Jurnal Impresi Indonesia (JII)
p-ISSN: 2828-1284 e-ISSN: 2810-062x
Vol. 5. No. 8. 2026

Clustering Analysis of Tourism Regions in Bali Province Using Similarity Matrix and Log-Log Regression Methods

Candhrika Padma Janardhani*, Nila Armelia Windasari
Institut Teknologi Bandung, Indonesia
Email: candhrikapadma@gmail.com*, nila.armelia@sbm-itb.ac.id

Keywords:
regional clustering, similarity index, log-log regression, GRDP, Bali tourism.

Abstract

The tourism accommodation sector in Bali Province has grown rapidly; however, its distribution across regencies/cities remains highly uneven, motivating the need for a rigorous clustering method to map accommodation-capacity similarities among regions. This study aims to group the nine regencies/cities of Bali Province based on the similarity of ten types of accommodation and tourism-support facilities (star hotels, non-star hotels, villas, apartments, homestays, camping grounds, restaurants, bars, spas, and water tourism facilities), and to examine their relationship with Gross Regional Domestic Product (GRDP). The method combines log-log regression to test the elasticity of star hotel capacity on GRDP, squared Euclidean distance and maximum distance (dmax) to compute a similarity index between every pair of regencies/cities, and the Kruskal-Wallis test to validate differences in GRDP across zones. The regression results show that the star hotel variable has a very strong and statistically significant effect on GRDP ($R^2 = 84.89\%$; $p = 0.00042$). The similarity matrix reveals extreme inequality: Badung Regency shows a similarity value that is essentially zero relative to every other regency/city, whereas Jembrana and Tabanan demonstrate a very high similarity value (0.9547). Based on tertile scores, the regions are classified into three zones: Zone I, representing high capacity (Badung and Denpasar); Zone II, representing medium capacity (Tabanan, Gianyar, and Buleleng); and Zone III, representing low capacity with the highest homogeneity (Jembrana, Klungkung, Bangli, and Karangasem). The Kruskal-Wallis test confirms that the GRDP differences across zones are statistically significant ($H = 7.00$; $df = 2$; $p = 0.0302$). These findings indicate that tourism-equalization policies in Bali should be designed differently for each zone rather than applied uniformly across all regions.

INTRODUCTION

The accommodation sector is one of the main pillars of Bali's tourism industry, contributing significantly to the region's Gross Regional Domestic Product (GRDP). As a critical component of the tourism value chain, the availability and quality of accommodation directly influence tourist satisfaction, length of stay, and overall expenditure, thereby shaping the economic landscape of a destination (Ayele & Singh, 2024; Barrera-Martínez et al., 2025; Cheunkamon et al., 2022; Iriqat, n.d.; Mbau, 2026; Santos, 2024; Yacoub et al., 2025). Over the past decade, the number and variety of accommodation facilities in Bali ranging from star hotels, non-star hotels, villas, and apartments to homestays have grown rapidly (Maulana et al., 2024; Widiantara et al., 2024; Wijayanthi et al., 2024). However, this growth has not been evenly distributed across regencies/cities. Data from the Bali Provincial Tourism Office reveal a stark concentration of accommodation units in the southern region, with Badung Regency

alone accounting for more than 15,000 units, which is over 100 times the capacity of Bangli Regency, which has only 130 units. This disparity is further underscored by the Gini coefficient for hotel distribution in Bali, which stands at 0.54, indicating a moderate-to-high level of inequality.

This spatial inequality has been widely discussed in the Bali tourism literature, which records that the concentration of tourism development in southern Bali (Kuta, Legian, Seminyak, Jimbaran, and Nusa Dua) has driven strong economic growth on the one hand while simultaneously creating an uneven distribution of tourism benefits across regions on the other (Tantra & Murthi, 2025). Studies have consistently highlighted how this tourism-driven development has resulted in significant land-use conversion, increasing property prices, and social challenges, often described as a form of "spatial colonization" (Wirawan et al., 2026). The rapid and often uncontrolled growth in areas such as Canggu and Seminyak has transformed traditional agricultural land into commercial properties (Suamba et al., 2022; Suhirman et al., 2026), displaced local communities, and created economic dependence on tourism, making these communities vulnerable to external shocks (Suyadnya et al., 2025). Such inequality risks producing ineffective policies if the government applies a uniform, one-size-fits-all tourism development strategy without considering the actual characteristics and capacities of each regency/city (Wiryawan et al., 2026).

