JURNAL EMACS e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.7 No.3 September 2025: 353-362 DOI: 10.21512/emacsjournal.v7i3.14112

Enhancing Tourism Demand Forecasting Accuracy Through Clustering Time Series: A Comparison MAPE Analysis of Indonesian Provincial Domestic Tourist Flows

Mohammad Dian Purnama

Department of Mathematics, Faculty of Mathematics and Natural Sciences,
University of Surabaya,
Surabaya, Indonesia 60231
mohammaddian.20053@mhs.unesa.ac.id

Correspondence: mohammaddian.20053@mhs.unesa.ac.id

Abstract – *The post-pandemic recovery period of the Indonesian tourism sector poses new challenges for accurate tourism demand forecasting across Indonesia's diverse provincial richness. This research aims to enhance the predictive accuracy of domestic tourism demand by comparing conventional single-provincial forecasting methods with clustering-based time series techniques. The Geometric Brownian Motion (GBM) model analyzed data regarding the monthly influx of domestic tourists to 34 provinces from January 2021 to May 2025. This study utilized average linkage agglomerative nesting (AGNES) clustering to discern structural similarities among provinces. Subsequently, silhouette analysis was employed to determine the optimal number of clusters. The findings demonstrate that the cluster-based forecasting approach markedly improved accuracy relative to the non-clustered model. The Mean Absolute Percentage Error (MAPE) for the traditional provincial forecasts was 16.48%. The first cluster-based model had an MAPE of 13.38% and the second cluster-based model had an MAPE of 6.54%. These findings indicate that grouping provinces with analogous temporal patterns enhances the model's ability to identify the underlying dynamics in domestic tourism flows. The work underscores the efficacy of combining stochastic models with hierarchical clustering to enhance evidence-based tourist planning and policy development. This study improves sustainable tourism management by providing an empirical foundation for enhanced forecasting precision, particularly in post-crisis recovery periods.*

Keywords: *Tourism; Forcasting; Geometric Brownian Motion; Clustering.*

I. INTRODUCTION

The global tourism industry, which contributed 10.3% to the world GDP and maintained 333 million jobs prior to the COVID-19 pandemic, experienced unprecedented disruption and corresponding recovery patterns that fundamentally transformed traditional paradigms of prediction (Shen et al., 2023). In the Indonesian case, the post-pandemic recovery period 2021-2025 presents a unique moment of analysis to investigate domestic tourist flows between provinces, when economic stabilization has created more regular patterns than in the pandemic rollercoaster years. This study bridges an important gap in provincial domestic tourist arrivals time series forecasting by investigating whether time series clustering of the provinces' domestic tourist arrivals will significantly improve forecasting accuracy over traditional aggregate procedures, measured through comparisons on Mean Absolute Percentage Error (MAPE).

Theoretical foundation of this research is in the postulation that the tourism destinations of the provinces exhibit heterogeneous temporal patterns due to different socioeconomic, geographical, and infrastructural characteristics. Recent advances in the use of machine learning algorithms in tourism forecasting have achieved appreciable improvements in predictive accuracy, with the literature citing MAPE reduction up to 50% when using sophisticated clustering methods

(Wu et al., 2021). But most of the available literature either addresses international tourist travel streams or single-destination studies, and it is such a significant research void to understand the performance of cluster algorithms used in domestic inter-provincial tourism travel streams of developing economies in post-crisis recovery.

The methodological novelty of this study is the systematic comparison of the predictive performance of the clustered and non-clustered approaches with identical temporal frameworks and evaluation metrics. While traditional tourism demand forecasting treats regional destinations as independent entities, recent research shows that clustering similar destinations by their temporal tourism profiles can reveal underlying structural similarities and improve predictive power (Li et al., 2022). This approach is particularly suitable for the heterogeneous provincial profile of Indonesia, where tourism flows might adhere to similar patterns based on accessibility, tourism facility development, or seasonal attractiveness, irrespective of geographic distance. One study that implemented Time Series Clustering was conducted by Wijaya & Ngatini (2020) on the development of a rice oil price forecasting model in western Indonesia with an average error accuracy rate of 2.05% at the cluster level and 3.52% at the individual level.

