves how the realization of the variable at time t relates to the realization of the variable discussed at some point in the future. ACF can be calculated by the following formula:

$$\tau S| = + \frac{E[y_t - E(y_t)][y_{t-s} - E(y_{t-s})]}{E[y_t - E(y_t)]^2} = \frac{\gamma_s}{\gamma_0}; s = 0, 1, 2, \ldots \tag{6}$$

The model selected from the estimation/estimation results can be AR(p), MA(q), ARMA(p,q), or ARIMA (p, d, q) models. Then the model obtained from the estimation results is then tested to get the best model based on several criteria. Akaike Information Criterion (AIC), Schwartz Information Criterion (SIC), and Hannan Quinn Criterion (HQIC). The order that has the smallest data value is the order that has the smallest value of the information criteria. There are no very superior information criteria, here we can use multiple criteria (orders selected by several information criteria) (Yusup and Purqon, 2015).

3. Arima Model Verification

Model verification is carried out to find out whether the model fits the observation data. The model verification includes the residual independence test and the residual normality test. A Residual independence test was conducted to see if there was a residual correlation between lags. The hypothesis for the residual independence test is: (Widyantomo et al., 2018).

H_0 : = There is no residual correlation between lags
H_1 : = There is a residual correlation between lag

The level of significance used is $\alpha = 0.05$, with test statistics:

$$Q = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}p_k^2 \tag{7}$$

The test criteria are H_0 is accepted if the probability value is $\geq \alpha$ dan H_1 is accepted if the probability value is $< \alpha$.

DOI: https://doi.org/10.30812/varian.v5i2.1543

The normality of the residuals was checked by the Jarque-Bera test measuring the difference between the skewness and kurtosis of the data from a normal distribution and included a measure of variance. The hypotheses for the residual normality test are:

H_0 : = Residuals are normally distributed

H_1 : = Residual distribution is not normal

With significance level $\alpha = 0.05$ and test statistics:

$$JB = N - K/6(S^2 + 1/4(k-3)^2) \tag{8}$$

Where:

JB = Jarque-bera

K = Kurtosis

S = Skewness

k = The number of estimated coefficients

n = Number of observations

The test criteria are H_0 accepted if the probability value $\geq \alpha$ dan H_1 is accepted if the probability value is $< \alpha$.

4. ARCH-GARCH

a. ARCH

The ARCH model is commonly used in modeling financial time series. The Autoregressive Conditional Heteroscedastic (ARCH) model is now commonly used to describe and predict changes in the volatility of financial data according. This Autoregressive Conditional Heteroscedastic (ARCH) model is used to overcome non-constant errors in time series data. The ARCH model was first introduced by Engle in 1982 (Bilondatu et al., 2019).

ARCH is the first model to provide a systematic framework for volatility modelling (Tsay, 2010). The basic concept of the ARCH model is that the residual variance depends on the square of the past residual. ARCH with order m (ARCH M) is model with the following equation (Tsay, 2010):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \alpha_{t-1}^2 + \cdots + \alpha_m \alpha_{t-m}^2 \tag{9}$$

Where:

σ_t^2 = residual variance value

α_0 = constant value

α_{t-1}^2 = last residual square

α_1 = constant value to $= (1, 2, \ldots)$

It is assumed that the model formed is ARCH (1), so to predict $t+1$ is as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1 \alpha_t^2 \quad \text{(Juliana et al., 2019)} \tag{10}$$

b. GARCH

GARCH is a process with a more general class, which can be used for a much more flexible lag structure. The GARCH method can be used for modeling heteroscedasticity data without eliminating the heteroscedasticity nature. In this modeling, the residual variance does not only depend on the square of the past residual but also the past residual variance. The general form of the GARCH model with the order m,s: (Tsay, 2010)

$$\sigma_t^2 = \alpha_0 + \alpha_i \alpha_{t-1}^2 + \sum_{j=1}^{S} \beta_j \sigma_{t-j}^2 \tag{11}$$

Where:

σ_t^2 = t-th data variance

α_0 = constant

α_j = ARCH Parameters

β_j = GARCH Parameters

α_{t-1}^2 = Residual to $t-i$

DOI: https://doi.org/10.30812/varian.v5i2.1543

It is assumed that the model formed is GARCH (1,1) then to predict t + 1 the formula is (Tsay, 2010).

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1 \alpha_t^2 + \beta_1 \beta_t^2 \quad \text{(Juliana et al., 2019)} \tag{12}$$

Order identification on GARCH can also be done by looking at the ACF and PACF patterns from the time series data (Azmi and Syaifudin, 2020).

5. Lagrange Multiplier Test (ARCH-LM Test)

The Lagrange Multiplier test is also often referred to as the ARCH-LM. This test is used to detect the effect of heteroscedasticity of the data, and the test also shows the effect of ARCH which is the subject of this study. Therefore, this study will use the heteroscedasticity test or the Lagrange Multiplier test (ARCH-LM test) to test the heteroscedasticity and the effect of ARCH. The test that can overcome the heteroscedasticity problem that can overcome the variance that is not constant in the time series developed by Engle is called the ARCH-LM test. The main idea of this test is that the residual variance is not only a function of the independent variable but depends on the residual squared in the previous period. Here are the steps in testing the ARCH effect:

Hypothesis:

H_0 : = $\alpha_0 = \alpha_1 = \ldots = \alpha_p = 0$ (There is no ARCH-GARCH effect in the residual)

H_1 : = there is at least one $\alpha \neq 1$ for $i = 1, 2, \ldots p$ (there is an ARCH-GARCH effect in the Residual)

Test Statistics

$$F = \frac{\dfrac{(SSR_0 - SSR_1)}{p}}{\left(\dfrac{SSR_1}{(T - 2p - 1)}\right)} \tag{13}$$

Where:

$$SSR_0 = \sum_{t=p+1}^{T} (\varepsilon - \omega)^2$$

$$\omega = \frac{\sum_{t=1}^{T} \varepsilon_t^2}{T}$$

$$SSR_1 = \sum_{t=p+1}^{T} W_t^2$$

with:

α = Significance level (0.05)

p = Number of independent variable

W_t^2 = Least square residual

ω = The average of T

Decision H_0 rejected if $F_{hit} > X_p^2(\alpha)$ or $p - value < \alpha$

6. Akaike Information Criterion (AIC) Test

AIC is an information standard that provides a measure of the information that can be found between the balance and measure of model goodness and model specifications. Models are used to select the best model. To get the best model, it can be seen from the smallest AIC value. The AIC formula is as follows:

$$AIC = \left(e^{\frac{2k}{n}}\right)\left(\frac{\sum e_i^2}{n}\right) = \left(e^{\frac{2k}{n}}\right)\left(\frac{SSE}{n}\right) \tag{14}$$

Where:

$SSSE$ = Sum square error $= \sum e_i^2$

$= \sum \left(\widehat{Z}_I - Z_i\right)^2$

K = Number of parameters in the model

n = Number of sample observations (Ermawati et al., 2018)

DOI: https://doi.org/10.30812/varian.v5i2.1543

7. Forecasting

As the demand for business planning and economic activity increases, so does the need for accurate predictions of future conditions. The development of computing technology has led to the development of various forecasting methods and techniques to predict future conditions that can meet these needs (Firdaus, 2020).

C. RESEARCH METHOD

The data required is the daily closing data of PT. Telekomunikasi Indonesia starting from January 2020 to January 2021 was obtained from the website www.finance.yahoo.com. The variable used in this study is the daily closing data variable for the stock price of PT. Telekomunikasi Indonesia (Z_t). The closing stock price is considered the most accurate assessment of stock to assess changes in stock prices from time to time.

In building the ARCH-GARCH method, a methodology is needed, namely the stages carried out in the use of the ARCH-GARCH model. The steps for implementing the ARCH-GARCH, method are as follows:

1. Make data tabulation.
2. Check the stationary of the data.
3. Identify the ARIMA model.
4. Estimation of model parameters and selection of the best model.
5. Verify the model.
6. Heteroscedasticity Test.
7. Parameter estimation and selection of the best ARCH/GARC model.
8. Verify the best model.
9. Do Forecasting.

D. RESULTS AND DISCUSSION

The data used in this study is the daily historical data closing the stock price of PT. Telekomunikasi Indonesia from 06 January 2020 to 06 January 2021 which was obtained online through finance.yahoo.com.

1. Stationery Test

Time series data cannot be separated from the stationarity test of the data to be studied. Therefore, a stationary test was carried out. Data testing can still be done by looking at the data graph and by going through the unit root test statistic test using the Augmented Dickey-Fulller (ADF) method. Based on the closing data of PT. Telekomunikasi Indonesia then obtained a time series plot as shown in Figure 1 below:

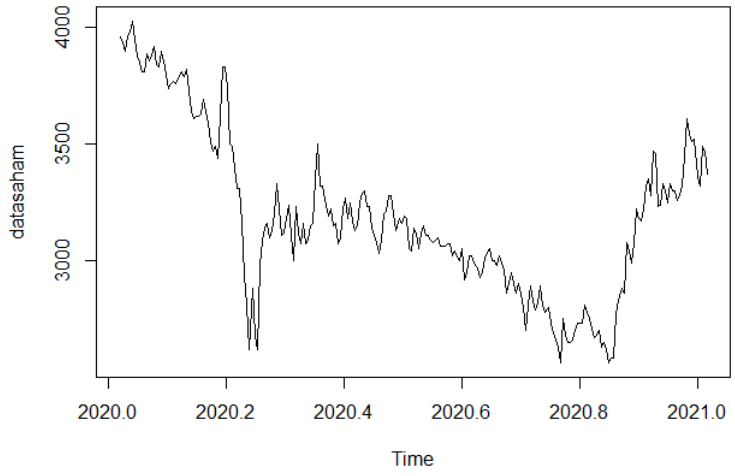


Figure 1. Stock Price Closing Data

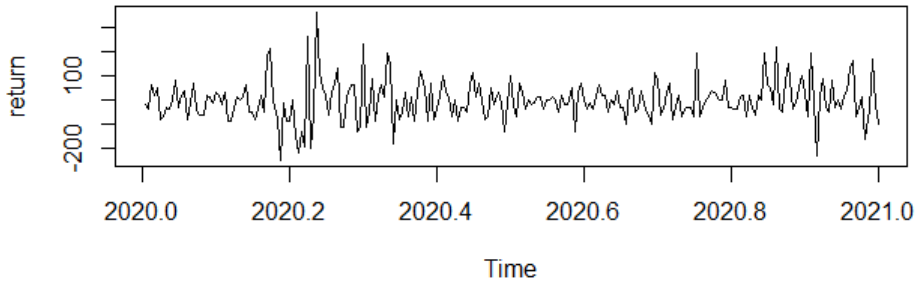


Figure 2. Data differencing 1 Stock Closing Data

Based on the stock data plot, it can be seen that there are differences in Figure 1, namely the actual stock data before the first difference, and Figure 2, which is the stock data after the first difference. In Figure 1 it can be explained that the plot of stock price data experienced unstable movements so that the data can be said to be not stationary on average. Figure 2 shows that the stock price data pattern is around a constant average value, which means the data is stationary. To strengthen the assumptions, it can be continued to the next stage.

Hypothesis:

H_0 : = There is a unit root so the data is not stationary

H_1 : = There is no unit root so the data is stationary

Significance level $\alpha = 0.05$

Table 1. Augmented Dickey-Fuller Before Differencing

	Lag	ADF	P-Vale
[1,]	0	-0.725	0.420
[2,]	1	-0.700	0.428
[3,]	2	-0.715	0.423
[4,]	3	-0.715	0.423
[5,]	4	-0.756	0.408

Based on Table 1 above, shows that the data is not stationary because all p-values are greater than $\alpha = 0.05$. So the data is said to be not stationary. To be stationary, order 1 differencing can be done.

Table 2. Augmented Dickey-Fuller After Differencing

	Lag	ADF	P-Vale
[1,]	0	-15.05	0.01
[2,]	1	-12.92	0.01
[3,]	2	-8.65	0.01
[4,]	3	-8.12	0.01
[5,]	4	-7.12	0.01

From Table 2. above, a p-value of 0.01 is obtained, which means it is smaller than α so H_1 is accepted, which means that the data is stationary and can be continued to the next stage, namely the identification of the ARIMA model.

2. Identify the ARIMA Model

To perform the identification stage of AR and MA models from a time series data, it can be done by looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots at various lags. The ACF diagram can be seen in Figure 3 and Figure 4 below.

DOI: https://doi.org/10.30812/varian.v5i2.1543

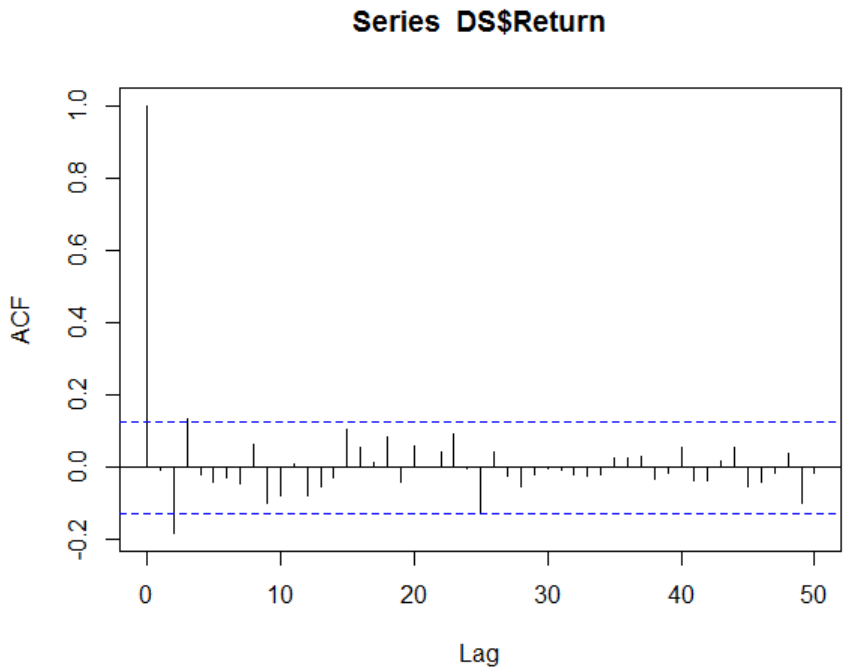


Figure 3. ACF

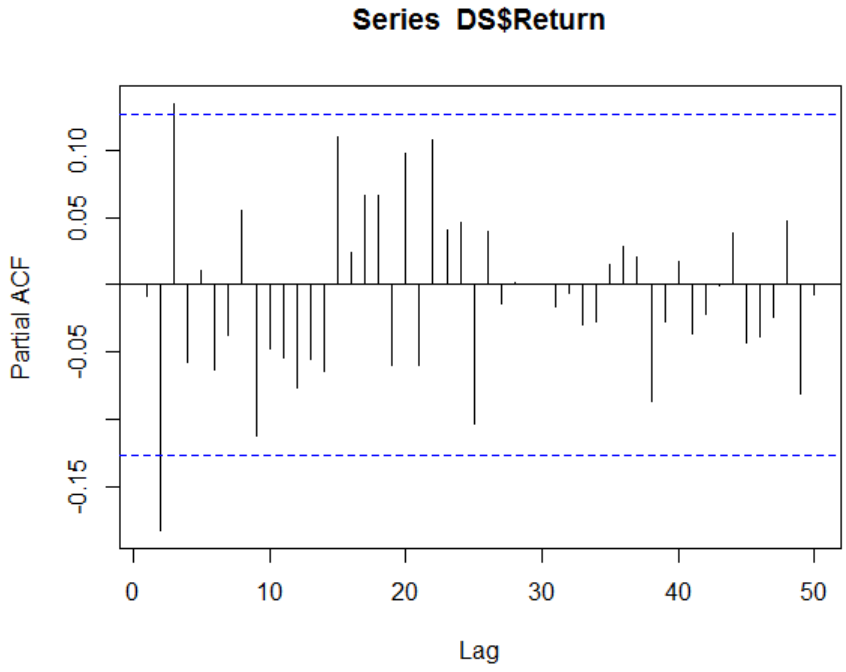


Figure 4. PACF

Based on Figures 3 and 4 above, it can be seen that lag 2 is slowly decreasing towards 0 so that the model that can be generated is AR(2), then the results on the ACF and PACF plots show that the ACF and PACF values are on the interval line (5%), lags that exceed the boundary line are identified as AR (based on ACF plots) and MA (based on ACF plots) levels.

DOI: https://doi.org/10.30812/varian.v5i2.1543

3. Estimation and Selection of ARIMA Model (p, d, q)

After the identification stage on the ARIMA model (p, d, q) is carried out, the ARIMA model parameter estimation process (p, d, q) is carried out. ARIMA model estimation can be seen in Table 3 below:

Table 3. Estimated ARIMA Model

Model	AIC
ARIMA (0,1,1)	2792.85
ARIMA (1,1,1)	2799.72
ARIMA (1,1,2)	2795
ARIMA (1,1,3)	2793.64
ARIMA (1,1,4)	2793.17
ARIMA (2,1,1)	2793.51
ARIMA (2,1,2)	2790.63
ARIMA (2,1,3)	2787.84
ARIMA (2,1,4)	2788.45
ARIMA (3,1,1)	2791.56
ARIMA (3,1,2)	2792.23
ARIMA (3,1,3)	2794.59
ARIMA (3,1,4)	2789.24

More than one model is formed, so in R, the best ARIMA model is obtained. Judging from the smallest AIC value, the best model is ARIMA (2,1,3) where AIC = 2787.84, so that the model can be used to the next stage.

4. ARIMA Model Verification

An examination of the adequacy of the model is carried out to prove that the model obtained is adequate, as can be seen in Table 4 below.

Table 4. ARIMA Modeling $(2, 1, 3)$

Parameter	Estimation	S.E	Z-Value	$Pr(>\|z\|)$
AR (1)	-1.489264	0.095727	-15.5574	$< 2.2e^{-16}$
AR (2)	-0.812673	0.065233	-12.4581	$< 2.2e^{-16}$
MA (1)	0.634092	0.100978	6.2795	$3.397e^{-10}$
MA (2)	-0.742370	0.063144	-11.7568	$< 2.2e^{-16}$
MA (3)	-0.100344	0.081988	-10.7466	$< 2.2e^{-16}$

Based on Table 4 above, the $p - values$ for all parameters AR(1), AR(2), MA(1), MA(2), MA(3) are less than 0.05 which indicates that they represent statistically significant estimates.

5. Heteroskedasticity Test

The heteroscedasticity test is a test conducted to see whether there is a heteroscedasticity effect on the best ARIMA (p, d, q) model. Previously, the best ARIMA model was obtained, namely, the ARIMA model $(2, 1, 3)$ using the ARCH-LM. Test. With the following hypothesis:

H_0 : = No ARCH-GARCH effect on residue

H_1 : = There is an ARCH-GARCH effect on the residual

The level of significance used is $= 0.05$, while the results of the ARCH-LM test are in the Table 5 as follows:

DOI: https://doi.org/10.30812/varian.v5i2.1543

Table 5. ARCH LM test results

Model	Order	LM	P-Value
[1,]	4	105.55	$0.00e^{+00}$
[2,]	8	45.37	$1.16e^{-07}$
[3,]	12	25.37	$8.05e^{-03}$
[4,]	16	16.37	$3.58e^{-01}$
[5,]	20	11.21	$9.17e^{-01}$
[6,]	24	8.07	$9.98e^{-01}$

Based on Table 5 above, it can be seen that the $p - value$ on the Lagrange-Multiplier test which is smaller than 0.05 means H_1 accepted which indicates that there is an ARCH effect on the estimated model.

6. Estimation and Selection of the Best Model

Based on the parameter estimation results of the ARCH-GARCH model. Selection of the best model based on the smallest significance (5%) of the AIC and BIC models. The estimation results of the ARCH-GARCH model can be seen in Table 6 below:

Table 6. Best ARCH-GARCH Model

company stock	Best Model (p,q)	AIC	BIC
PT. Telekomunikasi Indonesia	ARCH(1,0)	11.558	11.689
	ARCH-GARCH(1,1)	11.441	11.586

Based on Table 6 above, the shares of PT. The best Indonesian telecommunications model is the ARCH-GARCH(1,1) model with the smallest AIC value of 11.558 and the smallest BIC value of 11.689. After the best model is found, a residual diagnostic test or model verification will be carried out to determine whether the model found does not contain the ARCH (Heteroscedasticity) effect, which can be seen in Table 7 below.

Table 7. ARCH-LM test

Company Stock	Best Model (p,q)	Lag ke	Statistic	P-Value
PT. Telekomunikasi Indonesia	ARCH(1) GARCH(1)	1	0.2697	0.6036
		2	5.2946	1.0000
		4	9.5231	0.8762

Based on Table 7, it can be seen that all models do not contain heteroscedasticity, where the p-value is greater than $= 0.05$, so the model is declared valid for use in forecasting. After obtaining a valid model used in forecasting.

The results of forecasting the share price of PT. Telekomunikasi Indonesia from 06 January 2020 to 6 January 2021, which is in Table 8 below.

Table 8. Forecasting Results for the Next 10 Days

Date	Forecasting (Idr)
07/01/2021	3374.884
08/01/2021	3379.617
11/01/2021	3378.305
12/01/2021	3376.610
13/01/2021	3380.050
14/01/2021	3376.372
15/01/2021	3379.071
18/01/2021	3377.964
19/01/2021	3377.515
20/01/2021	3379.002

Based on Table 8 above, the forecasting results for the next 10 days are 07/01/2021, 08/01/2021, 11/01/2021, 12/01/2021, 13/01/2021, 14/01/2021, 15/01/2021, 18/01/2021, 19/01/2021, 20/01/2021 is 3374.884, 3379.617, 3378.305, 3376.610,

3380.050, 3376.372, 3379.071, 3377.964, 3377.515, 3379.002. These results are close to factual data so they are worth considering in investing.

E. CONCLUSION AND SUGGESTION

The ARIMA model applied to the stock data of PT.Telekomunikasi Indonesia is the ARIMA model (2,1,3) with an AIC value of 2787.84. In this study, the best ARCH-GARCH model was selected on the stock data of PT. Telekomunikasi Indonesia is the ARCH-GARCH (1.1) model with an AIC value of 11.441 and a BIC of 11.586. and the model is declared valid, the residual diagnostic test or model verification is found to no longer contain the ARCHGARCH (Heteroskedatisits) effect. Forecasting results for the next 10 days are IDR 3374.884, IDR 3379.617, IDR 3378.305, IDR 3376.610, IDR 3380.500, IDR 3376.372, IDR 3379.071, IDR 3377.964, IDR 3377.515, IDR 3379.002.

ACKNOWLEDGEMENT

Thank you to All those who have helpedini this research.

REFERENCES

Angraeny, N. (2019). Penerapan Metode Arch Garch untuk Analisis Peramalan Nilai Ekspor Indonesia.

Arumningtyas, F., Prahutama, A., and Kartikasari, P. (2021). Value-At-Risk Analysis Using ARIMAX-GARCHX Approach For Estimating Risk of Bank Central Asia Stock Returns. *Jurnal Varian*, 5(1):71–80.

Azmi, U. and Syaifudin, W. H. (2020). Peramalan Harga Komoditas dengan Menggunakan Metode Arima-Garch. *Jurnal Varian*, 3(2):113–124.

Bakhtiar, S. M., Syata, I., Alwi, W., Ibnas, R., Anugrawati, S. D., et al. (2021). Estimation of Average Value at Risk (AVaR) on Sharia Joint-Stock Index Using Glosten, Jaggnathan and Runkle (GJR) model. In *1st International Conference on Mathematics and Mathematics Education (ICMMEd 2020)*, pages 143–149. Atlantis Press.

Bilondatu, R., Nurwan, N., and Isa, D. (2019). Model ARCH(1) dan GARCH(1,1) pada Peramalan Harga Saham PT. Cowell Development Tbk. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 13(1):9–18.

Ermawati, E., Basri, N., and Nurfadilah, K. (2018). Pemodelan Beban Puncak Energi Listrik Menggunakan Model GJR-GARCH. *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, 6(1):27–27.

Firdaus, M. (2020). *Aplikasi Ekonometrika dengan E-Views, Stata dan R*. PT Penerbit IPB Press.

Heriyanto (2022). *Analisa Saham. Sahamn TLKM: Perseroan Fokus Mengembangkan Layanan Digital*. Ajaib.co.id.

Juliana, A., Hamidatun, and Muslima, R. (2019). *Modern Forecating Teori dan Aplikasi (GARCH, Artifical Neural Network, Neuro-Garch) (1st ed.)*. Deepublish Publisher.

Krispin, R. (2019). *Hands-On Time Series Analysis with R: Perform Time Series Analysis and Forecasting using R*. Packt Publishing Ltd.

Larasati, E. N., Hendikawati, P., and Zaenuri, Z. (2016). Analisis Volatility Forecasting Sembilan Bahan Pokok Menggunakan Metode Garch Dengan Program R. *UNNES Journal of Mathematics*, 5(1):90–99.

Prasetya, B. D., Pamungkas, F. S., and Kharisudin, I. (2020). Pemodelan dan Peramalan Data Saham dengan Analisis Time Series menggunakan Python. In *PRISMA, Prosiding Seminar Nasional Matematika*, volume 3, pages 714–718.

Soeksin, S. D. and Fatanah, C. (2020). Peramalan Harga Saham PT. Bumi Serpong Damai Tbk. Dengan Metode Garch. *JAMAN (Jurnal Aplikasi Manajemen dan Akuntansi)*, 1(01):64–71.

Tsay, R. S. (2010). *Analysis of Financial Time Series (third edit)*. John wiley & sons.

Widyantomo, R. P., Hoyyi, A., and Widiharih, T. (2018). Pemodelan Volatilitas Return Portofolio Saham menggunakan Feed Forward Nerural Network (Studi Kasus: PT Bumi Serpong Damai Tbk. dan PT HM Sampoerna Tbk.). *Jurnal Gaussian*, 7(2):200–211.

Yusup, M. and Purqon, A. (2015). Aplikasi Ekonofisika Menggunakan Metode ARCH–GARCH pada Analisis Beberapa Saham Index LQ45. *Prosiding SKF 2015*, pages 512–521.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 137–148
e-ISSN: 2581-2017

Defuzzification Methods Comparison of Mamdani Fuzzy Inference System in Predicting Tofu Production

Grandianus Seda Mada[1], Nugraha Kristiano Floresda Dethan[2], Andika Ellena Saufika Hakim Maharani[3]

[1,2]Departement of Mathematics, Timor University, Indonesia
[3]Departement of Computer Science, Universitas Bumigora, Indonesia

Article Info

Article history:

Received : 03-05-2022
Revised : 04-26-2022
Accepted : 04-30-2022

Keywords:

Fuzzy Inference System;
Mamdani;
Defuzzification;
Tofu

ABSTRACT

One of the tofu-producing companies in Kupang City is Bintang Oesapa. With the Covid-19 pandemic, the factory needs to reconsider the amount of production by taking into account the unpredictability of demand and resources to minimize losses due to excessive accumulation or shortages of supplies. In determining the amount of production, Mamdani's Fuzzy Inference System (FIS) can be used, which is a method for the analysis of an uncertain system. This method has three stages in the process of decision making, namely fuzzification, inferencing and defuzzification. In the defuzzification stage, the FIS Mamdani has five methods, namely Centroid, Bisector, Mean of Maximum (MOM), Smallest of Maximum (SOM), and Largest of Maximum (LOM). This study discusses an application of FIS Mamdani with five defuzzification methods for determining daily tofu production. The purpose of this study is to offer a solution by first comparing the five defuzzification methods in assessing the amount of tofu production at the Bintang Oesapa factory and then determining that which is most appropriate. The input variables used in this research are the amount of demand and the amount of available stock, while the amount of production is our variable of interest. The results showed that the best defuzzification method was the MOM method with an accuracy level of 94.73% and a small error value, 5.27%. The MOM defuzzification is expected to aid decision makers in determining the best amount of production during the pandemic.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1816

Corresponding Author:

Grandianus Seda Mada,
Department of Mathematics, Timor University
Email: grandianusmada@gmail.com

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The current Covid-19 pandemic has had adverse effects on the national economy and hence East Nusa Tenggara as well. The pandemic can precipitate a large economic crisis, which has been marked by the cessation of production activities at various companies, the decline in public consumption levels and consumer confidence index.

The Covid-19 pandemic has urged Indonesian government to implement a Large-Scale Social Restriction (LSSR) policy. This policy will cause the economy in Indonesia to decline, and especially has a big impact on Micro, Small and Medium Enterprises (MSMEs) (Bahtiar, 2021). One of the MSMEs operating under Covid-19 is the Bintang Oesapa Tofu Factory MSMEs which is located in Kupang City, East Nusa Tenggara.