A quantitative and objective clustering approach is therefore needed to group Bali's regencies/cities according to the similarity of their accommodation characteristics (Kurniawati et al., 2023; Meliyana et al., 2025; Tantra & Murthi, 2025). This approach is essential for moving beyond anecdotal observations toward a data-driven understanding of regional disparities. A distance-based clustering approach enables each pair of regions to be measured using a similarity scale ranging from 0 to 1, allowing policymakers to identify regions that are genuinely comparable and those that represent structural outliers. While previous studies have applied clustering algorithms to tourism data in Bali—such as the use of K-Means and X-Means algorithms for clustering tourist attractions—they have primarily focused on attractions themselves or patterns of tourist movement. In contrast, this study employs an integrated approach combining log-log regression and similarity matrix analysis to cluster regions based on the composition of their accommodation supply, providing a novel perspective for tourism planning. By understanding the current regional characteristics, stakeholders can ensure that all regions and communities benefit from Bali's thriving tourism economy.

A review of existing literature reveals a significant research gap. Many studies on Bali's tourism sector have focused on the impacts of overtourism in southern Bali or on the development of specific accommodation types. However, limited research has conducted a comprehensive quantitative analysis that clusters all nine regencies/cities based on the full spectrum of accommodation and tourism-support facilities while statistically linking these clusters to economic performance (GRDP). This study addresses this gap by providing a systematic, data-driven classification of Bali's tourism regions. This research is particularly relevant because, as emphasized by Coordinating Minister Agus Harimurti Yudhoyono, Bali's development requires a focus on priority clusters to prevent excessive concentration of tourism activities and environmental degradation. The findings provide a timely and evidence-based framework for supporting differentiated tourism development strategies. The urgency of this research is further strengthened by the evolving nature of Bali's tourism landscape, which has

shifted from a relatively stable two-cluster structure toward a more fragmented and diversified network, requiring adaptive management approaches.

The novelty of this research lies in its methodological integration. Unlike previous studies that have applied clustering techniques to tourism data, this study combines a log-log regression model to establish the economic significance of star-hotel capacity with a similarity matrix based on a comprehensive set of ten accommodation and tourism-support variables. This dual approach not only identifies which regions share similar characteristics but also provides an economic rationale for the importance of clustering in policymaking. Furthermore, the use of a similarity matrix that normalizes distances based on a region-specific maximum distance (dmax) provides a robust and intuitive approach for comparing regions, particularly in identifying structural outliers such as Badung Regency. This study also examines whether accommodation capacity—specifically star-hotel capacity as a proxy for formal accommodation supply and quality—has a statistically significant relationship with GRDP, thereby providing an economic justification for promoting more balanced accommodation development.

The purpose of this study is to group the nine regencies/cities of Bali Province based on the similarity of ten accommodation and tourism-support facility types and to examine their relationship with GRDP. The specific objectives are: (1) to determine the elasticity of star-hotel capacity with respect to GRDP using a log-log regression model; (2) to calculate a similarity index between every pair of regencies/cities using squared Euclidean distance and the maximum distance approach; and (3) to validate the resulting regional clusters using the Kruskal-Wallis test to confirm statistically significant differences in GRDP among clusters. The contribution of this research is to provide empirical evidence supporting a differentiated, zone-specific tourism development strategy for Bali, moving beyond a uniform policy approach. The benefit of this research is to provide policymakers with a clear quantitative basis for allocating resources and designing targeted interventions to promote more balanced and sustainable tourism growth across the province.

METHOD

This study used a quantitative approach based on secondary data obtained from the Bali Provincial Tourism Office (2024), including accommodation and tourism-support facility data for nine regencies/cities, and the Bali Provincial Statistics Agency (BPS), including regency/city Gross Regional Domestic Product (GRDP) data at current prices. The study population comprised accommodation and tourism-support facilities across Bali Province, while the sample consisted of nine regencies/cities selected purposively based on the availability and completeness of official data. The analysis was conducted in three main stages using statistical software, namely Microsoft Excel for data processing and calculations and R for statistical testing.