Critical examination of existing forecasting techniques recognizes several limitations that are sought to be addressed by this research. First, most tourism forecasting studies employ univariate techniques that fail to exploit cross-sectional data from similar destinations with potential exclusion of useful structural relationships in data (Song & Li, 2008). Second, evaluation of clustering effectiveness in tourism forecasting has been largely anecdotal with minimal systematic comparison of enhancement in forecast accuracy. Third, the post-pandemic period presents unique challenges, but also opportunities, for tourism forecasting because recovery patterns may have fundamentally altered historical relationships, making traditional forecasting approaches less useful. Recent studies have shown that machine learning approaches, including clustering-based approaches, outperform traditional econometric models in handling disrupted data patterns (Gunter & Önder, 2023).

The significance of this research extends from methodological development to practical policy importance to tourism planning and resource management. By demonstrating whether the application of clustered forecasting techniques yields improved accuracy in predicting domestic tourism flows, this study provides empirical justification for tourism planners to make more informed decisions regarding infrastructure investment, marketing initiatives, and capacity planning across provincial destinations. Furthermore, the use of MAPE as the primary evaluation metric adheres to industrial practice in forecast accuracy measurement, ensuring pragmatic usability of the findings. The period of 2021-2025 captures Indonesia's tourism recovery pattern, providing insights into forecast performance throughout the critical period of industry stabilization and growth, thereby contributing to the broader literature on tourism resilience and recovery forecast methodologies.

II. METHODS

This study adopts Geometric Brownian Motion (GBM) as the core method for forecasting tourism demand across Indonesian provinces. The primary objective is to evaluate and compare the predictive accuracy of models applied individually to each province versus those applied to province clusters, using Mean Absolute Percentage Error (MAPE) as the metric for evaluation. Monthly data on domestic tourist arrivals from January 2021 to May 2025 across all 34 provinces were sourced from Central Statistics Agency (BPS) of Indonesia. To ensure consistency throughout the analysis, provinces that experienced administrative restructuring were merged back with their original administrative units.

The forecasting procedure begins with splitting the dataset into training and testing subsets. GBM parameters are estimated through maximum likelihood estimation applied to the logarithmic returns of the time series data. For the baseline model, MAPE is computed for each province, and the average of these values is used to assess overall performance. Next, the provinces are grouped using Agglomerative Nesting (AGNES) clustering with average linkage, based on the similarities in their tourism demand patterns. The optimal number

of clusters is identified using Silhouette analysis. Within each cluster, the monthly tourist arrival figures are averaged to form a representative time series for the group, which is then subjected to the same GBM modeling steps to calculate MAPE for each cluster.

The final stage of analysis involves a systematic comparison of forecasting results. Specifically, the average MAPE from the individual provincial forecasts is compared with that from the clustered models. Figure 1 provides an illustration of these experimental steps.

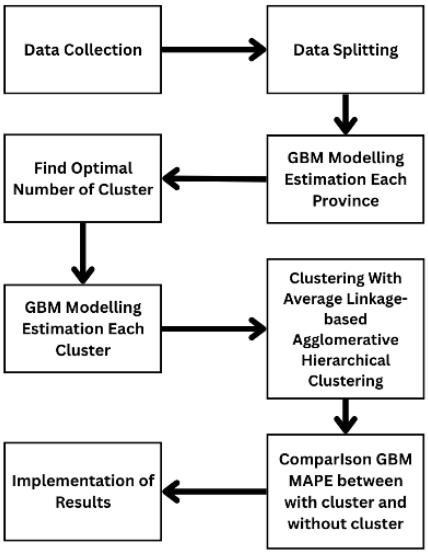

Figure 1. Steps of experiment

2.1 Data Collection

The data collection phase encompasses monthly domestic tourist arrival statistics for all 34 Indonesian provinces covering the period from January 2021 to May 2025, sourced from the Central Statistics Agency (Badan Pusat Statistik). In 2023, there was an expansion of territory in Papua. To address data consistency issues arising from provincial administrative changes and territorial expansion during the study period, provinces that underwent territorial division or administrative restructuring were recombined with their parent provinces. This maintains temporal continuity and ensures a comparable historical basis. The data concerning domestic tourists is displayed in Figure 1.

