Exploring Influencing Factors on International Tourist Arrivals in the Philippines Using K-Means Clustering and Negative Binomial Regression

Fengyi An[1]* and Frumencio F. Co
Department of Mathematics and Statistics, De La Salle University
**Corresponding Author: fengyi_an@dlsu.edu.ph*

Abstract: K-means clustering algorithm is a commonly-used clustering algorithm with many advantages, such as simple understanding, realizing quickly, and processing large datasets conveniently. Count data often applies to many fields, such as medicine, sociology, and psychology. It is an essential statistical data type. Count data is analyzed using some frequently-used models, such as the Poisson regression and the negative binomial regression models. The negative binomial regression model has the phenomenon of overdispersion, wherein the variance is greater than the mean, that exists in the count data. As a consequence, overdispersion data analysis has become a crucial statistical issue. This paper focuses on studying the application of K-means clustering and the negative binomial regression model in an overdispersed inbound tourism data of the Philippines from 2009 to 2018. The K-means method is used to cluster 58 countries or regions by purpose of travel in the Philippines. The negative binomial regression model is performed for each cluster to identify the determinants of foreign tourist arrivals in the Philippines. Results showed that only the pattern of the number of tourist arrivals for holiday purpose has a trend stationarity. The number of tourists for holiday purpose is expected to improve the development of tourism. Influencing factors were found to vary among the different clusters.

Key Words: Inbound Tourism; K-means; Negative Binomial Regression; Count Data; Overdispersion

1. INTRODUCTION

Tourism has become an indispensable part of the world economy, especially for Southeast Asia, which has developed rapidly as a tourist destination. At present, tourism has become the most extensive and robust industry, playing a pivotal role, inseparable from people's lives, and having a more significant impact on economic development. It has become one of the fastest-growing industries globally and has entered the era of mass consumption (Hitchcock et al., 2018).

At this stage, global tourism demand continues to increase, and the status of the Asian region continues to rise. Because of this incredible trend of tourist arrivals, Philippine tourism has become one of the primary national incomes for increasing the national economy. Since the tourism reform in 2009, the Philippine tourism industry has developed rapidly. The Philippines hosts more and more tourists from all over the world, which gains vital benefits for the country, such as improving direct and indirect employment and increasing the country's whole economy. In 2017, the tourism sector directly contributed 13 percent of the employment in the country, which means 5.3 million jobs for Filipinos (Vera, 2019).

Inbound tourism is conducive to promoting employment and increasing foreign exchange income. At the same time, it can enhance Philippine popularity and national influence. In 2017, the income of Philippine international tourism (foreign exchange) reached 1.929 trillion Philippine pesos. It is 12.2% of the Philippine Gross Domestic Product (GDP), ranking six globally. The income of 1.929 trillion

Philippine pesos is a dramatic increase. It is too much higher than the income in 2009 (466 billion Philippine pesos). In 2009, the tourism sector only contributed 5.8% to the Philippine GDP. This incredible growth of the Philippine tourism industry has happened since 2009. In 2018, around seven million international visitors came to the Philippines, compared to three million arrivals from other countries in 2009. The foreign exchange income of international tourism rises year by year, and the growth rate is increasing (The Philippine Star, 2019).

However, due to the covid-19 pandemic, the Philippine tourism industry is affected negatively. Travel of foreign nationals to the Philippines is banned since March 2020. There are still some inbound and outbound travel restrictions like cutting more flights and providing proof of vaccination. For the national economy, recovering tourism is necessary in the Philippines after the pandemic. The Philippine government has been doing many national policies and projects to protect the environment and attract more tourists from all over the world in the future, like closing the island of Boracay area for six months because of the local environment problem. Some top corporations are also doing projects to boost tourism, like building new international airports in the Philippines to host more visitors. Tourism is so important to the Philippines that its potential factors are interesting to explore. International tourist arrival is one of the critical indicators of international tourism. Facing the increasing complexity of the inbound tourism market and increasing uncertainty, it has a practical meaning to timely summarize the influencing factors. It provides a basis for the Philippine tourism sector to formulate policies for relevant tourism companies to make decisions.

Two main statistical methods were used for this study, cluster analysis (CA) and negative binomial regression (NBR). Mishra and Bansal (2017) have clustered inbound tourists based on travel time, travel distance, while clustering of tourist arrivals by purpose of travel has not been studied.