Stage 1 — Log-Log Regression Model: The first stage examined the statistical significance of accommodation variables in relation to GRDP using a log-log regression model with the general form: $\ln(GRDP) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon$, where X_i represents a candidate accommodation variable (in this study, the average number of star hotels from 2021 to 2023). The log-log model allowed the coefficient β_1 to be interpreted as elasticity. Among the tested

variables, the variable with the highest and statistically significant coefficient of determination (R^2, $p < 0.001$) was selected for further analysis.

Stage 2 — Maximum Distance and Relative Distance Formulation: To measure similarity between regions across ten accommodation dimensions simultaneously, the squared Euclidean distance between each pair of regencies/cities (r and s) was calculated using the formula: $d(r,s) = \Sigma_k (X_{rk} - X_{sk})^2$. For each region r, the maximum distance (dmax) from other regions was identified. This value was used as a normalization basis to ensure that regional distances could be compared on a relative scale. The similarity value S(r,s) was calculated using the formula: $S(r,s) = 1 - [d(r,s) / d\_max(r)]$, where S(r,s) ranged from 0 to 1.

Stage 3 — Regional Classification and Statistical Validation: The validity and reliability of the analysis were supported by the use of official government data sources and a systematic, replicable quantitative procedure. Secondary data were collected from published official reports, followed by data cleaning and transformation before analysis. The regions were then grouped using an agglomerative stepwise approach based on the highest similarity values, with the merging process terminated when similarity values showed a substantial decline. Each region's composite score was subsequently classified into three zones based on tertile values derived from average similarity rankings and GRDP values. To validate whether the three-zone classification represented actual economic differences rather than being solely a consequence of the grouping procedure, the Kruskal-Wallis test was applied because it is suitable for comparing more than two independent groups without assuming normally distributed data.

RESULTS AND DISCUSSION

Log-Log Regression Analysis

Among the candidate variables tested, the model ln(GRDP) ~ ln(Star Hotels, average 2021–2023) was selected as the best model, yielding the highest and statistically significant R^2.

Table 1. Selected log-log regression results

Model	R^2	p-value	Notes
ln(GRDP) ~ ln(Star Hotels, average 2021–2023)	0.8489 (84.89%)	0.00042	Significant at α = 0.001; supply-side dimension (formal quality)

Source: Processed by the authors using secondary data obtained from the Bali Provincial Tourism Office (2024) and the Bali Provincial Statistics Agency (BPS Bali Province), analyzed using Microsoft Excel and R statistical software

An R^2 of 84.89% means variation in the number of star hotels explains nearly 85% of the variation in GRDP across Bali's regencies/cities, while a p-value of 0.00042 (well below the 0.001 threshold) indicates that this relationship is highly statistically significant and unlikely to be due to chance. This finding is consistent with several Bali tourism-economics studies showing that formal, high-quality accommodation (star hotels) is strongly associated with regional economic performance, as it reflects large-scale investment and greater appeal to international visitors, generating higher value-added than informal accommodation.

Accommodation and GRDP Data by Regency/City

Table 2. Total accommodation units and GRDP by regency/city in Bali Province (Source:

Regency/City	Total Accommodation Units	Accommodation Rank	GRDP (IDR Billion)	GRDP Rank
Badung	15,779	1	75,098	9
Denpasar	3,256	2	65,300	8
Buleleng	2,231	3	41,387	7
Gianyar	1,820	4	33,046	6
Klungkung	1,453	5	11,015	2
Tabanan	696	6	27,649	5
Jembrana	330	7	16,845	3
Karangasem	307	8	20,210	4
Bangli	130	9	8,465	1

Source: Bali Provincial Tourism Office. (2024). Statistical Data on Accommodation and Tourism-Support Facilities of Bali Province; and Bali Provincial Statistics Agency (BPS Provinsi Bali). (2024). Gross Regional Domestic Product of Regencies/Municipalities in Bali Province by Industry, 2019–2023.

