APPLICATION OF K-MEANS CLUSTERING TO CLASSIFICATION OF INBOUND TOURISTS TO CHINA

[1]YUEYI CHEN, [2]NAPAT HANPORNCHAI

[1,2]Faculty of Economics Chiang Mai University Thailand
Email: [1]13577186557@163.com, [2] napateconstudent@gmail.com

Abstract - The number of inbound tourists received by China is among the top in the world, which makes China's inbound tourist market very huge. Obtaining sub-tourist markets by segmenting the tourist market is beneficial to better meet the needs of tourists, helps travel agencies make market decisions and develop markets. This paper classifies inbound tourist of China and get the sub-tourist markets by K-means clustering method. Using K-means clustering method to classify tourist samples requires collecting data on different attributes of the sample.This paper use K-means clustering to cluster the inbound tourist samples to China, which containing multiple attribute data, the clustering results are sub-tourist markets and provide appropriate policy suggestions to the tourism agencies basis the characteristics of each sub-tourist market.

Keywords - Inbound Tourist, Tourism, K-means Clustering, China

I. INTRODUCTION

The motivation of this paper can be divided into three parts. The first motivation for this study is based on the meaning of inbound tourist segmentation in China. Imagine that if the number of tourists in the market is too little, then the benefits of tourist segmentation will be less than the cost, so tourist segmentation is not necessary. If there is only one type of tourists in Chinese inbound tourist market, all the policies of the travel agency and the government will only need to meet the needs of this type of tourists, then the tourist segmentation will not be necessary. If all types of tourists have the same behaviors, choices and preferences, then the segmentation of tourists is meaningless, because the purpose of tourist segmentation is to study the behavior of different types of tourists.Therefore, for the inbound tourist market in China, the necessity of inbound tourist segmentation supposes come from: 1) large number of tourists. 2) Various types of inbound tourists. 3) Different types of inbound tourists have different behavior. According to the China Inbound Tourism Development Report 2019, in 2018, there were more than 1.4 billion person-times of overseas tourism. Among them, the number of Chinese international tourists reached 62.9 million, increase by 3.6% compare with 2017, ranking first in East Asia and fourth in the world. These tourists come from different countries, have different ages, occupations and incomes, and because of the above information, tourists have different choices in accommodation, dining, and transportation, which makes the research of this article necessary.

The second motivation for this study is based on the necessity of tourist segmentation in Chinese inbound tourist market. The main necessity are: 1) Tourist market segmentation make foreigner tourists easier to meet their needs. In sub-tourist markets, information is easier to understand and feedback. Once the needs of tourists change, travel agencies can quickly change their marketing strategies, formulate corresponding countermeasures to adapt to changes in market demand, and continue to meet the needs of tourist. 2) Help travel agencies make market decisions and develop markets. In the current market economy, indifferent marketing become more and more inefficient due to trying to please everyone's needs. Market segmentation is to identify the differences in consumer demand and select the appropriate sales target as their target market. Through market segmentation, travel agencies can analyze and compare the purchase potential, satisfaction, competition, etc. of each market segment, and explore market opportunities that are conducive to the company, so that they can make timely sales decisions, grasp market initiative, and improve enterprise resilience and competitiveness. 3) enables the resources of travel agencies to be most effectively allocated. The manpower, material resources and funds of any travel agency are limited. Through market segmentation, travel agencies have chosen the market segment that suits them, avoiding the blindness of sales, and can concentrate resources such as human, financial, and material resources on a certain market segment, gaining local market advantages.

The third motivation for this study is based on the finding that most of the other clustering studies on Chinese inbound tourists only select two to three attributes, and do not distinguish between attributes. For example, Li(2008) and He (2015)select nationality and purpose of as attributes to study the motivation patterns of tourists from various countries. Zhang (2011) selects tourists' communication preferences and involvement types to study the involvement of tourists with different levels of communication. Cai (2011) select nationality and annual tourist expenditure as attribute to classify inbound tourist in Fujian. Different from the above research, this study has 27 attributes, and can be divided into two categories of information attributes and behavior

attributes, which can systematically segment inbound tourists.

