Thailand Statistician
July 2025; 23(3): 481-500
http://statassoc.or.th
Contributed paper

Data-Driven Approach in Provincial Clustering for Sustainable Tourism Management in Thailand

Pichit Boonkrong [a], Teerawat Simmachan [b,c], Roumporn Sittimongkol [b] and Rattana Lerdsuwansri [b]*

[a] College of Biomedical Engineering, Rangsit University, Pathum Thani, Thailand
[b] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand
[c] Thammasat University Research Unit in Statistical Theory and Applications, Thammasat University, Pathum Thani, Thailand
*Corresponding author; e-mail: rattana@mathstat.sci.tu.ac.th

Received: 6 August 2024
Revised: 2 December 2024
Accepted: 17 April 2025

Abstract

Since tourism industry is one of the most important contributors to Thailands GDP, this study aims to gain insight into the structure of tourism data in Thailand for suitable administration. In machine learning framework, there were 13 predictors from Thailand's tourism dataset and the average revenue was assigned as response variable which was transformed into multi-level categorical. The ordinal logistic regression (OLR) was implemented for clustering of 77 Thailands provinces in 8 different scenarios designed by varying the number of clusters to be 2, 3, 4 and 5 together with outlier adjustment technique. Evaluating models' performance, the numerical results show that the most suitable number of provincial clusters is three and the number of primary, secondary, and tertiary provinces are 18, 29 and 30, respectively. The significant factors are number of foreign occupancy, Thai visitors, and their average expense. Based on available infrastructures and tourism resources in each cluster, it is challenging for Thailand to recover the foreign tourist arrivals and promote domestic tourism after the COVID-19 pandemic. Through the strategic management in resource allocation and enhancement of marketing efficacy via provincial clustering, this study comprehensively addresses a multifaceted framework of tourism strategies, integrating guidelines, policies, and best practices that span national initiatives, leverage digital marketing, and reinforce soft power. The framework actively involves the participation of small and medium-sized enterprises (SMEs) and aligns with the overarching objectives of sustainable development goals (SDGs), thus fostering a holistic approach to sustainable tourism growth in Thailand.

Keywords: Classification, clustering, ordinal logistic regression, sustainable tourism, Thailand.

1. Introduction

Thailand stands as a globally renowned tourist hubs, celebrated for its picturesque beaches, natural beauty, warm hospitality, and historical landmarks. Predictably, the tourism sector emerges as a pivotal economic force, wielding substantial influence on the nations GDP [Fakfare et al. (2022); Rasool et al. (2021); Sharafuddin (2015); The World Bank]. Strategic planning, policy development,

and effective management are crucial for Thailands tourism sector to navigate challenges, capitalize on opportunities, and sustain growth [Hashim (2023); Soh et al. (2021); Alnajim and Fakieh (2023)].

Travel and tourism (T&T) exports totaled 9.2 trillion USD in 2019, 4.7 trillion USD in 2020, 1.9 trillion USD in 2021, 7.7 trillion USD in 2022, and 9.5 trillion USD in 2023, making it one of the most substantial contributors to global GDP [Office of the National Economic and Social Development Council; UN Tourism; United Nations; World Tourism Organization and Global Tourism Economy Research Centre; Wu et al. (2023)]. Many countries are trying to restore international tourism to pre-epidemic levels amid the COVID-19 pandemic. Thailand is a top tourist destination due to its beautiful coastlines, delicious food, awe-inspiring temples, stunning scenery, rich history, and unique cultural and archaeological treasures. Tourism, which accounts for 17.79% of Thailands GDP in 2019, boosts economic growth and employment [Ministry of Tourism and Sport; Piboonrungroj et al. (2023)]. This proportion shows how important T&T are to the nation's economy. Thailand has increased its tourism numbers and profits, ranking seventh in 2019 [World Tourism Organization and Global Tourism Economy Research Centre]. Despite rising tourist numbers and spending, the COVID-19 pandemic, global economic developments, changing consumer behaviours, and digital technology are affecting both global and Thai tourism [Alnajim and Fakieh (2023); Borges-Tiago et al. (2021); Liu (2023); Mueller and Sobreira (2024)]. To foster enduring visitor experiences, the Tourism Authority of Thailand (TAT) emphasizes the kingdom's 5F soft power foundations: Food, Film, Festival, Fight, and Fashion [Thailand Center]. These underpin efforts to enrich tourism offerings and create meaningful engagements. Thailand's appeal is further enhanced by wellness retreats, emerging urban destinations, and world-class events. TAT has established a revenue target of three trillion Baht for the fiscal year 2024. This projection is comprised of an estimated 1.92 trillion Baht from international tourism and 1.08 trillion Baht from domestic tourism [TAT Newsroom]. Fostering Thailand's tourism success, tourism policies and strategies are the blueprints guiding the development and management of a destination's tourism sectors.

This paper introduces the novel clustering method to divide 77 provinces in Thailand towards the suitable tourism management. Using available datasets such as number of Thai and foreign visitors, their length of stay, average expenses of tourists, number of occupancy and room booked as the predictors, their tourism revenues were analyzed and each province was allocated into its suitable cluster by ordinal logistic regression model. The clustering method is able to enhance the strategic planning for both local and national tourism promotion, inform digital marketing efforts, and support SMEs in the tourism sector. The resulting classification provides a data-driven framework for tailoring tourism policies and initiatives to the unique characteristics and potentials of each provincial cluster, enabling more effective and targeted tourism development across Thailand. This approach not only enhances Thailands tourism appeal but also promotes sustainable practices by considering carbon emission trading and environmental conservation as core components of tourism strategy. By creatively managing and promoting tourism resources within their specific contexts, Thailand can attract international visitors while minimizing environmental impact.

2. Related Works

Tourism promotion success requires clustering tourist sites and creating strategic plans for each cluster. Thailand's provinces can be clustered by tourism resources, attractiveness, and competitiveness for customized strategic planning, resource investment, and policy making. In the context of tourism development, a cluster-based approach allows for strategic prioritization of resources. For instance, infrastructure enhancement in one cluster may take precedence over destination marketing efforts in another. The implementation of cluster-specific tourism policies and programs can significantly augment both efficiency and overall impact. This approach leverages the clustering problem, wherein tourism data is grouped based on similarity metrics.

The clustering methodology offers several advantages, including the identification of latent patterns, data summarization, the creation of representative prototypes, anomaly detection, and the facilitation of further analytical processes [Mahfuz and Yusoff (2019); Melnykov (2013); Putman and

Carbone (2014)]. Comparing the concept in clustering and classification, clustering considers all traits, which may not be relevant, and the results are often ambiguous whereas classification methods are chosen for grouping because they can identify significant features during training with clear decision criteria and relevance scores [Akarajarasroj et al. (2023); O'Connell (2006); Yilmaz and Demirhan (2023)]. Literally, the parametric OLR model is one of the most popular classification methods in education, economics, management, healthcare, tourism, and other fields various applications such as education, economic, management, healthcare, tourism, etc. [Al Abri et al. (2023); Alrumaidhi and Rakha (2022); O'Connell (2006); Duan (2020); Liu and Koirala (2012); Shui et al. (2022); Wang et al. (2022)]. According to Abri et al. (2022), local and international tourists differ in demographic, psychological, and perceived political risk factors of attitudinal and behavioral allegiance to Oman as a tourism destination [Al Abri et al. (2023)]. Duan (2020) used OLR to evaluate a survey of 2025 space tourism expectations by gender, age, education, and income to assist data mining for product design [Duan (2020)]. Education, wealth, and age greatly affected projected involvement, but gender did not. Shui et al. (2022) used survey data and OLR to examine how household livelihood capital affects tourism engagement in 60 Tibetan households in Jiaju Village, China [Shui et al. (2022)]. Physical, human, and social capital greatly affected tourism participation, while financial capital did not. Despite danger assumptions, overseas tourists have established attitudinal and behavioral devotion to Oman.