Bali Provincial Tourism Office, 2024; BPS Bali Province)

Table 2 reveals a sharp inequality: Badung Regency holds 15,779 accommodation units (rank 1) — more than 100 times the capacity of Bangli Regency, which has only 130 units (rank 9). Notably, the accommodation-unit ranking does not perfectly align with the GRDP ranking (for instance, Klungkung has a larger total accommodation count than Tabanan, yet a lower GRDP), suggesting that the composition of accommodation types — particularly the share of star hotels — matters more for economic contribution than the sheer number of accommodation units.

Similarity Matrix Analysis

The similarity matrix S(r,s) was computed for all 36 regency/city pairs (a 9 × 9 region matrix). Table 3 presents the full similarity matrix.

Table 3. Similarity matrix S(r,s) between regencies/cities (0–1 scale)

Regency/City	Jembrana	Tabanan	Badung	Gianyar	Klungkung	Bangli	Karangasem	Buleleng	Denpasar
Jembrana	—	0.9547	0.0000	0.8732	0.9248	0.9747	0.9735	0.8620	0.7583
Tabanan	0.9547	—	0.0168	0.9088	0.9183	0.9453	0.9513	0.8934	0.7502
Badung	0.0000	0.0168	—	0.0943	0.0489	-0.0202	-0.0143	0.1110	0.1734
Gianyar	0.8732	0.9088	0.0943	—	0.8841	0.8585	0.8652	0.9507	0.7728
Klungkung	0.9248	0.9183	0.0489	0.8841	—	0.9117	0.9127	0.8783	0.8165
Bangli	0.9747	0.9453	-0.0202	0.8585	0.9117	—	0.9812	0.8438	0.7439
Karangasem	0.9735	0.9513	-0.0143	0.8652	0.9127	0.9812	—	0.8524	0.7422
Buleleng	0.8620	0.8934	0.1110	0.9507	0.8783	0.8438	0.8524	—	0.7722
Denpasar	0.7583	0.7502	0.1734	0.7728	0.8165	0.7439	0.7422	0.7722	—

Source: Calculated and processed by the authors using the Squared Euclidean Distance and Maximum Distance (dmax) similarity matrix based on secondary data from the Bali Provincial Tourism Office (2024).

Two key findings emerge from Table 3. First, the Jembrana–Tabanan pair shows a very high similarity of 0.9547, meaning the two regions' accommodation-capacity composition is nearly identical across all ten measured dimensions. Other pairs with extremely high similarity

include Bangli–Karangasem (0.9812) and Jembrana–Bangli (0.9747), indicating a cluster of small-to-medium-scale regencies that are relatively homogeneous.

Second, and more striking, Badung Regency shows a similarity of 0.0000 with Jembrana, and values close to zero or even negative with nearly every other regency (Tabanan 0.0168; Klungkung 0.0489; Bangli -0.0202; Karangasem -0.0143). These near-zero or negative similarity values occur because Badung's distance from other regions is nearly always the maximum distance (d_max) itself — in other words, Badung is a structural outlier whose accommodation profile is not comparable to any other regency/city in Bali. This finding quantitatively confirms the accommodation-dominance pattern widely discussed in the Bali tourism literature, which highlights how the concentration of tourism development in southern Bali (Kuta, Legian, Seminyak, Jimbaran, Nusa Dua) creates structural inequality relative to other regions.

Tertile-Based Zone Classification

Table 4. Zone classification of regencies/cities based on GRDP tertiles and similarity patterns

GRDP Rank	Regency/City	GRDP (IDR Billion)	Zone
1	Bangli	8,465	III
2	Klungkung	11,015	III
3	Jembrana	16,845	III
4	Karangasem	20,210	III
5	Tabanan	27,649	II
6	Gianyar	33,046	II
7	Buleleng	41,387	II
8	Denpasar	65,300	I
9	Badung	75,098	I

Source: Compiled by the authors based on similarity matrix analysis, agglomerative clustering results, and GRDP data from the Bali Provincial Statistics Agency (BPS Bali Province, 2024).