II. LITERATURE REVIEW

2.1Research on information attributes in this study

Kozak (2002) asserts that tourist motivations are different regarding to their nationalities. Sussmann, Rashcovsky (1997) confirmed the influence of nationality on tourist vacation travel pattern and attitudes towards the destination. The Yukse (2003) survey shows that female tourists are more likely to perceive services that are intangible and difficult to express in language or behavior, so female tourists are more willing to spend on this aspect to enhance the travel experience.David (2008) surveyed 2000 individuals living in major metropolitan areas and classify observations into the family life cycle: age, marital status, and presence/absence of children. He found that young and single people (age18-34) want excitement and adventure in a vacation, and the activity they mainly choose is nightclub.Shen (2018) Researched the relationship between family tourism in China and the education level of family members, and found that the structure of family tourism consumption was most affected by the highest academic qualifications of family members. The higher the education level of the family, the greater the expenditure on tourism.Song and Witt (2000) find that personal disposable income or private consumption is a better proxy for income when leisure tourism is investigated.

2.2 Research on China's inbound touristsegmentation

Zhang, Ma and Li (2010) conducted a random sample survey of inbound tourists in Xi'an and collected 968 inbound tourist samples. The tourist attributes they select are gender, age, and education level. At the same time, suppose two index as internal driving force index and attractiveness index. The method used is the one-way variance method.Zhao (2011) obtained second-hand data on inbound foreign tourists through the travel yearbook. The selected tourist attributes are gender, age, occupation, entry method and daily expenditure. Their research method is shift-share analysis (SSM), which calculates the competitive advantage of each source in the inbound tourism market in Yunnan, and classifies tourists according to their different competitive advantages.

2.3 Research on application of k-means clustering method in the tourist segmentation

Park and Yoon (2006) collected the data from rural tourists who visited rural tourism villages in Korea over a period of 12 months between September 2004 and August 2005. These villages, which were funded by the Rural Traditional Theme Village (RTTV) and Green Rural Experience Village (GREV) programs to stimulate rural tourism, were chosen nationwide to ascertain general rural visitors' behavior.Pulido-Ferna´ndez and Sa´nchez-Rivero (2009) focuses on a subset of 2,983 respondents who traveled to one of these medium-sized towns of Andalusia. The survey was conducted between February and November of 2008. With this sample size, the maximum sampling error for a 95% confidence level and the worst case ($p = q = 0.50$) is ±1.8%.Yotsawat and Srivihok (2014) used secondary data accumulated from the Department of Tourism, Ministry of Tourism and Sports, Thailand. The total number of inbound tourists in the data files is 83,402 who arrived to Thailand in a period from 2008 through 2010.

III. DATACOLLECTION

The data of this study collect by questionnaires. The questionnaire surveyed were overnight foreign tourists who entered China. This research plan prepares 300 questionnaires, but because some questionnaires is invalid, the actual sample size is 237. The questionnaire consists of 27 questions, which corresponds to 27 attributes of inbound tourists. The 27 attributes are: nationality, gender, age, marital status, education level, occupation, income(annual/USD), Last time entry, companion, trip type, purpose of entry, length of stay, day type, entry times, daily expenditure (USD), consumption ratio on food, consumption ratio on accommodation, consumption ratio on transportation, consumption ratio on sightseeing, Consumption ratio on entertainment, Consumption ratio on shopping, preference of food, preference of accommodation, preference of transportation, preference of sightseeing, preference of entertainment, preference of shopping.