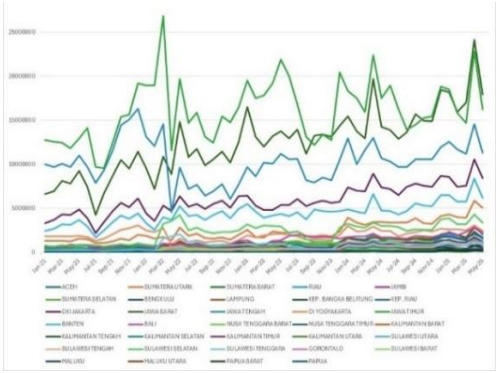

Figure 2. Data of Domestic Tourism

2.2 Geometric Brownian Motion

Geometric Brownian Motion (GBM) is the stochastic process that originated in modeling the price of financial assets but is equally applicable to model time-series data with random variations and overall exponential trends like tourism demand (Sinha, 2021). GBM was further extended in the application of forecasting where the drift and diffusion terms are estimated apart and recursively to illustrate the flexibility beyond the conventional financial applications (Sinha, 2024). Geometric Brownian motion is the combination of a Wiener process with exponential growth to capture the innate randomness and long-term upward direction noted on monthly tourist arrivals. Geometric Brownian motion as the stochastic modeling approach utilizes the window-rolling to estimate the components of the drift and diffusion for generating one-period-ahead forecasts (Chen et al., 2024). For the applications on tourism, Brownian motion concepts were used to interpret mathematical models on the distribution of the random motion based on environmental conditions and psychological choice where the tourist flow dynamics are like the collision Brownian particle effects of the molecule (Li et al., 2020). In this research, GBM is utilized with the assumption that the variations of the number of tourists domestically over time can be viewed as the random process guided by the stochastic differential equation where there is the aspect of both volatility and the dynamics of the growth on the behavior of tourists. The GBM model is expressed in equation 1.

$$dS = \mu S dt + \sigma S dw \quad (1)$$

Where dS is the total number of tourists domestically change, dt is the time interval

between observations, dW is the change in the Wiener process, μ is the drift, and σ is the volatility. The volatility (σ) formula is expressed in equation 2.

$$\sigma = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_k - \bar{R})^2}}{\sqrt{t}} \quad (2)$$

Where n is the sum of return total, R_k is the return value of k, $\bar{R}$ is the average of return, and t is the total of time. The formula for calculating returns is expressed in equation 3. (Maulana et al, 2025).

$$R_{it} = \frac{P_{it} - P_{it-1}}{P_{it-1}} \quad (3)$$

Where R_{it} is the return of tourists domestically total i on month t, P_{it} is the total of tourists domestically i on month t, and P_{it-1} is the total of tourists domestically i on month $t-1$. Then, the formula of drift (μ) is expressed in equation 4.

$$\mu = \frac{\bar{R}}{t} + \frac{\sigma^2}{2} \quad (4)$$

The model in equation 1 can also be written in equation 5.

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t) \quad (5)$$

If equation 5 is written in terms of long-time intervals between consecutive values, the following equation is obtained in equation 6 and equation 7.

$$\frac{dS(t)}{s(t)} = d\big(lnS(t)\big) = \ln\left(\frac{s(t)}{s(t-1)}\right) \quad (6)$$

$$\ln\left(\frac{s(t)}{s(t-1)}\right) = \mu dt = \sigma dW(t) \quad (7)$$

The above equation is obtained from the application formula, which will be explained in general. For each function $G(S,t)$ from S and t where X satisfies the stochastic differential in equation 8.

$$dX = adt + bdW(t) \quad (8)$$

For some constants, a, b, and $dW(t)$ are the Brownian motions. Then, the Ito formula itself is defined in equation 9.

$$dG = \left(\frac{\partial G}{\partial s}a + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial s^2}b^2\right)dt + \frac{\partial G}{\partial s}dW \quad (10)$$

Then, to determine $G(t,S) = ln(S(t))$ in order to fulfil the GBM form, the equation is derived and entered into the Ito formula obtained:

$$d\big(\ln(S(t))\big) = \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW(t) \quad (11)$$

$$\frac{s(t)}{s(t-1)} = \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW(t) \quad (12)$$

If simplified, we will get a stock price estimation model in GBM in equation 13.

$$S_{t+1} = S_t.\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma Wt\right) \quad (13)$$

2.3 Mean Absolute Percentage Error

The study uses the Mean Absolute Percentage Error (MAPE) to serve as the model evaluation metric for the degree of accuracy. It is normal to use the MAPE to set up evaluation predictions against the actual values effect (Purnama, 2025). The formula for the MAPE is calculated in equation 14

$$MAPE = \frac{\sum\left|\frac{Y_t - F_t}{Y_t}\right|}{n} \times 100\% \quad (14)$$

Where Y_t is the value of testing data at time t, F_t is the value of estimation data at time t, and n is the total of testing data. The smaller the MAPE value, the more accurate the model.