From the aforementioned evidences, OLR has been conformed as a well-known and effective clustering technique. Thus, it was chosen as the major research tool for this study. Regarding the available tourism data in Thailand and analysis tools, the research questions for this study are:

- *What extent can machine learning algorithms, particularly OLR, effectively categorize 77 provinces of Thailand into distinct groups based on tourism-related data?*
- *How can provincial clusters, derived from average tourism revenue data, be used to develop an optimal budget allocation strategy for Thailand's tourism sector?*
- *What policy framework can be developed to allocate tourism resources, manage destinations, and implement marketing strategies at national and local levels to promote sustainable tourism development in Thailand?*

To address these research questions using tourism data in Thailand, the data should first be preprocessed by cleaning and defining relevant features, including visitor numbers, revenue, and spending patterns. Then, the OLR model should be applied to categorize provinces based on tourism characteristics, resulting in distinct clusters. Subsequently, these clusters can be analyzed to identify optimal budget allocation strategies that align resources with the tourism potential of each cluster. Finally, the clusters should be interpreted to develop targeted policy frameworks, enabling resource-efficient marketing, destination management, and sustainable development strategies at both local and national levels.

3. Data Description

The tourism data used in this study was retrieved from Open Government Data of Thailand (2020) provided by the Department of Tourism, Ministry of Tourism and Sports of Thailand in 2018 (Digital Government Development Agency). The dataset consists of tourism information from all 77 provinces with 14 variables including one response and 13 predictors. To examine the factors influencing tourism revenue across Thailand's 77 provinces, the analysis incorporates a comprehensive set of predictors, including visitor demographics, length of stay, and expenditure patterns. Specifically, the predictors encompass the number of Thai and foreign visitors, their average length of stay, and the average daily expenses for various categories of visitors (tourists and excursionists) from both domestic and international origins. Additionally, the model considers accommodation-related factors such as Thai and foreign occupancy rates and the total number of available rooms. The description

and descriptive statistics including min, max, mean, and standard deviation of each predictor were given in Table 1. By exploring these diverse variables, the research aims to identify key drivers of tourism revenue and provide insights for strategic decision-making in Thailand's tourism sector. It is seen that the distribution of average tourism revenues from 77 provinces was highly skewed to the right. Coping with provincial clustering, the average revenue was converted to 8 response categorical variables generated by a combination of numbers of levels and outlier adjustment schemes (OAS) as presented in Table 2. Running OLR models, the values of 13 predictors were not transformed, but the response variable (average tourism revenues) was transformed from scale to categorical variables which is more suitable for provincial clustering.

Table 1 Descriptions and descriptive statistics of predictors

Predictor	Description	Min	Max	Mean	SD
TH.no.visitor	Number of Thai visitors (people)	146,222.00	41,682,963.00	2,958,105.62	4,979,063.72
FR.no.visitor	Number of foreign visitors (people)	1,849.00	23,851,318.00	977,208.82	3,182,590.58
TH.los	Average length of stay for Thais (days)	1.59	4.12	2.24	0.38
FR.los	Average length of stay for foreigners (days)	1.41	6.61	2.70	0.94
TH.visitor.exp	Average expenses of Thai visitors (Baht/person/day)	670.81	4,950.54	1,711.60	886.31
FR.visitor.exp	Average expenses of foreign visitors (Baht/person/day)	849.39	8,388.38	2,283.24	1,401.93
TH.tourist.exp	Average expenses of Thai tourists (Baht/person/day)	784.29	5,023.82	1,901.67	923.58
FR.tourist.exp	Average expenses of foreign tourists (Baht/person/day)	1,108.10	8,432.18	2,561.25	1,415.61
TH.exc.exp	Average expenses of Thai excursionists (Baht/person/day)	421.77	2,899.65	1,124.77	532.37
FR.exc.exp	Average expenses of foreign excursionists (Baht/person/day)	472.39	3,609.91	1,376.17	672.65
TH.no.occupied	Number of Thai occupancy (people)	95,705.00	14,089,493.00	1,395,895.04	1,914,479.62
FR.no.occupied	Number of foreign occupancy (people)	385.00	21,721,074.00	769,119.77	2,893,944.26
No.room	Total numbers of rooms (rooms)	379.00	152,616.00	9,693.51	21,202.59

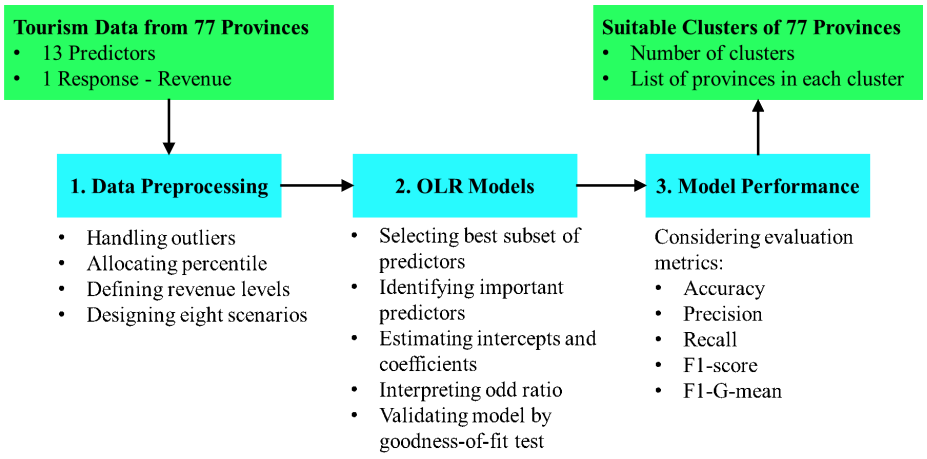

Figure 1 Flow diagram demonstrating the key steps for the proposed clustering framework

4. Machine Learning Framework

Classifying 77 provinces in Thailand into certain clusters, the average revenue levels are generated as the number of clusters. In machine learning, classification, a supervised approach, proves more practical than clustering. It uses predefined labels to train models, creating robust rules for future data categorization. This method excels with limited datasets. Classification's predetermined structure ensures more reliable grouping, outperforming unsupervised clustering when category knowledge is available. Instead of traditional clustering techniques, OLR is selected as the main tool for research objectives including the suitable number of clusters and significant factors. As there were 8 scenarios for provincial clustering, the comparison of their suitability was made through their performances. Figure 1 demonstrates the flowchart of our proposed clustering technique in three main steps. In this section, data pre-processing, overview of OLR model and evaluation of model performance are illustrated.

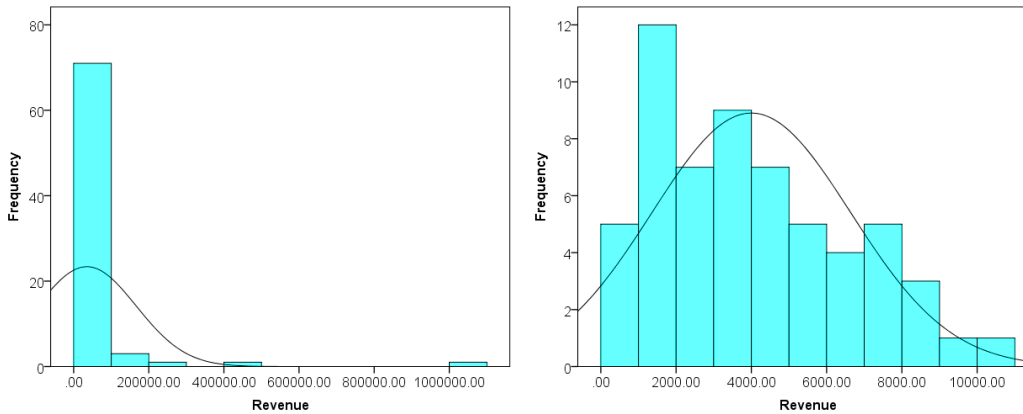

Figure 2 Histograms of the average revenue without (left) and with OAS (right)

4.1. Data pre-processing

For provincial clustering, the OLR model was utilized to examine Thailand's tourism data in 2018. Designing the trials, the average revenue was coded as categorical response variables with 2, 3, 4, and 5 levels in two different scenarios: one with and one without outlier correction. Any data point outside the range Q1−1.5*IQR and Q3+1.5*IQR was considered as an anomaly. Visualizing the distribution of average revenues, Figure 2 shows the two scenarios' average revenue histograms. The 18 provinces were outliers from 77 because their average revenue surpassed the top bound. The distribution of average revenues from all 77 provinces was substantially skewed. After excluding the 18 provinces, the distribution of the remaining 59 provinces was slightly skewed on the right, but it remained normal (Kolmogorov-Smirnov = 0.104 and p-value = 0.181). Then, there were two different approaches in generating multi-level category of average tourism revenues from all 77 provinces. Without OAS, all 77 provinces were directly grouped by level and percentile. With OAS, the 18 provinces were placed into one class, while the others were sorted by level and percentile. For instance, the 33.33rd and 66.67th percentiles were applied in the three-class case, i.e., the province whose average revenues below the 33.33rd percentile, 33.33rd 66.67th percentiles and above 66.67th percentile were denoted as levels 1, 2 and 3, respectively. In OAS scenario, the 18 provinces were classified as 3, then the 50th percentile was computed from the remaining provinces. The description and descriptive statistics for the response variables are summarized in Table 2. Observing the effect of creating levels for classification, there were 4 scenarios without OAS (Y2, Y3, Y4, Y5) and 4 scenarios with OAS (Y2*, Y3*, Y4*, Y5*). Simultaneously, IQR was used to find an outlier for all predictors, like the response variable. It was found that there were outliers in all predictors. Finding meaningful predictors by OLR models, the transformation of those outliers were required. As

logarithms are often employed in regression modeling to eliminate outliers, all predictors in Table 1 were log transformed before developing predictive models. Then, best subset selection was utilized to identify some meaningful predictors. The final model was chosen since all predictors had to be statistically significant.