IV. STUDYFRAMEWORK

1) Collect the tourist sample set X of foreign tourists entering China through questionnaires. This paper has 237 samples.
2) Convert categorical data to numeric data. In this study, the data collected through the questionnaire survey is categorical data, but the k-means clustering method is only used for numeric data, so data preprocessing is required.
3) The k value is the number of clusters in a clustering process.Since different k values will result in very different clustering results, the selection of k value has a great influence on the K-means clustering. This paper using Gap evaluation to determine the optimal number of clusters$\hat{k}$.
4) The centroid is the center point of each cluster, and the initial centroids are the center points of the cluster selected for the first time when performing the k-means clustering algorithm. Usually, the selection of the initial centroid is random, but the random selection of centroid can sometimes cause a large error in the clustering result, so this article uses K-means++

method to select the optimal initial centroid μ^*.
5) The k-means clustering method classifies the sample points by minimizing the distance between each attribute data of each sample point to the centroid of the attribute. This paperusing the K-means clustering method to cluster tourist sample set X.
6) Propose policies and suggestions based on the characteristics of clustering results.

V. DETERMINE THE OPTIMAL K VALUE BY GAP EVALUATION

After encoding the collected 176 visitor data by label encoding, this paper uses the gap evaluation method to calculate the optimal K value.The gap statistics method is published in the paper: Estimating the number of clusters in a data set via the gap statistic （Tibshirani, Walther and Hasile, 2003）.

Tourist data set X, $X = \{x_{if}\}$, where $i = 1, 2, \ldots 237$ represent the data point in each data set.frepresentsthe attributes contained in the sample set. First, randomly choose a k value and a set of initial centroids μ clustered on the data set X.Let d_{ii^*} denote the distance between i and i^*, expressed using Euclidean distance as:

$$d_{ii^*} = \sum_f (x_{if} - x_{i^*f})^2 \qquad (1)$$

Then clustered each of the data set into k clusters $C_1, C_2, \ldots C_k$, with C_r denoting the indices of observation in cluster r, and $n_r = |C_r|$, represent the sample size of C_r. Calculate $D_r$as:

$$D_r = \sum_{ii^* \in C_r} d_{ii^*} \qquad (2)$$

D_r is the sum of the pairwise distances for all point in cluster r. Calculate $W_k$as:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r (3)$$

If the distance d is the squared Euclidean distance, then W_k is the pooled within-cluster sum of squares around the cluster mean. The sample size n is suppressed in this notation.

The idea of this approach is to standardize the graph of $\log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of the data in each data set. Estimate the optimal number of clusters is then the value of k for which $\log(W_k)$ falls the farthest below this reference curve. Hence define:

$$Gap_m(K) = E_m^*(\log W_K) - \log(W_K) \quad (4)$$

Were E_m^* denotes expectation under a sample of size m from the reference distribution. The estimate $\hat{k}$ will be the value maximizing $Gap_m(k)$ after take the sampling distribution into account. Therefore, $\hat{k}$represent the optimal kvalue in each cluster.

NumObservations	176
InspectedK	[1 2 3 4 5 6 7 8]
CriterionValues	
[0.6745 0.7010 0.8175 0.8621 0.8834 0.8790 0.8905 0.8894]	
OptimalK5	

Table 1: The result of gap evaluation

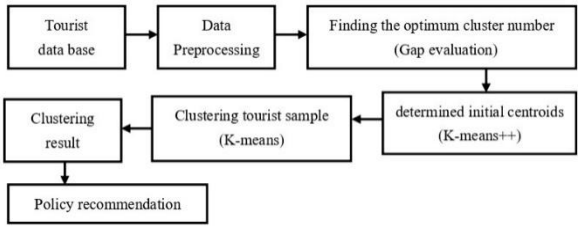
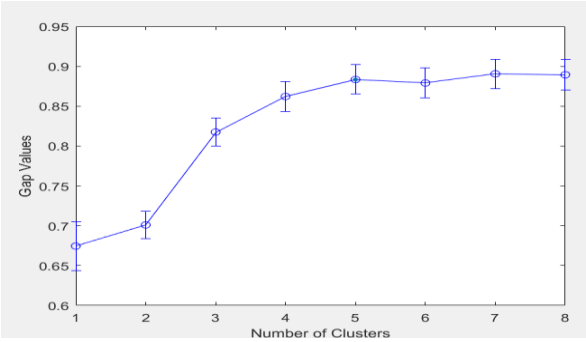


Figure 1: The relationship between number of clusters(K) and Gap values

According to table 1,assuming that the optimal k value is between 1 and 8, based on the gap criterion, the optimal number of clusters is five.