2.4 Average Linkage-based Agglomerative Hierarchical Clustering

Clustering is one of the basic unsupervised learning methods that separate objects into disjoint sets along lines of internal differences as well as similarities (Purnama, 2025). A set of clusters holds data that are very close together within a set but clearly separable with members of other sets (Wang et al., 2024). There are different approaches to identify the best number of clusters and one of those approaches is using the silhouette approach. The approach subjects the clusters to ruggedness together with efficiency checking whether there is suitable fit for every individual item within a given cluster or not. The approach uses cohesion as well as separation approaches to come up with its judging (Purnama, 2025). The equation of this approach is in equation x.

$$s(i) = \frac{(\min d(i,C)) - \left(\frac{1}{|A|-1}\sum_{j\in A, j\neq i} d(i,j)\right)}{max\left(\left(\min d(i,C)\right), \left(\frac{1}{|A|-1}\sum_{j\in A, j\neq i} d(i,j)\right)\right)} \quad (15)$$

where $b(i)$ is the minimum average distance between object i and all objects in another cluster C, and $a(i)$ is the average distance between the i-th object and all objects within the same cluster.

2.5 Average Linkage-based Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) is one of those exceptional algorithms that may belong to one of the various methods of utilizing hierarchical clustering (Bateni et al., 2021). The methodology to utilize this algorithm includes a bottom-up approach with all individual data being considered as individual sets at the onset, then iteratively seeking and marrying the most proximal pair to form ever-larger associations until one whole hierarchy is achieved. The work pattern by the Agglomerative Hierarchical Clustering algorithm starts with approximating an overall distance matrix between all sets of data with the metric of utilizing Euclidean distance (Purnama, 2025). The mathematical definition of using Euclidean distance is indicated in equation 16.

$$d(y_i, y_j) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \qquad (16)$$

Where $d(y_i, y_j)$ is Euclidean distance between object i and object j, n is total of variable, x_{ik} is value of object i in variable k, x_{jk} is value of object j in variable k. Then two subsequent clusters with minimum distance by predefined proximity measures are found and joined by algorithm. Distance matrix gets automatically updated to reflect proper spatial relations between new combinations of resulted achieved clusters by themselves and singles ones. Most typical four of such agglomerative linkage schemes that are present within process to create clusters are Single Linkage, Average Linkage, Complete Linkage, and Ward's Method (Mustafidah & Purnama, 2024). Research conducted by Purnama (2025) demonstrates that among these four approaches, the Average Linkage method exhibits superior performance characteristics for hierarchical clustering applications. The Average Linkage approach advances with computation of means of data points distances between all included within clusters resulted as mathematically defined within equation 17.

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j), A \in y_i, B \in y_j \quad (17)$$

Where the total number of $yidi\ A\ dan\ yj$ in B. At each step of average linkage, $D\ (A, B)$ is used to find the combination of the two clusters with the smallest distance (Purnama, 2025).

III. RESULTS AND DISCUSSION

Prior to applying predictive modeling provincial application to local tourist arrival data were under a systematic divide process. The divide plan created two mutually exclusive data sets, namely training corpus and validation corpus. The training corpus was used to foster model development as well as application of clustering analysis, and similarly, the validation corpus was used to foster performance evaluation as well as accuracy estimation of inherent modeling. The data of tourist arrival between January 2021 to July 2023 comprised the training dataset, while those between August 2023 to May 2025 comprised the testing dataset. Application of modeling commenced with Aceh Province being a preliminary case application. As accuracy for prediction encompasses simplification, return computation from dataset could not be ignored by sheer stochastic nature of home country tourism data with huge volatility by means of sharp up-down fluctuations.

Prior to carrying out modeling estimation process, there were required to estimate drift parameter, as well as parameter and volatility coefficient within GBM model structure. The coefficients obtained include drift (μ) is 0.056 and volatility (σ) is 0.285. Such coefficients that had been thus obtained upon such estimation were then formulated to be inserted into modeling of volume of domestic tourists for subsequent intervals of time. The GBM model specified under equation 15 translated into adjusted GBM specification be developed:

$$S_{t+1}^{Aceh} = S_t^{Aceh} . \exp\left(\left(0.056 - \frac{1}{2}0.285^2\right)dt + \sigma Wt\right)$$

When equation that model operated upon 10000 times by simulation and then compared with testing dataset for verification, graphical results obtained are as follows as Figure 3 and Figure 4.