Table 2 Description of multi-level categorical variables and their frequencies

Variable	Revenue Level	Range (Million THB)	Frequency (%)	Variable	Revenue Level	Range (Million THB)	Frequency (%)
Y2	1	≤4,610.14	39 (50.65)	Y2*	1	≤10,841.24	59 (76.62)
	2	>4,610.14	38 (49.35)		2	>10,841.24	18 (23.38)
Y3	1	≤3,100.69	25 (32.47)	Y3*	1	≤3,470.96	30 (38.96)
	2	3,100.70 7,488.73	27 (35.06)		2	3,470.97 10,841.24	29 (37.66)
	3	>7,488.73	25 (32.47)		3	>10,841.24	18 (23.38)
Y4	1	≤2,176.32	19 (24.68)	Y4*	1	≤2,216.25	19 (24.68)
	2	2,176.33 4,610.14	20 (25.96)		2	2,216.26 4,994.14	21 (27.26)
	3	4,610.15 10,294.05	19 (24.68)		3	4,994.15 10,841.24	19 (24.68)
	4	>10,294.05	19 (24.68)		4	>10,841.24	18 (23.38)
Y5	1	≤1,618.98	15 (19.48)	Y5*	1	≤1,599.91	15 (19.48)
	2	1,618.99 3,549.56	16 (20.78)		2	1,599.92 3,470.96	15 (19.48)
	3	3,539.57 6,516.35	15 (19.48)		3	3,470.97 5,687.87	15 (19.48)
	4	6,516.36 19,127.96	16 (20.78)		4	5,687.88 10,841.24	14 (18.18)
	5	>19,127.96	15 (19.48)		5	>10,841.24	18 (23.38)

4.2. OLR model

The concept of OLR model was employed to determine the relationship between the set of predictors $(X_1, X_2, \ldots, X_p)$ and the ordinal response variables (Y). Ordinal variables are categorical variables with meaningful categories and perhaps uneven gaps. OLR is an extension of binary logistic regression that works for response variables with more than two ordered categories (Abuzaid and Nae'l (2024); Agresti (2012); Akarajarasroj et al. (2023); Liang et al. (2020); Liu (2009); Liu and Koirala (2012)). The binary logistic regression (BLR) assigns Y as a binary random variable with values of 0 or 1 to estimate the probability of belonging to a response variable category based on a set of predictors. The independent events give Y as a binomial distribution. Let π be the probability of the event of interest or $P(Y = 1)$, and it depends on the predictors. The BLR model can be expressed as follows:

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{1}$$

Like a linear regression model, β_0 is an intercept term and β_i is the partial regression coefficient of the corresponding ith predictor $(i = 1, 2, \ldots, p)$. The ratio $\pi/(1-\pi)$, is called odds, and the logarithm of odds is known as the logit. The equation (1) can be directly converted into terms of π as

$$\pi = P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \tag{2}$$

In OLR, Y is an ordinal response variable with J categories. In this study, Y is the tourism revenue level, and each category or level represents each cluster. $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category where $j = 1, 2, \ldots, J$ (Liang et al. (2020); O'Connell (2006); Wang et al. (2022)). The odds of being less than or equal a particular category can be written as

$$Odds = \frac{P(Y \leq J)}{P(Y > J)}. \tag{3}$$

Since $P(Y > J) = 0$, dividing by zero is undefined, and $P(Y \leq J) = 1$. In this case, we can obtain only $J - 1$ valid odds. From equation (3), the resulting logit function is defined by

$$\ln(Odds) = \text{logit}\left(P(Y \leq J)\right) = \beta_{0j} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{4}$$

where β_{0j}s are $J-1$ constants. Each predictor has the same effect across response variable categories. This implies that any predictor affects model chances of every category equally (Agresti (2012); Liang et al. (2020); O'Connell (2006)). Because of the assumption of parallel lines, constant terms vary by category, but slopes are uniform (Liu and Koirala (2012); O'Connell (2006)). Generally, the Odds Ratio (OR) is popularly employed to make comparison between the odds of groups (events) A and B. Determining OR in OLR model, the value of OR is evaluated by

$$OR = \exp(\beta_i) \tag{5}$$

where β_i denotes the coefficient of X_i. The values $OR > 1$ indicates the higher chance to obtain the focused event whereas $OR < 1$ indicates the less one. If $OR = 1$, both events are able to occur equally. Considering the study's response variables, the hierarchical tourism revenue levels from 8 scenarios in Table 2 were analyzed. The best subset selection method was used to find important predictors, statistically significant to choose the right model. To examine the assumption of parallel lines, two goodness of fit tests including likelihood ratio test (LRT) and Pearson Chi-square (χ^2) were employed. The LRT was applied to determine if model complexity enhances accuracy. The null hypothesis says the simpler model is best, while the alternative hypothesis recommends the complex model. In our case, the null hypothesis shows the model with intercepts only, while the alternative hypothesis shows predictors included in the model. If the null hypothesis is rejected, it indicates that the more complex model is much better than the simpler one [Liu (2009); MacKenzie et al. (2017)]. The LRT is defined as the difference between the two log-likelihood function values: $\ln(L_0)$ and $\ln(L_1)$ under the null and alternative hypotheses, respectively. Thus, the LRT statistics is expressed as

$$LRT = -2\left[\ln(L_0) - \ln(L_1)\right]. \tag{6}$$

Its asymptotically χ^2 distribution has degrees of freedom equal to the differences between the number of parameters from two models. Another goodness-of-fit statistics is χ^2, which measures the difference between observed and expected frequencies. In OLR, a larger χ^2 value indicates poor model fitting. If it is not significant, the model fits the data well. It is noted that the test is sensitive to larger sample sizes. However, it is suitable for this dataset as the sample size of 77 is not considerably large. The χ^2 is typically written as

$$\chi^2 = \sum_{i=1}^{J}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{7}$$

where O_{ij} is the observed frequency in category i of the jth group, and E_{ij} is the expected frequency in category i of the jth group. Typically, expected frequencies are determined using the OLR model. Probabilities for each category are estimated by the model and used to calculate predicted frequencies. The degrees of freedom are the difference between observed frequencies and numbers of estimated parameters.

4.3. Evaluation criteria

The confusion matrix for classification tasks was employed as a common evaluation tool [Akarajarasroj et al. (2023); Na Bangchang et al. (2023); Simmachan et al. (2023); Yilmaz and Demirhan (2023)]. It assigns n labels into J classes or categories. Its columns and rows refer to the actual classes and predicted classes, respectively. Let n_{ij} represents the number of labels in class i and predicted to be in class j where $i, j = 1, 2, \ldots, J$. The row and column totals are determined by frequencies in the dataset, denoted by $n_{i\cdot}$ and $n_{\cdot i}$, respectively. Since evaluation metrics are computed using the probability notion, the probability associated with the specific cell is denoted as $p_{ij} = n_{ij}/n$. Determining the predictive performance of the classification models, n_{ii} on diagonal cells indicates the number of correctly predicted labels for any class. Practically, there are five basic evaluation metrics including accuracy, precision, recall, F1-scores with arithmetic and geometric means [Simmachan et al. (2023); Yilmaz and Demirhan (2023)]. Each evaluation metric is computed as follows:

- **Accuracy**: A provincial clustering algorithm's accuracy measures its ability to correctly group similar provinces together. It reflects how well the algorithm identifies meaningful patterns and relationships among provinces based on chosen features, minimizing errors in cluster assignments. It is a popular metric defined by the ratio of accurately predicted labels to total labels, i.e.,

$$\text{Accuracy} = \frac{\sum_{i=1}^{j} n_{ii}}{n} \tag{8}$$

  In classification problems, accuracy can be misleading, especially with imbalanced datasets. It simply measures the proportion of correctly predicted labels to total labels but does not account for the distribution of classes. When accuracy is misleading, especially in cases of class imbalance, other alternatives of evaluation metrics should be considered to measure the model performance.