The Bintang Oesapa Tofu Factory is one of the many factories that produce tofu in Kupang City. In the tofu production process, the factory produces according to market demand. Based on the interview process with the factory, it is known that on average, it can produce approximately 40,000 pieces in a day. However, due to the pandemic, production has been drastically reduced to merely 20,000-30,000 pieces per day. This is because the remaining stock is around 1000-6000 pieces per day. The Bintang Oesapa Tofu

Factory needs a plan to determine the amount of production such that market demand can be met and that profits remain optimal.

Maximum profit is achieved from maximum sales. Maximum sales means being able to meet existing demand. If the amount of goods/service provided by the company is less than the demand, the company will lose the opportunity to get maximum profit. Conversely, if amount of goods/service offered is much more than demand, the company will experience losses. Therefore, planning the amount of production in a company is very important in order to meet market demand and in the right amount. Factors that need to be considered in determining the number of products include: the amount of leftovers from previous period and the estimated demand for the next period (Abrori and Primahayu, 2015). During this pandemic, of course, these factors become uncertain or erratic.

Fuzzy set theory is a method for the analysis of uncertain systems which has more than one method in calculating the estimation of a case. In estimating the amount of tofu production at the Bintang Oesapa tofu factory, no prediction method was used previously, especially the fuzzy set theory method. The decision-making process using fuzzy set theory is called the Fuzzy Inference System (FIS). There are 3 FIS methods that can be used in determining the final decision of a problem, namely the Tsukamoto Method, the Mamdani Method and the Sugeno Method (Priyo, 2017). The three methods generally has the same stages, namely fuzzification, inferencing and defuzzification, but have a different calculation method for each stage.

(Rahakbauw, 2015) applies the FIS Sugeno method to predict the amount of bread production by paying attention to inventory data and requested data and an accuracy rate of 86.92%. Then (Rahakbauw et al., 2019) applied the Mamdani FIS method to predict the amount of rubber production by taking into account the availability of data and the requested data and an accuracy rate of 87.83%. Meanwhile, (Mada et al., 2021) compared the predictions for the amount of brick production with the Mamdani FIS and Tsukamoto FIS methods and obtained an accuracy of 99.96% for the Mamdani FIS method and 99.92% for the Tsukamoto FIS method.

In the decision-making process, especially at the defuzzification stage, the Mamdani FIS method has five methods that can be used, namely Centroid, Bisektor, Mean of Maximum (MOM), Smallest of Maximum (SOM), Largest of Maximum (LOM) (Sutikno, 2011). So far, the most frequently used defuzzification method is the Centroid method. There are several previous studies such as (Wardani et al., 2017) in predicting the amount of palm oil production, (Santya et al., 2017) in predicting the amount of banana chips production, (Sari, 2021) in predicting the amount of salt production, (Susetyo et al., 2020) in predicting t-shirt production, (Sahulata et al., 2020) and (Rianto and Manurung, 2022) in predicting the amount of bread production, all using the Centroid defuzzification method of the Mamdani. Hence, we may ask why in production always use the Centroid Mamdani type in defuzzification method? What about the other four Mamdani defuzzification methods?

This research aims to compare the amount of production calculated by the five Mamdani defuzzification methods. To determine the best method, the magnitude of error from the prediction results of the five methods will be calculated using the Mean Absolute Percentage Error (MAPE) formula. The smaller the MAPE value obtained, the better the prediction results (Yusuf et al., 2017).

B. LITERATURE REVIEW

1. Fuzzy Set

In (Rahakbauw, 2015), it has been explained that the concept of fuzzy set was first introduced by Prof. Lofti A. Zadeh of UC Berkley, USA in 1965 with the following definition.

Definition 1. *Let X be a set. A fuzzy subset A of X is a subset of X whose membership is defined through a function*

$$\mu_A : X \to [0, 1] \tag{1}$$

which associates an element $x \in X$ to a real number $\mu_A(x)$ in $[0, 1]$. The value $\mu_A(x)$ indicates the degree of membership of x in A. A fuzzy set A is written as follows:

$$A = \{(x, \mu_A(x)|x \in X)\} \tag{2}$$

The pair $(x, \mu_A(x)$ reads x has the degree of membership $\mu_A(x)$.

2. Membership Function

In this research, we used three types of membership functions;

a. Linearly Increasing Membership Function

Definition 2. (Mada et al., 2021) *A membership function μ is said to be linearly increasing (on (a, b)) if it can be represented as the following*

$$\mu(x) = \begin{cases} 0 & ; \quad x \leq a \\ \frac{x-a}{b-a} & ; \quad a \leq x \leq b, \\ 1 & ; \quad x \geq b \end{cases} \tag{3}$$

For more details, the geometric shape of this function can be seen in Figure 1(a)

b. Linearly Decreasing Membership Function

Definition 3. (Mada et al., 2021) *A membership function μ is said to be linearly decreasing (on (a, b)) if it satisfies the following*

$$\mu(x) = \begin{cases} 1 & ; \quad x \leq a \\ \frac{b-x}{b-a} & ; \quad a \leq x \leq b, \\ 0 & ; \quad x \geq b \end{cases} \tag{4}$$

For more details, the geometric shape of this function can be seen in Figure 1(b)

c. Triangular Membership Function

Definition 4. (Mada et al., 2021) *A membership function μ is said to be triangular (on (a, b)) if it can be written as*

$$\mu(x) = \begin{cases} 0 & ; \quad x \leq a \text{ or } x \geq c \\ \frac{x-a}{b-a} & ; \quad a \leq x \leq b, \\ \frac{c-x}{c-b} & ; \quad b \leq x \leq c, \\ 1 & ; \quad x = b \end{cases} \tag{5}$$

For more details, the geometric shape of this function can be seen in Figure 1(c)

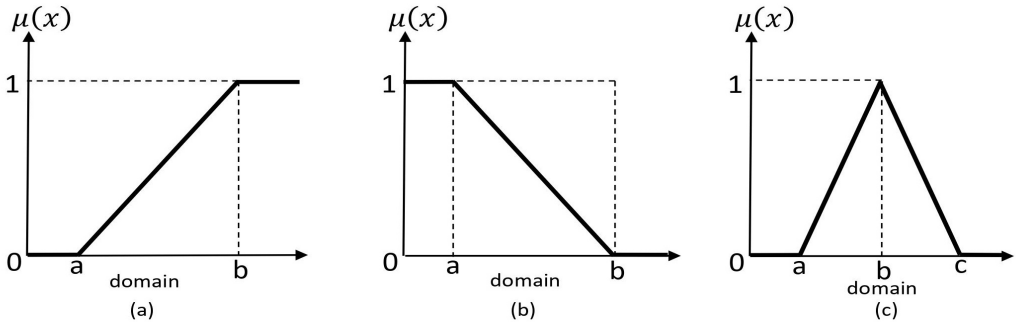


Figure 1. (a) Linearly Increasing, (b) Linearly Decreasing, and (c) Triangular Membership Function

A more detailed explanation regarding applications of the three membership functions can be found in chapter results and discussion.

3. Fuzzy Inference System (FIS)

Fuzzy inference is the process of mapping a given input space into an output space using fuzzy set theory. From this mapping process, the basis of the decisions made or the existing patterns can be seen.

According to (Priyo, 2017), in general, there are several stages needed for designing a fuzzy system, namely:

a. Fuzzification is a process of changing non-fuzzy variables (numeric variables) into fuzzy variables (linguistic variables) using appropriate membership functions.
b. Inferencing Stage (Rule Base). At this stage, rules are made which will be used as a reference in determining the output of the fuzzy system. The rules produced are in the form of If...And...Then... which is a combination of input variables and output variables.
c. Defuzzification is the process of converting fuzzy data into numerical data as the final decision.

There are 3 FIS methods, namely the Tsukamoto, Mamdani and Sugeno methods. In this study, only the Mamdani method will be discussed.

4. Mamdani Fuzzy Inference System

This method was first introduced by Ebrahim Mamdani in 1975. The creation of this method was motivated by the work of Lotfi Zadeh (1973) on the Fuzzy algorithm for complex systems and was used in the decision-making process (Sari, 2021).

The Mamdani method is a type of fuzzy inference where the fuzzy sets resulted from each rule are combined using aggregation operator and produces a fuzzy set which is then defuzzified to produce a certain output from a system. The output requires several stages, including:

a. Generating The Fuzzy Sets

In the Mamdani method, both input and output variables are divided into one or more fuzzy sets (Sari, 2021).

b. Applying Implication Function

(Sari, 2021) states that in the Mamdani method, the function used is the Min function can be written as:

$$\mu_{A\cap B} = \min\{\mu_A(x), \mu_B(x)\} \tag{6}$$

c. Composing The Rules

Unlike monotonic reasoning, if the system consists of several rules, then inference is obtained from the collection and correlation between rules. There are 3 methods used in performing fuzzy system inference, namely max, additive and probabilistic OR (probor).

The Max (Maximum) method takes the solution of the fuzzy set obtained by taking the maximum value of the rule, then using it to modify the fuzzy area, and applying it to the output using the OR (union) operator. If all propositions have been estimated, then the output will contain a fuzzy set that reflects the contribution of each proportion (Sari, 2021). In general it can be written as:

$$\mu_{A\cup B} = \max\{\mu_A(x), \mu_B(x)\} \tag{7}$$

d. Defuzzification

There are several defuzzification methods on MAMDANI composing rules (Sutikno, 2011);

– Centroid (Composite Moment) Method

In the centroid method, the crisp solution is obtained by taking the center point of the fuzzy area. In general, it can be written as:

For continuous universe of discourse (Generally for nonlinear cases),

$$z^* = \frac{\int_{z_l}^{z_u} z\mu(z)dz}{\int_{z_l}^{z_u} \mu(z)dz}; z_l \le z \le z_u \tag{8}$$

For discrete universe of discourse,

$$z^* = \frac{\sum_{i=1}^{n} z_i\mu(z_i)}{\sum_{i=1}^{n} \mu(z_i)} \tag{9}$$

DOI: https://doi.org/10.30812/varian.v5i2.1816

– Bisector Method

In the bisector method, the crisp solution is obtained by taking the domain which has a value from the number of membership values in the fuzzy area. Generally this is written

$$z = \frac{1}{2}\sum_{i=1}^{n} z_i \mu(z_i) \tag{10}$$

– Mean of Maximum (MOM) Method

In the MOM method, the crisp solution is obtained by taking the average value of the domain that has the maximum membership value

– Largest of Maximum (LOM) Method

In the LOM method, the maximum solution is obtained by taking the largest value from the domain that has the maximum membership value.

– Smallest of Maximum (LOM) Method

In the SOM method, the crisp solution is obtained by taking the smallest value from the domain that has the maximum membership value.

5. Mean Absolute Percentage Error (MAPE)

According to (Wardani et al., 2017), forecasting techniques are not always appropriate because they are not necessarily in accordance with the nature of the data. Therefore, it is necessary to evaluate forecasting so that it can be known whether the forecasting technique used is appropriate or not, so that a more precise forecasting technique can be selected and determined by setting tolerance limits for deviations that occur.

In principle, forecasting evaluation is done by predicting the results of what actually happened. The use of forecasting techniques that produce deviations is the most suitable forecasting technique to use.

The magnitude of the forecast error is calculated by subtracting the real data from the estimated size. In calculating the forecast error, the Means Absolute Percentage Error (MAPE) is used, which is the average absolute percentage of a forecast:

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{X_t - \widehat{X}_t}{X_t} \right|}{N} \times 100\% \tag{11}$$

where N = the number of forecasting periods, X_t = the true value at time t, $\widehat{X}_t$ = the forecasting value at time t.

According to (Azmi et al., 2020), the forecasting ability is very good if it has a MAPE value of less than 10% and has good forecasting ability if the MAPE value is less than 20%.

C. RESEARCH METHOD

1. Place and Funding Sources

The research was conducted for 2 months (January 2022 – February 2022). The type of data used in this research is secondary, obtained from the Bintang Oesapa Tofu Factory. Data regarding demand and supply taken is daily data starting from January 02, 2022 to February 10, 2022 and several days for data processing.

2. Data Identification

Data identification was carried out to determine the variables and the universe of discourse for calculations and problem analysis. This process is detail explained in the results and discussion chapter.

3. Data Processing

Data processing was carried out using the five Mamdani FIS defuzzification methods with Matlab R2013a software. Subsequently, the magnitude of the error for each defuzzification method is calculated using MAPE and the accuracy of each defuzzification method is determined.

4. Drawing Conclusion

The FIS method and the model are said to be adequate if it has the smallest predicting error value. The error rates are calculated based on Mean Absolute Percentage Error (MAPE).

D. RESULT AND DISCUSSION

Based on the data collection process regarding the data on the amount of tofu production that is influenced by demand and supply, the data obtained are as presented in Table 1.

Table 1. Tofu Production Data at Bintang Oesapa Tofu Factory on January 2, 2022 – February 10, 2022

No.	Date	Demand	Stock	Production	No.	Date	Demand	Stock	Production
1	Jan 2, 2022	18,400	2,200	22,400	21	Jan 22, 2022	25,200	4,500	27,500
2	Jan 3, 2022	21,000	3,600	24,200	22	Jan 23, 2022	28,800	1,400	31,100
3	Jan 4, 2022	22,400	4,600	29,200	23	Jan 24, 2022	18,200	2,600	32,800
4	Jan 5, 2022	27,200	1,600	28,600	24	Jan 25, 2022	23,600	2,800	23,600
5	Jan 6, 2022	20,400	600	29,400	25	Jan 26, 2022	20,400	3,200	29,600
6	Jan 7, 2022	18,800	1,400	22,400	26	Jan 27, 2022	26,400	2,800	26,400
7	Jan 8, 2022	22,800	1,000	21,200	27	Jan 28, 2022	25,000	1,600	30,800
8	Jan 9, 2022	25,600	3,200	27,000	28	Jan 29, 2022	20,200	4,400	31,000
9	Jan 10, 2022	17,600	3,800	32,600	29	Jan 30, 2022	18,800	1,600	26,200
10	Jan 11, 2022	21,600	4,800	26,200	30	Jan 31, 2022	24,400	2,400	22,800
11	Jan 12, 2022	19,200	4,400	30,800	31	Feb 01, 2022	23,150	3,200	30,000
12	Jan 13, 2022	19,600	5,400	29,000	32	Feb 02, 2022	24,342	6,400	32,750
13	Jan 14, 2022	24,400	3,200	28,200	33	Feb 03, 2022	19,200	1,800	32,542
14	Jan 15, 2022	24,000	800	28,400	34	Feb 04, 2022	17,200	1,200	22,200
15	Jan 16, 2022	20,464	4,200	29,000	35	Feb 05, 2022	18,000	5,200	23,600
16	Jan 17, 2022	24,800	6,380	31,044	36	Feb 06, 2022	24,400	3,600	26,800
17	Jan 18, 2022	18,800	1,000	32,180	37	Feb 07, 2022	21,200	3,000	31,000
18	Jan 19, 2022	25,200	2,800	22,600	38	Feb 08, 2022	23,000	4,800	29,000
19	Jan 20, 2022	20,400	4,400	32,400	39	Feb 09, 2022	21,800	2,200	30,000
20	Jan 21, 2022	23,200	3,600	28,400	40	Feb 10, 2022	25,200	3,400	27,400

1. Fuzzification

It can be seen from Table 1 that from 2 January 2022 to 10 February 2022, the highest demand was on 23 January with a total of 28,800 pieces, while the least was on 4 February with a mere 17,200 pieces. Then the highest supply was on 2 February with 6,400 pieces in total, and the least supply was on 6 January with only 600 pieces. As for productions, the highest was on 24 January with a total of 32,800 pieces, while the minimum was on 8 January with 21,200 pieces. Hence, based on the data of least and most of each of the variables above, the universe of discourse for these variables can be defined as shown in Table 2. Then, for each of the variables, three fuzzy sets are constructed; Low, Medium, and High, with their respective domains also presented in Table 2.

Table 2. Fuzzy Sets

Function	Variable	Universe of Discourse	Fuzzy Set	Domain
Input	Demand	$[17,200,\ 28,800]$	Low	$[17,200,\ 23,000]$
			Medium	$[20,100,\ 25,900]$
			High	$[23,000,\ 28,800]$
	Stock	$[600,\ 6,400]$	Low	$[600,\ 3,500]$
			Medium	$[2,050,\ 4,950]$
			High	$[3,500,\ 6,400]$
Output	Production	$[21,200,\ 32,800]$	Low	$[21,200,\ 27,000]$
			Medium	$[24,100,\ 29,900]$
			High	$[27,000,\ 32,800]$

Based of equations (3), (4) and (5), the membership functions for demand, supply and production are defined as follows:

DOI: https://doi.org/10.30812/varian.v5i2.1816

$$\mu_{Low-DEMAND}(x) = \begin{cases} 1 & ; \quad x \leq 17,000 \\ \frac{23,000-x}{5,800} & ; \quad 17,000 \leq x \leq 23,000, \\ 0 & ; \quad x \geq 23,000 \end{cases} \tag{12}$$

$$\mu_{MEDIUM-DEMAND}(x) = \begin{cases} 0 & ; \quad x \leq 20,100 \; or \; x \geq 25,900 \\ \frac{x-20,100}{2,900} & ; \quad 20,100 \leq x \leq 23,000 \\ \frac{25,900-x}{2,900} & ; \quad 23,000 \leq x \leq 25,900 \\ 1 & ; \quad x = 25,900 \end{cases} \tag{13}$$

$$\mu_{HIGH-DEMAND}(x) = \begin{cases} 0 & ; \quad x \leq 23,000 \\ \frac{x-23,000}{5,800} & ; \quad 23,000 \leq x \leq 28,800 \\ 1 & ; \quad x \geq 28,800 \end{cases} \tag{14}$$

The geometric representation of the membership function of the three fuzzy sets for demand is presented in Figure 2(a). Furthermore, the membership function for the stock variable is defined in the following way:

$$\mu_{LOW-STOCK}(x) = \begin{cases} 1 & ; \quad x \leq 600 \\ \frac{3,500-x}{2,900} & ; \quad 600 \leq x \leq 3,500 \\ 0 & ; \quad x \geq 3,500 \end{cases} \tag{15}$$

$$\mu_{MEDIUM-STOCK}(x) = \begin{cases} 0 & ; \quad x \leq 2,050 \; or \; x \geq 4,950 \\ \frac{x-2,050}{1,450} & ; \quad 2,050 \leq x \leq 3,500 \\ \frac{4,950-x}{1,450} & ; \quad 3,500 \leq x \leq 4,950 \\ 1 & ; \quad x = 4,950 \end{cases} \tag{16}$$

$$\mu_{HIGH-STOCK}(x) = \begin{cases} 0 & ; \quad x \leq 3,500 \\ \frac{x-3,500}{2,900} & ; \quad 3,500 \leq x \leq 6,400 \\ 1 & ; \quad x \geq 6,400 \end{cases} \tag{17}$$

The geometric representation of the membership functions of the three fuzzy sets for supply is presented in Figure 2(b). Next, the membership function for the production variable is defined as follows:

$$\mu_{LOW-PRODUCTION}(x) = \begin{cases} 1 & ; \quad x \leq 21,200 \\ \frac{27,000-x}{5,800} & ; \quad 21,200 \leq x \leq 27,000 \\ 0 & ; \quad x \geq 27,000 \end{cases} \tag{18}$$

$$\mu_{MEDIUM-PRODUCTION}(x) = \begin{cases} 0 & ; \quad x \leq 24,100 \; or \; x \geq 29,900 \\ \frac{x-24,100}{2,900} & ; \quad 24,100 \leq x \leq 27,000 \\ \frac{29,000-x}{2,900} & ; \quad 27,000 \leq x \leq 29,900 \\ 1 & ; \quad x = 29,900 \end{cases} \tag{19}$$

$$\mu_{HIGH-PRODUCTION}(x) = \begin{cases} 0 & ; \quad x \leq 27,000 \\ \frac{x-27,000}{5,800} & ; \quad 27,000 \leq x \leq 32,800 \\ 1 & ; \quad x \geq 32,800 \end{cases} \tag{20}$$

And the geometric representation of the membership functions of the three fuzzy sets for the productions is presented in Figure 2(c).

As an illustration, take the data from Table 1 on 2 January 2022, where there were 18,400 pieces of demand, 2,200 pieces of supply and 22,400 pieces of production. Then it can be seen from Table 2 that this demand is part of the low-demand fuzzy set

DOI: https://doi.org/10.30812/varian.v5i2.1816

category, the supply is in of low-and-medium-stock fuzzy set, while the production is in the low-production fuzzy set category. By using the membership functions (12), (15), (16) and (18), the degrees of membership for each of these numbers have also been obtained. Their geometric representations can be seen in Figure 2.

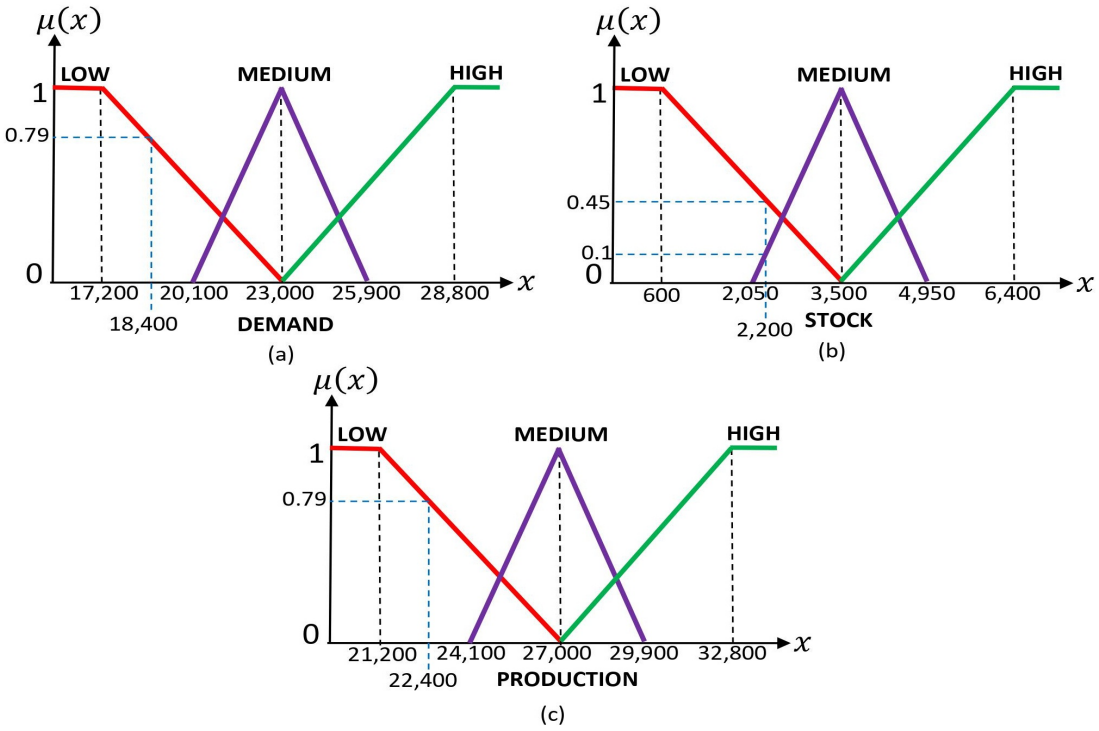


Figure 2. Fuzzification Stage

2. Inferencing (Rule based)

At this stage, rules are made which will be used as a reference in determining the output of the fuzzy system. The rule in question is in the form of If...And...Then... which is a combination of 3 variables (2 input variables and 1 output variable). At the stage of forming fuzzy sets, it is known that each variable has 3 fuzzy sets, namely Low, Medium and High fuzzy sets. So the number of rules that can be made in this case is obtained from $3^3 = 27$ rules. However, to simplify the calculations and to make the rules more representative, the rules are based on the data that has been obtained (Table 1). So, based on the data in Table 1, the previous 27 rules were reduced to 18 rules that represent every situation that might occur at the Bintang Oesapa Tofu Factory.

With the help of Matlab R2013a software, the 18 rules are as follows:

$[R1]$: If (DEMAND is **L**) and (STOCK is **L**) then (PRODUCTION is **L**)
$[R2]$: If (DEMAND is **L**) and (STOCK is **L**) then (PRODUCTION is **M**)
$[R3]$: If (DEMAND is **L**) and (STOCK is M) then (PRODUCTION is **L**)
$[R4]$: If (DEMAND is **L**) and (STOCK is M) then (PRODUCTION is **M**)
$[R5]$: If (DEMAND is **L**) and (STOCK is **H**) then (PRODUCTION is **L**)
$[R6]$: If (DEMAND is **L**) and (STOCK is **L**) then (PRODUCTION is **L**)
$[R7]$: If (DEMAND is **M**) and (STOCK is **L**) then (PRODUCTION is **M**)
$[R8]$: If (DEMAND is **M**) and (STOCK is **L**) then (PRODUCTION is **H**)
$[R9]$: If (DEMAND is **M**) and (STOCK is **M**) then (PRODUCTION is **M**)
$[R10]$: If (DEMAND is **M**) and (STOCK is **M**) then (PRODUCTION is **H**)
$[R11]$: If (DEMAND is **M**) and (STOCK is **H**) then (PRODUCTION is **L**)
$[R12]$: If (DEMAND is **M**) and (STOCK is **H**) then (PRODUCTION is **M**)
$[R13]$: If (DEMAND is **H**) and (STOCK is **L**) then (PRODUCTION is **M**)
$[R14]$: If (DEMAND is **H**) and (STOCK is **L**) then (PRODUCTION is **H**)
$[R15]$: If (DEMAND is **H**) and (STOCK is **M**) then (PRODUCTION is **M**)

DOI: https://doi.org/10.30812/varian.v5i2.1816

$[R16]$: If (DEMAND is **H**) and (STOCK is **M**) then (PRODUCTION is **H**)
$[R17]$: If (DEMAND is **H**) and (STOCK is **H**) then (PRODUCTION is **L**)
$[R18]$: If (DEMAND is **H**) and (STOCK is **H**) then (PRODUCTION is **M**)
where **L** = Low, **M** = Medium and **H** = High.

For example, at the fuzzification stage, the degree of membership of low-demand for $x = 18,400$ is 0.79, the degree of membership of low-stock for $x = 2,200$ is 0.1, the degree of membership of medium-stock for $x = 2,200$ is 0.45 and The degree of membership of the low-production for $x = 22,400$ is $0,79$. Based on the fuzzy set that corresponds to each data, there are 2 possible rules, namely $[R1]$ and $[R3]$. Next, the $\alpha - predicate$ value will be searched for each rule using the Min function in the equation 6.

$[R1]$: If (DEMAND is **L**) and (STOCK is **L**) then (PRODUCTION is **L**)

$$\begin{aligned} \alpha - predicate_1 &= \mu_{LOW-DEMAND \cap LOW-STOCK} \\ &= \min\{\mu_{LOW-DEMAND}(18,400), \mu_{LOW-STOCK}(2,200)\} \\ &= \min(0.79, 0.45) \\ &= 0.45 \end{aligned}$$

Because the output of $[R1]$ is Low-Production then by using the equation (18) we get $z_1 = 24,390$. The geometric interpretation of the calculation $[R1]$ can be seen in Figure 3.

$[R3]$: If (DEMAND is **L**) and (STOCK is M) then (PRODUCTION is **L**)

$$\begin{aligned} \alpha - predicate_2 &= \mu_{LOW-DEMAND \cap MEDIUM-STOCK} \\ &= \min\{\mu_{LOW-DEMAND}(18,400), \mu_{MEDIUM-STOCK}(2,200)\} \\ &= \min(0.79, 0.1) \\ &= 0.1 \end{aligned}$$

Because the output of $[R3]$ is Low-Production then by using the equation (18) we get $z_2 = 26,420$. The geometric interpretation of the calculation $[R3]$can be seen in Figure 3.