Figure 1: The relationship between number of clusters and gap values



According to figure 1, the maximum value of the gap criterion occurs at seven clusters. However, the value at five clusters is within one standard error of the maximum, so the suggested optimal number of clusters is five.

VI. OPTIMIZEINITIAL CENTROID SELECTION BY K-MEANS ++

The K-Means++ algorithm is a method for selecting the optimal initial centroid.

First randomly select a column vector from the tourist data set X as the first cluster center μ_1. For each x_i in the data set, calculate the distance from the nearest cluster center in the selected cluster center:$D(x_i) = \arg\min \sqrt{\sum_{f=1}^{p} (x_i - \mu_r)^2}, r = 1,2, \ldots \hat{k}_{selected}$ (5)

The value of $\hat{k}$ is determined by gap statistics above, which is 5.Select a new vector μ_r as the new

clustering center, the vector with larger $D(x)$, the probability of being selected as the clustering center is bigger. The probability is:

$$P = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \quad (6)$$

Cluster	Centroid	Centers corresponding
1	112	1.2500
2	5	1.0000
3	18	1.0000
4	24	1.2857
5	174	6.5000

Table 2: The optimal initial centroids by K-means++

According to table 2, the five point with the biggest probability are selected as the five centroids of the five clusters.

VII. K-MEANSCLUSTERINGPROCESS

K-means clustering algorithm proposed by MacQueen (1967), which belongs to nonhierarchical clustering algorithms and portions the data into k number of clusters specified by the researcher. K-means clustering algorithm maximizes the distance between clusters and minimizes the distance of observations from one another in a cluster. The reason for choosing this method in this paper is that the K-means clustering method is simple in calculation and the time complexity is low, which is O(tKmn). Where t is the number of iterations, K is the number of clusters, m is the number of records, n is the number of dimensions (features).

Substituting the optimal K value, the optimal centroids and the maximum number of iterations N into the K-means clustering algorithm can get the clustering result.

For $n = 1,2, \dots, N$,Initialize C to $C_j = \emptyset j = 1,2 \dots \hat{k}$. Calculate the weighted Euclidean distance d_{ir} between attributes value of each data point x_i and each centroid vector $\mu_r$in X.

$$d_{ir} = \sqrt{\sum_{f=1}^{p} W(x_i - \mu_r)^2} (7)$$

Classify x_i into the category λ_i that minimizes d_{ir} and update C_{λ_i}

$$C_{\lambda_i} \cup \{x_i\} (8)$$

Recalculate the centroid of the sample in C_j and get new vector μ_r ,where $\mu_r$is the mean vector of cluster C_j.

$$\mu_{rf} = \frac{1}{|C_j|} \sum_{x_{if} \in C_j} x_{if} (9)$$

Repeat the above step until all k centroids no longer change, the number of iterations is n, then proceed to step.Output final cluster partition

$$C_j = \{C_1, C_2, \dots C_{\hat{k}}\} (10)$$

Repeat the above step until k cluster centroids are selected, get the initial centroids $\mu_r.\mu_r$is the optimal initial centroid $\mu^*$of tourist sample set.