- **Macro-average precision (M.Precision)**: Precision in provincial clustering algorithms measures the accuracy of grouping similar provinces together. It quantifies how often provinces placed in the same cluster truly belong together, based on predefined criteria. It is calculated as the arithmetic mean of precisions over J classes. Class precision is calculated in the confusion matrix column direction. The macro-average precision is shortly called precision in the rest of the study.

$$\text{Precision}_i = \frac{n_{ii}}{n_{\cdot i}} \quad \text{and} \quad \text{M.Precision} = \frac{\sum_{i=1}^{J} \text{Precision}_i}{J} \tag{9}$$

- **Macro-average recall (M.Recall)**: Recall in provincial clustering measures the algorithm's ability to correctly identify and group data points belonging to the same province. It is computed as the arithmetic mean of recalls for all classes. The recall for each class is calculated in the row direction of the confusion matrix.

$$\text{Recall}_i = \frac{n_{ii}}{n_{i\cdot}} \quad \text{and} \quad \text{M.Recall} = \frac{\sum_{i=1}^{J} \text{Recall}_i}{J} \tag{10}$$

- **F1-score**: F1-score is determined by assessing the overall performance of the model for the positive class by considering both precision and recall simultaneously. F1_i is computed as the harmonic mean of the precision and recall for each class. Then, the classical F1-score is determined by the arithmetic mean of F1_i over J classes, i.e.,

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad \text{and} \quad \text{F1-score} = \frac{\sum_{i=1}^{J} \text{F1}_i}{J} \tag{11}$$

- **F1-G-mean**: The new evaluation metric is proposed as the geometric mean of the F1-score for all classes. Geometric mean accounts for compounding effects, unlike arithmetic mean which is sensitive to outliers. Since the geometric mean is advantageous for ratios and rates, it is more suitable to apply F1-G-mean as the new indicator for this classification problem.

$$\text{F1-G-mean} = \sqrt[J]{\prod_{i=1}^{J} \text{F1}_i}. \tag{12}$$

The classification of Thailand's 77 provinces extends beyond analyzing tourism-related data using OLR models. It also requires qualitative validation and exploration. This dual approach ensures a comprehensive understanding of provincial tourism characteristics, combining statistical insights with in-depth qualitative assessments.

5. Numerical Results

Running OLR models to classify 77 provinces into a certain number of clusters, 8 scenarios based on the tourism revenues in Table 2 were carried out using IBM SPSS Statistics 26 software (Corp, 2019). The results include important predictors, model performance, best OLR model and list of provinces in each cluster. That is, key predictors influencing cluster formation were identified and evaluated. Model performance is assessed to determine the optimal OLR model for accurate and meaningful cluster delineation towards implementation.

5.1. Important predictors

The primary investigative approach employed in this study is OLR model as illustrated in Section 4.2. To identify salient predictors, the best subset selection method was utilized. The selection criteria for the optimal model stipulated that all predictors must demonstrate statistical significance. The LTR was employed to assess whether increased model complexity resulted in improved predictive accuracy. Considering all 13 predictors for each scenario as mentioned in Table 1, the best subset selection approach was firstly utilized to investigate the important predictors. Findings indicate that the alternative hypothesis supports the adoption of the more complex model, incorporating the identified predictors. Each scenario has different important factors, i.e.,

- Y2: {TH.no.occupied, FR.no.occupied},
- Y2*: {TH.no.occupied, TH.visitor.exp, FR.los},
- Y3: {TH.no.visitor, TH.visitor.exp, TH.los },
- Y3*: {TH.no.visitor, TH.visitor.exp, FR.no.occupied},
- Y4: {TH.no.visitor, TH.visitor.exp, FR.no.occupied, No.room},
- Y4*: {TH.no.visitor, TH.visitor.exp, FR.no.occupied, No.room},
- Y5: {TH.no.visitor, TH.visitor.exp, FR.no.occupied, No.room} and
- Y5*: {TH.no.visitor, TH.visitor.exp, FR.no.occupied, No.room}.

The removed tourism predictors were highly correlated with other variables in the model, namely multicollinearity. Considering the important predictors based on OAS, both scenarios Y2 and Y2* simultaneously recruited TH.no.occupied, but the rest predictors were different. Y3 and Y3* recruited TH.no.visitor and TH.visitor.exp while TH.los was in Y3 and FR.no.occupied was in Y3* only. Predictor sets for scenarios Y4 and Y4* were identical, as were those for Y5 and Y5*. Analysis of tourism data suggests that the predictor variable "TH.no.visitor", representing the number of Thai visitors, consistently emerges as a significant factor in provincial clustering algorithms. Since there were only $2-4$ predictors in each scenario, it is more convenient for model formulation, less computation time, reflecting actual patterns in the data, avoiding overfitting risk, and enhancing model performance. However, it remains to consider the performance of each scenario through different evaluation matrices.

5.2. Model performance

To visualize the performance of OLR models in a classification task, confusion matrices were presented for all scenarios as shown in Figures 3 and 4. It is seen that there are more correct predictions on diagonal cells in OLR models with OAS than OLR models without OAS. The more incorrect predictions on off-diagonal cells appeared in OLR models without OAS. Obviously, the response variables with OAS (Y2*, Y3*, Y4*, Y5*) consistently outperformed their counterparts in accurately predicting provincial clustering. Notably, diagonal elements increase, indicating more correct predictions, while off-diagonal errors decrease, showing that OAS enhances model classification balance across classes. Considering the number of clusters based on OLR models with OAS, 77 provinces were classified into 2, 3, 4 and 5 clusters as 59:18, 30:29:18, 19:21:19:18 and 15:15:15:14:18, respectively. For Y2*, Y3*, Y4* and Y5* scenarios, it is noted that the number of provinces with the highest revenue level was always 18 as they are the upper outlier. For the clusters with lower revenue levels, the rest 59 provinces were classified according to their similarities in important predictors. However, it remains to investigate the model performance for each scenario and determine how many clusters are the most suitable.

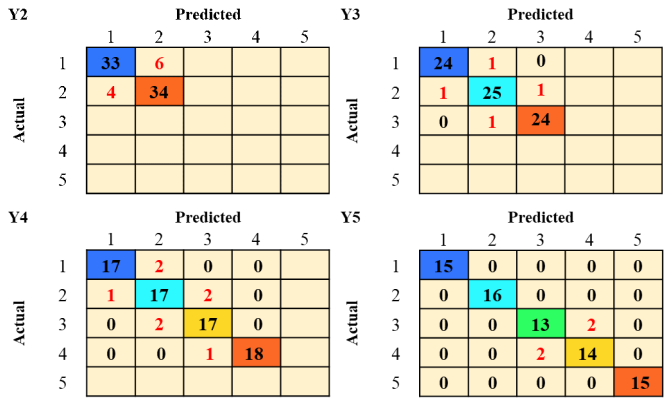

Figure 3 Confusion matrices for 4 scenarios without OAS.

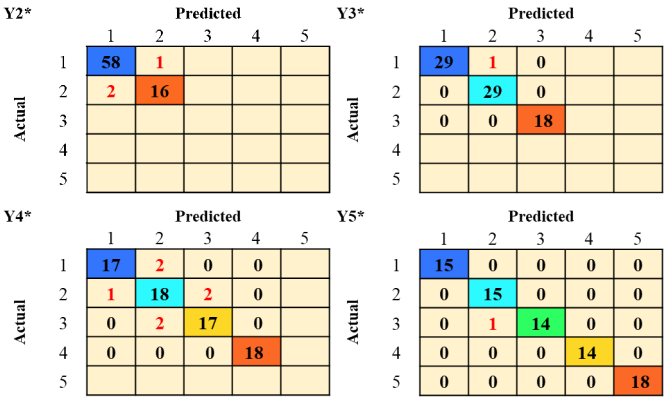

Figure 4 Confusion matrices for 4 scenarios with OAS.