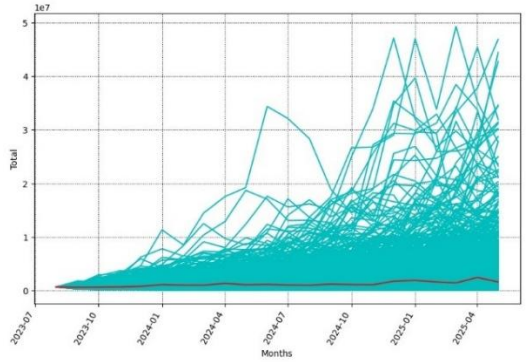

Figure 3. GBM Simulation with 10000 iterations in Aceh Province

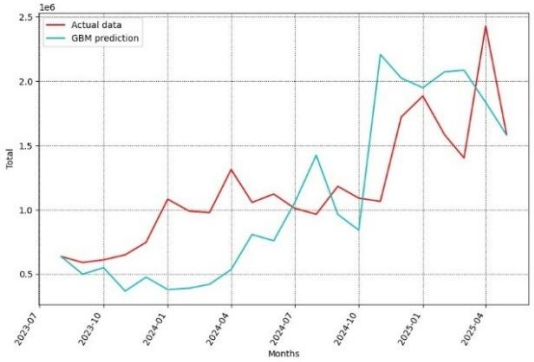

Figure 4. Comparison of actual and GBM Prediction in Aceh Province

From figure 4, it can also be seen that the MAPE model is 14.76%. The accuracy of the model can be seen from the MAPE value. The lower the MAPE value, the more accurate the model. Then, with the same GBM model, $S_{t+1}^{province} = S_t^{province}.\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma Wt\right)$, the GBM modeling for each province can be seen in Table 1.

Table 1. Estimation of Parameters

Province	μ	σ	MAPE (%)
N. Aceh D.	0.056	0.285	14.76
Sumatera Utara	0.050	0.269	10.70
Sumatera Barat	0.051	0.272	11.54
Riau	0.040	0.229	17.08
Jambi	0.041	0.236	11.27
Sumatera Selatan	0.035	0.215	14.09
Bengkulu	0.054	0.282	10.92
Lampung	0.041	0.239	13.33
Kep. Bangka Belitung	0.037	0.203	30.44
Kep. Riau	0.039	0.214	24.82
DKI Jakarta	0.037	0.221	10.98
Jawa Barat	0.046	0.250	17.17
Jawa Tengah	0.050	0.278	6.82
D.I. Yogyakarta	0.041	0.250	7.41
Jawa Timur	0.036	0.227	5.38
Banten	0.039	0.229	14.97
Bali	0.041	0.224	33.35
Nusa Tenggara Barat	0.067	0.287	60.67
Nusa Tenggara Timur	0.074	0.332	11.56
Kalimantan Barat	0.044	0.241	13.78
Kalimantan Tengah	0.029	0.184	10.96
Kalimantan Selatan	0.038	0.220	16.95
Kalimantan Timur	0.045	0.224	52.39
Kalimantan Utara	0.031	0.189	36.62
Sulawesi Utara	0.029	0.186	13.48
Sulawesi Tengah	0.149	0.497	17.16
Sulawesi Selatan	0.200	0.583	20.70
Sulawesi Tenggara	0.750	1.774	33.14
Gorontalo	0.028	0.189	9.12
Sulawesi Barat	0.207	0.595	25.31
Maluku	0.032	0.207	12.80
Maluku Utara	0.036	0.214	19.92
Papua Barat	0.033	0.232	6.42
Papua	0.047	0.264	12.39

As Table 1 indicates, the mean MAPE of the provinces is 18.48%. The time-series will then undergo clustering, then the initial step in the identification of the number of clusters. The number of clusters optimally best to use is through the Silhouette method, also the most used method used in literature. The approach is through the assessment of the quality of clusters using the mean value in such a way that the larger the value, the better the clusters. A plot graph was generated using the Silhouette method, as in Figure 5.