The stepwise agglomerative merging process based on highest similarity produces three zones consistent with the GRDP distribution. Zone III (High Homogeneity) comprises Bangli, Karangasem, Jembrana, and Klungkung, with similarity remaining above 90% throughout their merging — reflecting a group of low-capacity but relatively uniform regencies. Zone II (Medium Homogeneity) comprises Tabanan, Gianyar, and Buleleng, with similarity ranging from 88% to 94%. The merging process stopped at an 80% threshold: Denpasar's similarity dropped to 77.81% at merge stage 7, and at stage 8 Badung's similarity plunged to 0.00%, so both Denpasar and Badung were separated out as Zone I (high/extreme capacity).

Statistical Validation: Kruskal-Wallis Test

Table 5. Summary of GRDP ranks per zone for the Kruskal-Wallis test

Zone	Members (n)	Sum of Ranks (R)
I	2	17
II	3	18
III	4	10

Source: Processed by the authors using the Kruskal–Wallis nonparametric test based on zone classification and GRDP data analyzed with R statistical software.

Based on Table 5, the H statistic is computed as follows:

$$H = [12 / (9\times10)] \times [(17^2/2) + (18^2/3) + (10^2/4)] - 3(10) = 0.1333 \times 277.5 - 30 = 36.99 - 30 = 7.00$$

With df = k − 1 = 2 (k = 3 zones), an H value of 7.00 compared against the Chi-Square distribution yields a p-value = 0.0302 (computed via the CHISQ.DIST.RT(7.00, 2) function). Since the p-value = 0.0302 < α = 0.05, the null hypothesis is rejected, demonstrating that the GRDP differences across zones are statistically significant and not due to chance. This result validates that the three-zone grouping based on accommodation similarity is not merely a spurious statistical pattern, but genuinely reflects real economic inequality between regions in Bali Province.

CONCLUSION

Based on the log-log regression, similarity matrix analysis, tertile classification, and Kruskal-Wallis test, this study drew several conclusions. First, star-hotel capacity was identified as a strong and statistically significant predictor of regency/city GRDP in Bali (R^2 = 84.89%; p = 0.00042), confirming the importance of high-quality formal accommodation as a driver of the tourism economy. Second, the similarity matrix revealed extreme structural inequality, with Badung Regency emerging as an absolute outlier, showing similarity values close to or equal to zero compared with every other regency/city. In contrast, regions such as Jembrana and Tabanan demonstrated very high similarity in their accommodation characteristics (0.9547).

Third, based on tertile classification and similarity-based merging patterns, Bali's regencies/cities were classified into three zones: Zone I—high capacity (Badung and Denpasar); Zone II—medium capacity (Tabanan, Gianyar, and Buleleng); and Zone III—low capacity with the highest homogeneity (Jembrana, Klungkung, Bangli, and Karangasem). Fourth, the Kruskal-Wallis test confirmed that GRDP differences across zones were statistically significant (H = 7.00; p = 0.0302), indicating that the zoning results reflected actual economic disparities rather than merely being a statistical artifact.

The policy implications of these findings suggest that the Bali Provincial Government should implement a differentiated, zone-specific tourism development strategy. For example, policies in Zone I (Badung and Denpasar) should focus on managing investment concentration and addressing overtourism-related pressures, while Zone III requires targeted incentives and infrastructure support to stimulate accommodation growth without reproducing existing spatial inequalities. Future research is recommended to incorporate additional dimensions, such as tourist arrivals, length of stay, and environmental carrying capacity, to further enhance the robustness of this clustering model.

REFERENCES

Ayele, L. A., & Singh, A. (2024). Tourism value chain, quality tourism experience and competitive advantage: Evidence from star-rated hotels in Ethiopia. *Geo Journal of Tourism and Geosites, 54*, 831–841.

Bali Provincial Statistics Agency (BPS Provinsi Bali). (2024). *Gross regional domestic product of regencies/municipalities in Bali Province by industry, 2019–2023*. BPS Provinsi Bali. https://bali.bps.go.id

Bali Provincial Tourism Office (Dinas Pariwisata Provinsi Bali). (2024). *Statistical data on accommodation and tourism-support facilities of Bali Province, 2024*. Dinas Pariwisata Provinsi Bali.

Barrera-Martínez, A. M., Santana-Talavera, A., & Parra-López, E. (2025). Destination

competitiveness through the lens of tourist spending: A case study of the Canary Islands. *Sustainability, 17*(7), 3262.