Confusion matrices are essential tools in evaluating classification models as they provide a detailed summary of a model's predictions across actual and predicted classes. Each entry in the matrix represents the frequency of predictions for specific actual versus predicted class combinations, making it easy to observe where the model excels or struggles. Based on the evaluation metrics in Equations (8) − (12), Table 3 presents the model performances including accuracy, precision, recall, F1 scores with arithmetic and geometric means from the OLR models in 8 scenarios.

Table 3 Overall performance of 8 different scenarios

Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	F1-G-mean (%)
Y2	87.01	87.09	87.04	87.01	87.01
Y2*	96.10	96.21	92.78	94.45	94.41
Y3	94.81	94.86	94.86	94.86	94.85
Y3*	**98.70**	**98.89**	**98.89**	**98.87**	**98.87**
Y4	89.61	90.10	89.67	89.82	89.66
Y4*	90.91	91.43	91.17	91.27	91.09
Y5	94.81	94.83	94.83	94.83	94.62
Y5*	**98.70**	98.75	98.67	98.67	98.65

Obviously, the OLR model with OAS and three clusters (Y3*) is outstanding among 8 scenarios, i.e., it showed the highest accuracy, precision, recall, F1-score and F1-G-mean indicating 98.70%, 98.89%, 98.89%, 98.87% and 98.87%, respectively. Following Y3*, the OLR model with OAS and five clusters (Y5*) stood as the second-best option for provincial clustering. Its accuracy, precision, recall, F1-score and F1-G-mean are 98.70%, 98.75%, 98.67%, 98.67% and 98.65%, respectively. The Y3* model considered Thai visitors, average Thai visitor expenses, and foreign occupancy as relevant predictors, while Y5* model added total rooms. Even if the Y5* model performed well, classifying 77 provinces into three clusters is more convenient to managerial implementation. The OLR models with two and four clusters had poor clustering accuracy, precision, recall, F1-score, and F1-G-mean. OLR models with OAS outperformed those without when clustering 77 provinces into 2, 3, 4, and 5 clusters. The strong diagnostic performance across confusion matrix-derived measures indicates good generalization and valuable revelatory capacity in the OLR model built for Thailand's provinces.

5.3. Best OLR model

The presentation of the best OLR model for the Y3* scenario is essential to understand the key factors influencing tourism revenue levels. By analyzing Thai visitors' spending, visitor quantities, and foreign occupancy, we gain insights into which variables significantly impact the likelihood of higher tourism revenue levels. This result demonstrates the model's suitability and robustness, evidenced by significant predictor coefficients and high model fit metrics. Highlighting this model allows for better-informed strategic planning in tourism by identifying critical drivers of revenue enhancement. Among the 8 scenarios, the Y3* scenario was outstanding so that it was further analyzed. Based on the results in Table 4, the estimated models for Y3* can be written as

$$\text{logit}\left(\hat{p}(Y \leq 1)\right) = 269.30 + 13.80x_1 + 10.67x_2 + 1.84x_3, \quad (13)$$

$$\text{logit}\left(\hat{p}(Y \leq 2)\right) = 284.41 + 13.80x_1 + 10.67x_2 + 1.84x_3 \quad (14)$$

where $\hat{p}$ is the estimated probability. The overall accuracy of the Y3* model is satisfactory. As mentioned in Equation (4) for $J = 1$ and $J = 2$, each predictor has the same effect across response variable categories while the constants (β_{0j}s) are different. Thus, the assumption of parallel lines or uniform slopes are satisfied. The LRT indicates that the model with predictors is more suitable than the model with only intercept terms. According to the low χ^2 value , the final model fits

well. For the model explanation, the Cox and Snell R-square statistic of 0.883 indicating the better model explanation (the log likelihood for the model compared to the log likelihood for a baseline model). With categorical outcome, the predictors and dependent variable were highly associated. Thus, it can be said that the three predictors hold statistical significance and the final model is suitably implementable. Interpreting the results in Table 4, the tourism revenue level is affected by Thai visitors' average spending, quantity of Thai visitors, and foreign occupancy in logarithmic order. The computed coefficient for TH.visitor.exp is 13.80, meaning that held other variables constant, the probability of shifting to a higher category are multiplied by $\exp(13.80) = 9.8 \times 10^5$ for a one-unit increase in Thai visitors' logarithmic average expenses. The very high value of odds ratio indicates that Thai visitors' normal spending is significant. Furthermore, the high odds ratios for TH.no.visitor and FR.no.occupied suggest that Thai visitors and foreign occupancy can increase the likelihood of moving up. For ordinal category threshold comparisons, the two determined intercept terms or thresholds give baseline log-odds. The first term, estimated at 269.30, represents the log-odds of the response variable being Secondary or higher (Primary) against Tertiary when all predictors are zero. The second term, 284.41, indicates the baseline log-odds of the response variable being Primary rather than Secondary or Tertiary without a predictor. Therefore, the greater intercepts result in higher categories.

Table 4 OLR model for Y3* scenario

Terms	Estimate	SE	Wald Test	P-Value	OR	95% CI for OR	
						Lower	Upper
Intercepts							
Tertiary|Secondary	269.30	93.74	8.25	0.001			
Secondary|Primary	284.41	98.48	8.34	0.001			
Coefficients							
TH.visitor.exp (x_1)	13.80	5.14	7.22	0.01	$9.8\text{x}10^5$	42.10	$2.3\text{x}10^{10}$
TH.no.visitor (x_2)	10.67	3.92	7.43	0.01	$4.3\text{x}10^4$	20.09	$9.3\text{x}10^7$
FR.no.occupied (x_3)	1.84	0.85	4.74	0.03	6.30	1.20	33.12

LRT: 165.52 and P-value: $<$0.001
Pearson Chi-square: 5.97 and P-value: 0.998
Cox and Snell pseudo-R-square: 0.883

5.4. Provincial clustering

Based on the model performances among 8 scenarios in Table 3, the Y3* scenario had the best performance. Subsequently, the 77 provinces of Thailand should be classified into 3 clusters and the number of tertiary, secondary and primary provinces are 30, 29 and 18, respectively. Considering the relative location of 77 provinces and their clusters based on the data analysis, Figure 5 exhibits the color map classifications for 2, 3, 4 and 5 provincial clusters under OAS, respectively. By analyzing their strategies and challenges, policymakers and tourism stakeholders can identify best practices and tailor them to the specific needs of each province. Based on Y3* model (the second one in Figure 5), the list of provinces in each clusters are as follows:

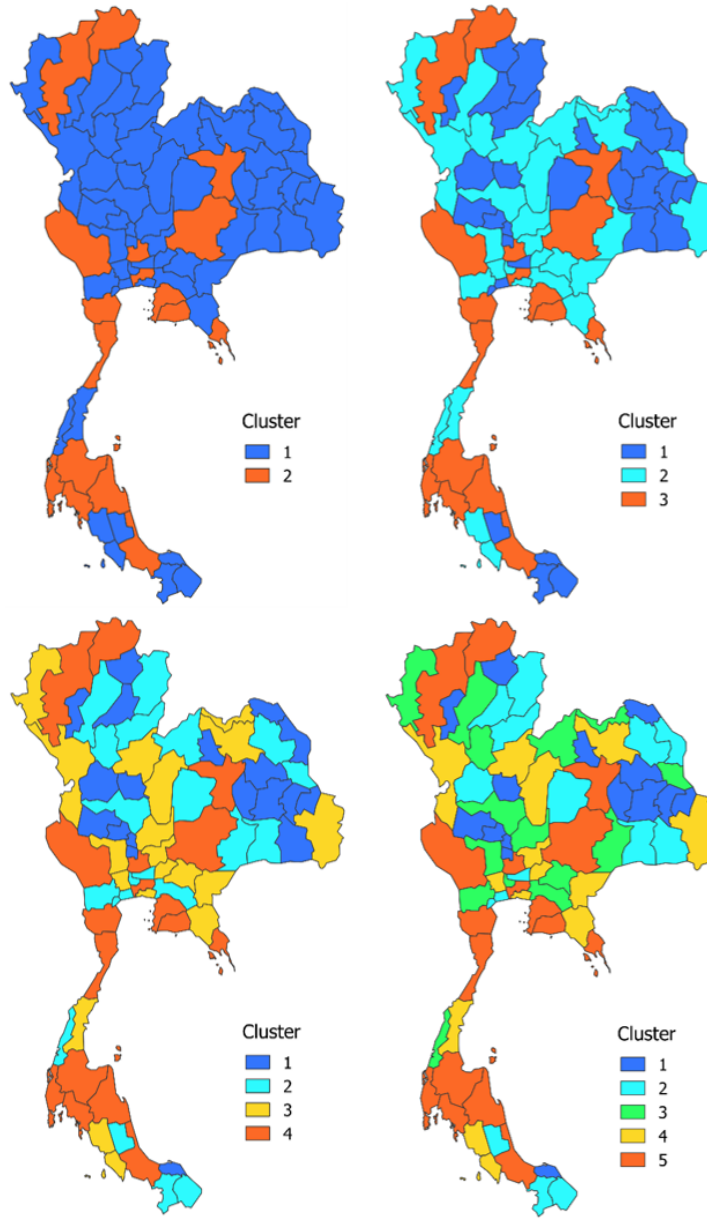

Figure 5 Color map classification for 2, 3, 4 and 5 clusters based on OLR models with OAS technique.