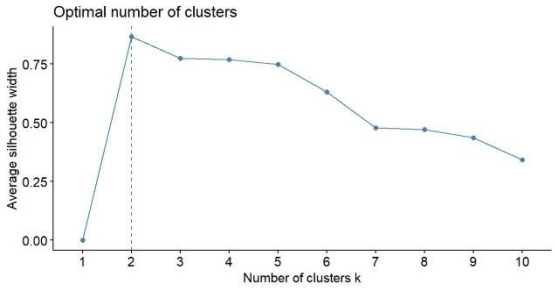

Figure 5. Optimal cluster results using silhouette method

The results of the Silhouette method in Figure 5 show that the optimal number of clusters is two. This can be seen from the vertical line that shows the highest average value among the clusters. According to the Silhouette method, the recommended number of clusters ensures the best data separation, thereby supporting the recommendation regarding the number of clusters to be used. Therefore, the recommended clustering approach is considered the best choice based on the Silhouette evaluation.

The clustering process is depicted using a dendrogram, which presents results from left to right and employs scale lines to indicate distances between merged clusters. Vertical

lines denote cluster mergers. Figure 6 illustrates the clustering outcomes with Average Linkage-based Agglomerative Hierarchical Clustering.

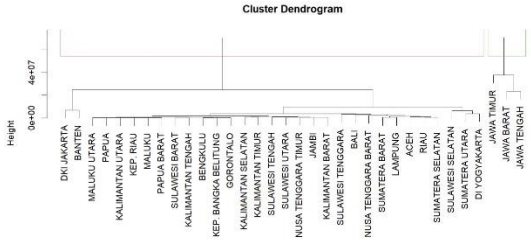

Figure 6. Cluster Results

The Silhouette outcome, as presented in Figure 1, shows that the number of clusters that is the best is two. The trend is presented in the vertical line of the maximum mean of the clusters. The Silhouette approach is a statistical way of obtaining the number of clusters that is the best, thus optimally separating the data and recommending the number of clusters that is to be used. The clustering approach that is thus presented is the solution that is the best, according to the Silhouette assessment.

Following clustering, the model estimation process was carried out, requiring the estimation of drift parameters, as well as volatility parameters and coefficients in the GBM model structure. The first cluster coefficients obtained include drift (μ) is 0.032 and volatility (σ) is 0.201. The GBM model specified under equation 15 translated into adjusted first cluster (fc) GBM specification be developed:

$$S_{t+1}^{fc} = S_t^{fc}.\exp\left(\left(0.032 - \frac{1}{2}0.201^2\right)dt + \sigma Wt\right)$$

When equation that model operated upon 10000 times by simulation and then compared with testing dataset for verification, graphical results obtained are as follows as Figure 7 and Figure 8.

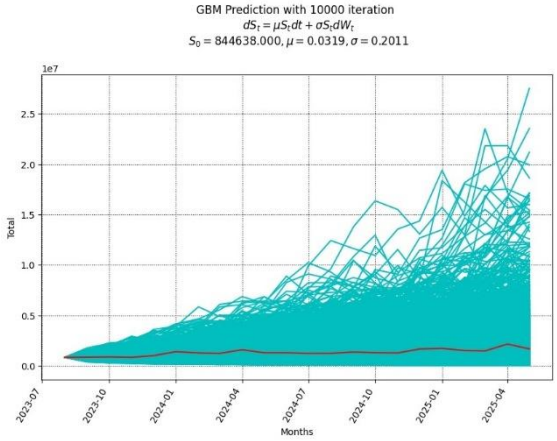

Figure 7. GBM Simulation with 10000 iterations in First Cluster

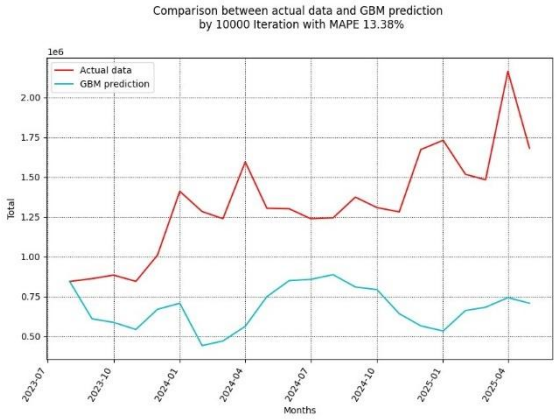

Figure 8. Comparison of actual and GBM Prediction in First Cluster

From figure 8, it can also be seen that the MAPE model is 13.38%. The accuracy of the model can be seen from the MAPE value.