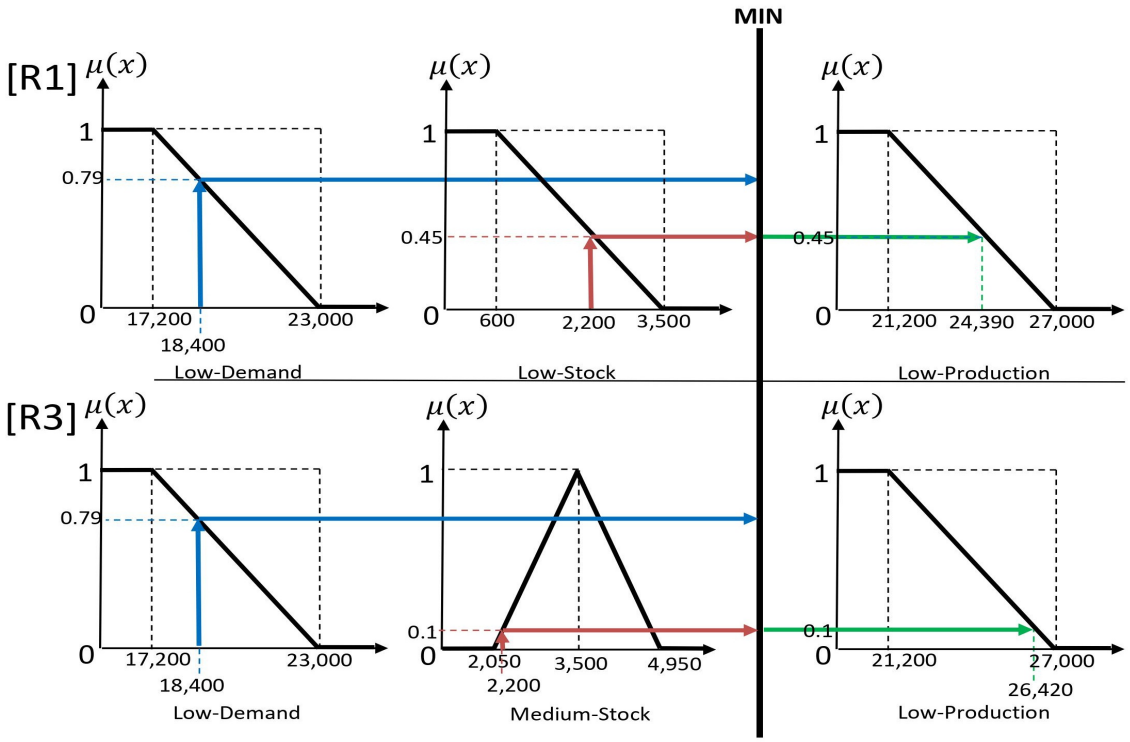


Figure 3. Inferencing Stage

From the calculation process, 2 values of z are obtained, then in the defuzzification stage a single value of z will be calculated which is the output of the data on January 2, 2022.

3. Defuzzification

In this stage, the final fuzzy data is transformed into crisp data. The defuzzification process is performed by using the five methods mentioned above. For example, based on the process of fuzzification and inferencing of data on January 2, 2022, we get 2 values of z namely $z_1 = 24,390$ with $\mu_{z_1} = 0,45$ and $z_2 = 2,420$ with $\mu_{z_2} = 0,1$. The defuzzification process generates $z* = 23,400$ with Centroid method, $z* = 23,200$ with Bisector method, $z* = 22,900$ with MOM method, $z* = 27,500$ with LOM method, and $z* = 18,400$ with SOM method. The crisp data obtained from the defuzzification is the compared to the actual data in Table 1. The comparison result is presented in Table 3.

Table 3. Comparison Between Calculations Obtained Through FIS Methods and The Actual Data

No.	Date	Actual Production	Mamdani FIS Method				
			Centroid	Bisector	MOM	LOM	SOM
1	Jan 2, 2022	22,400	23,400	23,200	22,900	27,500	18,400
2	Jan 3, 2022	24,200	23,500	23,400	23,200	27,900	18,400
3	Jan 4, 2022	29,200	27,800	27,800	28,100	32,800	23,400
4	Jan 5, 2022	28,600	28,000	27,800	28,700	32,800	24,400
5	Jan 6, 2022	29,400	30,000	30,100	30,900	32,800	28,900
6	Jan 7, 2022	22,400	23,200	23,400	22,400	26,500	18,400
7	Jan 8, 2022	21,200	23,100	23,400	22,200	26,000	18,400
8	Jan 9, 2022	27,000	27,800	27,800	28,300	32,800	23,700
9	Jan 10, 2022	32,600	30,400	30,600	32,100	32,800	31,400
10	Jan 11, 2022	26,200	23,400	23,400	22,900	27,500	18,400
11	Jan 12, 2022	30,800	29,900	29,900	30,600	32,800	28,300
12	Jan 13, 2022	29,000	27,900	27,800	28,500	32,800	24,200
13	Jan 14, 2022	28,200	27,900	27,800	28,300	32,800	23,900
14	Jan 15, 2022	28,400	25,600	27,800	28,700	32,800	24,400
15	Jan 16, 2022	29,000	27,800	27,800	28,200	32,800	23,600
16	Jan 17, 2022	31,044	29,900	29,900	30,600	32,800	28,300
17	Jan 18, 2022	32,180	30,300	30,500	31,900	32,800	30,900
18	Jan 19, 2022	22,600	23,400	23,400	23,100	27,800	18,400
19	Jan 20, 2022	32,400	29,900	29,900	30,600	32,800	28,300
20	Jan 21, 2022	28,400	28,100	27,800	29,600	32,800	25,500
21	Jan 22, 2022	27,500	27,700	27,800	28,000	32,800	23,300
22	Jan 23, 2022	31,100	25,600	25,600	25,600	25,600	25,600
23	Jan 24, 2022	32,800	27,700	27,800	28,000	32,800	23,200
24	Jan 25, 2022	23,600	23,300	23,400	22,900	27,300	18,400
25	Jan 26, 2022	29,600	30,000	30,100	30,900	32,800	28,900
26	Jan 27, 2022	26,400	27,900	27,800	28,300	32,800	23,900
27	Jan 28, 2022	30,800	29,800	29,900	30,500	32,800	28,200
28	Jan 29, 2022	31,000	23,400	23,400	23,100	27,800	18,400
29	Jan 30, 2022	26,200	28,000	27,800	28,700	32,800	24,400
30	Jan 31, 2022	22,800	25,600	25,700	25,600	32,800	18,400
31	Feb 1, 2022	30,000	30,400	30,600	32,100	32,800	31,400
32	Feb 2, 2022	32,750	30,100	32,200	31,100	32,800	29,500
33	Feb 3, 2022	32,542	30,100	30,400	31,400	32,800	29,900
34	Feb 4, 2022	22,200	23,200	23,400	22,400	26,300	18,400
35	Feb 5, 2022	23,600	23,300	23,400	22,700	27,000	18,400
36	Feb 6, 2022	26,800	27,900	27,800	28,300	32,800	23,900
37	Feb 7, 2022	31,000	29,900	29,900	30,600	32,800	28,300
38	Feb 8, 2022	29,000	23,400	23,400	22,900	27,500	18,400
39	Feb 9, 2022	30,000	30,000	30,100	30,900	32,800	28,900
40	Feb 10, 2022	27,400	27,800	27,800	28,100	32,800	23,400

Based on the calculation comparison process in Table 3, several comparisons of the five Mamdani FIS defuzzification methods were obtained based on the level of accuracy, error in analysis, manual calculation time and calculation complexity level. The comparison is presented in Table 4.

DOI: https://doi.org/10.30812/varian.v5i2.1816

Table 4. Comparison of The Five Mamdani FIS Defuzzification Methods

No.	Indicator	Centroid	Bisector	MOM	LOM	SOM
1	Accuracy in analyzing	$94,17\%$	$76,83\%$	$94,73\%$	$86,61\%$	$85,27\%$
2	Error in analyzing	$5,83\%$	$23,17\%$	$5,27\%$	$13,39\%$	$14,73\%$
3	Time	The manual calculation process for the centroid defuzzification method takes a long time.	The manual calculation process for the bisector defuzzification method does not take too long, when compared to the centroid.	The manual calculation when process for the MOM defuzzification method requires less time, compared to the centroid and bisector.	The manual calculation process for the SOM defuzzification method requires less time, when compared to centroids, bisector and MOM.	The manual calculation process for the LOM defuzzification method takes the same time as the SOM method.
4	Calculation	The calculation of this method is quite complicated.	The calculation of this method is quite complicated.	The calculation of this method is somewhat simpler when compared to the centroid and bisector.	The calculation of this method is the simplest when compared to centroids, bisectors, and MOM.	The calculation of this method is as simple as the SOM method.

The level of accuracy presented in Table 4 was calculated using MAPE formula in equation (11). Based on the comparison result in Table 4, for the case of determining the amount of tofu production at the Bintang Oesapa Tofu Factory by considering the stock and demand obtained that the MOM method has a higher level of accuracy than the Centroid method with the difference in accuracy between the two is $0,56\%$. Because each case has its own characteristics, it is better if the Mamdani method does not use the centroid method directly as the defuzzification method, but instead uses the centroid method as a defuzzification method it is necessary to review other defuzzification methods first. Based on the results of the calculations in Table 4 it can be said that that for this case, the accuracy of the calculation results of the Mamdani method ranges from $76\% - 95\%$ which can be said to be very good or very similar to the decision making by the Bintang Oesapa Tofu Factory. This is in line with the statement (Kaur and Kaur, 2012) in their research, Mamdani method is widely accepted for capturing expert knowledge. It allows us to describe the expertise in more intuitive, more human-like manner.

E. CONCLUSION AND SUGGESTION

From data processing and relevant journals, we conclude that the best FIS Mamdani defuzzification method to predict the amount of tofu production at the Bintang Oesapa Tofu Factory was the Mean of Maximum (MOM) defuzzification method. The first reason is that the accuracy value of the other 5 methods is 94.73% and the value of error or small error is 5.27%. Bisides that, the manual calculations require a substantial amount of time and are not complicated.

Thus, when carrying out the process of determining the production of tofu using the Mamdani fuzzy inference system method, it is highly recommended to use the MOM method in the defuzzification process. Not only in the case of tofu production, this method is also highly recommended for determining the production of other goods.

ACKNOWLEDGEMENT

We would like to thank the leaders of Timor University and Bumigora University for their support. We would also like to thank Bintang Oesapa Tofu Factory's Manager who made this research possible by sharing us their data.

REFERENCES

Abrori, M. and Primahayu, A. (2015). Aplikasi Logika Fuzzy Metode Mamdani dalam Pengambilan Keputusan Penentuan Jumlah Produksi. *Kaunia: Integration and Interconnection Islam and Science*, 12(2):91–99.

Azmi, U., Hadi, Z. N., and Soraya, S. (2020). ARDL METHOD: Forecasting Data Curah Hujan Harian NTB. *Jurnal Varian*, 3(2):73–82.

Bahtiar, R. (2021). Dampak Pandemi Covid-19 Terhadap Sektor Usaha Mikro, Kecil, dan Menengah serta Solusinya. *Info Singkat: Kajian Singkat Terhadap Isu Aktual Dan Strategis*, 13(10):19–24.

Kaur, A. and Kaur, A. (2012). Comparison of Mamdani-Type and Sugeno-Type Fuzzy Inference Systems for Air Conditioning System. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(2):323–325.

Mada, G., Lobo, M., and Pangaribuan, R. (2021). Analisis Perbandingan Fuzzy Inference System Mamdani dan Fuzzy Inference System Tsukamoto Dalam Penentuan Jumlah Produksi Pada Ud. Batako "Cabang Farmasi" Kupang. *Prosiding Seminar Nasional Pendidikan Matematika Universitas Timor*, 1:1–9.

Priyo, W. (2017). Penerapan Logika Fuzzy Dalam Optimasi Produksi Barang Menggunakan Metode Mamdani. *Jurnal Ilmiah: SoulMath*, 5(1):14–21.

Rahakbauw, D. (2015). Penerapan Logika Fuzzy Metode Sugeno untuk Menentukan Jumlah Produksi Roti Berdasarkan Data Persediaan dan Jumlah Permintaan (Studi Kasus: Pabrik Roti Sarinda Ambon). *Barekeng: Jurnal Ilmu Matematika dan Terapan*, 9(2):121–134.

Rahakbauw, D., Rianekuay, F., and Lesnussa, Y. (2019). Penerapan Metode Fuzzy Mamdani Untuk Memprediksi Jumlah Produksi Karet (Studi Kasus: Data Persediaan Dan Permintaan Produksi Karet Pada Ptp Nusantara XIV (Persero) Kebun Awaya, Teluk Elpaputih, Maluku-Indonesia). *Jurnal Ilmiah Matematika dan Terapan*, 16(1):119–127.

Rianto, E. and Manurung, J. (2022). Total Prediction Decision Support System Bakery and Cake Production Using Mamdani Fuzzy Method. *Jurnal Mandiri IT*, 10(2):74–79.

Sahulata, E., Wattimanela, H. J., and Noya Van Delsen, M. S. (2020). Penerapan Fuzzy Inference System Tipe Mamdani Untuk Menentukan Jumlah Produksi Roti Berdasarkan Data Jumlah Permintaan Dan Persediaan (Studi Kasus Pabrik Cinderela Bread House Di Kota Ambon). *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 14(1):079–090.

Santya, L., Miftah, M., and Mandala, V. (2017). Penerapan Metode Fuzzy Mamdani untuk Pendukung Keputusan Penentuan Jumlah Produksi Lantak Si Jimat. *Jurnal Rekayasa Teknologi Nusaputra*, 2(2):27–38.

Sari, Y. R. (2021). Penerapan Logika Fuzzy Metode Mamdani dalam Menyelesaikan Masalah Produksi Garam Nasional. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 8(1):341–356.

Susetyo, J., Asih, E. W., and Raharjo, H. (2020). Optimalisasi Jumlah Produksi Menggunakan Fuzzy Inference System Metode Min-Max. *JURNAL REKAYASA INDUSTRI (JRI)*, 2(1):8–14.

Sutikno (2011). Perbandingan Metode Defuzzifikasi Aturan Mamdani Pada Sistem Kendali Logika Fuzzy (Studi Kasus Pada Pengaturan Kecepatan Motor DC). *Jurnal Masyarakat Informatika*, 2(2):27–38.

Wardani, A. R., Nasution, Y. N., and Amijaya, F. D. T. (2017). Aplikasi Logika Fuzzy Dalam Mengoptimalkan Produksi Minyak Kelapa Sawit Di PT. Waru Kaltim Plantation Menggunakan Metode Mamdani. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, 12(2):94.

Yusuf, A., Widayat, E., and Hatip, A. (2017). Penerapan Logika Fuzzy Dalam Memperkirakan Jumlah Produksi Telur Terhadap Permintaan Pasar. *Limits: Journal of Mathematics and Its Applications*, 14(2):169–193.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 149–160
e-ISSN: 2581-2017

Expansion of Stock Portfolio Risk Analysis Using Hybrid Monte Carlo-Expected Tail Loss

Wisnowan Hendy Saputra[1], **Ika Safitri**[2]
[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia
[2]Department of Actuarial Science, Institut Teknologi Sepuluh Nopember, Indonesia

Article Info

Article history:

Received : 02-28-2022
Revised : 04-21-2022
Accepted : 04-26-2022

Keywords:

ARIMA-GARCH;
Expected Tail Loss;
Monte Carlo;
Multi-Objective Optimization;
Optimized Stock Portfolio.

ABSTRACT

Monte Carlo-Expected Tail Loss (MC-ETL) is the new expansion method that combines simulation and calculation to measure investment risk. This study models US stock prices using ARIMA-GARCH and forms an optimized portfolio based on Multi-Objective that aims to analyze the portfolio investment return. The next portfolio return will be simulated using the Monte Carlo (MC) method, measured based on the Expected Tail Loss (ETL) calculation. The optimized portfolio comprises 5 US stocks from 10 years of data, with the biggest capitalization market on February 25, 2021. MSFT has the most considerable weight in the optimized portfolio, followed by GOOG, AAPL, and AMZN, whereas TSLA shares have negligible weight. Based on the simulation result, the optimized portfolio has the smallest ETL value compared to its constituent stocks, which is ± 0.029 or about 2.9%. This value means that the optimized portfolio is concluded as an investment choice for investors with a low level of risk.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1813

Corresponding Author:

Wisnowan Hendy Saputra
Department of Statistics, Institut Teknologi Sepuluh Nopember.
Email: wisnowanwhs@gmail.com

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The state of a country's economy will be directly proportional to the state of the capital market. A capital market is a meeting place for investors as parties who have excess funds and issuers as parties who need funds (Wardiyah, 2017). At the end of 2020, International Financial Statistics stated that the United States was one of the largest and most stable economies. Thus, these conditions significantly affected the price stability of financial instruments in the United States capital market. One of the financial instruments in the United States capital market that has price stability is the type of stock. In addition, stocks are the most sought-after financial instrument by investors (Sururi et al., 2021). That is most likely why many investors are interested in investing in US stocks.

Investment activities certainly cannot be separated from risks. When viewed from the level of risk, investment in the capital market is known to have a higher level than other instruments such as deposits, bonds, mutual funds, and others (Bellofatto et al., 2018). In investing, investors aim to get high returns with low risk (Mayuni and Suarjaya, 2018). In actual practice, this objective contradicts the relationship between return and risk, which states that the higher the expected return, the higher the risk to be faced. Therefore, it is necessary to invest in a model that can describe the volatility of returns.

The return volatility of a stock will explain the risk of a given stock return. Research related to investment volatility modeling has been conducted by (Azmi and Syaifudin, 2020). This study applies the ARIMA-GARCH model to estimate commodity prices, concluding that the model obtained has an error rate of less than 2% for all types of commodities used (Azmi and Syaifudin, 2020). The process of minimizing the volatility of stock returns can be done by forming an investment portfolio (Salim and Rizal, 2021).

Stock Portfolio Optimization Approach using Multi-Objective Optimization has been proposed by Chen et al. (2018). The study uses a real dataset to test the effectiveness of the proposed approach (Zheng and Chen, 2013).

The Expected Tail Loss (ETL) method was used to estimate risk on the Malaysian market index by Nguyen and Huynh (Nguyen and Huynh, 2019), with the result that the ETL risk value was 2.26%. Then a similar study using ETL to compare cryptocurrency risk by Feng et al. (Feng et al., 2018) resulted in 7.92%, 14.12%, and 1.86% ETL values for Bitcoin, Ethereum, and S&P500, respectively. Conditional Value at Risk (CVaR), another name for ETL, is used by Arif and Sohail (Arif and Sohail, 2020) to estimate the risk of the Pakistan stock exchange with an optimal investment ETL value of 7.52%. Carried out the ETL method during the covid-19 pandemic on Nordic stock markets, which stated that the ETL value was over 20% on OMXS30 data and over 15% on OMXH25 and OMXS30 during the Covid-19 crisis.

Even though the investment portfolio has been formed, the risk of loss will inevitably remain. This risk can be predicted by obtaining the characteristics and distribution of investment returns. This study has an update that adds a simulation process using Monte Carlo in optimized portfolio returns to get a stable return. After iteration is done, Expected Tail Loss (ETL) is used in the optimized portfolio return to measure investment risk.

B. LITERATURE REVIEW

1. US Big-Cap Stocks

A stock market is a meeting place for sellers and buyers involved in stock economic transactions. The stock market is also an integral and inseparable part of a country's economy. The stock market is also an integral and inseparable part of a country's economy. The United States is one of the countries with the largest and most developed financial market, one of the instruments is stocks (Zheng and Chen, 2013).

There are three major stock market indexes in the United States: the DJIA, S&P 500, and NASDAQ. The DJIA is one of the most important economic indices globally, with 30 of the largest blue-chip multinational companies' stocks included in its components. However, the DJIA index is narrower or less comprehensive in scope than the S&P 500 index (Jareño et al., 2016). The S&P 500 index uses a market capitalization methodology. It has covered 500 companies in large-cap market sectors, so the S&P 500 index is usually used as a benchmark for the performance of large-cap stocks and the influence of US stock prices on other countries.

On the other hand, the NASDAQ stock index includes many small, high-growth stocks. This makes the NASDAQ stock more volatile when compared to the other two stock indices (Zheng and Chen, 2013). Companies that are members of the NASDAQ stock index are mostly in the technology sector. The NASDAQ stock index is compiled based on the capitalization method, which determines the weight of each stock. From this explanation, the stock index with large capitalization is the NASDAQ stock index.

2. Portfolio

A portfolio is a term for a combination of some securities. Portfolios mainly deal with the problem of how to allocate one's capital to a large number of securities so that investments can bring about the most profitable returns (Chandra, 2017). Portfolio analysis is a quantitative method for selecting the optimized portfolio to maximize returns and minimize risk in various uncertain environments. To choose the optimized portfolio, we must first answer the questions "what is the portfolio's rate of return" and "what is the risk of the portfolio". Not only about how large the capital is, but preliminary evidence to suggest that the portfolio's overall performance may be favorable because the experienced low correlation of the impact of investments to traditional markets reduces portfolio risk and increases sustainability (Brandstetter and Lehner, 2015).

Let the asset weight vector $\vec{x} = [x_1, x_2, \ldots, x_n]^T$ with x_i as the weight of asset i in the portfolio and the expected return for each asset in the portfolio is expressed in the vector form $\vec{r} = [r_1, r_2, \ldots, r_n]^T$ with r_i as the mean return of asset i in the portfolio. The portfolio's expected return is the weighted average of the returns on individual assets is expressed as

$$x_p = \vec{r}^T \vec{x} = \sum_{i=1}^{n} x_i r_i \tag{1}$$

The variance and covariance of individual assets are characterized by the variance-covariance matrix V =

$\begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$ with σ_{ii} is the variance of asset i and $\sigma_{ij} = \sigma_{ji}$ is the covariance between asset i and j. The portfolio variance is expressed as

$$\sigma_p^2 = \vec{x}^T V \vec{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{ij} \tag{2}$$

where n is the number of assets in the portfolio (Giemza, 2021).

For the most part, investors or companies, and financial managers (whether they control capital projects or financial services) may be responsible for many investments. Therefore, extending the analysis to the return of portfolios with more than two assets is important. For a multi-asset portfolio, the number of assets equals n, and x_i represents the proportion of funds invested in each. The portfolio return is as follows

$$R_p = \sum_{i}^{n} x_i R_i \tag{3}$$

where R_i is the i-th asset return and R_p is the portfolio return (Kulali, 2016).

3. ARIMA-GARCH Model

Let $\{z_t\}$ represent the observed time series observation data at spaced times t, also let $\tilde{z}_t = z_t - \mu$ be the series of deviations from μ. In addition, let $\{a_t\}$ represent the unobserved white noise series, i.e., a sequence of identically distributed independent random variables with zero mean. In most studies, the assumption of independence can be replaced by a weaker assumption that $\{a_t\}$ is an uncorrelated random variable. The Mixed Autoregressive Moving Average Model (ARMA) assumes that the time series observation data is partly autoregressive and partly moving average. In general, the ARMA (p, q) model is stated as follows.

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{4}$$

or

$$\theta(B)\tilde{z}_t = \theta(B) a_t \tag{5}$$

where B is the backward shift operator such that $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, ϕ and θ are Autoregressive (AR) and Moving Average (MA) parameter. The model uses $p + q + 2$ unknown parameters such us $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_a^2$, that are estimated from the observation data (Box et al., 2015).

Many series encountered in an industry or business (e.g., stock prices) exhibit nonstationary behavior and typically do not vary about a fixed average. The process provides a robust model for describing stationary and nonstationary time series and is called an Autoregressive Integrated Moving Average process, order (p, d, q), or ARIMA(p, d, q) process. The process can be written as follows

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{6}$$

where $w_t = (1 - B)^d z_t$ and note to change w_t to $z_t - \mu$ when $d = 0$ (Box et al., 2015).

For a stationary ARIMA process, the unconditional mean of the series is constant over time, while the conditional mean of $E[z_t | F_{t-1}]$ varies as a function of the previous observations. Parallel to this, the ARCH model assumes that the unconditional variance of the error process is constant over time but allows the conditional variance of a_t to vary as a function of the past squared error. Let $\sigma_t^2 = Var(a_t | F_{t-1})$ denote the conditional variance, given the last F_{t-1}, the basic model ARCH(s) can be formulated as

$$a_t = \sigma_t e_t \tag{7}$$

where $\{e_t\}$ is a sequence of iid random variables with mean zero and variance one. So that

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_s a_{t-s}^2 \tag{8}$$

where α is ARCH parameter with $\alpha_0 > 0, \alpha_i \geq 0$ for $i = 1, 2, \cdots, s - 1$, and $\alpha_s > 0$ (Box et al., 2015).

DOI: https://doi.org/10.30812/varian.v5i2.1813

The ARCH model has the disadvantage of often requiring a sequence of high lags to describe the evolution of volatility over time adequately. An extension of the ARCH model is called generalized ARCH or GARCH. The GARCH(s, r) is given by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{s} \alpha_i a_{t-i}^2 + \sum_{j=1}^{r} \beta_j \sigma_{t-j}^2 \tag{9}$$

where α and β are GARCH parameter with $\alpha_0 > 0, \alpha_i \geq 0$ for $i = 1, 2, \cdots, s-1, \alpha_s \geq 0, \beta_j \geq 0$ for $j = 1, 2, \cdots, r-1$, and $\beta_r > 0$ (Box et al., 2015).

4. Multi-Objective Optimization

The concept of multi-objective optimization was first introduced by the French-Italian economist Pareto (Lampio, 2018). This theory combines all objectives into one objective function, and the standard solution method for minimizing the total objective is applied as follows.

$$\min F(x) = [f_1(x), f_2(x), \cdots, f_m(x)] \tag{10}$$

subhect to:

$$G(x) = [g_1(x), g_2(x), \cdots, g_k(x)] < 0 \tag{11}$$

$$H(x) = [h_1(x), h_2(x), \cdots, h_l(x)] = 0 \tag{12}$$

where $F(x)$ is a vector of m objective functions with $f_i(x)$ is the objective function i for $i = 1, 2, \ldots, m, G(x)$ is a vector of k inequality constraints with $g_i(x)$ is the inequality constraint i for $i = 1, 2, \cdots, k, H(x)$ is a vector of l equality constraints with $h_i(x)$ is the equality constraint i for $i = 1, 2, \cdots, l,$ and x is a vector of decision variables with $x = [x_1, x_2, \cdots, x_n]$ (Liang et al., 2016).

5. Monte Carlo Simulation

Suppose the simulation in this study aims to estimate an unknown quantity l based on the $\widehat{l}$ estimator, which is a function of the data generated by the simulation. The general condition is when l is the expectation of the output variable Y from the simulation. For example, a simulation experiment that is run repeatedly produces independent copies of $Y_1, \cdots, Y_N$ of Y. A statistical estimator of l then the sample mean is as follows

$$\widehat{l} = \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{13}$$

This estimator is unbiased in the sense that $E\left[\widehat{l}\right] = l$. Moreover, according to the law of large numbers, $\widehat{l}$ converges to l with probability one as $N \to \infty$. Note that the estimator is viewed as a random variable. Certain results or observations are called estimates (numbers), often denoted by the same letter (Ling and Rubinstein, 1983).

6. Expected Tail Loss

The Expected Tail Loss (ETL), also called Conditional Value at Risk (CVaR), interprets the expected loss (in present value terms) given that the loss exceeds the Value at Risk (VaR). VaR indicates the maximum loss of an asset or portfolio of assets in a specific confidence interval (Siswono et al., 2021). The ETL risk metric is more informative than VaR because VaR does not measure the extent of extraordinary losses. VaR states the level of loss that we believe will not be exceeded: it does not tell us how much could be lost if VaR is exceeded. However, the ETL tells us how much loss we can expect, given that the VaR is exceeded. ETL provides a complete picture of portfolio risk than simply reporting VaR alone. This means that ETL is a better risk metric for regulatory and economic capital allocation. Suppose the distribution function F_X for a certain probability, the VaR value of X denoted $VaR(F_X, p)$ can be calculated as follows:

$$VaR(F_X, \alpha) = VaR_\alpha(X) = -F_X^{-1}(1-\alpha) \tag{14}$$

DOI: https://doi.org/10.30812/varian.v5i2.1813

Thus, the ETL value, which is expressed as a measure of portfolio risk, can be written as follows:

$$ETL_\alpha = -E\left[X|X < -VaR_\alpha(X)\right] \tag{15}$$

$$ETL_\alpha = -\frac{\int_{\infty}^{VaR} xF_X(x)\,dx}{F_X(-VaR(X))} \tag{16}$$

where α is the quantil of the distribution of X (Airouss et al., 2018).

C. RESEARCH METHOD

This study takes five US stocks that have the largest market capitalization on February 25, 2021, namely Apple (AAPL), Microsoft (MSFT), Alphabet/Google (GOOG), Amazon (AMZN), and Tesla (TSLA). The research was conducted using ten years of data from January 2, 2012, to December 31, 2021, obtained through finance.yahoo.com. The entire analysis was carried out using R software, especially the PerformanceAnalytics packages, to perform Monte Carlo simulation and calculate the Expected Tail Loss (MC-ETL).

The steps in the research are as follows:

1. Analyze descriptive statistics to see the distribution of each stock data.
2. Modeling the daily closing price of each stock using ARIMA-GARCH.
3. Calculating the mean-return of each stock based on the ARIMA-GARCH model.
4. Forming an optimized portfolio using the Multi-Objective Optimization Method.
5. Perform a Monte Carlo simulation on optimized portfolio returns.
6. Calculate the ETL value of the optimized portfolio for each simulation.
7. Calculate the mean value of ETL from all simulations that have been carried out.
8. Calculating the individual ETL value of each portfolio stock.
9. Comparing the optimal ETL portfolio value with the individual ETL of each constituent stock.
10. Interpret the results obtained and draw conclusions.