- *Tertiary provinces*: The tertiary provinces with low tourism revenue presents a fascinating case study in regional economic development. This provincial cluster includes 1) Amnat Charoen, 2) Ang Thong, 3) Bueng Kan, 4) Chai Nat, 5) Chaiyaphum, 6) Kalasin, 7) Kamphaeng Phet, 8) Lamphun, 9) Maha Sarakham, 10) Nakhon Phanom, 11) Nan, 12) Narathiwat, 13) Nongbua Lumphoo, 14) Pathum Thani, 15) Pattani, 16) Phatthalung, 17) Phayao, 18) Phichit, 19) Phrae, 20) Roi Et, 21) Sakon Nakhon, 22) Samut Sakhon, 23) Samut Songkhram, 24) Sisaket, 25) Sing Buri, 26) Surin, 27) Uthai Thani, 28) Uttaradit, 29) Yala, and 30) Yasothon. To increase tourism revenue in these provinces, several strategies can be considered to attract significant tourist numbers. As these provinces have limited tourism infrastructure, rural and local experiences, appealing to eco-tourists and adventure seekers should be highlighted.

- *Secondary provinces*: The secondary provinces have the moderate level of tourism revenue based on their commonalities and potential. This cluster includes 1) Buriram, 2) Cha Choeng Sao, 3) Chanthaburi, 4) Chumphon, 5) Lampang, 6) Loei, 7) Lop Buri, 8) Mae Hong Son, 9) Mukdahan, 10) Nakhon Nayok, 11) Nakhon Pathom, 12) Nakhon Sawan, 13) Nong Khai, 14) Nonthaburi, 15) Phetchabun, 16) Phitsanulok, 17) Prachin Buri, 18) Ranong, 19) Ratchaburi, 20) Sa Kaew, 21) Samut Prakan, 22) Saraburi, 23) Satun, 24) Sukhothai, 25) Suphan Buri, 26) Tak, 27) Trang, 28) Ubon Ratchathani and 29) Udon Thani. These provinces have moderate tourism appeal, often catering to domestic tourists and niche markets with natural landscapes and cultural festivals. As the goal is to identify potential areas for development to boost tourism in these provinces, a detailed analysis of each province would require extensive data.

- *Primary provinces*: The primary provinces are 1) Bangkok, 2) Chiang Mai, 3) Chiang Rai, 4) Chon Buri, 5) Kanchanaburi, 6) Khon Kaen, 7) Krabi, 8) Nakhon Ratchasima, 9) Nakhon Si Thammarat, 10) Phang Nga, 11) Phetchaburi, 12) Phra Nakhon Si Ayutthaya, 13) Phuket, 14) Prachuap Khiri Khan, 15) Rayong, 16) Song Khla, 17) Surat Thani and 18) Trat. These provinces have significant tourism resources, often featuring popular attractions like cultural sites, beaches, and historical landmarks. Sustainable tourism, infrastructure development, tourism diversification, human capital development, safety and security should be focused.

Even the number of clusters have been varied and the performance of OLR models were investigated, the scenario Y3* was claimed to be the best one among all 8 scenarios. In term of T&T management, it is more convenient for the government to utilizes tourism resources. By employing clustering techniques, policymakers and tourism managers can develop more targeted and effective strategies, optimizing resource allocation and maximizing the potential for sustainable tourism development across diverse geographical and thematic clusters.

6. Discussion and Implication

OLR can assist for clustering Thailands provinces for targeted T&T promotion. This allows evidence-based grouping into clusters with customized tourism development and promotion strategies tailored for their estimated visitor profiles and requirements (Crotts et al. (2022); Muazir and Hsieh (2019)). Regarding the results obtained in this study, 77 provinces were suitably classified into three clusters based on their revenue levels and the most appropriate proportion of primary:secondary:tertiary provinces is 18:29:35. The sustainable tourism framework in Thailand emphasizes balancing growth with ecological and cultural preservation. By targeting regional clusters, promoting local experiences, and applying personalized marketing strategies, the initiative seeks to optimize resource allocation and address diverse tourist preferences. This approach supports long-term economic resilience, enhances local engagement, and promotes Thailand as a sustainable destination, attracting both luxury and eco-conscious travelers. Implementing the research findings and observing the tourism similarities within cluster, significant factors, tourism promotion and marketing are additionally discussed in the following subsections.

6.1. Empirical implication

The best of 8 scenarios was Y3*. Number of Thai tourists, average spending, and foreign occupancy matter. That is, this study highlights the impact of different tourist segments on average spending, suggesting a need to balance quantity with quality of visitors, i.e., either luxury or budget travelers. Firstly, more domestic tourists could boost tourism and average revenue. Since international travel is difficult, domestic tourists travel closer to home, increasing micro-tourism. Thus, local tourism output for consumption will increase, boosting the local economy to meet demand. New experiences, population exchange, and novel service design are possible with micro-tourism. Micro-tourism rejuvenating local communities will greatly contribute to UN SDGs 8, 11, and 12 (Centre for SDG Research and Support; United Nations; World Tourism Organization and Global

Tourism Economy Research Centre). Domestic T&T may foster inclusive and sustainable economic growth. Secondly, Thai tourist numbers directly affect typical expenses. Large numbers of domestic visitors may not generate more revenue than fewer high-end foreign visitors, but sustainable development, especially in secondary and tertiary provincial clusters, prioritizes using local resources wisely and promoting community strength. Thirdly, long-stay foreign occupancies (1 month or more) in important Thai tourism sites may boost average tourism earnings (Shekari et al. (2022)). That is, longer international visitor stays allow fixed costs to be spread across more room nights and generate more food, facilities, transport, etc. In the key provincial clusters like Bangkok, Phuket, Chon Buri, Chiang Mai, Krabi, and others, digital nomads, expats, retirees, and remote workers are likely to book apartment rentals or condos rather than hotels. They pay for living expenses, local services, medical care, and recreation, unlike shorter-term foreign tourists. Luxury overseas travelers spend more per night/trip than budget travelers. Even if aggregate tourism expenditure rises, a substantial increase in backpacking or budget foreign tourists could lower average income per tourist. Secondary and tertiary clusters must create optional tourism resources and packages for domestic and foreign tourists.

6.2. National tourism guideline

Thailand aims to balance tourism growth with sustainability by focusing on resilience, quality, enriching experiences, and environmental protection as outlined in the four main pillars of the National Tourism Development Plan (Ministry of Tourism and Sport). Under the 20-year National Strategy (2018-2037), Thailand aims to uphold its position as a premier global tourism destination by improving its tourism infrastructure holistically (Office of the National Economic and Social Development Council). The focus shall be on attracting high-value tourists while diversifying tourism offerings to meet varied traveler interests. Simultaneously, Thailand will preserve its cultural heritage and natural assets. Considering the readiness and potential, the provinces in primary cluster presents their outperformance due to their tourism resource, infrastructure, and transportation. In contrast, the provinces in secondary and tertiary clusters seem to have more challenges and they need more promotion in many aspects (The Nation Thailand; The Nation Thailand). Thus, most tourism income has been concentrated in the primary provincial cluster such as Bangkok, Phuket, Chonburi, Chiang Mai, Krabi and Surat Thani. In 2018, the revenue proportion of the primary to secondary and tertiary provinces was 91.26%: 8.74%. Comparing the revenue between Bangkok and the rest of Thailand, the proportion was 38.56%: 61.44%. The proportion does not just include the tourism income that is concentrated in primary provinces, but also comprises the government budgets allocated to those provinces. To implement the national tourism guideline effectively and promote lesser-known locales, data-driven tourism strategies or AI-driven personalized marketing can significantly enhance destination and visitors' unique experiences that emphasize its culture, cuisine, and hospitality. With VR/AR technology effectively showcasing attractions and sustainable practices appealing to eco-conscious travelers, regional clustering fosters multi-destination visits, optimizing resource allocation, balancing luxury and budget offerings, and attracting long-term visitors to support resilient tourism growth.