Attribute	Cluster 1	2	3	4	5	Total	Attribute	Cluster 1	2	3	4	5	Total
Nationality							Mantal status						
Belgium	0	0	0	1	0	1	Single (No children)	58	28	41	1	8	146
Argentina	0	0	0	1	0	1	Single (Have children)	1	0	2	0	1	4
Italy	0	0	0	7	0	7	Married (No children)	1	22	1	10	5	39
Romania	0	0	0	11	0	11	Married (Have children)	1	3	6	18	20	48
Germany	0	0	0	8	0	8	Education level						
Russia	0	0	0	7	0	7	Primary school	0	0	0	0	0	0
Finland	0	0	0	1	0	1	Middle school	0	0	0	0	0	0
France	0	0	0	1	0	1	High school	18	8	8	1	1	36
Portugal	0	0	0	1	0	1	Bachelor degree	38	37	35	11	9	130
Czech	0	0	0	1	0	1	Master degree	5	8	6	24	23	66
Poland	1	0	1	0	0	2	Doctor degree	0	0	1	3	1	5
United Kingdom	0	1	2	0	0	3	Occupation						
Latvia	0	0	1	0	0	1	Government officer	3	2	5	4	1	15
Mauritius	0	0	1	0	0	1	Employee	2	31	2	29	26	90
India	0	0	1	0	0	1	Sale/Self employee	0	1	2	0	1	4
Japan	3	12	3	0	0	18	Farmer	1	0	0	0	0	1
Thailand	38	19	29	0	0	86	Worker	1	1	13	1	0	16
Laos	8	3	1	0	0	12	Student	52	15	27	5	6	105
Nepal	4	3	0	0	0	7	Housewife or stay home husband	0	1	1	0	0	2
Korea	0	10	1	0	0	11	Non-worker	2	2	0	0	0	4
Myanmar	3	3	10	0	0	16	Income(annual/USD)						
Sri Lanka	1	0	0	0	0	1	Under 1000	49	18	15	8	2	92
Vietnam	2	2	0	0	0	4	1000-4000	9	32	14	8	2	65
Cambodia	1	0	0	0	0	1	4000-8000	1	1	16	1	1	20
Indonesia	0	0	0	0	1	1	8000-12000	1	1	0	17	12	31
Philippines	0	0	0	0	1	1	Over 12000	1	1	5	5	17	29
New Zealand	0	0	0	0	3	3	Last time entry						
Australia	0	0	0	0	2	2	2020	16	11	5	5	0	37
United States	0	0	0	0	24	24	2019	24	35	5	24	23	111
Canada	0	0	0	0	3	3	2018	9	5	5	2	5	26
Gender							2017	3	2	3	2	2	12
Male	13	24	18	20	16	91	Other years	9	0	32	6	4	51
Female	48	29	32	19	18	146	Companion						
Age							Alone	19	38	11	9	25	102
Under 15	0	0	0	0	0	0	Family	9	2	11	16	6	34
15-24	53	18	29	3	5	108	Friend	16	6	24	9	5	68
25-44	8	33	17	2	16	97	Lover	1	1	0	1	0	3
45-64	0	2	4	13	13	32	Colleague/classmate	16	6	4	4	0	30
Over 64	0	0	0	0	0	0							