D. RESULTS AND DISCUSSION

1. Data Characteristics

Data used daily closing price data for five US stocks with the largest market capitalization. The data characteristics of the data, as a sample, are presented in descriptive statistics as follows, presented in Table 1.

Table 1. Characteristics Data of The Daily Closing Price of Stock (USD)

Stock	AAPL	MSFT	GOOG	AMZN	TSLA
Market Cap	2.69T	2.229T	1.778T	1.565T	837B
Sample	2516	2516	2516	2516	2516
Mean	48.99	97.92	988.1	1245.7	140.941
Median	32.36	62.69	798.2	818.7	49.783
Minimum	13.95	26.37	278.5	175.9	4.558
Maximum	180.33	343.11	3014.2	3731.4	1229.91
Standard Deviation	38.65549	78.15647	628.0657	1052.281	237.3425
Skewness	1.567776	1.340117	1.476066	0.9001429	2.417873
Kurtosis	4.329771	3.855755	4.870672	2.538728	7.884726

From the table above, it can be seen that each stock has different statistical values or characteristics. The difference in value can lead to differences in the return and risk of each stock. As an illustration, the price movement of each stock is shown in Figure 1.

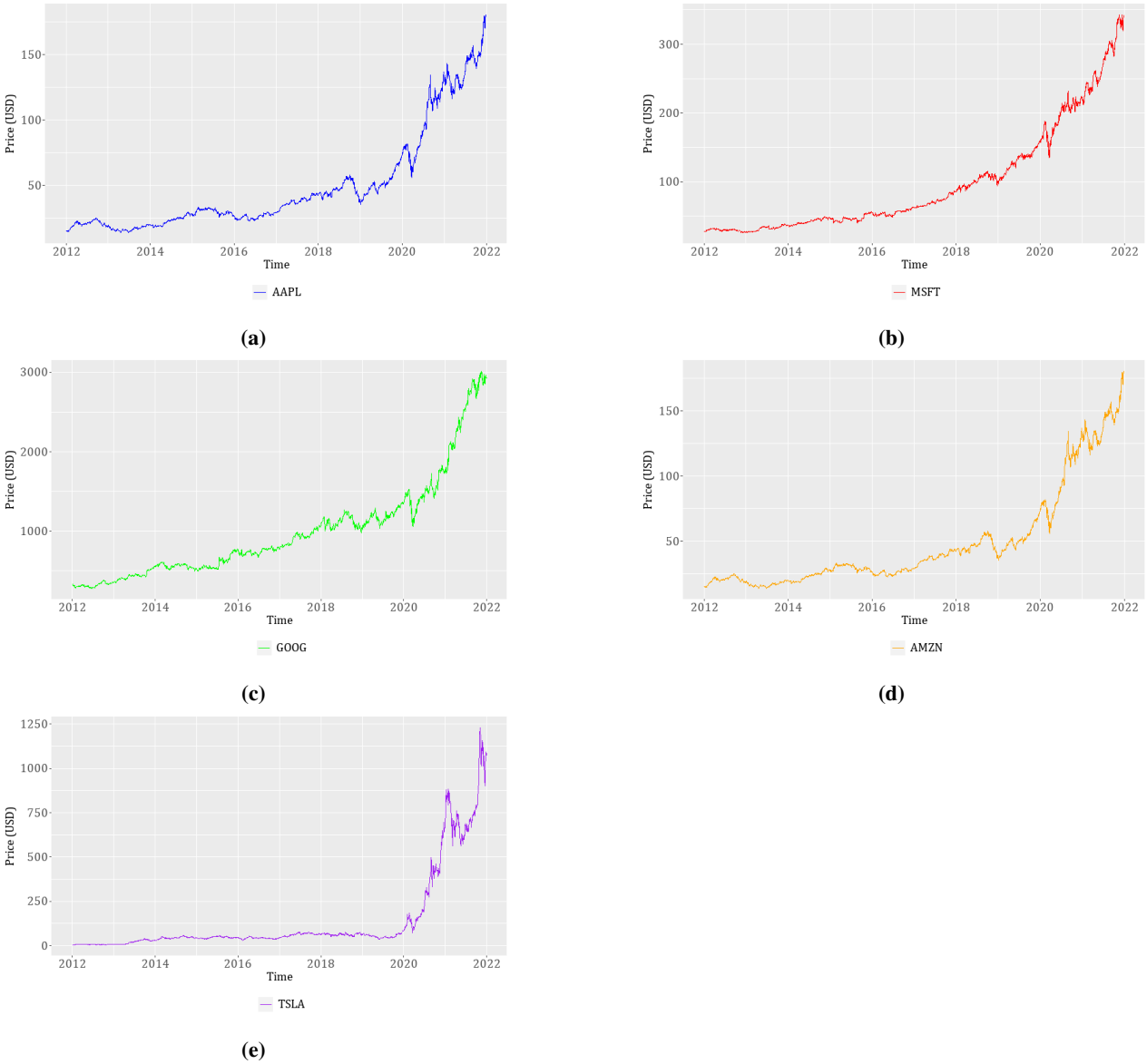


Figure 1. The stock price movement of (a) AAPL, (b) MSFT, (c) GOOG, (d) AMZN, (e) TSLA

The existence of a downward movement or graph of stock prices in Figure 1 defines that the value of a stock is experiencing a decline. These events are events that investors must avoid. Therefore, we will try to measure how much impact the price decline has on the investment value of an investor, of course, based on all price movements and volatility.

2. ARIMA-GARCH Estimation

Before doing the modeling, the first step that must be done is to test the data to obtain stationarity information from the data. The test was carried out using the Augmented Dickey-Fuller Unit Root Test. If the data is stationary, then the modeling can be continued immediately, but if the data is not stationary, a differencing process is needed to make the data stationary. With the null hypothesis that the data is not stationary, the results of the stationarity test are shown in Table 2.

Table 2. Stationary Test of Each Data

Stock	t-Statistic	P-value
AAPL	3.022902	1.0000
MSFT	3.433418	1.0000
GOOG	2.385097	1.0000
AMZN	0.345716	0.9806
TSLA	2.356797	1.0000

Table 2 indicates that there is no stationary data. Therefore, differencing is carried out before estimating the model parameters to get the stationary one.

First, the model parameters can be estimated after obtaining stationary data through the differencing process. For the best ARIMA-GARCH model based on AIC, the significant parameter model of each stock are stated in Table 3. According to formula (5) and (8). The full model can be written as

- AAPL

$$w_t = 0.050904 - 0.602387w_{t-1} + 0.152217w_{t-2} + a_t - 0.728637a_{t-1}$$
$$\sigma_t^2 = 1.039583 + 0.289830a_{t-1}^2$$

- MSFT

$$w_t = 0.103251 - 0.956880w_{t-1} - 0.774892w_{t-2} - 0.092009w_{t-3} + a_t - 0.886725a_{t-1} - 0.680193a_{t-2}$$
$$\sigma_t^2 = 3.099956 + 0.15a_{t-1}^2 + 0.05a_{t-2}^2$$

- GOOG

$$w_t = 0.507973 - 0.670364w_{t-1} + 0.064923 + a_t - 0.704250a_{t-1}$$
$$\sigma_t^2 = 2.265719 + 0.739152a_{t-1}^2$$

- AMZN

$$w_t = 1.085834 - 1.254674w_{t-1} - 0.853132w_{t-2} + a_t - 1.320794a_{t-1} - 0.902022a_{t-2}$$
$$\sigma_t^2 = 5.989786 + 0.812942a_{t-1}^2$$

- TSLA

$$w_t = 0.062952 - 0.244092w_{t-1} + 0.102440w_{t-2} + a_t - 0.162214a_{t-1}$$
$$\sigma_t^2 = 7.151434 + 1.5004a_{t-1}^2$$

DOI: https://doi.org/10.30812/varian.v5i2.1813

Table 3. Estimated Parameter of The Best ARIMA-GARCH Model for Each Stock

Stock	ARIMA-GARCH model	Parameter	Parameter Coefficient	P-Value	Parameter	Parameter Coefficient	P-Value
AAPL	ARIMA (2,1,1) -GARCH (0,1)	Constant	0.050904	0.0493	Constant	1.039583	0.0000
		AR(1)	-0.602387	0.0000	RESID2(1)	0.289830	0.0000
		AR(2)	0.152217	0.0000			
		MA(1)	0.728637	0.0000			
MSFT	ARIMA (3,1,2) -GARCH (0,2)	Constant	0.103251	0.0082	Constant	3.099956	0.0000
		AR(1)	-0.956880	0.0000	RESID2(1)	0.150000	0.0000
		AR(2)	-0.774892	0.0000	RESID2(2)	0.050000	0.0000
		AR(3)	-0.092009	0.0000			
		MA(1)	0.886725	0.0000			
		MA(2)	0.680193	0.0000			
GOOG	ARIMA (2,1,1) -GARCH (0,1)	Constant	0.507973	0.0774	Constant	2.265719	0.0000
		AR(1)	-0.670364	0.0000	RESID2(1)	0.739152	0.0000
		AR(2)	0.064923	0.0000			
		MA(1)	0.704250	0.0000			
AMZN	ARIMA (2,1,2) -GARCH (0,1)	Constant	1.085834	0.0430	Constant	5.989786	0.0000
		AR(1)	-1.254674	0.0000	RESID2(1)	0.812942	0.0000
		AR(2)	-0.853132	0.0000			
		MA(1)	1.320794	0.0000			
		MA(2)	0.902022	0.0000			
TSLA	ARIMA (2,1,1) -GARCH (0,1)	Constant	0.062952	0.7460	Constant	7.151434	0.0000
		AR(1)	-0.244092	0.0000	RESID2(1)	1.500400	0.0000
		AR(2)	0.102440	0.0000			
		MA(1)	0.162214	0.0028			

3. Optimized Portfolio

Portfolio optimization is obtained after the information on the mean return of each stock is known. The estimate of the mean stock return is obtained based on the ARIMA-GARCH model that has been formed, which is listed in Table 4.

Table 4. Estimated Mean Return

Stock	Mean Return
AAPL	0,000967466
MSFT	0,000936567
GOOG	0,000171679
AMZN	0,001068245
TSLA	0,001403081

Based on Table 4, information is obtained that TSLA has the greatest mean return with a 0,14% value. In the Multi-Objective method, the formation of each asset's weight is determined through the mean return and by considering risk or volatility. In addition, this method also considers the risk aversion value of investors. The greater investors' risk-aversion indicates that investors tend to maintain safe conditions to avoid investment risks that cause high losses. The movement of the results of calculating the weight of each stock in the optimized portfolio based on the risk aversion value is stated in Figure 2.

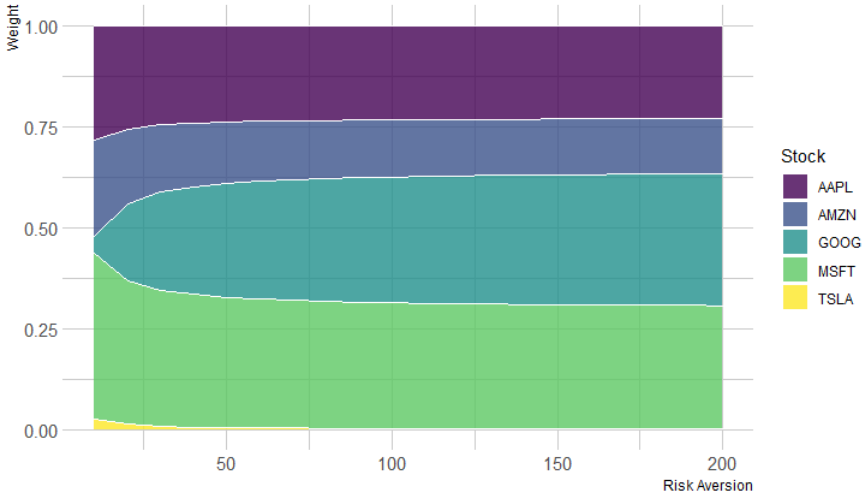


Figure 2. Porfolio Weight Based on Risk Aversion Value

We assume that an investor does not want to incur losses of more than USD 100 (USD 100 risk aversion) on this portfolio investment. Then the weight of each stock is obtained for the optimized portfolio, which is stated in Table 5.

Table 5. Weight of Each Stock in Portfolio

Stock	Weight
AAPL	0.231769937
MSFT	0.311375499
GOOG	0.311218321
AMZN	0.141717391
TSLA	0.003918852

Based on the table above, MSFT shares have the most considerable weight in the portfolio, whereas TSLA shares have negligible weight. The movement of the optimized portfolio value according to the weight is illustrated according to Figure 3.

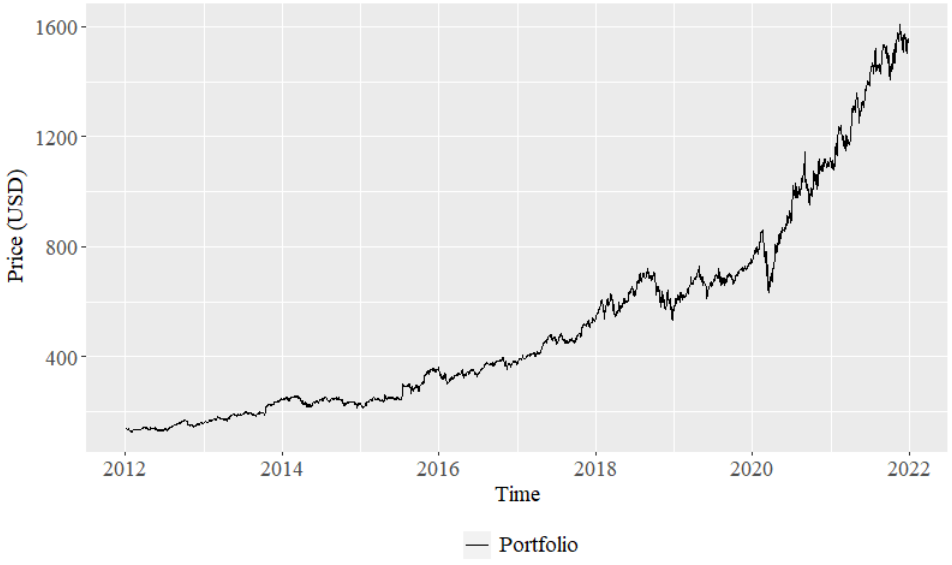


Figure 3. Portfolio Price Movement

The process of forming the optimized portfolio weight based on the Multi-Objective method shows that the weight of each

DOI: https://doi.org/10.30812/varian.v5i2.1813

stock is not much different; it is just that TSLA has a relatively minimal weight compared to other stocks. This occurrence could be due to the high volatility in TSLA shares. To clarify the allegations, obtained through the following analysis, measurement of investment risk using Expected Tail Loss (ETL).

4. Hybrid Monte Carlo-Expected Tail Loss

At this stage, the first step is to get a return from the optimized portfolio, not forgetting to calculate the mean and standard deviation. A Monte Carlo simulation was carried out based on the mean and standard deviation of optimized portfolio returns to obtain statistical stability. After just calculating Expected Tail Loss (ETL) using the historical method, namely the movement of past returns. This study uses 1000 iterations to calculate the mean; the ETL value is considered stable. The simulation results for calculating ETL are illustrated according to Figure 4.

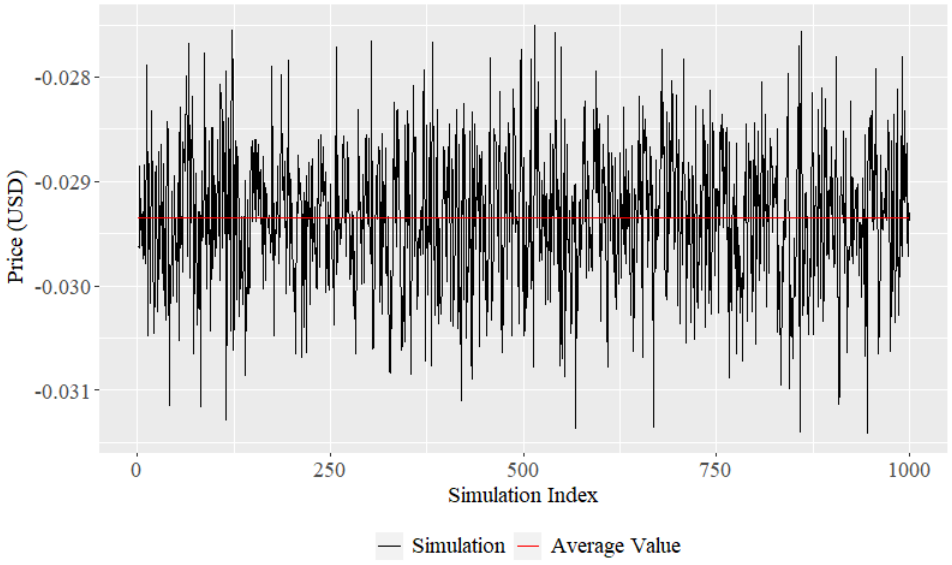


Figure 4. ETL Calculation Simulation Results

According to the picture above, it can be seen that each iteration will produce a different calculation value. Therefore, the average value of these results is considered a reasonable and stable value to measure the investment risk of the optimized portfolio.

After the 1000th iteration, ETL's mean or average value is in the range of 0.029. From these results, it can be concluded that the ETL value of the optimized portfolio is 2.9%. This value shows the maximum expected loss experienced by investors the next day after the research period. In other words, suppose an investor invests US$ 1000 of capital, then with a five percent probability, the total expected daily loss in the optimized portfolio will equal or exceed US$ 29 (2.9% of US$ 1000).

After getting the ETL value from the optimized portfolio, we will then compare it with the ETL value of each stock. The comparison results are displayed as shown in Figure 5. Based on the illustration in the figure, it can be seen that the optimized portfolio has the lowest ETL value. These results conclude that investing in an optimized portfolio provides more security for losses arising from investment risk. Thus, this study proves that investment risk modeling and analysis can obtain an optimal portfolio that provides the lowest level of risk so that the expected losses that investors can experience can be minimized.

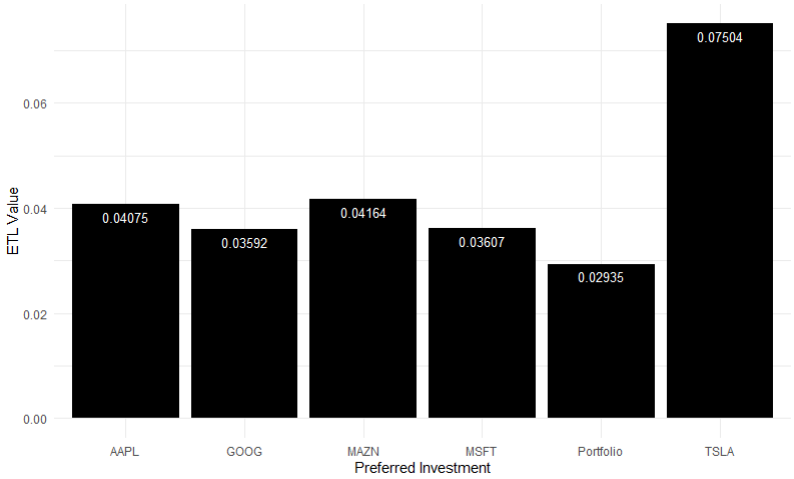


Figure 5. ETL Value of Porfolio and Each Stock

E. CONCLUSION AND SUGGESTION

Based on the results of modeling and analysis of the optimized portfolio of US big-cap Stocks based on Multi-Objective Optimization (MOO) using Autoregressive Integrated Moving Average-Generalized Autoregressive Conditional Heteroscedasticity (ARIMA-GARCH) and Monte Carlo-Expected Tail Loss (MC-ETL), several conclusions can be drawn. There are 3 variations of the ARIMA-GARCH model, including ARIMA(2,1,1)-GARCH(0,1) for AAPL, GOOG, and TSLA, ARIMA(3,1,2)-GARCH(0,2) for MSFT , ARIMA(2,1,2)-GARCH(0,1) for AMZN. Based on this model, TSLA has the highest mean return, while GOOG has the lowest mean return. Calculation of optimized portfolio weight based on MOO has been obtained, and the assumption of risk aversion value is taken to obtain a static optimized portfolio weight. Then the optimized portfolio return volatility is analyzed using ETL to obtain a measure of investment risk. The Monte Carlo simulation on the ETL calculation yields a value of 2.9% for the optimized portfolio. This is the smallest value compared to all the constituent stocks' ETL values. Thus, the optimized portfolio is concluded as an investment choice for investors with a low level of risk.

For the next research, we suggest using other models to model stock prices. Portfolio formation methods and other risk measurements can also be used to obtain comparisons. Finally, use the calculation of the level of accuracy to find out how much the model can estimate the actual data.

REFERENCES

Airouss, M., Tahiri, M., Lahlou, A., and Hassouni, A. (2018). Advanced Expected Tail Loss Measurement and Quantification for the Moroccan All Shares Index Portfolio. *Mathematics*, 6(3):38.

Arif, U. and Sohail, M. T. (2020). Asset Pricing With Higher Co-Moments and CVaR: Evidence from Pakistan Stock Exchange. *International Journal of Economics and Financial Issues*, 10(5):243.

Azmi, U. and Syaifudin, W. H. (2020). Peramalan Harga Komoditas Dengan Menggunakan Metode Arima-Garch. *Jurnal Varian*, 3(2):113–124.

Bellofatto, A., DHondt, C., and De Winne, R. (2018). Subjective Financial Literacy and Retail Investors Behavior. *Journal of Banking & finance*, 92:168–181.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting And Control*. John Wiley & Sons.

Brandstetter, L. and Lehner, O. M. (2015). Opening the Market for Impact Investments: The Need for Adapted Portfolio Tools. *Entrepreneurship Research Journal*, 5(2):87–107.

Chandra, P. (2017). *Investment Analysis and Portfolio Management*. McGraw-Hill Education.

Feng, W., Wang, Y., and Zhang, Z. (2018). Can Cryptocurrencies Be A Safe Haven: A Tail Risk Perspective Analysis. *Applied Economics*, 50(44):4745–4762.

Giemza, D. (2021). Ranking of Optimal Stock Portfolios Determined on The Basis of Expected Utility Maximization Criterion. *Journal of Economics and Management*, 43(1):154–178.

Jareño, F., Negrut, L., et al. (2016). US Stock Market and Macroeconomic Factors. *Journal of Applied Business Research (JABR)*, 32(1):325–340.

Kulali, I. (2016). Portfolio Optimization Analysis with Markowitz Quadratic Mean-Variance Model. *European Journal of Business and Management*, 8(7):73–79.

Lampio, K. (2018). Optimization of Fin Arrays Cooled by Forced or Natural Convection.

Liang, J. J., Yue, C., and Qu, B.-Y. (2016). Multimodal Multi-Objective Optimization: A Preliminary Study. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 2454–2461. IEEE.

Ling, R. F. and Rubinstein, R. Y. (1983). Reviewed Work: Simulation and the Monte Carlo Method. *Journal of the American Statistical Association*, 78(382):511–512.

Mayuni, I. A. I. and Suarjaya, G. (2018). Pengaruh ROA, FIRM SIZE, EPS, dan PER tTrhadap Return Saham pada Sektor Manufaktur di BEI. *E-Jurnal Manajemen Unud*, 7(8):4063–4093.

Nguyen, S. P. and Huynh, T. L. D. (2019). Portfolio Optimization from a Copulas-GJRGARCH-EVT-CVAR Model: Empirical Evidence from ASEAN Stock Indexes. *Quantitative Finance and Economics*, 3(3):562–585.

Salim, D. F. and Rizal, N. A. (2021). Portofolio optimal Beta dan Alpha. *Jurnal Riset Akuntansi dan Keuangan*, 9(1):181–192.

Siswono, G. O., Saputra, W. H., Pricila, V., and Lina, Y. A. (2021). Application of Holt-Winter and Grey Holt-Winter Model in Risk Analysis of United States (US) Energy Commodities Futures Using Value at Risk (VaR). In *International Conference on Global Optimization and Its Applications 2021*, volume 1, pages 126–126.

Sururi, W., Yahya, I., and Abubakar, E. (2021). Analysis of the Effect of Financial Performance, Company Size on Stock Prices with Dividend Policy as Moderating Variable in Pharmaceutical Companies Listed on the Indonesia Stock Exchange.

Wardiyah, M. L. (2017). Manajemen pasar uang dan pasar modal.

Zheng, X. and Chen, B. M. (2013). *Stock Market Modeling and Forecasting*, volume 442. Springer.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 161–170
e-ISSN: 2581-2017

Modified Hungarian Method for Solving Balanced Fuzzy Transportation Problems

Fried M. Allung Blegur[1], **Nugraha K. F. Dethan**[2]
[1,2]Department of Mathematics, Timor University, Indonesia

Article Info

Article history:

Received : 04-20-2022
Revised : 04-29-2022
Accepted : 04-30-2022

Keywords:

Fuzzy Transportation;
Hungarian Method;
Trapezoidal Fuzzy Number;
Robusts Ranking.

ABSTRACT

This paper discusses how to solve a balanced transportation problem, with transportation costs in the form of trapezoidal fuzzy numbers. Fuzzy costs are converted into crisp costs using Robust's method as a ranking function. A new modified approach of the Hungarian method has been applied to solve the balanced fuzzy transportation problem with the number of sources distinct from destinations. The analysis that we carry out comes from various literature studies and begins with examining the problem of fuzzy transportation, then collecting and connecting theories related to the problem of fuzzy transportation. The Hungarian method for the assignment problem was modified by adding some steps involving the principles of the transportation problem, such as that all that can be supplied will be supplied, in order to meet demand. This principle allows for one source to send products to multiple destinations and one destination can receive supplies from multiple sources. This is the basic concept of building this new method. This approach solves the fuzzy transportation problem in one optimization stage and produces the same results as other methods that solve the problem in two stages.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1865

Corresponding Author:

Fried M. Allung Blegur
Department of Mathematics, Timor University.
Email: allung.friedblegur@gmail.com

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

The transportation problem is part of a wider class of linear problems, namely network flow. This problem can be solved using a transportation model based on the characteristics that a product is transported from a number of sources to a certain number of destinations with a minimum cost and optimized demand fulfilment. The basic model of the transportation problem assumes that each source can supply a number products and that each destination has a fixed amount demand (Taylor et al., 2013).

The Hungarian method proposed by (Kuhn, 1955) is one of the tools for solving a special type of transportation problem, namely the assignment problem. This method is more efficient in the iteration process than other methods. A disadvantage of this method is that it can only solve a balanced assignment problem, that is, the machines and the assignments have the same number. In order to solve the unbalanced assignment problem, it is necessary to add dummy machines which we will then ignore the work assigned to these machines. Given that the dummy is a pseudo activity, so the duration (cost) of the dummy activity (machines) is zero (Razi and Yudiarti, 2020). (Rabbani et al., 2019) proposed a different concept in solving the problem of unbalanced assignments, a modified Hungarian method which does not involve dummy workers (machines).

The Hungarian method uses deterministic data, which renders it unreliable for solving real problems that do not have definite and complete information. Fuzzy method becomes the best tool for solving problems with ambiguous information. This method is built based on the concept of fuzzy sets proposed by (Zadeh, 1965) and the concept of decision making involving fuzzy numbers (Bellman and Zadeh, 1970). Incorporating fuzzy numbers into the assignment problem provides a more realistic solution. (Kar et al., 2021) proposed a new approach which can solve fuzzy assignment problems using the Hungarian method.

Solving a more general fuzzy assignment problem, namely fuzzy transportation, has been carried out by several researchers such as (Patil and Chandgude, 2012), (Malini and Kennedy, 2013), and (Hunwisai and Kumam, 2017). These researchers solved the fuzzy transportation problem in two stages, namely determining a feasible solution and then ending with determining the optimal solution. This paper proposes a new approach to solve the fuzzy transportation problem in one optimization stage. This approach is built by modifying the Hungarian method.

B. LITERATURE REVIEW

1. Earlier Research

A modified method has been developed by (Kumar, 2006) to deal with unbalanced assignments. The unbalanced assignment cost matrix was split into two balanced parts and then solved using the Hungarian method. A similar approach was done by (Yadaiah et al., 2016) using the Lexi-search approach. (Betts et al., 2016) revised the numerical example provided by (Yadaiah et al., 2016) by retaining the original matrix of assignment costs and adding dummy rows to balance the assignments. The solution is carried out using the Hungarian method. (Younis and Alsharkasi, 2019) compare the use of the Hungarian method and the VAM method in solving transportation problems with the number of sources not the same as destinations. They added one dummy source so that it could be solved using the Hungarian method.

(Rabbani et al., 2019) modified the Hungarian method for solving unbalanced assignments without dummy variables. The results obtained are better than the modified Hungarian method applied by (Kumar, 2006), (Yadaiah et al., 2016), and (Betts et al., 2016) for the same problem. (Evipania et al., 2021) used a modified Hungarian method proposed by (Kumar, 2006) in solving unbalanced assignments for employees of Mitra Tex Convection.