6.3. Promotion policy and best practice

To foster sustainable growth and harness the potential of different regions, the government should implement promotion policies and best practices tailored to three distinct provincial clusters [Crotts et al. (2022); Muazir and Hsieh (2019); Soh et al. (2021); Sumanapala and Wolf (2021)]. From the numerical results, the tourism promotion of Thailand can be discussed in the following issues.

- Firstly, major primary provinces encounter very high in both domestic and foreign tourist flows. Dealing with overcrowding issues, the primary provinces should support international and do-

mestic travel as well as expand accommodation options. Preserving local culture and ecosystems, sheer volume of tourists at certain sites should be managed.

- Secondly, primary province tourism funding should be transferred to intermediate and tertiary provinces to promote activities that smoothly integrate the tourism supply chain from origin to destination. The budget can maintain infrastructure and transportation to improve secondary and tertiary province connectivity. To sustain T&T in secondary or tertiary provinces, assess preparation and potential inequities, focusing on local communities.

- Thirdly, multi-destination packages and web marketing should promote secondary and tertiary provinces' less-visited sites. Micro-tourism can create new routes and promote secondary and tertiary provinces through intra-cluster and inter-cluster collaboration. Thailand is an agricultural country with abundant natural resources and cultural richness, so micro-tourism can lead to agro-tourism, eco-tourism, gastro-tourism, cultural tourism, lifestyle, and spiritual tourism [Mangrove Resources Promotion and Development Subdivision; The Nation Thailand]. Additionally, the tourists key demographics should be targeted for special tour packages.

- Fourthly, secondary or tertiary provinces should become health tourism hubs to attract affluent, long-term visitors [Thomas and van Nieuwerburgh (2021); Sumanapala and Wolf (2021)]. These provinces should improve amenities and tourist attractions based on tourism and health safety requirements like the Amazing Thailand Safety and Health Administration (SHA) [The Nation Thailand; Suphanida (2023); Thailand Incentive and Convention Association]. This strategy targets high-quality tourists and increases secondary and tertiary provincial cluster appeal.

- Finally, the brand awareness should be boosted for visibility in Thai secondary and tertiary cities through targeted local events, partnerships with community influencers, and strategic digital marketing. The culturally relevant campaigns may be utilized to resonate with the local audience, emphasizing the brands unique value or authenticity.

6.4. Tourist behavior and digital marketing

Regarding tourist behavior for sustainable development, particularly in Thailand, three main aspects including the number of Thai tourists, average spending, and foreign occupancy indicates the shift towards micro-tourism due to international travel difficulties caused by the outbreak of COVID-19. To recover and potentially surpass international tourist numbers while fostering sustainable tourism growth in post-COVID-19 period, some possible data-driven and digital marketing are recommended as follows:

- Data analytics can revitalize local economies by identifying trending destinations and experiences, enabling targeted digital campaigns to showcase unique local offerings [Borges-Tiago et al. (2021); Liu (2023)]. Big data analysis reveals underexplored destinations with rich cultural heritage and natural beauty. Simultaneously, digital content promotion of these locales can redistribute tourist flows, fostering sustainable tourism and supporting local SMEs through social media marketing. This approach promotes economic equity and preserves destination authenticity while alleviating pressure on oversaturated areas.

- Leveraging food tourism to capitalize on Thailand's world-famous cuisine, digital content showcasing cooking classes, street food tours, and farm-to-table experiences in rice paddies or by the beach can be amplified. Furthermore, VR or AR can immerse viewers in Thailand's cultural sites, traditions, and stunning natural landscapes-from mountain ranges to beaches. This approach not only caters to those who cannot travel immediately but also piques interest and encourages future visits, solidifying Thailand's position as a top food and cultural tourism destination [Al Abri et al. (2023); Piboonrungroj et al. (2023)].

- Marketing strategies should focus on attracting long-term visitors such as digital nomads and retirees. Data-driven campaigns can highlight Thailand's suitability for extended stays, showcasing relevant accommodations, services, and activities tailored to these demographics' preferences. Highlighting new experiences for both domestic and international tourists, novel service designs and experiences should be creatively developed and marketed. The social media and influencer partnerships have to focus on these unique Thai experiences, emphasizing food, culture, and local hospitality [Altinay and Taheri (2019)].

- Data analytics facilitates balanced luxury and budget tourism in Thailand through expenditure-based segmentation, enabling targeted marketing strategies. High-end travelers are offered exclusivity, while budget tourists receive affordable authenticity. Identification of eco-conscious travelers aligns with sustainability trends. This approach optimizes resources, diversifies the tourist base, and enhances resilience, maintaining Thailand's broad appeal and potentially fostering sustainable growth across market segments.

- Personalized marketing in tourism leverages AI and machine learning to unlock the potential of tourist data [Alnajim and Fakieh (2023); Borges-Tiago et al. (2021); Velentza and Metaxas (2023)]. By analyzing this data, which can encompass demographics, travel history, and online behavior, AI can generate tailored recommendations and itineraries for individual tourists. These personalized experiences can be delivered through various channels, including chatbots, mobile apps, and targeted email campaigns, leading to a more satisfying and engaging travel experience.

To encourage visitors to explore multiple destinations during their stay, the regional tourism board is clustering complementary provincial attractions into joint tourism routes and campaigns. Inter-regional branding expands access and aligns underused tourism capacities to spread economic benefits.

7. Conclusions and Recommendations

Countries that contribute more to GDP from T&T, like Thailand, emphasize its relevance for economic development. Revenue-based provincial classification improves strategic planning, resource distribution, and policymaking. This study added outlier correction to OLR model for provincial clustering. The 8 scenarios include 2, 3, 4, and 5 cluster models with and without OAS. This study found that the OLR models with OAS show better predictive performance for all number of clusters. Notably, the OLR model with OAS and three clusters (Y3*) had the highest accuracy, precision, recall, F1-score, and F1-G-mean among 8 scenarios. The number of Thai tourists, average Thai visitor spending, and foreign occupancy were authorized for Y3* due to their high OR values. The numerical results showed that 77 provinces were appropriately clustered into 18 primary, 29 secondary, and 30 tertiary provinces. Thailands economy relies on tourism, however Bangkok benefits more than other regions. The numerical results in this study can provide insights that can be applied to tourist behavior analysis and data-driven digital marketing strategies for Thailand's tourism recovery. The empirical implications of national tourism guidelines, promotion policies, and best practices can be achieved in the context of both large-scale national tourism initiatives and smaller-scale SMEs, providing a comprehensive analysis of their effectiveness across varied operational scopes. Furthermore, this study delves into the intricate relationship between tourist behavior and digital marketing strategies, elucidating how these factors influence and are influenced by the implementation of tourism policies at different scales, thereby offering valuable insights for policymakers and industry stakeholders alike. This holistic approach facilitates the development of nuanced, data-driven policies that balance national objectives with local needs, ultimately enhancing the competitiveness and sustainability of the tourism sector across various scales.

Acknowledgments

The authors express sincere gratitude to the referees for their valuable comments and suggestions. This research was made possible through the generous support of the Thammasat University Research Fund (Contract No. TUFT 047/2563).

References

Abuzaid AH Nae'l AA. Comparative Study on Outliers-Detection Procedures in Binary Logistic Regression Model. Thail Stat. 2024; 22(1): 180-191.

Agresti A. Categorical data analysis. New York: John Wiley & Sons; 2012.

Akarajarasroj T, Wattanapermpool O, Sapphaphab P, Rinthon O, Pechprasarn S, Boonkrong P. Feature Selection in the Classification of Erythemato-Squamous Diseases using Machine Learning Models and Principal Component Analysis. 15th Biomedical Engineering International Conference (BMEiCON). IEEE 2023: 1-5.

Al Abri I, Alkazemi M, Abdeljalil Waed, Al Harthi H, Al Maqbali F. Attitudinal and behavioral loyalty: Do psychological and political factors matter in tourism development? Sustainability. 2023; 15(6): 5042.