Attribute	Cluster 1	2	3	4	5	Total	Attribute	Cluster 1	2	3	4	5	Total
Trip type							40%-60%	22	4	36	6	13	81
Self-trip	41	24	12	5	28	130	60%-80%	0	3	1	2	2	8
Organization trip	6	5	4	25	1	41	Over 80%	0	0	0	0	0	0
Tour agency	10	1	30	6	2	49	Consumption ratio on transportation						
Self-trip with agency	4	23	4	3	3	37	Under 20%	26	39	13	13	2	93
Purpose of entry							20%-40%	20	8	35	19	31	123
Meeting/Business	0	0	6	2	24	32	40%-60%	11	2	1	5	1	20
Sightseeing and leisure	0	32	9	21	18	60	60%-80%	1	2	0	1	0	4
Medical treatment	0	0	7	2	2	11	Over 80%	3	2	1	1	0	7
Visiting relatives or friends	2	2	26	2	0	32	Consumption ratio on sightseeing						
Religious activities	0	0	1	0	0	1	Under 20%	21	11	25	11	28	96
Working	2	5	1	5	0	13	20%-40%	20	9	11	12	1	53
Study abroad	57	14	0	7	0	78	40%-60%	17	27	13	15	2	74
Length of stay							60%-80%	2	5	1	1	3	12
Under 3	9	0	8	3	0	20	Over 80%	1	1	0	0	0	2
3-7	8	2	27	12	29	78	Consumption ratio on entertainment						
7-15	2	2	8	23	1	36	Under 20%	34	35	15	22	10	116
Over 15	42	49	7	1	4	103	20%-40%	14	8	25	14	2	63
Day type							40%-60%	10	9	9	2	21	51
Weekend	14	5	4	7	1	31	60%-80%	3	1	1	1	0	6
Holiday	20	31	40	8	7	106	Over 80%	0	0	0	0	1	1
Workday	27	17	6	24	26	100	Consumption ratio on shopping						
Entry times							Under 20%	8	29	20	22	19	98
1 time	10	5	11	16	27	69	20%-40%	12	9	7	9	12	49
2-3 times	16	30	16	11	2	75	40%-60%	22	8	19	8	1	58
4-6 times	16	5	18	6	1	46	60%-80%	12	6	3	0	2	23
Over 6 times	19	13	5	6	4	47	Over 80%	7	1	1	0	0	9
Daily expenditure (USD)							Preference of food						
Under 130	44	41	10	10	6	111	Fast food	12	2	21	2	1	38
130-260	13	7	26	10	4	60	Night market	20	8	9	3	4	44
260-390	4	2	9	16	13	44	Home cooking	13	32	8	2	1	56
Over 390	0	3	5	3	11	22	restaurant (Chinese food)						
Consumption ratio on food							Home cooking	2	5	4	24	1	36
Under 20%	13	6	7	5	4	35	restaurant (Food of your own country)						
20%-40%	14	34	30	22	17	127	Fine restaurant	10	5	8	4	22	49
40%-60%	20	11	12	7	13	63	(Chinese food)						
60%-80%	12	1	0	4	0	17	Fine restaurant	4	1	0	4	5	14
Over 80%	2	1	1	1	0	5	(Food of your own country)						
Consumption ratio on accommodation													
Under 20%	17	9	5	6	3	40							
20%-40%	22	37	8	25	16	108							

Table 3: The clustering result

Attribute	Cluster 1	2	3	4	5	Total	Attribute	Cluster 1	2	3	4	5	Total
Preference of accommodation							City view	25	10	29	8	2	74
One-star to two-star Hotel	7	22	4	2	1	36	Religious building	3	0	5	1	0	9
Three-star Hotel	15	8	8	19	18	68	Museum/exhibition	1	2	1	2	0	6
Four-star to five-star Hotel	8	3	24	4	11	50	Folk culture experience place	3	2	1	0	0	6
Guesthouse	4	4	7	6	3	24	Theme park	3	0	0	1	0	4
Resort	2	7	3	1	0	13	Preference of entertainment						
Apartment	25	9	4	7	1	46	Outdoor sport/Adventure	32	11	4	24	10	81
Preference of transportation							Indoor sport	9	6	7	3	1	26
Airplane	21	0	13	4	26	64	Watch a show/match	9	4	24	4	0	41
Train	28	0	7	7	5	47	Hot spring/spa	4	3	5	3	21	36
Taxi	12	2	21	18	1	54	Bar/Night show	7	29	10	5	2	53
Coach	0	0	3	1	0	4	Preference of shopping						
Ferry	0	0	0	0	0	0	Shopping center	36	42	21	9	10	118
Bus	0	45	4	8	2	59	Souvenir Shop	2	0	7	8	11	28
On foot	0	6	2	1	0	9	Duty free shop	3	0	2	13	0	18
Preference of sightseeing							Luxury store	1	0	1	0	11	13
Natural scenery	15	32	6	5	6	64	Supermarket or	19	11	19	9	2	60
Historic sites	11	7	8	22	26	74	Convenient store						

Table 4: The clustering result (continue)

According to table 3-4, five clusters of inbound tourists could be distinguished.