The Hungarian Fuzzy approach in solving transportation problems with the same number of sources as the destinations was carried out by (Patil and Chandgude, 2012). The type of fuzzy numbers used was triangular fuzzy numbers (TFN) and the problem was solved using the MODI method. (Khalifa, 2020) solved transportation problems with heptagonal fuzzy numbers using the Goal Programming approach. In the same year (Srinivasan et al., 2020) solved the fuzzy transport problem for transporting material by utilizing a ranking function (beta distribution). (Manimaran and Ananthanarayanan, 2012) used Yager ranking on the fuzzy assignment problem using LINGO 9.0.

The solution to transportation problems with the number of sources distinct from the number of destinations was carried out by (Saman et al., 2020) using the Fuzzy Analytical Hierarchy Process (AHP) with triangular fuzzy numbers. (Dhanasekar et al., 2017) used the zero-point method and MODI with triangular fuzzy numbers, (Bisht and Srivastava, 2019) used One-Point Conventional, while (Aini et al., 2021) used the zero-suffix method with triangular fuzzy numbers.

2. Fuzzy Number

Definition 1. (Zimmermann, 1978) Let X be a set. A fuzzy set $\tilde{A}$ in X is a set of ordered pairs $\tilde{A} = \{(x, \mu_{\tilde{A}}(x))|x \in X\}$, with $\mu_{\tilde{A}}(x)$ represents the degree of membership of x in $\tilde{A}$ in the interval $[0, 1]$.

Definition 2. (Bector et al., 2005) $\tilde{A}$ be a fuzzy set in R. Then $\tilde{A}$ is called a fuzzy number if:

1. $\tilde{A}$ is a convex, that is

$$\mu_{\tilde{A}}(\lambda\mu + (1-\lambda)v) \geq \min\left(\mu_{\tilde{A}}(u), \mu_{\tilde{A}}(v)\right), \forall u, v \in R, \lambda \in [0, 1] \tag{1}$$

2. $\tilde{A}$ is normal
3. $\mu_{\tilde{A}}$ is upper semicontinuous, and
4. $\tilde{A}$ has bounded support.

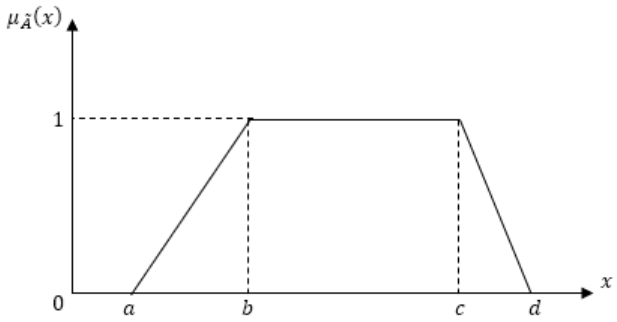


Figure 1. $TrFN\tilde{A} = (a, b, c, d)$

Definition 3. (Sakawa, 2013) Let α be a real number in $(0, 1)$ and $\tilde{A}$ be a fuzzy set. The α-level set of the fuzzy set $\tilde{A}$ is the set $\tilde{A}_\alpha = \{x \in X \;\; \mu_{\tilde{A}}(x) \geq \alpha\}$.

Definition 4. $TrFN\tilde{A} = (a, b, c, d)$ is a special fuzzy set in R, with membership function defined as the following (Figure 1):

$$\mu_{\tilde{A}}(x) = \begin{cases} 1 & , b \leq x \leq c \\ \dfrac{x-a}{b-a} & , a \leq x \leq b \\ \dfrac{d-x}{d-c} & , c \leq x \leq d \\ 0 & , \text{otherwise} \end{cases} \tag{2}$$

where $a \leq b \leq c \leq d$.

3. Robusts Ranking Technique

Definition 5. The Robusts ranking index for a convex fuzzy number $\tilde{A}$ is defined as

$$R\left(\tilde{A}\right) = \frac{1}{2}\int_0^1 \left[\tilde{A}_\lambda^L - \tilde{A}_\lambda^U\right] \, d\lambda \tag{3}$$

as for $TrFN\left[\tilde{A}_\lambda^L - \tilde{A}_\lambda^U\right] = (a + (b-a)\lambda) + (d - (d-c)\lambda)$

4. Balanced Fuzzy Transportation Problems

A fuzzy transportation problem with the number of sources distinct from the number of destinations is presented in Table 1.

Table 1. A fuzzy transportation problem with $m \neq n$

	$\boldsymbol{D_1}$	$\boldsymbol{D_2}$	...	$\boldsymbol{D_n}$	**Supply**
$\boldsymbol{S_1}$	$\tilde{C}_{11}$	$\tilde{C}_{12}$	...	$\tilde{C}_{1n}$	SS_1
$\boldsymbol{S_2}$	$\tilde{C}_{21}$	$\tilde{C}_{22}$	...	$\tilde{C}_{2n}$	SS_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$\boldsymbol{S_m}$	$\tilde{C}_{m1}$	$\tilde{C}_{m2}$	...	$\tilde{C}_{mn}$	SS_2
Demand	SD_1	SD_2	...	SD_n	$\sum_{i=1}^{m} SS_i \sum_{j=1}^{n} SD_j$

where

S_i = the i^{th} source,
D_j = the j^{th} destination,
$\tilde{C}_{ij}$ = fuzzy transportation cost from source i to destination j,
SS_i = the maximum number of products can be transported from source i,
SD_j = demand from destination j.

A mathematical model for solving fuzzy transportation problem in Table 1 is

$$\text{Minimize } \tilde{Z} = \sum_i \sum_j \tilde{C}_{ij} X_{ij} \tag{4}$$

Subject to

$$\begin{aligned} \sum_j X_{ij} &\leq SS_i, \quad \forall i \\ \sum_i X_{ij} &\geq SS_j, \quad \forall j \\ X_{ij} &\geq 0 \qquad \forall i, j \end{aligned} \tag{5}$$

where X_{ij} is the number of products transported from source i to source j.

C. RESEARCH METHOD

The method used in this research is the analysis of theories relevant to fuzzy transportation problems with the number of sources not equal to the number of destinations. This analysis is sourced from various literature studies. This research begins by examining the problem of fuzzy transportation, then collecting and connecting theories related to the problem of fuzzy transportation. Furthermore, a new approach was developed to solve the fuzzy transportation problem in one optimization stage. The new approach in question is to modify the Hungarian method into a new method that can be applied to fuzzy transportation problems where the number of sources is not the same as the number of destinations. This new approach is then compared to the method used by (Hunwisai and Kumam, 2017). This new approach only uses one optimization stage, while (Hunwisai and Kumam, 2017)use two optimization stages, namely the initial basic feasible solution (IBFS) and the modified distribution method (MODIM).

D. RESULTS AND DISCUSSION

1. Algorithm

Given an unbalanced fuzzy transportation problem with m sources $S = \{S_1, S_2, \ldots, S_m\}$ and n destinations $D = \{D_1, D_2, \ldots, D_n\}$, where $m \neq n$. Let the fuzzy transportation cost from source i to destination j is $\tilde{C}_{ij}$. The rows of the table represent the m sources and the columns represents the n destinations.

Step-1 : Create an allocation table for the m sources and n destinations.

Step-2 : Insert the fuzzy transportation costs $\tilde{C}_{ij}$ into the allocation table according to their rows and columns.

Step-3 : Apply the Robust ranking technique to transform the fuzzy costs into crisp costs.

Step-4 : If there are more columns than rows, subtract the values in each column by the minimum cost in the column, then subtract the values in each row by the minimum cost in the row. Conversely, if there are more rows than columns, subtract the values in each row by the minimum cost in the row, then subtract each column by the minimum cost in the column. This Step-is to ensure that delivery is made on the lowest cost basis.

Step-5 : Test whether the ideal transportation has been achieved. Do this by determining that the minimum number of lines covering all zeros is equal to $\max\{m, n\}$ and that there is at least one intersecting lines (to ensure that at least one source can deliver the product to more than one destination or at least one destination can receive the product from multiple sources). If these criteria are met then proceed to Step-8, otherwise continue to step-6.

Step-6 : If the number of lines is less than $\max\{m, n\}$ or there are no intersecting lines, then choose the smallest cost that is not covered by a line then subtract each cost that is not covered by a line with the smallest cost that is not covered by a line. If any of the lines intersect at 0, add the smallest cost not covered by the 0 line.

Step-7 : Repeat Step-6 and 7 until the number of lines equals $\max\{m, n\}$ and there is at least one intersecting lines.

Step-8 : Test whether the supply in each row is no more than the demand in the columns containing zero for each row (to ensure that all products at each source can be delivered to the destination). If met, go to step-11, otherwise go to step-9.

Step-9 : If there is supply in a row that is less than the demand in the zero-loaded columns in that row, subtract each nonzero cost by the smallest nonzero cost.

Step-10 : Repeat Step-8 and 9 until the supply in each row is no less than the demand in the zero column for each of the same rows.

Step-11 : Delete all nonzero cells.

Step-12 : Find the row that contains only one zero, say row i, and find the column that contains that zero, say column j. Choose that zero and replace it with the number of products that can be delivered, i.e. $\min\{SS_i, SD_j\}$, then subtract SS_i and SD_j by the number of products sent (to find out the number of products that have not been delivered and the amount of demand that have not been fulfilled). If there is more than one row that contains only one zero, perform this Step-according to the largest supply. Do the same for columns that only have one zero. Delete the row if the supply in that row has been met and delete the column if the demand in that column has been met.

Step-13 : If Step-12 generates another row or column containing only one zero, repeat Step-12, if not, proceed to Step-14.

Step-14 : Select the row with the most supply (SS_i) that has not been transported and the column that has the most demand (SD_j) that has not been fulfilled. Put the value $\min\{SS_i, SD_j\}$ in the cell where supply and demand intersect. If there is more than one row that has the same remaining supply, select the row that has the least cost. If there is more than one column that has the same remaining demand, select the column containing the largest cost in row instead of i. Delete the row if the supply in that row has been met and delete the column if the demand in that column has been met.

Step-15 : Repeat Step-12 and 14 until every supply is fulfilled.

2. Numerical Example

1. Example 1

In this example, we analyze a fuzzy transportation problem put forward by (Hunwisai and Kumam, 2017) using the algorithm explained earlier. This transportation problem involves four sources and three destinations. The transport costs are in the form of trapezoidal fuzzy numbers as presented in Table 2.

Table 2. The Fuzzy Transportation Problem from Example 1

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1	$(2, 5, 8, 15)$	$(2, 3, 4, 7)$	$(3, 7, 9, 15)$	25
S_2	$(3, 6, 9, 12)$	$(4, 7, 9, 11)$	$(4, 8, 10, 13)$	35
S_3	$(3, 7, 10, 16)$	$(5, 6, 12, 16)$	$(4, 6, 8, 14)$	50
S_4	$(3, 4, 6, 9)$	$(4, 5, 7, 9)$	$(5, 8, 11, 13)$	10
Demand (SD_j)	30	40	50	120

This is a balanced transportation problem since the quantity supplied equals that of the demand, namely 120.

Step-1 : Create the allocation table for $m = 4$ sources and $n = 4$ destinations (Table 2).

Step-2 : Fill the $\tilde{C}_{ij}$ values in the table accordingly (Table 2).

Step-3 : Apply the Robust ranking technique to transform the fuzzy costs into crisp costs.

The fuzzy costs to transport a product from S_1 to D_1 in Table 2 are $(2, 5, 8, 15$. Since $[\tilde{C}^L_{11\lambda} - \tilde{C}^U_{11\lambda}] = (2+3\lambda)+(15-7\lambda) = 17-4\lambda$, we have

$$R\left(\tilde{C}\right) = (2, 5, 8, 15) = \frac{1}{2}\int_0^1 [\tilde{C}^L_{11\lambda} - \tilde{C}^U_{11\lambda}]\, d\lambda = \frac{1}{2}\int_0^1 (17-4\lambda)\, d\lambda = 7.5 \tag{6}$$

The remaining $\tilde{C}_{ij}$ values are calculated in similar fashion, hence we obtain $R(\tilde{C}_{12}) = 4$, $R(\tilde{C}_{13}) = 8.5$, $R(\tilde{C}_{21}) = 7.5$, $R(\tilde{C}_{22}) = 7.75$, $R(\tilde{C}_{23}) = 8.75$, $R(\tilde{C}_{31}) = 9$, $R(\tilde{C}_{32}) = 9.75$, $R(\tilde{C}_{33}) = 8$, $R(\tilde{C}_{41}) = 5.5$, $R(\tilde{C}_{42}) = 6.25$, and $R(\tilde{C}_{43}) = 9.25$. This gives us Table 3:

Table 3. Fuzzy Transportation After Ranking

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1	7.5	4	8.5	25
S_2	7.5	7.75	8.75	35
S_3	9	9.75	8	50
S_4	5.5	6.25	9.25	10
Demand (SD_j)	30	40	50	120

Step-4 : Since there are more rows than columns, we subtract each row by the minimum cost in the row, thereafter subtract each column by minimum cost in the column (Table 4).

Table 4. Step-4

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1	2.5	0	3.25	25
S_2	0	0	0	35
S_3	0	1.5	0	50
S_4	0	0.5	2.5	10
Demand (SD_j)	30	40	50	120

Step-5 : An ideal transport has been achieved since the number of minimum lines covering all zeros is four, the same as $\max\{m, n\}$. There are also intersecting lines (Table 5).

Table 5. Step-5

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1	2.5	0	3.25	25
S_2	0	0	0	35
S_3	0	1.5	0	50
S_4	0	0.5	2.5	10
Demand (SD_j)	30	40	50	120

We then proceed to **Step-8**, as suggested by the algorithm. From Table 5, supplies in each row are no more than demands in columns containing zeros for each row. This implies all products from each source can be delivered to the destinations and hence we can go to **Step-11**.

Step-11 : Delete all nonzero cells (Table 6).

Table 6. Step-11

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		0		25
S_2	0	0	0	35
S_3	0		0	50
S_4	0			10
Demand (SD_j)	30	40	50	120

Step-12 : We can see there are two rows that have one zero, namely S_1 and S_4 with SS_1 having the biggest demand, namely 25. The demand that contains 0 is SD_2, namely 40. Replace that 0 with $\min\{SS_1, SD_2\} = 25$ and repeat this on row S_4. We see that SS_1 become $25 - 25 = 0$ and SD_1 become $40 - 25 = 15$ respectively (Table 7).

Table 7. Step-12

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		25		0
S_2	0	0	0	35
S_3	0		0	50
S_4	10			0
Demand (SD_j)	20	15	50	85

Step-13 : Step-12 leaves column D_2 with only one zero. Using the same treatment as in Step-12 we obtain Table 8.

Table 8. Step-14

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		25		0
S_2	0	15	0	20
S_3	0		0	50
S_4	10			0
Demand (SD_j)	20	0	50	70

Step-14 : Notice that row S_2 and S_3 have more than one zero. Choose S_3 since it has the most demand, namely $SS_3 = 50$.

DOI: https://doi.org/10.30812/varian.v5i2.1865

SS_3 can deliver a product to D_1 and D_3. Choose D_3 since it has the most demand, namely $SD_3 = 50$. Replace the value in the cell of intersection between S_3 and D_3 with $\min\{50, 50\} = 50$ then subtract SS_3 and D_3 by that value (Table 9).

Table 9. Step-14

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		25		0
S_2	0	15		20
S_3			50	0
S_4	10			0
Demand (SD_j)	20	0	0	20

Step-15: Only one zero left, namely in the cell of intersection between S_2 and D_1. Replace that 0 with $\min\{SS_2, SD_1\} = 20$ and subtract SS_2 and SD_1 by 20 (Table 10). We see that SS_i and SD_j becomes zero, which means supplies and demand have been met.

Table 10. Step-15

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		25		0
S_2	20	15		0
S_3			50	0
S_4	10			0
Demand (SD_j)	0	0	0	0

Table 11 shows the best transport decision to deliver all products from every source and meets demand from every destination.

Table 11. The Best Transport Decision

Source	D_1	D_2	D_3	**Supply** (SS_i)
S_1		4(25)		100
S_2	7.5(20)	7.75(15)		266.25
S_3			8(50)	400
S_4	5.5(10)			55
Demand (SD_j)	205	216.25	400	825.25

The optimal fuzzy transport cost is $4(25) + 7.5(20) + 7.75(15) + 8(50) + 5.5(10) = 821.25$, which agrees with that which obtained by (Hunwisai and Kumam, 2017).

2. Example 2

This time we take a transportation problem put forward by (Balasubramanian and Subramanian, 2018), where there are more destinations than sources. The transport costs, supply and demand are in the form of trapezoidal fuzzy number and presented in Table 12.

Table 12. Fuzzy Transportation Problem

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	(1,2,3,4)	(1,3,4,6)	(9,11,12,14)	(5,7,8,11)	(1,6,7,12)
S_2	(0,1,2,4)	(-1,0,1,2)	(5,6,7,8)	(0,1,2,3)	(0,1,2,3)
S_3	(3,5,6,8)	(5,8,9,12)	(12,15,16,19)	(7,9,10,12)	(5,10,12,17)
Demand (SD_j)	(5,7,8,10)	(1,5,6,10)	(1,3,4,6)	(1,2,3,4)	

Results obtained from Robusts ranking technique (Step-3) can be found in Table 13.

Table 13. Step-3

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	2.5	3.5	11.5	7.5	6.5
S_2	1.75	0.5	6.5	1.5	1.5
S_3	5.5	8.5	15.5	9.5	11
Demand (SD_j)	7.5	5.5	3.5	2.5	19

Step-4 : Since the number of columns is more than the number of rows, subtract each column by the minimum cost in the column, then subtract each row by the minimum cost in the row (Table 14).

Table 14. Step-4

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	0	2.25	4.25	5.25	6.5
S_2	0	0	0	0	1.5
S_3	0	4.25	5.25	4.25	11
Demand (SD_j)	7.5	5.5	3.5	2.5	19

This transportation is not ideal, so we revert to Step-6 and 7, and obtain Table 15.

Table 15. Step-6 and 7

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	0	0	0	1	6.5
S_2	2.25	2	0	0	1.5
S_3	0	2	1	0	11
Demand (SD_j)	7.5	5.5	3.5	2.5	19

Step-8: Row SS_3 is bigger than $SD_1 + SD_3$, hence we proceed to Step-9 and obtain Table 16.

Table 16. Step-8 and 9

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	0	0	0	1	6.5
S_2	2.25	2	0	0	1.5
S_3	0	1	0	0	11
Demand (SD_j)	7.5	5.5	3.5	2.5	19

Continuing to Step-11, we delete all nonzero cells and do Step-12, which gives Table 17.

Table 17. Step-11 and 12

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1	0	5.5	0		1
S_2			0	0	1.5
S_3	0		0	0	11
Demand (SD_j)	7.5	0	3.5	2.5	13.5

Proceed to Step-14 and we get Table 18.

Table 18. Step-14

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1		5.5	0		1
S_2			0	0	1.5
S_3	7.5		0	0	3.5
Demand (SD_j)	0	0	3.5	2.5	6

Since row S_1 contains only one zero, we perform Step-12 for this row and get Table 19.

Table 19. Step-12

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1		5.5	1		0
S_2			0	0	1.5
S_3	7.5		0	0	3.5
Demand (SD_j)	0	0	2.5	2.5	5

DOI: https://doi.org/10.30812/varian.v5i2.1865

We continue to Step-14. Since S_3 has the most supply, it becomes a priority. Since $SD_3 = SD_4$ we choose $D_4(9.5)$ since it has the least transport cost in S_3 (see Table 13). We then obtain Table 20.

Table 20. Step-14 again

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1		5.5	1		0
S_2			0		1.5
S_3	7.5		0	2.5	1
Demand (SD_j)	0	0	2.5	0	2.5

Again, we perform Step-12 and obtain the optimal transportation table (Table 21).

Table 21. Step-12 again

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1		5.5	1		0
S_2			1.5		0
S_3	7.5		1	2.5	0
Demand (SD_j)	0	0	0	0	0

The best transportation decision is presented in Table 22.

Table 22. The Best Transportation Decision

Source	D_1	D_2	D_3	D_4	**Supply** (SS_i)
S_1		3.5(5.5)	11.5(1)		30.75
S_2			6.5(1.5)		9.75
S_3	5.5(7.5)		15.5(1)	9.5(2.5)	80.5
Demand (SD_j)	41.25	19.25	36.75	23.75	121

The optimal fuzzy transportation cost is 121, which agrees with that which obtained by (Balasubramanian and Subramanian, 2018).

E. CONCLUSION AND SUGGESTION

The modified Hungarian method can solve the fuzzy transportation problem with the number of sources not equal to the number of destinations. The optimal solution is obtained in one step. This approach yields the same results as other methods that solve the problem in two stages. Future studies can develop this approach to solve fuzzy transportation problems without transforming fuzzy costs into crisp costs. In addition, it is necessary to develop methods for fuzzy transportation probslems involving fuzzy supply and fuzzy demand.

REFERENCES

Aini, A. N., Shodiqin, A., and Wulandari, D. (2021). Solving Fuzzy Transportation Problem Using ASM Method and Zero Suffix Method. *Enthusiastic: International Journal of Applied Statistics and Data Science*, 1(1):28–35.

Balasubramanian, K. and Subramanian, S. (2018). An Approach for Solving Fuzzy Transportation Problem. *International Journal of Pure and Applied Mathematics*, 119(17):1523–1534.

Bector, C., Chandra, S., et al. (2005). *Fuzzy Mathematical Programming and Fuzzy Matrix Games*, volume 169. Springer.

Bellman, R. E. and Zadeh, L. A. (1970). Decision-Making in A Fuzzy Environment. *Management science*, 17(4):B–141.

Betts, N., Vasko, F. J., et al. (2016). Solving The Unbalanced Assignment Problem: Simpler Is Better. *American Journal of Operations Research*, 6(04):296.

Bisht, D. C. and Srivastava, P. K. (2019). One Point Conventional Model to Optimize Trapezoidal Fuzzy Transportation Problem. *Int J Math, Eng Manag Sci*, 5(4):1251–1263.

Dhanasekar, S., Hariharan, S., and Sekar, P. (2017). Fuzzy Hungarian MODI Algorithm to Solve Fully Fuzzy Transportation Problems. *International journal of fuzzy systems*, 19(5):1479–1491.

Evipania, R., Gandhiadi, G., and Sumarjaya, I. W. (2021). Optimalisasi Masalah Penugasan Tidak Seimbang Menggunakan Modified Hungarian Method. *E-Jurnal Mat*, 10(1):26.

Hunwisai, D. and Kumam, P. (2017). A Method for Solving A Fuzzy Transportation Problem Via Robust Ranking Technique and ATM. *COGENT mathematics*, 4(1):1283730.

Kar, R., Shaw, A., and Mishra, J. (2021). Trapezoidal Fuzzy Numbers (TrFN) and its Application in Solving Assignment Problems by Hungarian Method: A New Approach. *Fuzzy Intelligent Systems: Methodologies, Techniques, and Applications*, pages 315–333.

Khalifa, H. A. E.-W. (2020). Goal Programming Approach for Solving Heptagonal Fuzzy Transportation Problem Under Budgetry Constraint. *Operations Research and Decisions*, 30.

Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Kumar, A. (2006). A Modified Method for Solving the Unbalanced Assignment Problems. *Applied mathematics and computation*, 176(1):76–82.

Malini, S. and Kennedy, F. C. (2013). An Approach for Solving Fuzzy Transportation Problem using Octagonal Fuzzy Numbers. *Applied Mathematical Sciences*, 7(54):2661–2673.

Manimaran, S. and Ananthanarayanan, M. (2012). A Study on Comparison Between Fuzzy Assignment Problems using Trapezoidal Fuzzy Numbers with Average Method. *Indian journal of science and technology*, 5(4):2610–2613.

Patil, A. and Chandgude, S. (2012). Fuzzy Hungarian Approach for Transportation Model. *International Journal of Mechanical and Industrial Engineering*, 2(1):77–80.

Rabbani, Q., Khan, A., and Quddoos, A. (2019). Modified Hungarian Method for Unbalanced Assignment Problem with Multiple Jobs. *Applied Mathematics and Computation*, 361:493–498.

Razi, F. A. and Yudiarti, W. W. (2020). Network Optimization of Packaging Water Factory" Aeta" by Using Critical Path Method (CPM) in Tirta Taman Sari Drinking Water Company, Madiun City. *Jurnal Varian*, 4(1):11–18.

Sakawa, M. (2013). *Fuzzy Sets and Interactive Multiobjective Optimization*. Springer science & business media.

Saman, M., Surarso, B., Irwanto, B., et al. (2020). Solving of Fuzzy Transportation Problem Using Fuzzy Analytical Hierarchy Process (AHP). In *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, pages 10–15. Atlantis Press.

Srinivasan, R., Karthikeyan, N., Renganathan, K., and Vijayan, D. (2020). Method for Solving Fully Fuzzy Transportation Problem to Transform the Materials. *Materials today: proceedings*.

Taylor, B. W., Bector, C., Bhatt, S., and Rosenbloom, E. S. (2013). *Introduction to Management Science*. Pearson Boston, MA, USA.

Yadaiah, V., Haragopal, V., et al. (2016). A New Approach of Solving Single Objective Unbalanced Assignment Problem. *American Journal of Operations Research*, 6(01):81.

Younis, A. A. A. and Alsharkasi, A. M. (2019). Using of hungarian method to solve the transportation problems. *EPH - International Journal of Applied Science*, 1(1):834845.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 8(3):338–353.

Zimmermann, H.-J. (1978). Fuzzy Programming and Linear Programming with Several Objective Functions. *Fuzzy sets and systems*, 1(1):45–55.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 171–178
e-ISSN: 2581-2017

Cluster Analysis of Inclusive Economic Development Using K-Means Algorithm

Riska Yanu Farifah[1], Dita Pramesti[2]
[1,2]Bachelor of Information System, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received : 04-05-2022
Revised : 04-25-2022
Accepted : 04-30-2022

Keywords:

Cluster analysis;
Inclusive economic development;
k-means algorithm;
Silhouette coefficient.

ABSTRACT

This study aims to cluster 38 Districts/Cities in East Java based on the 10 forming indicators of inclusive economic development and to determine the inclusive economic growth of Districts/Cities above or below the total average. 10 indicators used in this study are GRDP per capita, GRDP by business field, Labor force participation rate, Unemployment rate, Gini ratio, Expenditure per capita, the number of poverty, Life expectancy, expectation years of schooling, and mean years of schooling. There are 3 scenarios in this study, namely 2 clusters, 3 clusters, and 4 clusters. The method of clustering in this study is using the K-means algorithm. This study uses the silhouette coefficient to evaluate the best cluster of 3 scenarios. The best k-means algorithm in this study is using 2 clusters with a silhouette coefficient of 0.87. There are 29 Districts/Cities included in cluster 1 with inclusive economic development below the total average and 9 Districts/Cities included in cluster 2 with inclusive economic development above the total average. The members of cluster 1 are mostly district areas and located in coastal or border areas and the members of cluster 2 are mostly urban or industrial areas.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1894

Corresponding Author:

Riska Yanu Farifah
Bachelor of Information System, Telkom University, Bandung, Indonesia
Email: riskayanu@gmail.com

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

Cluster analysis is part of the statistical method, which is included in the family of multivariate analysis. Cluster analysis is part of grouping analysis which aims to determine the results of separating objects from a population based on quantitative comparisons of several characteristics used (Webster, 2021). In other words, cluster analysis aims to partition the data into several groups based on the similarity of the measured characteristics. The algorithm in cluster analysis works to find new groups based on the pattern of the data used (Jain, 2010). Broadly speaking, cluster analysis has 3 main objectives, namely 1) the formation of basic structures used to identify prominent features or groups, 2) natural classification used to identify the degree of similarity of organisms (phylogenetic relationships), 3) compression used to organize or summarize the data through the prototype cluster (Jain, 2010).

There are two methods in the cluster analysis, namely hierarchical and non-hierarchical. The stages of the hierarchical method are starting from a large cluster and correcting cluster members, then merging similar cluster members into a new group. In other words, this hierarchical method creates clusters from a large group into several smaller groups (Jain, 2010), (Xie et al., 2019). In the non-hierarchical method, the number of clusters is determined first, then looking for cluster members based on the distance, which has the same characteristics (Xie et al., 2019). The most popular hierarchical method is average linkage, and the most popular non-hierarchical method is the k-means (Jain, 2010).