The second cluster coefficient obtained include drift (μ) is 0.036 and volatility (σ) is 0.227. The GBM model specified under equation 15 translated into adjusted second cluster (sc) GBM specification be developed:

$$S_{t+1}^{sc} = S_t^{sc}.\exp\left(\left(0.036 - \frac{1}{2}0.227^2\right)dt + \sigma Wt\right)$$

When equation that model operated upon 10000 times by simulation and then compared with testing dataset for verification, graphical results obtained are as follows as Figure 9 and Figure 10.

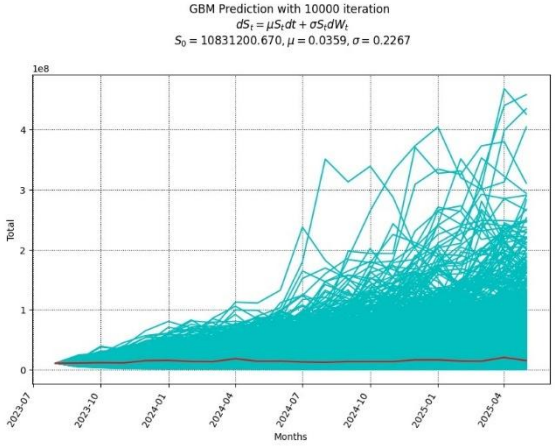

Figure 3. GBM Simulation with 10000 iterations in Second Cluster

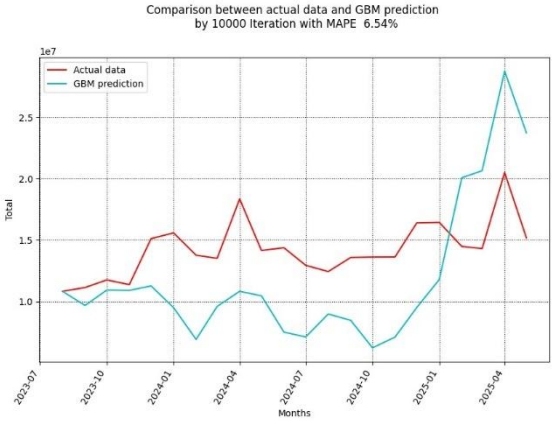

Figure 10. Comparison of actual and GBM Prediction in Second Cluster

From figure 10, it can also be seen that the MAPE model is 6.54%. The accuracy of the model can be seen from the MAPE value.

Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Jakarta, Yogyakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Gorontalo, Central Sulawesi, West Sulawesi, South Sulawesi, Southeast Sulawesi, Maluku, North Maluku, Papua, and West Papua, all of them with the same number of domestic tourists, are in the same group, that is, the first group. Furthermore, the provinces of East Java, Central Java, and the province of West Java also have the same number of domestic tourists, hence reflecting their similarity in the second group. However, we find a considerable difference when the two clusters are compared: the first group indicates a different trend of the data in comparison to the second group. The degree of the model's accuracy for each group could be determined through the value of mean absolute percentage error (MAPE) in Table 6.

Table 6. Comparison of MAPE GBM Models for Each Province and Cluster

Type	Cluster	MAPE (%)
Without Cluster	Without Cluster	16.48
With Cluster	First Cluster	13.38
	Second Cluster	6.54

The value of the mean absolute percentage error (MAPE) for the without clustering is identified through taking the mean MAPE value for each ruling province of Indonesia. As we see in Table 6, the GBM model with the use of the cluster shows that the MAPE is smaller than that of the GBM model for each of the provinces. The range of accuracy assessment of the forecast is as follows, as set through the model. MAPE less than 10% indicates Very strong, 11% to 20% indicates Good, 21% to 50% indicates Reasonable, greater than 51% indicates Inaccurate (Wijaya & Ngatini, 2020). The employment of cluster-level modeling, thus, could be claimed to be a more accurate approach when we compare that with its provincial version. The number of domestic tourists for the year that follows, thus, could be predicted in a more efficient manner using the time series GBM cluster model.