Cluster 1: International students (25.74% of total)
This cluster represent mainlystudent (85.25%)who

study abroad (93.44%). From the perspective of age group, most of the clusters are young people aged 15-24 (86.89%) who are unmarried and have no child (95.08%). In terms of income, most of the clusters are low-income earners whose annual income are under 1000 USD (80.33%). They came to China alone (31.15%) by self-trip (67.21%)to and the last time they came to China was 2019 (39.34%).This cluster is mostly long-term multiple-entrytravelerswho stay at China over 15 days (68.85%) and came to China more than 6 times (31.15%).This cluster is mostly low consumerswhose daily expenditure is under 130 USD (72.13%). Among all expenditures, tourists in this cluster spend the most on food, reaching 40%-60%(32.79%).In terms of consumer preferences, they prefer to eat at night market (32.79%), live at apartment (40.98%), travel by train (45.90%) and shopping in the shopping center (59.02%). It can be seen that the consumption preferences of international students are very close to those of local residents.

Cluster 2: Young tourist for sightseeing (22.36% of total)

This cluster represent mainly Asian (98.11%) young and middle-agedpeople aged 25-44 (62.26%) who mainly purpose to came China is to go sightseeing (60.38%). Those tourists are single and no children (52.83%). They are employee whose annual income are 1000-4000 USD (60.38%). Most tourists in this cluster choose to travel alone (71.70%) by self-trip with agency (43.49%), This kind of travel agency arranges hotels and booked air tickets, but the way tourists make their own itinerary is more and more popular with young people.

This group of tourists are also long-term travelers who stay at China over 15 days (92.45%) but they come to China less frequently than foreign students, they came to China 2-3 times (56.60%). This group of tourists are also low consumerswith daily expenditure under 130 USD (77.36%), this group of tourists spend the most expenditure (40%-60%) on sightseeing (50.94%). In consumer preferences, they prefer to eat at home cooking restaurant which sale Chinese food (60.38%), live at one-star to two-star hotel (41.51%), travel by bus (84.91%), see the natural scenery (60.38%), go to a bar (54.72%) and shopping in the shopping center (79.25%).

Cluster 3: Relative or friend visitor (21.10%)

This cluster represent mainly tourist who came to China to visiting relatives or friends (52%). Most of this group of tourists come to China with friends (48%) by tour agency (60%). Among this group of tourists, the number of people who came to China in 2017 is the largest (64%), and there are fewer people who have been to China recently.

These tourists are mostly short- and medium-term tourists, staying inChina for 3-7 days (54%) and came to China 4-6 times (36%). Thisgroup of tourists spends more than the above two groups with daily expenditure as 130-260 USD (52%), Most of their expenses (40%-60) are on accommodation (72%). They prefer to eat fast food (42%), live at four-star to five-star hotel (48%), travel by taxi (42%), see the city scenery (58%), watch a show (48%) and shopping in the supermarket (38%).

Cluster 4: Elderlytourist for sightseeing (16.46% of total)

This cluster represent mainly European middle-agedpeople aged 45-64 (33.33%) whose mainly purpose to entry China are to go sightseeing (53.85%). The tourist from this cluster are mostmarried people and have children (46.15%). Their occupationis employee (74.36%) whose annual income are 8000-12000 USD (43.59%) which is significantly more than the second group of young tourists.

They most came to China with family (41.03%) by organization trip (64.10%).They stay at China 7-15 days (58.97%) and most people came to China just 1 time (41.03%). Their daily expenditure is pretty high as260-390 USD (41.03%),the largest proportion is sightseeing as 40%-60% (38.46%). They prefer to eat at home cooking restaurant which salefood fromtheir own country (61.53%), live at three-star hotel (48.72%), travel by taxi (46.15%), visit the historic site (56.41%), do outdoor sports (61.54%) and shopping in the duty-free shop (33.33%).

Cluster 5: High-income business people (14.36% of total)

This cluster represent mainly the people who came to China to attend meetings or conduct business negotiations(70.59%). In terms of personal information, their young and middle-agedpeople aged 25-44 (47.06%) who are married and have children (58.82%). They are employee (76.47%) with master's degree (69.81%) whose income are over 12000 (50%), which is the highest among a total of 5 groups. They came to China alone (73.53%) by self-trip (82.35%) to and the last time they came to China was 2019 (67.65%). They stay at China 3-7 days (85.29%) and came to China 1 time (79.41%). In term of the consumption structure, their daily expenditure is 260-390 USD (38.24%), the largest proportion is which has structure as 20%-40% at entertainment (61.76%). They prefer to eat at fine restaurant which sale Chinese food (64.71%), live at three-star hotel (52.94%), travel by airplane (76.47%), see the historic sites (76.47%), do spa (61.76%) and shopping in the luxury store (32.35%).