Based on the descriptions above, the k-means algorithm is the most widely used. Several previous studies have shown that the k-means algorithm provides ease of implementation, simple analysis, efficiency, and has good performance (Jain, 2010), (Xie et al., 2019), (Qureshi and Ahamad, 2018). The k-means algorithm has been used in several disciplines. In its development, cluster analysis

has been used in information technology and information systems, for example, it has been applied to cloud computing (Sharma and Bala, 2020) and cellular network site management (Gbadoubissa et al., 2020). In health, a k-means algorithm has been used to group mutations of the coronavirus (Hozumi et al., 2021). In ecology, a k-means algorithm has been used to determine the spatial pattern and the relationship between controlling factors and toxic elements in the topsoil (Xu et al., 2021). In the economy, a k-means algorithm has been used to partition the open unemployment rate in the South Sulawesi Province (Akramunnisa and Fajriani, 2020).

This study uses the k-means algorithm to cluster inclusive economic development data in East Java. In contrast to the study that has been done by (Akramunnisa and Fajriani, 2020), this study uses 3 scenarios to get the best cluster and has an additional stage that is used to evaluate the cluster results. The method to evaluate the best cluster in this stage is using the silhouette coefficient (Naghizadeh and Metaxas, 2020).

Inclusive economic development aims to create equitable access and opportunities for all levels of society, improve welfare, and reduce the gap between regions. Inclusive economic development is divided into 3 Pillars, namely Pillar 1 concerning economic growth and development, Pillar 2 relates to income distribution and poverty reduction, and Pillar 3 is an opportunity and expanding access. These 3 pillars are used to measure and monitor the level of inclusiveness of development in Indonesia, both on a national and regional scale (Agency, 2021). The higher the percentage of achievement of these 3 pillars, the more inclusive the development will be, thus the more prosperous the population (Statistics-East Java, 2020), (Hapsari, 2019), (Setianingtias et al., 2019). The inclusive economic development index is one of the benchmarks for the success of a region in the welfare of its population. It is known that in 2019, East Java was generally above the national inclusive economic development index. Pillar 1 index of inclusive economic development is 5.72 and national is 5.48, Pillar 2 has an index of 6.56 and national of 6.57, Pillar 3 has index of 7.28 and national is 6.09 (Agency, 2021). However, not all Districts/Cities in East Java have a level of development above the national inclusive development. Therefore, it is necessary to group Districts/Cities based on indicators forming the East Java inclusive development index to provide an overview of which Districts/Cities have high and low inclusive economic development indexes. The formation of this cluster is expected to provide useful information and can be used as a basis for making efforts to improve the quality and quantity of the indicators forming the inclusive development index in East Java Province, especially for Districts/Cities with indicators lower than the global achievements of East Java Province.

B. LITERATURE REVIEW

K-means algorithm is one of the most frequently used partition-based algorithms for clustering (Jain, 2010). K-Means algorithm studies each object in the data and forms partitions called clusters, representation the members in each cluster having similar characteristics. If the data used is continuous, each cluster is represented by a centroid which is the mean of the cluster members. In categorical data, each cluster is represented by a medoid, which is the object that occurs most frequently. K-Means uses squared Euclidean distances as a measure of similarity for cluster membership (Patel and Kushwaha, 2020). The formula of squared Euclidean distances is:

$$d(x_i, x_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 \tag{1}$$

x_ij is the i_{th} object in the jth variable, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, p$ (the data dimension is $n \times p$), and x_{kj} in the k-means clustering is the value of k_{th} centroid, $k = 1, 2, \cdots, r$ (Gbadoubissa et al., 2020). So x_{kj} can be replaced with c_{kj} and the formula is (Kakushadze and Yu, 2017):

$$c_{kj} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \tag{2}$$

The goal of the k-means algorithm is to minimize the sum of the square errors (SSE) of all clusters formed (Jain, 2010), (Gbadoubissa et al., 2020). SSE can be written as follows:

$$SSE = \sum_{i=1}^{n} \sum_{k=1}^{r} w_{i,j} \parallel x_i - c_k \parallel^2 \tag{3}$$

$w_{i,j}$ has 2 values, 1 if x_i is in the k_{th} centroid and 0 if x_i is not in the k_{th} centroid (Patel and Kushwaha, 2020). The k-means algorithm has the following stages (Gbadoubissa et al., 2020): Input : $X_{n \times p}$ and $r, r \geq 2$ Output : r clusters $(c_1, c_2, \cdots, c_r)$ with members of each cluster

DOI: https://doi.org/10.30812/varian.v5i2.1894

1. Select r cluster centroids randomly $(c_1, c_2, \cdots, c_r)$
2. **Repeat**
3. For each x_i in $X_{n\times p}$ do
4. For all centroids $c_k, 1 \leq c_k \leq r$ do
5. If $\| x - c_k \| < \| x - c_l \|, k \neq l, 1 \leq k, l \leq k$ then
6. Assign x in c_k
7. **End if**
8. **End for all**
9. **End for each**
10. Calculate the latest centroids for each cluster
11. Until the objective function (SSE) is minimized

Based on the pseudocode above, the iteration of cluster formation will stop when the objective function (SSE) is minimum so that the cluster formation process is followed by drinking SSE (Jain, 2010). Evaluation of cluster formation from the k-means algorithm can use a silhouette coefficient. The silhouette coefficient is used in the optimization process of cluster formation to get the best number and members of the cluster (Naghizadeh and Metaxas, 2020). The algorithm of the silhouette coefficient has the following stages:

1. Calculating the mean distance between objects in the same cluster

$$a_k = \frac{1}{n_j - 1} \sum_{\substack{i=1 \\ ii \neq i}}^{n_j} (x_i - x_{ii})^2 \tag{4}$$

2. Calculating the mean distance between objects in different clusters and find the minimum

$$b_k = \min \frac{1}{n_j} \sum_{\substack{i=1 \\ x_i \in c_k \\ x_{ii} \neq c_k}}^{n_j} (x_i - x_{ii})^2 \tag{5}$$

3. Calculating the silhouette coefficient

$$s = \begin{cases} \dfrac{b_k - a_k}{\max(a - k, b_k)} & c_k > 1; \\ 0 & c_k = 1 \end{cases} \tag{6}$$

4. Determining the mean of the silhouette coefficients obtained.

C. RESEARCH METHOD

This study will group Districts/Cities in East Java based on the inclusive economic development index. This study uses 2 sub-pillars for each pillar with details, 4 indicators on pillar 1, 3 indicators on pillar 2, and 3 indicators on pillar 3. The data used is secondary data obtained from BPS East Java. The details of the data used are as follows:

Table 1. Data of The Study

Pilar	Sub Pilar	Variable	Indicator
Economic Growth and Development	Economic Growth	X_1	GRDP per capita
		X_2	GRDP by business field
	Employment Opportunity	X_3	Labor force participation rate
		X_4	Unemployment rate
Income Equity and Poverty Reduction	Inequality	X_5	Gini ratio
		X_6	Expenditure per capita
	Poverty	X_7	The number of poverty
Expansion of Access and Opportunities	Health	X_8	Life expectancy Mean years of schooling
	Education	X_9	Expectation years of schooling
		X_{10}	Mean years of schooling

This study uses the k-means algorithm to form clusters of Districts/Cities based on the indicators in Table 1. The stages in this analysis are as follows:

DOI: https://doi.org/10.30812/varian.v5i2.1894

1. Specifies the number of clusters. This study uses 3 scenarios, namely using 2, 3, and 4 clusters
2. Choosing an initial centroid randomly
3. Calculating Euclidean distance.
4. Defining new group members.
5. Calculating new centroid
6. Repeating process c to e until there is no change in the members of each cluster and the minimum SSE is obtained
7. Generating cluster members in each scenario
8. Evaluating the cluster formed based on the silhouette coefficient
9. The best scenario is chosen based on point h.

D. RESULTS AND DISCUSSION

1. The Results K-means Algorithm

This study uses 3 scenarios to get the optimal and best cluster based on the objective function criteria and the silhouette coefficient. The first scenario uses 2 clusters, the second uses 3 clusters, and the third uses 4 clusters. Before analyzing, the 10 indicators of inclusive economic development are standardized first because those indicators have different units. The first stage is to select the centroids randomly. A selection of the centroids can be seen in Table 2 below.

Table 2. Selection of Initial Centroids from Each Cluster

Scenario	District/City	Variable									
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1 (2 clusters)	Pacitan	26999,6	14968,4	79,55	0,95	71,52	12,61	7,19	75,86	0,35	8527
	Surabaya	186738,9	538845,5	68,61	5,87	73,98	14,78	10,46	130,55	0,36	17157
2 (3 clusters)	Pacitan	26999,6	14968,4	79,55	0,95	71,52	12,61	7,19	75,86	0,35	8527
	Surabaya	186738,9	538845,5	68,61	5,87	73,98	14,78	10,46	130,55	0,36	17157
	Bangkalan	24360,9	23846,7	63,11	5,84	69,94	11,58	5,33	186,11	0,29	8393
3 (4 clusters)	Surabaya	186738,9	538845,5	68,61	5,87	73,98	14,78	10,46	130,55	0,36	17157
	Sampang	19726,4	19105,4	66,42	2,81	67,79	11,76	4,36	202,21	0,26	8569
	Kediri City	447215,8	127716,8	64,6	4,22	73,8	14,96	9,91	20,54	0,33	11976
	Batu city	76004,9	15640,9	71,01	2,48	72,37	14,04	8,77	7,89	0,33	12466

Based on Table 2, it is known that Surabaya is selected in all scenarios and Pacitan is selected in scenario 1 and scenario 2. The next stage is creating clusters and determining the members of each cluster based on Euclidean distance. The results of the k-means algorithm of 10 indicators of inclusive economic development are:

Table 3. Membership of Each Cluster

Scenario	Cluster	Amount	Maximum Iteration
1 (2 clusters)	Cluster 1	29	6
	Cluster 2	9	
2 (3 clusters)	Cluster 1	16	4
	Cluster 2	5	
	Cluster 3	17	
3 (4 clusters)	Cluster 1	1	5
	Cluster 2	15	
	Cluster 3	1	
	Cluster 4	21	

The maximum iteration is based on the value of the objective function in the iteration. If SSE is used as an objective function and produces the smallest value, the cluster formation process is stopped (Gbadoubissa et al., 2020). Based on Table 3, scenario 2 reaches the optimum in the fourth iteration, and this is the least number of iterations of all the scenarios. Meanwhile, the first scenario has the most iterations, namely optimum in the sixth iteration. Based on the results of the iteration in Table 3, scenario 2 is the best.

The results of k-means clustering of inclusive economic development for all of the scenarios can be seen in the Figure 1.

DOI: https://doi.org/10.30812/varian.v5i2.1894

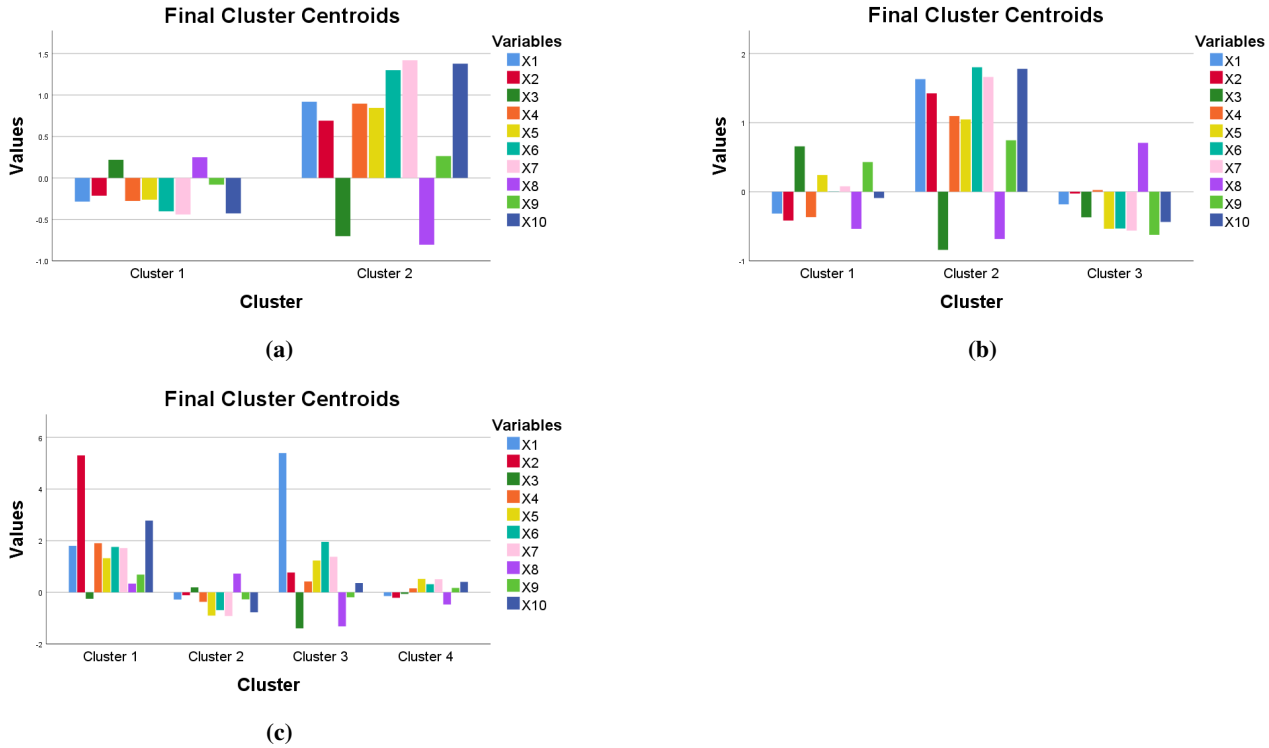


Figure 1. Results of the Cluster Centroids

Based on Figure 1, information about the last centroid of the 10 indicators of inclusive economic development can be obtained. If a centroid is negative, then the cluster members are below the total average and if a centroid is positive, then the members of a cluster are above the total average. From Figure 1, it is known that the last centroids between cluster 1 and cluster 2 of scenario 1 don't have the same criteria. In other words, each indicator in cluster 1 and cluster 2 have a different value. For example, GRDP per capita in cluster 1 has a negative value, but in cluster 2 has a positive value. This result is not the same in scenario 2 and scenario 3. In scenario 2 and scenario 3, the last centroids have the same criteria. For example in scenario 2, GRDP per capita of cluster 1 and cluster 3 have the same criteria (both values are negative). The details of the final result for the cluster centroids are shown in Table 4.

Table 4. Results of the Cluster Centroids

Scenario	Cluster	The Value of Indicators
1 (2 clusters)	Cluster 1	$X_1\downarrow, X_2\downarrow, X_3\uparrow, X_4\downarrow, X_5\downarrow, X_6\downarrow, X_7\downarrow, X_8\uparrow, X_9\downarrow, X_{10}\downarrow$
	Cluster 2	$X_1\uparrow, X_2\uparrow, X_3\downarrow, X_4\uparrow, X_5\uparrow, X_6\uparrow, X_7\uparrow, X_8\downarrow, X_9\uparrow, X_{10}\uparrow$
2 (3 clusters)	Cluster 1	$X_1\downarrow, X_2\downarrow, X_3\uparrow, X_4\downarrow, X_5\uparrow, X_6-, X_7\uparrow, X_8\downarrow, X_9\uparrow, X_{10}\downarrow$
	Cluster 2	$X_1\uparrow, X_2\uparrow, X_3\downarrow, X_4\uparrow, X_5\uparrow, X_6\uparrow, X_7\uparrow, X_8\downarrow, X_9\uparrow, X_{10}\uparrow$
	Cluster 3	$X_1\uparrow, X_2\downarrow, X_3\downarrow, X_4\uparrow, X_5\downarrow, X_6\downarrow, X_7\downarrow, X_8\uparrow, X_9\downarrow, X_{10}\downarrow$
3 (4 clusters)	Cluster 1	$X_1\uparrow, X_2\uparrow, X_3\downarrow, X_4\uparrow, X_5\uparrow, X_6\uparrow, X_7\uparrow, X_8\uparrow, X_9\uparrow, X_{10}\uparrow$
	Cluster 2	$X_1\downarrow, X_2\downarrow, X_3\uparrow, X_4\downarrow, X_5\downarrow, X_6\downarrow, X_7\downarrow, X_8\uparrow, X_9\downarrow, X_{10}\downarrow$
	Cluster 3	$X_1\uparrow, X_2\uparrow, X_3\downarrow, X_4\uparrow, X_5\uparrow, X_6\uparrow, X_7\uparrow, X_8\downarrow, X_9, X_{10}\uparrow$
	Cluster 4	$X_1\downarrow, X_2\downarrow, X_3\downarrow, X_4\uparrow, X_5\uparrow, X_6\uparrow, X_7\uparrow, X_8\downarrow, X_9\uparrow, X_{10}\uparrow$

From Table 4 and Figure 1, in the first scenario, the centroid values of the 10 indicators creating clusters in scenario 1 with scenario 2 are in different ranges. In the second scenario, there are several components whose centroid values are in the same range or intersect. For example, the centroid values for the Life Expectancy indicator in cluster 1 and cluster 2 are both negative. However, the centroid values are not the same. Likewise, with scenario 3, there are several indicators in each cluster that have a centroid value that is in one range, below the total average or above the total average. For example, the GRDP per capita indicator in cluster 1 and cluster 2 are both above the total average. To determine the type or name of each group in each scenario, it can calculate the average of the resulting centroids in each indicator. Determination of the best scenario of k-means clustering in the grouping of Districts/Cities in East Java Province based on 10 indicators of inclusive economic development, the silhouette

coefficient can be used.

2. Cluster evaluation with silhouette coefficient

The silhouette coefficient is one of the measurement criteria to determine the best number of clusters used in clustering (Corporal-Lodangco et al., 2014), (Naghizadeh and Metaxas, 2020). The results of the k-means clustering evaluation using the silhouette coefficient in the 3 scenarios used are:

Table 5. Silhouette Coefficients of Each Scenario

Scenarios	Silhouette Coefficient
1 (2 clusters)	0.87
2 (3 clusters)	0.72
3 (4 clusters)	0.79

The silhouette coefficient has a value from -1 to 1. There are 3 levels in the silhouette coefficient, namely $s < 0.2$ is a poor category, $0.2 \leq s \leq 0.5$ is a fair category, and $s > 0.5$ is a good category (Naghizadeh and Metaxas, 2020), (Mooi et al., 2011). Table 5 shows that the silhouette coefficients generated from the three scenarios used in the k-means clustering analysis are more than 0.5, meaning that all scenarios produce a good number of clusters. Of the three scenarios, scenario 2 is the best k-means clustering analysis on the grouping of Districts/Cities in East Java based on an inclusive economic development index. The silhouette coefficient is 0.87.

Based on the results listed in Table 5, the k-means clustering analysis on the grouping of the inclusive economic development index in East Java is using 2 clusters. The results are as follows:

Table 6. Membership of 2 Clusters in the First Scenario

Cluster	Amount	Districts/Cities
1	29	Pacitan, Ponorogo, Trenggalek, Tulunganggung, Blitar, Kediri, Malang, Lumajang, Jember, Banyuwangi, Bondowoso, Situbondo, Probolinggo, Pasuruan, Mojokerto, Jombang, Nganjuk, Madiun. Magetan, Ngawi, Bojonegoro, Tuban, Lamongan, Bangkalan, Sampang, Pamekasan, Sumenep, Probolinggo City, Batu City
2	9	Sidoarjo, Gresik, Kediri City, Blitar City, Malang City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City

Table 6 shows that 76% of Districts/Cities in East Java are in cluster 1. Based on Figure 1 and Table 4, cluster 1 is a group with most of the indicators that have a negative value centroid. Indicators in cluster 1 are GRDP per capita, GRDP by business field, unemployment rate, Gini ratio, expenditure per capita, the number of poverty, Expectation years of schooling, and Mean years of schooling. The low unemployment rate and the number of poverty indicate that unemployment and poverty in this cluster are better than in cluster 2. However, 6 of the other indicators that are lower than the total average suggest that inclusive economic growth is lower than cluster 2. Most of the Districts/Cities in cluster 1 are coastal and border areas in East Java, where most of the population work as farmers, fishermen, or laborers (Agency, 2021), (Statistics-East Java, 2020). Determination of the cluster's name or type can see in the last centroids that result (Corporal-Lodangco et al., 2014), (Clayman et al., 2020). This study shows that the overall centroids in cluster 1 are lower than in cluster 2. Cluster 1, the average of all the indicators is lower than the total average of inclusive economic development indexes in East Java. Districts/Cities in cluster 1 are Pacitan, Ponorogo, Trenggalek, Tulunganggung, Blitar, Kediri, Malang, Lumajang, Jember, Banyuwangi, Bondowoso, Situbondo, Probolinggo, Pasuruan, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Bojonegoro, Tuban, Lamongan, Bangkalan, Sampang, Pamekasan, Sumenep, Probolinggo City, Batu City.

Based on the description in the paragraph above, cluster 1 in scenario 1 is a group where the average inclusive economic development index of the 10 categories used is below the total average. Then it can be said that cluster 2 is above the total average. Cluster 2 consists of big cities or industrial areas, namely Sidoarjo, Gresik, Kediri City, Blitar City, Malang City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City.

E. CONCLUSION AND SUGGESTION

This study represents a situation of clustering Districts/Cities in East Java Province based on the economic inclusive development data. 76% of Districts/Cities are in cluster 1 and 24% are in cluster 2. Cluster 1 is a group of Districts/Cities with an inclusive economic development index below the total average of inclusive economic development index in Jawa Timur and cluster 2 is a

group of Districts/Cities with an inclusive economic development index above the total average. In other words, cluster 2 shows a higher inclusive economic growth than cluster 1.

Based on the descriptions above, k-means results can describe the spread of inclusive economic development in East Java. The results of this clustering can be used as a reference for local and provincial governments as a basis for policy/decision making in improving the quality and quantity of indicators for inclusive economic development, particularly in Districts/Cities with indicators of inclusive economic development below the total average.

REFERENCES

Agency, N. D. P. (2021). *Indeks Pembangunan Ekonomi Inklusif*. http://inklusif.bappenas.go.id/indeks.

Akramunnisa, A. and Fajriani, F. (2020). K-Means Clustering Analysis pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. *Jurnal Varian*, 3(2):103–112.

Clayman, C. L., Srinivasan, S. M., and Sangwan, R. S. (2020). K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes. *Procedia Computer Science*, 168:97–104.

Corporal-Lodangco, I. L., Richman, M. B., Leslie, L. M., and Lamb, P. J. (2014). Cluster analysis of north atlantic tropical cyclones. *Procedia Computer Science*, 36:293–300.

Gbadoubissa, J. E. Z., Ari, A. A. A., and Gueroui, A. M. (2020). Efficient K-Means Based Clustering Scheme for Mobile Networks Cell Sites Management. *Journal of King Saud University-Computer and Information Sciences*, 32(9):1063–1070.

Hapsari, W. R. (2019). Analisis Pertumbuhan Ekonomi Inklusif Kabupaten/Kota di Provinsi Jawa Tengah. *Jurnal Litbang Sukowati: Media Penelitian dan Pengembangan*, 3(1):11–11.

Hozumi, Y., Wang, R., Yin, C., and Wei, G.-W. (2021). UMAP-Assisted K-Means Clustering of Large-Scale SARS-CoV-2 Mutation Datasets. *Computers in biology and medicine*, 131:104264.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern recognition letters*, 31(8):651–666.

Kakushadze, Z. and Yu, W. (2017). *K-Means and Cluster Models for Cancer Signatures. *Biomolecular detection and quantification*, 13:7–31.

Mooi, E., , and Sarstedt, M. (2011). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics (1st ed.)*. Springer-Verlag Berlin Heidelberg.

Naghizadeh, A. and Metaxas, D. N. (2020). Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. *Procedia Computer Science*, 176:205–214.

Patel, E. and Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, 171:158–167.

Qureshi, M. N. and Ahamad, M. V. (2018). An Improved Method for Image Segmentation using K-Means Clustering with Neutrosophic Logic. *Procedia computer science*, 132:534–540.

Setianingtias, R., Baiquni, M., Kurniawan, A., et al. (2019). Pemodelan Indikator Tujuan Pembangunan Berkelanjutan di Indonesia. *Jurnal Ekonomi Dan Pembangunan*, 27(2):61–74.

Sharma, V. and Bala, M. (2020). An Improved Task Allocation Strategy in Cloud using Modified K-Means Clustering Technique. *Egyptian Informatics Journal*, 21(4):201–208.

Statistics-East Java (2020). *Perkembangan beberapa Indikator Utama Sosial Ekonomi Provinsi Jawa Timur 2020*.

Webster, M. (2021). *Cluster Analysis*. https://www.merriam-webster.com/dictionary/cluster analysis.

Xie, H., Zhang, L., Lim, C. P., Yu, Y., Liu, C., Liu, H., and Walters, J. (2019). Improving K-means Clustering with Enhanced Firefly Algorithms. *Applied Soft Computing*, 84:105763.

Xu, H., Croot, P., and Zhang, C. (2021). Discovering Hidden Spatial Patterns and Their Associations with Controlling Factors for Potentially Toxic Elements in Topsoil using Hot Spot Analysis and K-Means Clustering Analysis. *Environment International*, 151:106456.

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 179–190
e-ISSN: 2581-2017

Mask Compliance Modeling Related COVID-19 in Indonesia Using Spline Nonparametric Regression

Citra Imama[1], M. Haykal Adriansyah[2], Hadi Prayogi[3], Ferdiana Friska Rahmana Putri[4], Naufal Ramadhan Al Akhwal Siregar[5], Alfredi Yoani[6], M. Fariz Fadillah Mardianto[7]
[1,2,3,4,5,6,7]Departement of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received : 04-05-2022
Revised : 04-14-2022
Accepted : 04-28-2022

Keywords:

COVID-19;
SDGs;
Mask Compliance;
Nonparametric Regression;
Spline Estimator.

ABSTRACT

Until now, Coronavirus disease (COVID-19) has become a concern for Indonesia because of its significant development and impact on various sectors of life and hampering the target of achieving Sustainable Development Goals (SDGs). The achievements targeted in the SDGs, such as reducing poverty, hunger, and many more are very difficult to realize in the current pandemic conditions. The uncertain conditions of the pandemic made the government need some new ideas for consideration in creating policies to encourage sustainable development in this situation. This article covers modeling the effect of achieving the second dose of vaccination and the total cases of COVID-19 cases, which are often considered the reason for general negligence in complying with health protocols, especially wearing masks. This research was conducted using spline nonparametric regression because of its flexibility to handle uncertain data patterns. The results of this study are truncated spline nonparametric regression with 3 knots that produce a R-sq equal to 69.952%. Based on the results, the second dose vaccination coverage variables and the total COVID-19 cases together affect mask compliance. This result is expected to be a benchmark for the government to handle COVID-19 and efforts to achieve the SDGs.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1895

Corresponding Author:

M. Fariz Fadillah Mardianto
Departement of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
Email: m.fariz.fadillah.m@fst.unair.ac.id

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. In 2020, COVID-19 was highly concerned in Indonesia because of its significant impact on various sectors of life. The COVID-19 pandemic has affected various sectors worldwide, including hampering the achievement of the Sustainable Development Goals (SDGs) targets. The achievements targeted in the SDGs, namely reducing poverty, hunger, achieving a healthy and prosperous life, quality education, and gender equality, are of course very difficult to realize in today's conditions. After experiencing a rapid increase in cases, the government made a policy to deal with the COVID-19 pandemic with some restrictions to the public. The initial step taken by the government was to enforce Large-Scale Social Restrictions (known as PSBB in Indonesia) and Community Activities Restrictions Enforcement (known as PPKM in Indonesia) that causes all activities in various sectors to be stopped, resulting in a decline in overall economic activity (Iskandar et al., 2020).

The Indonesian government has made a policy to prevent the spread of COVID-19 by urging the public to implement health protocols properly (Indonesia Cable News Network, 2020). Health protocols include wearing masks properly as Coronavirus can spread from the mouth or nose of an infected person in small particles when they cough, sneeze, talk, sing or breathe (World Health Organization, 2021a). In addition, health protocols also need to be supported by physical distancing, washing hands regularly, avoiding crowds, reducing mobility, and vaccinations. Vaccination is a government-provided facility to protect the public from easily

infected with COVID-19 (Ministry of Health, 2021a). In July 2021, The Central Bureau of Statistics released the publication of survey results which stated that some people that hesitated to get vaccinated were 4.2% refused and 15.8% were still not sure. The reasons behind refusal and doubt about vaccines are diverse, such as feeling unnecessary because health protocols are sufficient, not being confident about vaccine safety, doubts about vaccine effectiveness, and fear of vaccine side effects (Central Bureau of Statistics, 2020), in addition to providing complete doses of primary vaccines for all citizens, the government also organizes booster vaccinations. Based on the study results, there has been a decrease in antibodies six months after receiving the complete primary dose of COVID-19 vaccination, so it is necessary to take the booster vaccine. The booster vaccine doses increase personal protection, especially in vulnerable groups (Ministry of State Apparatus Utilization and Bureaucratic Reform, 2022).