Alnajim RA, Fakieh B. A Tourist-Based Framework for Developing Digital Marketing for Small and Medium-Sized Enterprises in the Tourism Sector in Saudi Arabia. Data. 2023; 8(12): 179.

Alrumaidhi M, Rakha HA. Factors affecting crash severity among elderly drivers: a multilevel ordinal logistic regression approach. Sustainability. 2022; 14(18): 11543.

Altinay L, Taheri B. Emerging themes and theories in the sharing economy: a critical note for hospitality and tourism. Int J Contemp Hosp M. 2019; 31(1): 180-193.

Borges-Tiago T, Silva S, Avelar S, Couto JP, Mendes-Filho L, Tiago F. Tourism and COVID-19: The show must go on. Sustainability. 2021; 13(22): 12471.

Centre for SDG Research and Support. Basic Information about SDGs — SDG Move — sdgmove.com. https://www.sdgmove.com/intro-to-sdgs/. [Accessed 29-07-2024].

Crotts JC, Magnini VP, Calvert E. Key performance indicators for destination management in developed economies: A four pillar approach. Ann Tour Res Empir Insights. 2022; 3(2): 100053.

Digital Government Development Agency. Datasets - Open Government Data of Thailand — data.go.th. https://data.go.th/en/dataset?q=tourism. [Accessed 01-01-2024]. 2020.

Duan W. Ordinal logistic regression analysis on influencing factors of space tourism expectation model. Journal of Physics: Conference Series. Vol. 1651. 1. IOP Publishing. 2020: 012066.

Fakfare P, Lee JS, Han H. Thailand tourism: A systematic review. J Travel Tour Mark. 2022; 39(2): 188–214.

Hashim S. Modeling to Forecast International Tourism Demand in Thailand. Sci Technol Asia. 2023; 28(1): 70–76.

Liang J, Bi G, Zhan C. Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. Ann Transl Med. 2023; 8(16): 982.

Liu CH, Horng JS, Chou SF, Yu TY, Huang YC, Lin JY. Integrating big data and marketing concepts into tourism, hospitality operations and strategy development. Qual Quan. 2023; 57(2): 1905–1922.

Liu X. Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. J Mod Appl Stat Methods. 2009; 8(2): 632–645.

Liu X, Koirala H. Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. J Mod Appl Stat Methods. 2012; 11(1): 242–254.

MacKenzie DI, Nichols JD, Royle J A, Pollock KH, Bailey L, Hines JE. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier, 2017.

Mahfuz NM, Yusoff M, Ahmad Z. Review of single clustering methods. IAES Int J Artif Intel. 2019; 8(3):221.

Mangrove Resources Promotion and Development Subdivision. Ecotourism. 2021 — dmcrth.dmcr.go.th. https://dmcrth.dmcr.go.th/manpro/detail/11698/. [Accessed 01-01-2024].

Melnykov V. Challenges in model-based clustering. WIRES Comput Stat. 2013; 5(2): 135–148.

Ministry of Tourism and Sport. Tourism Development Plan No. 3. https://secretary.mots.go.th/more_news.php?cid=60. [Accessed 01-01-2024].

Muazir S, Hsieh HC. Urban network in strategic areas in Indonesia case study: Sambas Regency, West Kalimantan. J Des Built Environ. 2019; 19(2): 14–29.

Mueller R, Sobreira N. Tourism forecasts after COVID-19: Evidence of Portugal. Ann Tour Res Empir Insights. 2024; 5(1): 100127.

Na Bangchang K, Wongsai S, Simmachan T. Application of Data Mining Techniques in Automobile Insurance Fraud Detection. Proceedings of the 2023 6th International Conference on Mathematics and Statistics. 2023: 48–55.

O'Connell AA. Logistic regression models for ordinal response variables. Sage: 146, 2006.

Office of the National Economic and Social Development Council. National Strategy — nscr.nesdc.go.th. http://nscr.nesdc.go.th/ns/. [Accessed 15-03-2024].

Piboonrungroj P, Wannapan S, Chaiboonsri C. The impact of gastronomic tourism on Thailand economy: under the situation of COVID-19 pandemic. Sage Open. 2023; 13(1): 21582440231154803.

Putman AI, Carbone I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecol Evol. 2014; 4(22): 4399–4428.

Rasool H, Maqbool S, Tarique Md. The relationship between tourism and economic growth among BRICS countries: a panel cointegration analysis. Futur Bus J. 2021; 7(1): 1-11.

Sharafuddin MA. Types of Tourism in Thailand. E-review of Tourism Research. 2015; 12(3): 210-219.

Shekari F, Ziaee M, Faghihi A, Jomehpour M. Nomadic livelihood resilience through tourism. Ann Tour Res Empir Insights. 2022; 3(1): 100034.

Shui W, Zhang Y, Wang X, Liu Y, Wang Q, Duan F, Wu C, Shui W. Does Tibetan household livelihood capital enhance tourism participation sustainability? evidence from Chinas Jiaju Tibetan village. Int J Env Res Pub He. 2022; 19(15): 9183.

Simmachan T, Manopa W, Neamhom P, Poothong A, Phaphan W. Detecting fraudulent claims in automobile insurance policies by data mining techniques. Thail Stat. 2022; 21(3): 552–568.

Soh AN, Chong MT, Puah CH. A novel look at Thailand's tourism from a tourism composite index. Int J Tour Policy. 2021; 11(4): 401–415.

Sumanapala D, Wolf ID. A wellbeing perspective of Indigenous tourism in Sri Lanka. Ann Tour Res Empir Insights. 2023; 4(2): 100099.

TAT Newsroom. Amazing Thailand Launches Your Stories Never End Campaign at Arabian Travel Market 2024 - TAT Newsroom — tatnews.org. https://www.tatnews.org/2024/05/amazing-thailand-launches-your-stories-never-end-campaign-at-arabian-travel-market-2024/. [Accessed 29-07-2024].

Thailand Center. Soft Power (5F) of Thailand attracts tourists — thailand.go.th. https://www.thailand.go.th/issue-focus-detail/001_02_070. [Accessed 29-07-2024].

Thailand Incentive and Convention Association. TICA - — Medical Tourism — tica.or.th. https://www.tica.or.th/why-thailand-categories/medical-tourism. [Accessed 29-07-2024].

Suphanida T. An Overview of Thailand's Medical Tourism Industry — pacificprime.co.th. https://www.pacificprime.co.th/blog/thailand-medical-tourism/. [Accessed 01-01-2024].

The Nation Thailand. Thailand issues free guidebook to 60 spiritual tourism sites — nationthailand.com. https://www.nationthailand.com/thailand/tourism/40031327. [Accessed 15-03-2024].

The Nation Thailand. Bright growth opportunities for Thai medical tourism sector — nationthailand.com. https://www.nationthailand.com/thailand/tourism/40032370. [Accessed 20-01-2024].

The Nation Thailand. World Development Indicators — The World Bank — wdi.worldbank.org. https://wdi.worldbank.org/table/6.14#. [Accessed 15-01-2024].

Thomas P, van Nieuwerburgh C. The lived experience of long-term overland travel. Ann Tour Res Empir Insights. 2022; 3(1): 100040.

UN Tourism. Tourism Grows 4% in 2021 but Remains Far Below Pre-Pandemic Levels — unwto.org. https://www.unwto.org/news/tourism-grows-4-in-2021-but-remains-far-below-pre-pandemic-levels. [Accessed 20-03-2024].

United Nations. World Tourism Day — United Nations — un.org. https://www.un.org/en/observances/tourism-day. [Accessed 20-03-2024].

Velentza A, Metaxas T. The role of digital marketing in tourism businesses: An empirical investigation in Greece. Businesses. 2023; 3(2): 272–292.

Wang H, Quintana FG, Lu Y, Mohebujjaman M, Kamronnaher K. How are BMI, nutrition, and physical exercise related? An application of ordinal logistic regression. Life. 2022; 12(12): 2098.

World Tourism Organization and Global Tourism Economy Research Centre. UNWTO/GTERC Asia Tourism Trends 2019 Edition. UNWTO, Madrid. 2019; 128.

Wu TP, Wu HC, Liu YT, Chu S, He Z. Tourism and economic growth in Asia: a symmetric and asymmetric causality. J Policy Res Tour Leis Events. 2023; 1–17.

Yilmaz AE, Demirhan H. Weighted kappa measures for ordinal multi-class classification performance. Appl Soft Comput. 2023; 134: 110020.