Cheunkamon, E., Jomnonkwao, S., & Ratanavaraha, V. (2022). Impacts of tourist loyalty on service providers: Examining the role of the service quality of tourism supply chains, tourism logistics, commitment, satisfaction, and trust. *Journal of Quality Assurance in Hospitality & Tourism, 23*(6), 1397–1429.

Iriqat, M. S. A. (n.d.). *Upgrading the tourism value chain for enhanced hotels' competitive advantage in East Jerusalem*. Al-Quds University.

Kurniawati, I. P., Pratiwi, H., & Sugiyanto, S. (2023). Indonesian territory clustering based on harvested area and rice productivity using clustering algorithm. *Journal of Social Science, 4*(1), 100–110.

Maulana, S. R., Nada, N. M., & Indrajaya, T. (2024). Potential of the hotel and resort accommodation industry in East Lombok District, West Nusa Tenggara Province. *Bali Journal of Hospitality, Tourism and Culture Research, 2*(1), 1–10.

Mbau, C. W. (2026). *Sustainable tourism practices and customer satisfaction of five-star hotels in Nairobi Kenya*. KCA University.

Meliyana, S. M., Dunggio, A. S. S., & Ahmar, A. S. (2025). Performance evaluation of the K-Means clustering method in grouping Indonesian provinces based on potential disaster impact. *JINAV: Journal of Information and Visualization, 6*(2), 123–131.

Santos, E. (2024). Sustainable tourism dynamics: Understanding the impact of tourist stays on regional revenue and development. *Sustainability, 16*(19), 8403.

Suhirman, Sofhani, T. F., Wulansari, T. R., Nurul Annafi, H. R., & Putri, T. A. (2026). Tradition & modernity: Land use change and land tenure pattern in Canggu Village, Bali, Indonesia. *Planning Malaysia, 24*(40). https://doi.org/10.21837/pm.v24i40.1956

Suamba, I. B. S., Wiryasa, N. M. A., Dwijendra, N. M. A., & Putra, I. D. G. A. D. (2022). Characteristics and deviation patterns of agricultural land use in tourism area of Canggu, Bali, Indonesia. *Caspian Journal of Environmental Sciences, 20*(2), 423–430. https://doi.org/10.22124/cjes.2022.5591

Suyadnya, I. W., Cahyadi, R., Tirtayani, L. A., & Nuryani, A. F. (2025). Overtourism sensitivity and tourism development in Canggu, Bali. *Cogent Social Sciences, 11*(1), Article 2521364. https://doi.org/10.1080/23311886.2025.2521364

Tantra, I. G. L. P., & Murthi, N. W. (2025). Tourism growth and investment in Bali's inclusive economic recovery in 2023: A spatial and econometric approach. *Central Community Development Journal, 5*(2), 96–118. https://doi.org/10.55942/ccdj.v5i2.1399

Widiantara, I. G. A. B., Wirya, I. M. S., & Susila, I. M. G. D. (2024). Identification of Badung Government Center (PUSPEM) as a leading family tourism attraction in Bali. *Jurnal Manajemen Pelayanan Hotel, 8*(2), 129–154.

Wijayanthi, P. N. P. M., Mudana, I. G., & Murni, N. G. N. S. (2024). Competitive advantage-oriented marketing with cloud-based PMS for small-scale accommodation in Nusa Penida. *International Journal of Glocal Tourism, 5*(2), 119–134.

Wirawan, K., Adikampana, I. M., Sunarta, I. N., & Adhika, I. M. (2026). The colonization of space in the use of village space in Tegallantang Village, Ubud. *European Modern Studies Journal, 10*(1), 62–88. https://doi.org/10.59573/emsj.10(1).2026.5

Wiryawan, I. W. G., Julianti, L., Permadhi, P. L. O., & Binti Abdullah, N. (2026). Integrated spatial governance for sustainable tourism in Bali. *Journal of Sustainable Development and Regulatory Issues, 4*(1), 1–29. https://doi.org/10.53955/jsderi.v4i1.135

Yacoub, L., ElHajjar, S., Zgheib, Y., & Al Maalouf, N. J. (2025). Understanding perceived value in tourism: Insights from destinations facing crises. *PLOS ONE, 20*(9), e0331144.