The people of Indonesia do not fully implement a health protocol. The Central Bureau of Statistics survey in September 2020 showed that the word lack of awareness was the word most often used as an excuse for not implementing health protocols. The survey said that 39% of the population did not implement the health protocol because no cases of COVID-19 had emerged (Central Bureau of Statistics, 2020). These results indicate that it is essential to know whether the total number of COVID-19 cases affects people's adherence to masks. In addition, citing by detik.com, it was reported that compliance with health protocols had decreased because many felt immune after receiving vaccinations. The National Disaster Management Agency broadcast also stated that compliance with health protocols had indeed reduced after observing a behaviour change monitor (Dwianto, A, 2021).

This study is to model the effect of the achievement of the second dose of vaccine and the total cases of COVID-19 on public compliance in using masks as a benchmark for government policy based on nonparametric regression. Nonparametric regression analysis is flexible because it is not limited by assumptions that need to be met, as in parametric regression. One of the methods is spline, a continuous segmented polynomial slice, so it has the advantage of overcoming data patterns that show sharp ups and downs with the help of knot points (Pratiwi, 2017). The resulting curve is relatively smooth. The best spline regression model depends on the optimal knot point. The methods to find the optimal knot points that are often used are Generalized Cross-Validation (GCV), Mean Squared Error (MSE), and the R-sq (R^2) (Sanusi et al., 2017). The optimal knot point is obtained from the minimum GCV and MSE values and the maximum R-sq.

This research has state of the art spline nonparametric regression implementation related to COVID-19. Research on regression modeling related to COVID-19 has been carried out by (Ogundokun et al., 2020) using linear regression and (Almalki et al., 2022) using linear regression with a spatial approach. Existing studies do not use a non-parametric regression approach, especially the spline estimator, and none has investigated compliance with the use of masks in Indonesia. So, the novelty of this study is modelling the effect of the second dose of vaccine achievement, and the total cases of COVID-19 cases on adherence to wearing masks using nonparametric spline regression analysis and the results can be used as a benchmark for the government in setting policies in the context of sustainable development related to COVID-19 prevention. The urgency of this research is that the unpredictable and unfinished COVID-19 pandemic greatly affects daily life, so new insights are needed as a benchmark for government policies in handling COVID-19 in the future. Especially considering that there are variants of the COVID-19 virus that continue to grow and even create new spikes, such as the case of the Omicron variant, which has developed since the end of December 2021 (Ministry of Health, 2021b).

B. LITERATURE REVIEW

1. Mask Use Compliance

Public understanding of compliance with masks and social distancing is important to control the spread of disease in the absence of a vaccine or if a large proportion of the population refuses to receive vaccinations. Various studies are modelling the spread of COVID-19 support the importance of wearing masks and maintaining physical distance from others (Cohen et al., 2022). One modelling study suggested that 80% compliance with wearing a mask would reduce deaths from COVID-19 by up to 45% (Eikenberry et al., 2020), and it has been suggested that masks can reduce inoculum size, leading to less severe infections (Gandhi and Rutherford, 2020).

Based on data from The Central Bureau of Statistics (Central Bureau of Statistics, 2021), compliance with masks in Indonesia is quite high. The data presented that 88.6% of the community adhered to using masks, 9.1% rarely used masks, and 2.3% of people ignored masks. In addition, based on The Task Force for Handling COVID-19 in Indonesia (Indonesian Task Force for Handling COVID-19, 2021a), most regions in Indonesia have a high level of mask compliance. There are 190 districts or cities with a compliance rate of wearing masks in the range of 91-100%. Nevertheless, 61 districts or cities in Indonesia have a low compliance rate of wearing masks of 75%.

The need to wear masks is likely to continue along with the entry of the Omicron variant in Indonesia. The first confirmed

Omicron infection was on November 9 2021 (World Health Organization, 2021b), and the first case in Indonesia on November 27 2021 (Indonesian Task Force for Handling COVID-19, 2021b). The use of masks in public places is most effective in stopping the spread of the virus when compliance is high (Howard et al., 2021). The use of masks for infected individuals without symptoms can potentially reduce the risk of infecting others when the individual wears a mask to protect himself (Li et al., 2020). Thus, the use of masks is expected to reduce cases of COVID-19 infection.

2. Vaccination of COVID-19

Vaccination aims to bring up a person's immune response to the attack of the COVID-19 virus so that the body can fight infection with the virus. Vaccines are required to reduce COVID-19 related morbidity and mortality, and multiple platforms have been implicated in the rapid development of vaccine candidates (Baden et al., 2020). Of course, the immune system against COVID-19 after being vaccinated does not necessarily form instantly. The health protocols launched by the government must still be implemented to provide maximum protection against COVID-19 attacks (Ministry of Health, 2021a). The Indonesian government, through the Minister of Health, stated that it had distributed 1.2 million doses of COVID-19 vaccine to 34 provinces in Indonesia as of January 7 2021, while the vaccination was planned to be carried out in the second week of January 2021, after the issuance of an Emergency Use Authorization by Food and Drug Supervisory Agency (known as BPOM in Indonesia).

3. Spline Nonparametric Regression

Regression analysis is a statistical method used to investigate and model the relationship between variables (Montgomery et al., 2021). Regression analysis has become one of the most widely used statistical tools for analyzing multivariable data, which provides a simple conceptual method for investigating functional relationships between variables (Oyeyemi et al., 2015).

Regression analysis relies on several assumptions, where the type of relationship between the dependent and independent variables is essential to know (Garba et al., 2021). Three approaches can be used in regression analysis to estimate the regression curve: parametric, nonparametric, and semiparametric. According to Erilli and Alakus (Erilli and Alakus, 2014), parametric regression relies on several assumptions. Meanwhile, estimates formed when the assumptions are not met will result in poor estimates so that to make better assumptions, nonparametric regression models can be used.

Many nonparametric regression approaches have been developed, including using splines. Spline regression is an alternative to polynomial regression, also called segmented regression, because the method focuses on fitting a set of models to various segments of the relationship between the response variable and the predictor variable (Darlington and Hayes, 2017). The spline model has high flexibility with an excellent ability to handle data with behaviour changes at certain sub-intervals (Sohibien et al., 2022). According to Sohibien (Sohibien et al., 2022), one of the advantages of the spline approach is that this model seeks its estimation by following the movement of data patterns.

In the nonparametric regression method, the relationship between the two data is paired $(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)$, $i = 1, 2, \ldots, n$ can be explained by the following Equation 1.

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{pi}) + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

Where ε_i is a random error which is assumed to be identical, independent, and normally distributed with and $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma^2$.

If the regression curve is an additive model, then the equation can be described as Equation 2

$$f(x_{1i}, x_{2i}, \ldots, x_{mi}) = f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_p(x_{pi}) = \sum_{j=1}^{p} f_j(x_{ji}), \quad i = 1, 2, \ldots, n \tag{2}$$

If the regression curve $f_j(x_j)$ is assumed to be contained in a spline space of order m with knot points $K_{1j}, K_{2j}, \ldots, K_{rj}, j = 1, 2, \ldots, p$, the general equation for the univariate spline nonparametric regression model is as Equation 3.

$$f_j(x_j) = \sum_{k=0}^{m} \beta_{jk} x_j^k + \sum_{u=1}^{r} \beta_{j(m+u)} (x_j - K_{ju})_+^m \tag{3}$$

So that the equation of the multivariate nonparametric regression model is obtained as follows :

$$y_i = \sum_{j=1}^{p}\sum_{k=0}^{m}\beta_{ji}x_j^k + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_j - K_{ju})_+^m + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

where

$f(x_i)$	:	spline regression function
K_{ju}	:	knot point
x_{ji}	:	predictor variable j to $i, i = 1, 2, \ldots, n$
β	:	constant
u	:	denotes knot point, $u = 1, 2, \ldots, r$

and $(x_i - k_i)^m$ declare a cut (truncated) function, which can be described as model 5 :

$$(x_j - K_{ju})_+^m = \begin{cases} (x_j - K_{ju})^m & , x_{ij} \geq K_{ju} \\ 0 & , x_{ij} < K_{ju} \end{cases} \tag{5}$$

If $m = 1, 2,$ and3, we get linear spline, quadratic spline, and cubic spline, and K_{ju} is the knot point.

The knot point is a joint point where there is a change in behaviour in the data. According to Sohibien (Sohibien et al., 2022), the knot point is the melting point where the function changes its pattern at different sub-intervals. The methods to find the optimal knot points that are often used are Generalized Cross-Validation (GCV) and $R - sq(R^2)$.

1. Generalized Cross-Validation (GCV)

   The criterion used as a performance measure for a good estimator is the Generalized Cross-Validation (GCV). The optimal value chosen was based on the smallest GCV value (Mardianto et al., 2021). In general, GCV is defined as Equation 6 :

$$GCV(\boldsymbol{K}) = \frac{MSE(\boldsymbol{K})}{(n^{-1}\text{trace}[\boldsymbol{I} - \boldsymbol{S_\lambda}]^2)} \tag{6}$$

   where

$\boldsymbol{I}$	:	identity matrix
n	:	number of observations
$\boldsymbol{S_\lambda}$	:	is a sized matrix nn with the equation $\boldsymbol{S_\lambda} = (\boldsymbol{I} + \boldsymbol{\lambda K})^{-1}$

   with

$$MSE(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right) \tag{7}$$

   The minimum MSE value from calculating Equation 7 indicates that the estimated value is close to the true value (Maharani and Saputro, 2021).

2. $R - Sq(R^2)$

   The $R - sq(R^2)$ is a tool to measure the proportion of variance or total variance around the mean, which can be explained by the regression model. Based on Mardianto (Mardianto et al., 2021), the best estimator is based on the smallest MSE and largest R^2 values. The formula can be written as Equation 8 :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \times 100\% \tag{8}$$

   with

$\hat{y}_i$	:	the estimated value of the i^th response variable
$\bar{y}$	:	average response variable
y_i	:	the value of the i^th response variable

C. RESEARCH METHOD

In this study, the data analyzed were the level of compliance to the use of masks based on the achievement of the second dose of vaccine in 34 provinces in Indonesia in the peak period of COVID-19 from 12th to 18th July 2021. The data used is secondary data obtained from the official government website covid.go.id, vaccine.kemkes.go.id, and kawalcovid19.id. The research variables used in this study consisted of the dependent and independent variables, which are shown in Table 1.

Table 1. Research Variable

Variable Type	Variable	Operational Definition	Scale
Independent	Achievement of the 2^{nd} Dose of Vaccination (X_1)	Percentage of second vaccine dose received by each province	Ratio
	Total COVID-19 Cases(X_2)	The number of positive cases of COVID-19 in each province	
Dependent	Compliance Rate Wearing Mask(Y)	The level of compliance with wearing masks based on The Central Bureau of Statistics data	

Data analysis steps in the study, the effect of the second dose of vaccination, and the total cases of COVID-19 cases on public compliance with using masks as a benchmark for government policy are as follows:

1. Conduct descriptive analysis to find out the description of the data used.
2. Perform analysis with spline regression
   (a) Produce modelling with nonparametric spline regression using three-knots point with the GCV method
   (b) Select the optimal knot point using the GCV method
   (c) Generate data modelling with spline nonparametric regression using optimal knot points from the GCV method
   (d) Produce a comparison of nonparametric spline regression models with optimal knot points obtained from the GCV method
   (e) Comparing the results and choosing the best model with the R-sq and MSE criteria

D. RESULTS AND DISCUSSION

1. Descriptive Statistics

Descriptive statistics are statistics used to describe data into clearer and easier-to-understand information that provides an overview of the research. Based on the analyzed data, the following descriptive statistics were obtained:

Table 2. Descriptive Statistical Analysis

	N	Minimum	Maximum	Mean
Mask Compliance	34	14.49	98.76	84.7221
Second Dose of Vaccine	34	4.10	24.63	8.2126
COVID-19 Cases	34	0.30	7.06	1.1160
Valid N	34			

Table 2 shows that the average level of mask compliance in 34 provinces in Indonesia is 84.72%. However, the difference between the maximum percentage (North Kalimantan) and the minimum (North Maluku) is still significant. This means that while around 97.33% of the top five provinces have succeeded in achieving these results, the same cannot be said about the condition in North Maluku as the province with the lowest level of compliance.

Table 3. Five Provinces with The Highest Mask Compliance Rate

No	Province	Mask compliance (%)	Average
1	North Kalimantan	98.76	
2	Bali	98.29	
3	Central Kalimantan	98.12	97.33%
4	West Sulawesi	97.06	
5	Yogyakarta	94.44	

Table 3 lists the five most mask-compliant provinces in Indonesia. This also shows the average of the top five, which is 97.33% and is higher than the average of all 34 provinces in Indonesia, as shown in Table 2.

Table 4. Five Provinces with The Highest Mask Compliance Rate

Variables	Minimum		Maximum	
	Value	Province	Value	Province
Mask Compliance Rate (Y)	14.49	North Maluku	98.76	North Kalimantan
Second Vaccination Coverage (X_2)	4.10	Lampung	24.63	Jakarta
COVID-19 Positive Cases in Indonesia (X_1)	0.301	North Sumatra	7.056	Jakarta

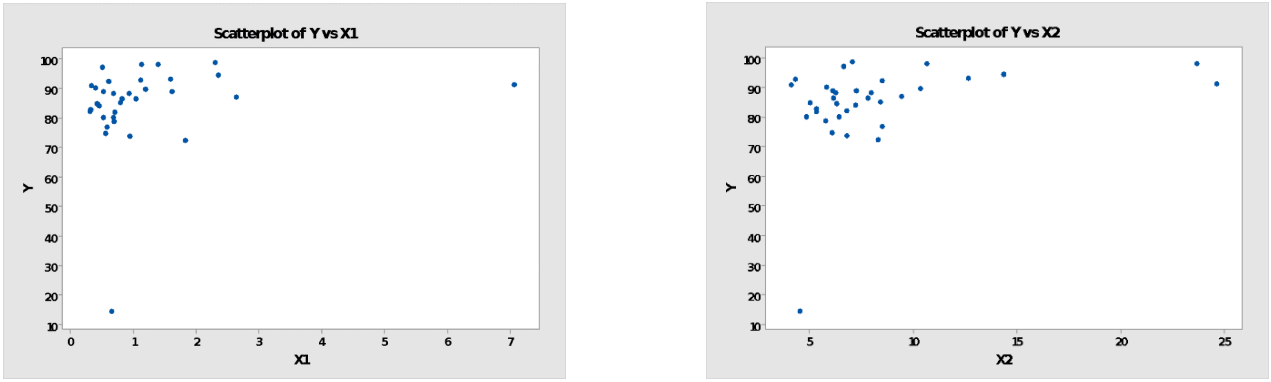
DOI: https://doi.org/10.30812/varian.v5i2.1895

Table 4 below shows more information about the maximum and minimum values of the variables. In this analysis, two types of variables were used, namely, the mask compliance rate (Y) as the dependent variable, the second vaccination coverage (X_1), and the total positive cases of COVID-19 (X_2) as independent variables.

2. Identification of Relationship Patterns Between Predictor Variables on Response

Identification of the pattern of relationships between predictor variables and response variables is the first thing that must be done to determine the proper method for modelling it. The relationship pattern can be seen visually through a scatterplot. The modelling can use a parametric approach if the scatterplot forms a quadratic, linear, or other pattern. Meanwhile, it can take a nonparametric approach if the scatterplot does not have a pattern.

It can be seen visually through Figure 1 that the scatterplot between the response variable, namely the percentage of mask compliance with the predictor variable X_1 namely, the percentage of achievement of the second dose of vaccination did not form a particular pattern. Therefore, the estimation of the model cannot be done using a parametric regression approach so that the variable is a nonparametric component.



(a) **(b)**

Figure 1. Variable Scatterplot Between Predictors to Response

It can be seen visually in Figure 1 that the scatterplot between the response variables, namely the percentage of mask compliance with the predictor variable X_2. The percentage of total positive cases of COVID-19 does not form a particular pattern. Therefore, the estimation of the model cannot be done using a parametric regression approach so that the variable X_2 is a nonparametric component.

Based on the explanation above, it can be concluded that the pattern of the relationship between the percentage of compliance wearing masks with the predictor factors does not form a particular pattern.

3. Spline Regression

The best spline nonparametric regression model is a model that has an optimal knot point. Node points or knot points are in a changing pattern of function behaviour. One of the methods commonly used to select the optimal knot point is the GCV method. The optimal knot point is obtained from the minimum GCV value. The knot points used in this study are limited to a one-knot point, two-knots point, and three-knots point.

After modelling nonparametric spline regression using one-knot point, two-knots point, and three-knots point (related parsimonious model), the minimum GCV values can be compared to select the best model. The model with the lowest GCV value will be chosen as the best model. The table below compares knots one, two, and three.

Table 5. The Best Model Selection

Model	Minimum GCV	R-sq
Knot 1	221.520	12.212
Knots 2	111.289	64.276
Knots 3	109.183	69.952

Table 5 shows the minimum GCV value of each knot. Because the three-knots point model has the lowest GCV value

DOI: https://doi.org/10.30812/varian.v5i2.1895

(109.183), it can be concluded that the best spline nonparametric regression model is the one that uses a three-knots point. The spline nonparametric regression model formed is as Equation 9

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)^1 + \hat{\beta}_3 (x_1 - K_2)^1 + \hat{\beta}_4 (x_2 - K_3) + \hat{\beta}_5 x_2 + \hat{\beta}_6 (x_2 - K_4)^1 + \hat{\beta}_7 (x_2 - K_5)^1 + \hat{\beta}_8 (x_2 - K_6)^1 \tag{9}$$

Below is a table containing GCV values for a spline nonparametric regression model with three-knots point.

Table 6. Selection of Optimum Knot Point with Three Knots

X_1	X_2	GCV
4.519	0.439	
4.938	0.577	112.385
6.195	0.990	
4.519	0.439	
4.938	0.577	109.183
6.614	1.128	
4.519	0.439	
4.938	0.577	113.055
7.033	1.266	

Based on Table 6, the minimum GCV value is 109.183, with the optimal knot point locations for each variable are as follows:

$$X_1 : (K_1 = 4.519 \quad K_2 = 4.938 \quad K_3 = 6.614)$$
$$X_2 : (K_4 = 0.439 \quad K_5 = 0.577 \quad K_6 = 1.128)$$

Table 7. Spline Nonparametric Regression Parameters

Variable	Parameter	Coefficient
	$\hat{\beta}_1$	-135.848137
X_1	$\hat{\beta}_2$	246.108305
	$\hat{\beta}_3$	-113.662566
	$\hat{\beta}_4$	4.545979
X_2	$\hat{\beta}_5$	53.922192
	$\hat{\beta}_6$	-95.869004
	$\hat{\beta}_7$	56.441824
	$\hat{\beta}_8$	-17.695708

Modelling the effect of the second dose of vaccination and the total cases of COVID-19 cases on community compliance in using masks as a benchmark for government policy with parameters according Tabel 7 as Equation 10 :

$$\hat{y} = 634.021 - 135.848 x_1 + 246.108 (x_1 - 4.519)_+^1 - 113.663 (x_1 - 4.938)_+^1 + 4.546 (x_2 - 6.614) + 53.922 x_2 - 95.869004 (x_2 - 0.439)_+^1 + 56.442 (x_2 - 0.577)_+^1 - 17.696 (x_2 - 1.128)_+^1 \tag{10}$$

So, based on the best spline nonparametric regression model using three-knots, the coefficient determination of the model is 69.952%, which means that the variable coverage of the second dose of vaccination and the total cases of COVID-19 cases affects the level of mask compliance for up to 69.952% and other factors influence the rest.

4. Significant Test

The significance test of the regression model parameters was carried out to determine whether the predictor variables significantly affected the mask compliance rate. There are two stages in testing the significance of the parameters. The first stage is conducting simultaneous tests. If the conclusion of the simultaneous test shows that there is at least one significant parameter, then proceed to the individual test.

DOI: https://doi.org/10.30812/varian.v5i2.1895

1. Simultaneous Testing
   Simultaneous testing is carried out to determine the significance of the regression model parameters together. The hypotheses of the simultaneous test are as follows:
   H_0 : $\beta_1 = \beta_2 = 0$
   H_1 : There is at least one , $\beta_j \neq 0; j = 1, 2$

Table 8. ANOVA Table Parameter Testing Simultaneously

Source of Variation	DF	SS	MS	F
Regression	8	4672.429	584.0537	.275054
Error	25	2007.042	80.2817	
Total	33	6679.472		

Table 8 shows the ANOVA table of simultaneous parameter tests. A decision is taken to reject Ho if the F-value is greater than the F-table, namely $F_{(0.05,8,25)}$. It is known that the F value is 7.275054, and the value of $F_{(0.05,2,31)}$ is 2.337, so a decision can be taken to reject Ho, which means that there is at least one parameter that has a significant effect on the percentage of compliance using masks. From this conclusion, it can be a continued individual or partial test.

2. Partial Test

Table 9. Partial Parameter Test Results

Variable	Parameter	Coefficient	t_{value}	P-value	Decision
	β_0	634.021275	5.8236166	0.000004	Reject H_0
	β_1	-135.848137	-5.3828794	0.000014	Reject H_0
X_1	β_2	246.108305	5.8574862	0.000004	Reject H_0
	β_3	-113.662566	-5.0520496	0.000032	Reject H_0
	β_4	4.545979	0.6641848	0.512651	Failed to reject H_0
	β_5	53.922192	1.1383173	0.265779	Failed to reject H_0
X_2	β_6	-95.869004	-1.2796018	0.212434	Failed to reject H_0
	β_7	56.441824	1.4022720	0.17313	Failed to reject H_0
	β_8	-17.695708	-2.1619903	0.040389	Reject H_0

Partial or individual testing is carried out if, at the same time, testing the model parameters, it is concluded that there is at least one significant parameter. It aims to determine which parameters that have or do not have a significant affect on the regression model. The hypothesis used in testing the individual parameters is as follows:
H_0 : $\beta_j = 0$
H_1 : $\beta_j \neq 0; j = 1, 2$
Table 9 is the result of testing the model parameters partially. If value if $|t_{\text{value}}|$ is greater than the value of $|t_{\text{table}}| = t_{(0.025,25)}$ is 2.060 and the P-value is less than $\alpha = 0.05$ the decision to reject H_0.
Based on Table 9, it is known that five decision parameters reject Ho, which means that these parameters are significant to the model. While four other parameters have a P-value of more than, the decision is to fail to reject Ho, which means that the parameter is not significant to the model. However, even though there are insignificant parameters, these variables are still used because there is one significant parameter at least in one variable. So that the predictor variable X_1 and X_2 has a significant effect on the percentage of compliance using masks. This model can be used to represent the effect of the second dose of vaccination achievement and total COVID-19 cases on mask compliance because it has an R-sq of 69.952%, which means that the second dose of vaccination achievement and the total cases of COVID-19 cases affects the level of mask compliance up to 69.952% and other factors influence the rest.

Using the analysis results, several recommendations regarding the public mask compliance for the Indonesian government were formulated, namely as follows.

1. The second dose of vaccination achievement percentage illustrates the uneven distribution of vaccines. Vaccine distribution is a challenge for the government, considering that there are still many difficult areas to reach. In this case, the government needs to allocate appropriate funding to implement vaccination evenly to areas that are difficult to reach because it is following the analysis results that vaccination achievement has a significant effect on public mask compliance.
2. Based on the analysis results, it can be seen that the total cases of COVID-19 have a significant effect on public mask compliance. In this case, transparency and ease in accessing accurate COVID-19 data are essential so that the public knows the

DOI: https://doi.org/10.30812/varian.v5i2.1895

current conditions and is aware of the risks.

3. In general, the level of mask compliance in Indonesia is quite good, with an average of 97.33%, but certain areas still have very low levels of compliance, such as North Maluku. This could be due to the uneven distribution of socialization regarding the importance of health protocols so that people do not have sufficient provisions to implement them. Therefore, the government needs to take a strategic approach to distribute information on the importance of implementing health protocols evenly.

These recommendations cannot be carried out by the governments alone. Indonesian society must work together to implement the protocols that have been established. Through this research, the central and regional governments are expected to continue to provide access to testing, tracing, and treatment transparently without any discrimination against the public and to ensure that the vaccination process is carried out quickly, evenly, and safely to immediately achieve group immunity so that it can control the spread of COVID-19 effectively. If strategic efforts can handle COVID-19, the target for achieving the SDGs can be realized.

E. CONCLUSION AND SUGGESTION

Public mask compliance is significantly affected by the second dose of vaccination achievement and the total number of COVID-19 cases. Based on the analysis and discussion results, it was found that the province with the highest percentage of compliance using masks was North Kalimantan Province at 98.76%, and the lowest was North Maluku Province at 14.49%. In addition, the best nonparametric spline regression model for modelling the percentage of community compliance in using masks in Indonesia is to use a three-knots point with a minimum GCV value of 109,183. Thus, it can be concluded that with the R-sq of the best model of 69.952%, it means that the achievement variable for the second dose of vaccination and the total cases of COVID-19 cases affects the level of mask compliance up to 69.952%, and the rest is influenced by other factors, each variable of which the achievement of the second dose of vaccination and the total cases of COVID-19 provides a significant effect on the percentage of compliance using masks. This research has contributed to a statistical study based on nonparametric spline regression modeling related to the factors that affect the level of mask compliance in Indonesia that are influenced by other factors vaccines dose 2 and coronavirus cases which previously did not exist. Recommendations from the research that have been presented at the end of the results and discussion section can have implications for the government and the general public if carried out. However, similar statistical modeling is still needed using other estimators in nonparametric regression so that a model that may be better than this result can be obtained.

REFERENCES

Almalki, A., Gokaraju, B., Acquaah, Y., and Turlapaty, A. (2022). Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues. In *Healthcare*, volume 10, page 324. MDPI.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., et al. (2020). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*.

Central Bureau of Statistics (2020). COVID-19. *Central Bureau of Statistics of Indonesia*.

Central Bureau of Statistics (2021). Community Behavior during PPKM. *Central Bureau of Statistics of Indonesia*. Jakarta. Retrieved February 4 2022.

Cohen, D. A., Talarowski, M., Awomolo, O., Han, B., Williamson, S., and McKenzie, T. L. (2022). Increased Mask Adherence After Important Politician Infected with COVID-19. *PloS one*, 17(1):e0261398.

Darlington, R. B. and Hayes, A. F. (2017). Regression Analysis and Linear Models. *New York, NY: Guilford*, pages 603–611.

Dwianto, A (2021). Prokes Compliance Drops Due to COVID-19 Vaccination Many Feel Immune? https://health.detik.com/berita-detikhealth/d-5512629. Retrieved February 4 2022.

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., and Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious disease modelling*, 5:293–308.

Erilli, N. A. and Alakus, K. (2014). Non-parametric regression estimation for data with equal values. *European Scientific Journal*, 10(4).

Gandhi, M. and Rutherford, G. W. (2020). Facial masking for Covid-19potential for "variolation" as we await a vaccine. *New England Journal of Medicine*, 383(18):e101.

Garba, N., Danchadi, N., and Abdulmumin, M. (2021). Evaluating the Performance of Ordinary Least Square and Polynomial Regression with Respect to Sample Size. *International Journal Of Science for Global Sustainability*, 7(4):25–31.

Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H.-M., von Delft, A., Price, A., Fridman, L., Tang, L.-H., et al. (2021). An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences*, 118(4).

Indonesia Cable News Network (2020). Survey: Only 64.8 Percent of Indonesian People Want to be Vaccinated Against Corona. Retrieved February 2 2022, https://www.cnnindonesia.com/nasional/20201031162756-20-564421.

Indonesian Task Force for Handling COVID-19 (2021a). National Level Health Protocol Compliance Monitoring. COVID-19 National Task Force. , https://covid19.go.id/. Retrieved February 4 2022.

Indonesian Task Force for Handling COVID-19 (2021b). The Origin of the First Omicron Variant Case in Indonesia COVID-19 National Task Force. COVID-19 National Task Force. , https://covid19.go.id/. Retrieved February 4 2022.

Iskandar, A., Possumah, B., and Aqbar, K. (2020). The Role of Islamic Economics and Social Finance during the Covid-19 Pandemic. *SALAM: Jurnal Sosial Dan Budaya Syar-I*, 7(7).

Li, T., Liu, Y., Li, M., Qian, X., and Dai, S. Y. (2020). Mask or no mask for COVID-19: A public health and market study. *PloS one*, 15(8):e0237691.

Maharani, M. and Saputro, D. (2021). Generalized Cross Validation (GCV) in Smoothing Spline Nonparametric Regression Models. In *Journal of Physics: Conference Series*, volume 1808, page 012053. IOP Publishing.