VIII. POLICYRECOMMENDATION

Based on the above clustering results, this article proposes three suggestions for travel agencies that provide inbound tourism services in China:

1) Young tourists generally do not like traditional group tours. For young tourists, travelagencies should provide semi-self-guided tours, that is, arrange hotels and transportation for tourists, but not arrange itineraries.
2) Marriage and childbirth have a great impact on tourists' travel methods and consumption structure. Tourists who are married and have families need more travel services, and their daily expenses are also higher. Therefore, travel agencies should provide these tourists with itineraries suitable for all ages.
3) International students and family visitors will frequently enter the country and their spending choices are close to those of locals. Travel agencies should provide visa and other document agency services for these tourists to reduce the trouble of multiple entries.

IX. CONCLUSION

Data clustering is a common method to understand the characteristics of homogeneous tourist groups. This study divides Chinese inbound tourists into five clusters, based on the sample of 237 inbound tourists from the questionnaire survey. K-means cluster analysis is an effective method to deal with demographic and behavioral variable. The experimental result identified the existence of five well-defined subject groups: “international students”,“young tourist for sightseeing”,“relative or friend visitor”,“elderlytourist for sightseeing”, and“high-income business people.

Cluster 1 “international students” is a group of students who study in China and behavior like local people. Cluster 2 “young tourist for sightseeing” is a group of young unmarried tourist who travel to China alone and spend little money. Cluster 3 “relative or friend visitor” is a group of people who come to China to visit their friends or relatives, their income and expenditure are both at low level. Cluster 4 “elderly tourist for sightseeing” is a group of elder married tourist who travel to China with their family, their income and expenditure are both higher than young tourist. Cluster 5 “high-income business people’ is a group of people who come to China for meeting or business activities, they have the highest income of total five group. The result of this study can provide useful information to travel agency to manage their business.

REFERENCE

[1] Davis, B. D., &Sternquist, B. (1987). Appealing to the elusive tourist: An attribute cluster strategy. Journal of Travel Research, 25(4), 25-31.

[2] Huang, J. H., Huang, C. T., & Wu, S. (1996). National character and response to unsatisfactory hotel service. International Journal of Hospitality Management, 15(3), 229-243.

[3] Kozak, M. (2002). Comparative analysis of tourist motivations by nationality and destinations. Tourism management, 23(3), 221-232.

[4] LI, H. J., & LIANG, L. K. (2010). Shift-Share Analysis on Competition of Inbound Tourism Destination in China. Geography and Geo-Information Science, 1.

[5] Pesonen, J., Laukkanen, T., &Komppula, R. (2011). Benefit segmentation of potential wellbeing tourists. Journal of Vacation Marketing, 17(4), 303-314.

[6] Pesonen, J., &Honkanen, A. (2014). Using cluster analysis to segment tourists: response-style effects. Matkailututkimus, 10(2), 7-22.

[7] Sussmann, S., &Rashcovsky, C. (1997). A cross-cultural analysis of English and French Canadian's vacation travel patterns. International Journal of Hospitality Management, 16(2), 191-208.

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

[9] Yotsawat, W., &Srivihok, A. (2015). Rules Mining Based on Clustering of Inbound Tourists in Thailand. In Advanced Computer and Communication Engineering Technology (pp. 693-705). Springer, Cham.

[10] ZHANG, C. H., ZHANG, H., BAI, K., & LIU, L. (2010). An Analysis of Industrial Structure and Competitiveness of Inbound Tourism in Shaanxi Province Based on Dynamic Shift-Share and Grey-relation Analysis [J]. In Tourism Forum (Vol. 1).

[11] Zhang, J. (2009). Spatial distribution of inbound tourism in China: Determinants and implications. Tourism and Hospitality Research, 9(1), 32-49.

★★★