IV. CONCLUSION

According to the research that was done, there is a noticeable increase in accuracy when comparing the MAPE values of models with and without clustering. Clustering successfully decreased the MAPE to 13.38% for the first cluster and 6.54% for the second cluster, but the model without clustering produced a MAPE value of 16.48%, which is in the good category. These findings demonstrate how the clustering technique can significantly improve the GBM model's predictive capabilities. Interesting results can also be found in the performance differences between clusters. For example, the second cluster, which includes the provinces of East Java, Central Java, and West Java, exhibits superior accuracy with a MAPE of 6.54% in the very strong category, while the first cluster achieved 13.38% in the good category. This

indicates that provinces in Java Island possess more homogeneous and consistent characteristics and patterns of domestic tourists, making them more predictable compared to other provinces in Indonesia.

REFERENCES

Bateni, M., Bhaskara, A., Lattanzi, S., & Mirrokni, V. (2021). Scalable hierarchical agglomerative clustering. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 65-75). ACM.

Chen, L., Mercurio, P. J., & Li, Y. (2024). Entropy corrected geometric Brownian motion. Scientific Reports, 14, 28234.

Gunter, U., & Önder, I. (2023). Improved tourism demand forecasting with CIR# model: A case study of disrupted data patterns in Italy. Tourism Review, ahead-of-print.

Li, H., Hu, M., & Li, G. (2022). Analysis of spatial patterns and driving factors of provincial tourism demand in China. Scientific Reports, 12(1), 2100.

Li, Y., Agbam, A., & Chen, J. (2020). Establishment of mathematical model of random motion distribution of Brownian motion based on dynamic management decision of tourists. Electronic Research Archive, 28(1), 217-232.

Maulana, D. A., Sofro, A. Y., Ariyanto, D., Romadhonia, R. W., Oktaviarina, A., & Purnama, M. D. (2025). Stock Price Prediction and Simulation Using Geometric Brownian Motion-Kalman Filter: A Comparison Between Kalman Filter Algorithms. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 19(1), 97-106.

Mustafidah, M. E., & Purnama, M. D. (2024). Grouping of Regencies/Cities in East Java Based on Dengue Fever Case Indicators Using Complete Linkage and Average Linkage (Pengelompokan Kabupaten/Kota Di Jawa Timur Berdasarkan Indikator Kasus Dbd Menggunakan Complete Linkage dan Average Linkage). MATHunesa: Jurnal Ilmiah Matematika, 12(2), 337-343.

Núñez, E., et al. (2024). Machine learning applied to tourism: A systematic review. WIREs Data Mining and Knowledge Discovery, 14(4), e1549.

Purnama, M. D. (2025). Average Linkage-based Agglomerative Hierarchical Clustering of East Java's Economic Development Indicators in 2022 (Average Linkage-based Agglomerative Hierarchical Clustering terhadap Indikator Pembangunan Ekonomi Jawa Timur 2022). Jurnal Sains dan Seni ITS, 12(6), D477-D482.

Purnama, M. D., Yulianto, I. P. R., & Sampoerna, R., (2025). Geometric Brownian Motion on the Prediction of Foreign Exchange Rate: A Study on Indonesian Rupiah Rate. Jurnal Matematika, 15(1), 13-22. https://doi.org/10.24843/JMAT.2025.v15.i01.p182

Shen, W., Li, H., & Zhang, Y. (2023). Smarter sustainable tourism: Data-driven multi-perspective parameter discovery for autonomous design and operations. Sustainability, 15(5), 4166.

Sinha, S. (2021). The reliability of geometric Brownian motion forecasts of S&P500 index values. Journal of Forecasting, 40(8), 1444-1456.

Sinha, S. (2024). Daily and weekly geometric Brownian motion stock index forecasts. Journal of Risk and Financial Management, 17(10), 434.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. Tourism Management, 29(2), 203-220.

Wang, L., Chen, Y., & Kim, S. (2024). Ensemble agglomerative hierarchical clustering with novel similarity measurements. Journal of King Saud University - Computer and Information Sciences, 36(3), 445-458.

Wijaya, S. U., & Ngatini, N. N. (2020). Development of Rice Price Modelling in Western Indonesia Using a Time Series Clustering Approach

(Pengembangan Pemodelan Harga Beras di Wilayah Indonesia Bagian Barat dengan Pendekatan Clustering Time Series). Limits: Journal of Mathematics and Its Applications, 17(1), 51-66.

Wu, Q., Law, R., & Xu, X. (2021). Tourism demand forecasting with time series imaging: A deep learning model. Annals of Tourism Research, 90, 103255.

Zhang, B., Huang, X., Li, N., & Law, R. (2021). Demand forecasting model using hotel clustering findings for hospitality industry. Information Processing & Management, 58(6), 102691.