Mardianto, M. F. F., Siti Maghfirotul Ulyah, S., Ardhani, B. A., Aprilianti, N. A., Rahmadina, R. F., et al. (2021). AMiBI: Application of Flood Mitigation in Indonesia based on the Results of Statistical Analyses of Causal Factors using Local Linear Estimators in Nonparametric Regression Model. *Journal of Southwest Jiaotong University*, 55(6).

Ministry of Health (2021a). Covid-19 Vaccination Protect Yourself Protect the Country. *Indonesia Ministry of Health*, 9:22–50.

Ministry of Health (2021b). Omicron Variant Detected in Indonesia. *Jakarta: Indonesia Ministry of Health*.

Ministry of State Apparatus Utilization and Bureaucratic Reform (2022). The Ministry of Health Issues a Circular on the Implementation of Booster Vaccinations. *Jakarta: Ministry of State Apparatus Utilization and Bureaucratic Reform of Indonesia.*

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Ogundokun, R. O., Lukman, A. F., Kibria, G. B., Awotunde, J. B., and Aladeitan, B. B. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, 5:543–548.

Oyeyemi, G. M., Bukoye, A., and Akeyede, I. (2015). Comparison of outlier detection procedures in multiple linear regressions. *American Journal of Mathematics and Statistics*, 5(1):37–41.

Pratiwi, L. P. S. (2017). PERBANDINGAN METODE CROSS VALIDATION DAN GENERALIZED CROSS VALIDATION DALAM REGRESI NONPARAMETRIK BIRESPON SPLINE. *Jurnal Varian*, 1(1):43–53.

Sanusi, W., Syam, R., and Adawiyah, R. (2017). Model regresi nonparametrik dengan pendekatan spline (studi kasus: Berat badan lahir rendah di rumah sakit ibu dan anak siti fatimah makassar). *JMathCos (Journal of Mathematics, Computations, and Statistics)*, 2(1):70–81.

Sohibien, G. P. D., Laome, L., Choiruddin, A., and Kuswanto, H. (2022). COVID-19 Pandemics Impact on Return on Asset and Financing of Islamic Commercial Banks: Evidence from Indonesia. *Sustainability*, 14(3):1128.

World Health Organization (2021a). Considerations in adjusting public health and social measures in the context of COVID-19. *World Health Organisation Interim Guidance, November, 113*. https://www.who.int/publications/i/item/considerations-in-adjusting-public-health-and-social-measures-in-the-context-of-covid-19-interim-guidance.

World Health Organization (2021b). Enhancing Readiness for Omicron (B.1.1.529): Technical Brief and Priority Actions for Member States. https://www.who.int/docs/default-source/coronaviruse/2021-12-23-global-technical-brief-and-priority-action-on-omicron.pdf?sfvrsn=d0e9fb6c_8. Retrieved February 3 2022.

Vol. 5, No. 2, April 2022, Hal. 179–190
DOI: https://doi.org/10.30812/varian.v5i2.1895

Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 191–198
e-ISSN: 2581-2017

K-Prototypes Algorithm For Clustering The Tectonic Earthquake In Sulawesi Island

Suwardi Annas[1], Irwan[2], Rahmat H.S[3], Zulkifli Rais[4]
[1,3,4]Statistics Department, Universitas Negeri Makassar, Indonesia
[2]Mathematics Department, Universitas Negeri Makassar, Indonesia

Article Info

Article history:

Received : 04-20-2022
Revised : 04-28-2022
Accepted : 04-30-2022

Keywords:

Algorithm K-Prototype;
Clustering;
Earthquake;
Magnitude.

ABSTRACT

Natural disasters related to the tectonic earthquakes frequently occur in Sulawesi Island, mainly in Central Sulawesi and West Sulawesi. This study aims to cluster the tectonic earthquakes occurrence in the range of 2017 to 2020. The variables used were magnitude, depth, and distance category. The characteristic of tectonic earthquakes produces a mixed type of objects between numeric and categorical type attribute. The method of k-prototypes algorithm was proposed for clustering the data because it can be used to handle on data mixed numeric scale and categorical scale. The study resulted four clusters in 2017, six clusters in 2018, five clusters in 2019, and six clusters in 2020. These clusters were formed based on the results cluster on a ratio of within-cluster distance against between-cluster distance. It can be related to the active fault on Sulawesi Island. The characteristics of clusters form each year are the greater magnitude. The result of study also showed that the used of k-prototype algorithm can properly classify the occurrence of tectonic earthquakes on the Sulawesi Island.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1908

Corresponding Author:

Suwardi Annas
Department of Statistics, Universitas Negeri Makassar.
Email: suwardi_annas@unm.ac.id

This is an open access article under the CC BY-SA license.

A. INTRODUCTION

Cluster analysis is one of the topics of multivariate statistical analysis or statistical learning, which is also known as unsupervised learning (Ansori Mattjik and Sumertajaya, 2011). Cluster analysis is the process of collecting n objects into k groups with k less than n (Ji et al., 2012). Objects with similar characteristics to each other are grouped into a group, while other objects are collected in different clusters. The group formed is called the cluster (Nooraeni et al., 2021). The similarity between objects is obtained based on the variables that characterize the observed objects. To measure the similarity, it is conducted by using the concept of distance. Mathematically, the smaller the distance between objects, the more similar the objects are and vice versa. The concept of distance that is commonly used is Euclidean distance (Dinh et al., 2021).

According (Pham et al., 2011) further mentioned that there are two main problems that need to be considered in non-hierarchical clustering, namely the number of clusters and the selection of cluster centre's because the clustering results depend on the selected centroids. Another challenge encountered is the type of variable that characterizes the objects (Li et al., 2019). Characteristics of objects consisting of numerical variables are measured by Euclid distance as in the k-means algorithm (Akramunnisa and Fajriani, 2020) (Annas et al., 2022). Furthermore, the characteristics of objects consisting of categorical variables can be measured using the mode, the smaller the value of the mode, the more similar objects are and vice versa. This concept is used in the k-modes algorithm, where the mode is the centroid of a cluster (Mau and Huynh, 2021).

When the object characteristics consist of numeric and categorical variables, the concept of distance that can be used is a combination of the concepts of k-means and k-modes distances (Kuo et al., 2021) (Nooraeni et al., 2021). In this case, k-prototype

method was proposed because the objects that are often encountered in real-world databases are mixed type objects between numeric and categorical (Kuo and Wang, 2022). Furthermore, this method can overcome the challenges of large-scale data compared to hierarchical-based method (Pham et al., 2011).

The characteristics of data resulted from the tectonic earthquakes events in Sulawesi Island are mixed type objects between numeric scale and categorical scale. Therefore, this study proposed the use of k-prototype algorithm for clustering the data. The study will cluster the areas on the island of Sulawesi related to the earthquake events base on the data of the strength and potential for regional earthquakes in South Sulawesi. So that the results of this study can be taken into consideration in the preparation of disaster mitigation policies.

B. LITERATURE REVIEW

The k-prototypes algorithm is one of the clustering methods based on partitioning (Pham et al., 2011) (Iriawan et al., 2018). This algorithm is the result of the development of the k-means algorithm (Mau and Huynh, 2021) (Ahmad and Dey, 2011) to handle clustering on data with mixed numeric and categorical type attributes (Dinh et al., 2021). The development carried out by Huang maintains the efficiency of the k-means algorithm in dealing with large data and can be applied to numerical and categorical data (Annas et al., 2022). The basic development of the k-prototypes algorithm is in measuring the similarity between the object and its centroid prototype (Pham et al., 2011). In general, the k-prototypes algorithm is divided into three main stages, (Sulastri et al., 2021), as follows: First, initialization of the prototype. In this process, several k-prototypes will be selected randomly from the X dataset according to the specified number of clusters.

Second, allocation of objects in X to the cluster with the closest prototype. Measure the object distance to all prototypes and place the object in the closest cluster. At this stage the k-prototype algorithm allocates all objects in the dataset to the cluster where the prototype of the cluster has the closest distance to the data object. Allocating all objects in data set X to the cluster that has the closest prototype distance to the object being measured. For each time object X has been allocated, the next step will be to calculate the related prototype cluster.

Third, reallocation of objects if there is a change in the prototype. After all objects in X have been allocated, the next step will be to re-measure the distance between all objects in X against all existing prototypes. If an object is found that is closer to another prototype, membership transfer will be carried out and then an update will be made on the old cluster prototype and the new cluster prototype. This process will continue until there are no more changes to the prototype or until the stopping criteria are met.

C. RESEARCH METHOD

The data used in this study was data on the occurrence of tectonic earthquakes on the island of Sulawesi from 2017 to 2020. The variables measured were the strength of the earthquake, the depth of the earthquake and the range of the earthquake. This type of data scale was a combination of numeric data and categorical data. This data was obtained from the Central Statistics Agency (BPS) of South Sulawesi, North Sulawesi, Central Sulawesi, West Sulawesi, Southeast Sulawesi, and Gorontalo.

The procedure for clustering the data by using k-prototype are as follows:

1. Data exploration is carried out in order to identify the relationship of variables by visualization using scatterplot and boxplot
2. Magnitude and depth earthquake is transformed by formula

$$x* = \frac{x - \bar{x}}{s} \tag{1}$$

   where, $x*$ is the transformed variable, x is the original variable, and $\bar{x}$ is the average, and s is the standard deviation
3. Implementation of k-prototype clustering as follows
   (a) Determining the centroid of the cluster as many as the k, where $k < n$, n is the number of samples as the starting point $C_1, C_2, \ldots, C_k$ on every variable $(X_1, X_2, \ldots, X_p)$;
   (b) Calculating the distance or similarity of data points on the data set against the centroid of the cluster, the data points are grouping into the cluster that has the closest distance to the centroid as follows:

$$d(X,Y) = \sum_{j=1}^{p}(x_{jn} - y_{jn})^2 + \gamma \sum_{j=p+1}^{m} d(x_{jc}, y_{jc}) = \begin{cases} 0 & , x_{jc} \neq y_{jc} \\ 1 & , x_{jc} = y_{jc} \end{cases} \tag{2}$$

   $d(X,Y)$ is distance or similarity of object X and Y, p and m are the number of numerical variables and categorical variable respectively, j is the jth variable, n and c is corresponding to numeric and category. The first term is Euclid distance

DOI: https://doi.org/10.30812/varian.v5i2.1908

for numerical characteristics and the second terms is frequency mismatch of level category for categorical characteristics where γ is a parameter that balances the variable scale difference.

(c) Calculating the new centroid of the cluster after all objects have been grouped into clusters, and then re-grouping all objects on the new centroids.

(d) The process would stop if there were not changing to the centroids, or it has been convergent.

4. The optimum cluster selection using diversity values. It is conducted by k optimum selection; Value of k is selecting by using ratio of variety of within-cluster distances (S_W) against variety of between-cluster distance (S_B). The ratio is plotted against the number of clusters (k) and the selected k is whose greatest changing of ratio proposed S_W and S_B of numerical variable is obtained by using Equation 3:

$$S_{W_n} = \frac{1}{K}\sum_{k=1}^{K} S_k, \quad S_{B_n} = \left[\frac{1}{K-1}\sum_{k=1}^{C}(\bar{x}_k - \bar{x})\right]^{\frac{1}{2}} \tag{3}$$

For categorical variable, proposed within and between sums of square are obtained by using Equation 4 as follows:

$$\begin{aligned} S_{Wc} &= (MSW)^{\frac{1}{2}}, \quad MSW = \frac{SSW}{(n-K)} = \frac{1}{(n-K)}\left[\frac{n}{2} - \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{m=1}^{M} n_{mk}^2\right] \\ S_{Bc} &= (MSB)^{\frac{1}{2}}, \quad MSB = \frac{SSB}{(K-1)} = \frac{1}{(K-1)}\left[\frac{1}{2}\left(\sum_{k=1}^{K}\frac{1}{n_k}\sum_{m=1}^{M} n_{mk}^2\right) - \frac{1}{2n}\sum_{m=1}^{M} n_m^2\right] \\ n &= \sum_{k=1}^{K} n_k = \sum_{m=1}^{M} n_m = \sum_{k=1}^{K}\sum_{m=1}^{M} n_{mk} \end{aligned} \tag{4}$$

D. RESULTS AND DISCUSSION

1. Data Description

There are 34 earthquake events with zero magnitude that occurrence in 2017 and 2019, so that the records are removed. The reason why it is removed that tectonic earthquake with zero magnitude, it is not categorized as tectonic earthquake but as an impact of human activities. Hence, the remaining earthquake events are 6493 where 678 in 2017, 2029 in 2018, 2024 in 2019, and 1762 in 2020.

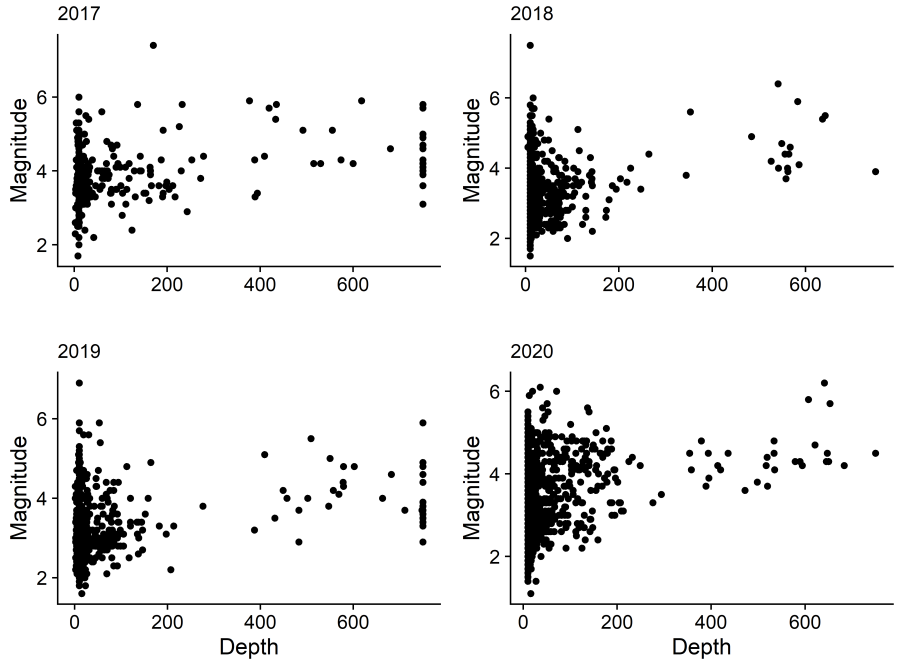


Figure 1. Relationship Between Magnitude and Depth From 2017 Until 2020

Relationship between earthquake magnitude and earthquake depth tends to directly weak relationship. It can be seen on Figure 1 which depicted that point patterns form positive pattern and spread enough for each year. The relationship between earthquake distance category and those numerical variables i.e., depth and magnitude are depicted on Figure 2. It seems like the relationship on Figure 1 (Li et al., 2019).

There is not significantly difference either magnitude against earthquake distance category or depth against earthquake distance category for each year. Therefore, it can be concluded that there is not a significance relationship among used variables. Finally, the scale of magnitude and depth of earthquake is difference as depicted on Figure 1 and Figure 2, so they are transformed using Equation 1.

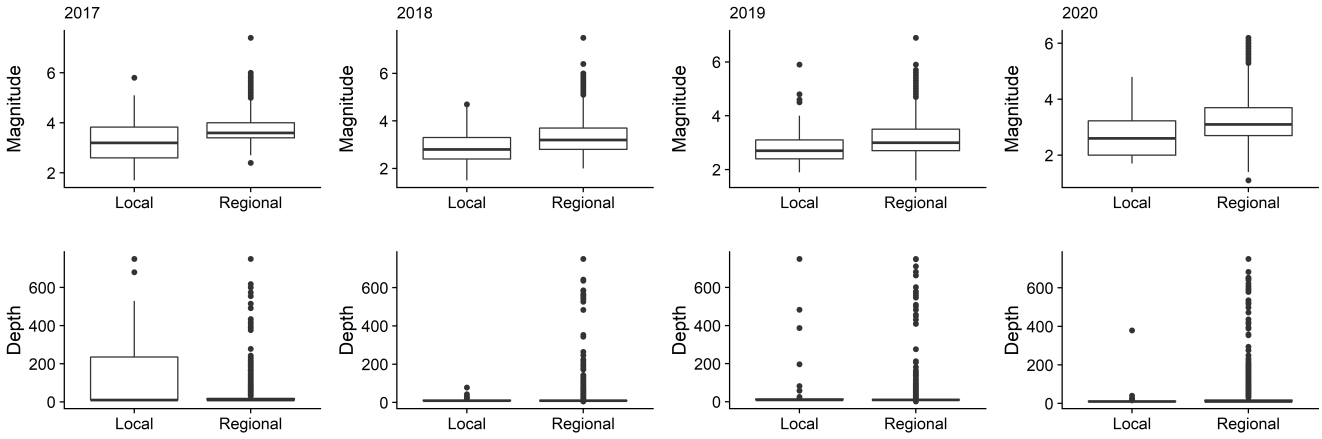


Figure 2. Magnitude and Depth Based on Tectonic Earthquake Distance Category From 2017 Until 2020

2. Clustering by K-Prototypes Algorithm

The clustering of the data was done every year. The values for each year 7.60, 11.43, 14.19, and 14.90. By using these balancing parameters, the optimum k-value for clustering tectonic earthquakes on Sulawesi Island is shown in Figure 3. In general, the value of the SW to SB ratio is fluctuating so that the k value is chosen based on the largest change in the ratio (Kuo and Wang, 2022). The optimum number of clusters with k-values are 4 in 2017, 6 in 2018, 5 in 2019, and 6 in 2020, respectively. The optimum number of clusters in each year is different because this is thought to be related to faults around the island of Sulawesi. The faults around the island of Sulawesi are the Kendari, Makassar, Luwuk, Matano, Palu Koro, Saddang, Walanae, and Poso faults (White et al., 2017).

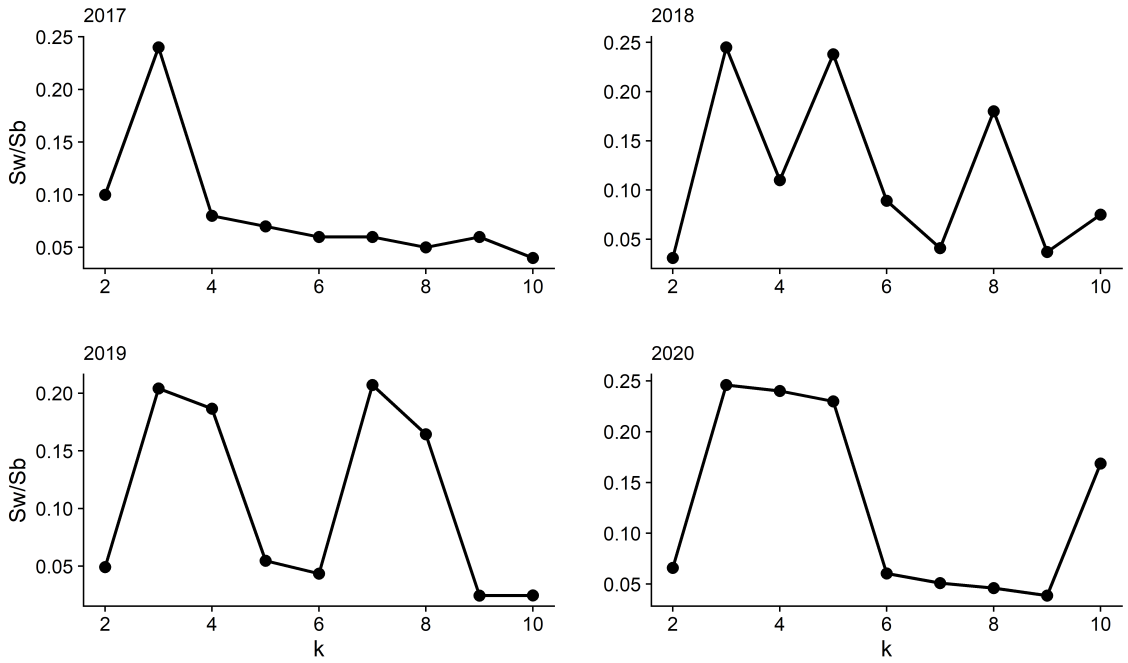


Figure 3. Ratio of S_W and S_B

3. Interpretation of Cluster

The number of elements of clusters for each year is shown on Table 1. Based on the number of elements of cluster, Cluster 3 is the cluster with the most elements for each year. Conversely, Cluster 6 tends to have the fewest elements for each year. This result means that the most tectonic earthquake occur in Cluster 3 followed by Cluster 2 and at least occur in Cluster 6.

Table 1. The Number of Elements of Clusters

Cluster	2017	2018	2019	2020
1	33	673	210	317
2	341	18	349	652
3	215	446	717	668
4	89	242	38	18
5		557	710	41
6		93		30

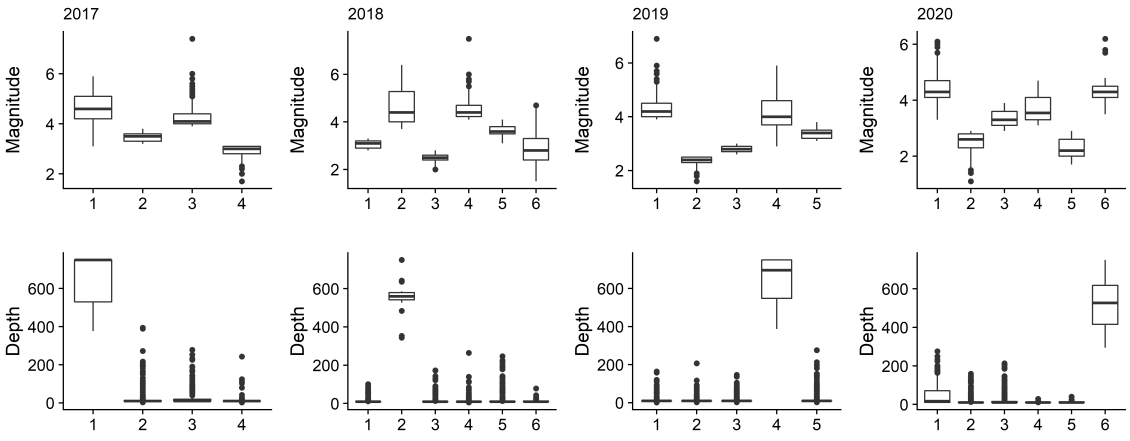


Figure 4. Magnitude and depth of clusters from 2017 until 2020

DOI: https://doi.org/10.30812/varian.v5i2.1908

Figure 4 shows magnitude and depth of tectonic earthquake each cluster. We can see that Cluster 1 and Cluster 2 contain greater magnitude of earthquake than others and Cluster 4 contains the least magnitude of earthquake in 2017. Nevertheless, Cluster 3 and Cluster 4 contain outliers. In 2018, Cluster 2 and Cluster 4 contain greater magnitude of earthquake than the others where the least magnitude of earthquake is Cluster 3. In 2019, Cluster 1 and Cluster 4 contain greater magnitude of earthquake than the others where the least magnitude of earthquake is Cluster 2. Finally, in 2020, Cluster 1 and Cluster 6 contain greater magnitude of earthquake than the others where the lowest magnitude of earthquake is Cluster 5. There is only one cluster contains the deepest for each year.

Furthermore, the depth of earthquake tectonic occurred in Cluster 1 in 2017, Cluster 2 in 2018, Cluster 4 in 2019, and Cluster 6 in 2020. Regarding to the distance category, most of tectonic earthquake from 2017 to 2020 is regional level. There is only a cluster contains local level that is in 2018. It can be related to the tectonic earthquake in Palu and Donggala on September 2018.

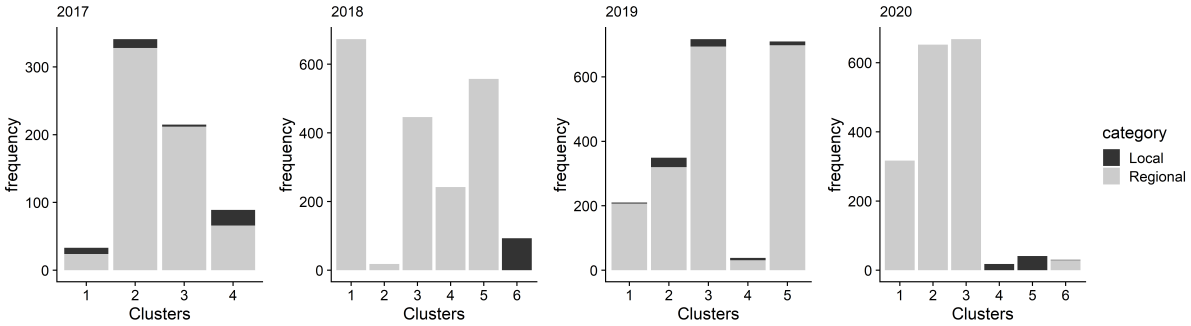


Figure 5. Earthquake distance category of clusters from 2017 until 2020

The tectonic earthquake in Figure 5 shows the number of earthquake events that are distinguished by local earthquakes and regional earthquakes. Earthquake events that occurred from 2017 to 2020, generally occurred in regional earthquakes. Where earthquakes that occur are generally based on fault or fault patterns in each region. In 2017, cluster 2 had the highest earthquake incidence, while in 2018 the highest earthquake occurred in cluster 1. In 2019, the cluster with the highest earthquake incidence was cluster 3, and cluster 6 had the highest earthquake occurrence in 2020. Cluster 2. In 2017, Cluster 1 in 2018, Cluster 3 in 2019 and Cluster 4 in 2020 were the highest because each cluster consisted of regions in the Central Sulawesi.

E. CONCLUSION AND SUGGESTION

The method of k-prototype algorithm was used for clustering the data of tectonic earthquake that occurred in Sulawesi Island in the range of 2017 until 2020. We concluded that this method could cluster the tectonic earthquakes according to depth, strength, and distance category. By implementing the k-prototype algorithm to cluster the Sulawesi Island tectonic earthquake data, the optimum cluster and the best number of clusters were produced. Therefore, it is easier to interpret the strength of tectonic earthquakes. This proposed method was also suitable for clustering the mixed type data between numeric scale and categorical scale, so that it can be used to analyse the same characteristics in another research.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of PNBP Post Graduate Program, Universitas Negeri Makassar. This dataset is based upon work supported by Central Statistics Agency (BPS) of Sulawesi Selatan, Sulawesi Utara, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Barat, Gorontalo, Indonesia.

REFERENCES

Ahmad, A. and Dey, L. (2011). A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters*, 32(7):1062–1069.

Akramunnisa, A. and Fajriani, F. (2020). K-Means Clustering Analysis pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. *Jurnal Varian*, 3(2):103–112.

Annas, S., Uca, U., Irwan, I., Safei, R. H., and Rais, Z. (2022). Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia. *Jambura Journal of Mathematics*, 4(1):167–176.

Ansori Mattjik, A. and Sumertajaya (2011). *Sidik Peubah Ganda Dengan menggunakan SAS*. IPB PRESS Edisi.

Dinh, D.-T., Huynh, V.-N., and Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, 571:418–442.

Iriawan, N., Fithriasari, K., Ulama, B., Suryaningtyas, W., Susanto, I., and Pravitasari, A. (2018). Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification. *Jurnal Ilmu Komputer dan Informasi*, 11(2).

Ji, J., Pang, W., Zhou, C., Han, X., and Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 30:129–135.

Kuo, R.-J., Zheng, Y., and Nguyen, T. P. Q. (2021). Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Information Sciences*, 557:1–15.

Kuo, T. and Wang, K.-J. (2022). A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification. *Computers & Industrial Engineering*, page 108164.

Li, C., Wu, X., Cheng, X., Fan, C., Li, Z., Fang, H., and Shi, C. (2019). Identification and analysis of vulnerable populations for malaria based on K-prototypes clustering. *Environmental research*, 176:108568.

Mau, T. N. and Huynh, V.-N. (2021). An LSH-based k-representatives clustering method for large categorical data. *Neurocomputing*, 463:29–44.

Nooraeni, R., Arsa, M. I., and Projo, N. W. K. (2021). Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering. *Procedia Computer Science*, 179:677–684.

Pham, D.-T., Suarez-Alvarez, M. M., and Prostov, Y. I. (2011). Random search with k-prototypes algorithm for clustering mixed datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2132):2387–2403.

Sulastri, S., Usman, L., and Syafitri, U. D. (2021). K-prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University. *Indonesian Journal of Statistics and Its Applications*, 5(2):228–242.

White, L. T., Hall, R., Armstrong, R. A., Barber, A. J., BouDagher Fadel, M., Baxter, A., Wakita, K., Manning, C., and Soesilo, J. (2017). The geological history of the Latimojong region of western Sulawesi, Indonesia. *Journal of Asian Earth Sciences*, 138:72–91.

Vol. 5, No. 2, April 2022, Hal. 191–198
DOI: https://doi.org/10.30812/varian.v5i2.1908


9 772581 201008