This paper explores how influencing factors affect the number of international tourists. In addition, does the impact of these factors differ based on different classifications? For the different clusters, influencing factors will be explored. Specifically, the particular objectives of the paper include the following: (1) To cluster the countries of the tourists' residence by purpose of travel to the Philippines; (2) To estimate NBR for the whole observations and to identify the determinants of foreign tourist arrivals in the Philippine setting; and (3) To build an NBR for each of the resulting clusters of countries, and assess whether the different purposes of travel within each cluster affect the relationship between international tourist arrivals and the selected factors.

Tourism is an indispensable part of the Philippine economy, which has been pursued since the 1970s. It is the backbone of the economy in the Philippines (Maguigad, 2013). In an empirical study of international inbound tourists, Mishra and Bansal (2017) clustered tourists according to the adjacency from the source region to the destination region. They also pointed out the direction for future research. Future studies can be conducted by subdividing international tourists according to the purpose of visit. Some tourism studies used the K-means technique to cluster data, while few studies segment foreign tourists according to the purpose of visit. This classification standard can provide valuable insights into the arrival of foreign tourists.

Count data is obtained by grouping observation units according to a specific attribute or category and separately summarizing each group's number of observation units. The number of tourist arrivals in the tourism data is a discrete random variable with a non-negative integer, which belongs to count data. The NBR model, an extension of the Poison regression model (PRM), can accommodate larger variations than the PRM. NBR model is more suitable for analyzing overdispersed data, so this paper mainly applies it in inbound tourism (Zhang, 2017).

2. METHODOLOGY

2.1 Data

This paper selects secondary data of tourists from 58 countries and regions the residents of which entered the Philippines for tourism from 2009 to 2018 annually. The data mainly came from the Philippine Statistical Yearbook (PSY) under the official website (https://psa.gov.ph) of the Philippine Statistic Authority. Some of the influencing factors came from World Bank Open Data (https://data.worldbank.org), and the information of the foreign trips of the presidents came from the Official Gazette (https://www.officialgazette.gov.ph).

The total number of tourists from each country was initially divided into eleven categories: holiday, business, official mission, convention, visiting friends and relatives, incentive, health medical, education, shop, others, and not reported. Since the categories of others and not reported are not necessary to interpret, and few people are recorded in the shop category, these three categories were removed.

Based on the different distributions of the proportion of tourists visiting the Philippines from different countries and regions, the 58 countries and regions are classified so that the countries from the same category have a similar distribution of the proportion of tourists visiting the Philippines. In contrast, the distributions of the proportion of tourists visiting the Philippines from different categories should not be similar. Since the dataset under PSY record the number of tourists, and this research was based on the similarity of the tourist percentage from different countries and regions, the number of tourists needs to be converted into a ratio. The total number of tourists for each travel purpose from 2009 to 2018 was divided by the total number of tourists. After cleaning up the air visitor arrival dataset divided by country of residence and travel purpose from 2009 to 2018, 58 observations (the different countries or regions) and eight valid variables (purpose of travel) were obtained. The eight variables in the K-means clustering are Holiday, Business, OfficialMission, Convention, VisitingFriendsandRelatives, Incentive, HealthMedical, and Education.

The dependent variable of the NBR is the annual number of inbound tourists from each country or region to the Philippines, which is the most commonly used indicator of inbound tourism demand. The dataset of the NBR was continued after the K-means clustering. This paper analyzed each cluster separately. Population, GDP, GDP per capita, Labor force, Labor productivity, and whether the president of the Philippines visited the country from 2009 to 2018 were used as factors that promote the annual number of inbound tourists from each country to the Philippines. This paper pre-selected the variables used for the NBR based on the qualitative analysis of the influencing factors of inbound tourism in the Philippines, combined with the relevant foundations of consumption and demand theory.

2.2 Statistical Methods

2.2.1 DF test and KPSS test

For judging whether a time series has a trend stationarity, researchers can observe the time series plot, but this is subjective. For the formal tests, researchers may use the unit root test or stationarity test. The commonly used unit root test methods are the augmented Dickey-Fuller (ADF) test and the Phillips-Perron (PP) test, but these two tests are inferior in the case of small sample sequences. In this situation, the Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test (H_0 of KPSS test: the series is trend stationarity vs H_1: presence of a unit root), performed together with the Dickey-Fuller (DF) unit root test (H_0 of DF test: the presence of a unit root vs H_1: trend stationarity). The R package used here are called "fUnitRoots" (Wuertz et al., 2017) and "tseries" (Trapletti and Hornik, 2019), where the only input required was the vector of the time series observations for each travel purpose.

2.2.2 K-means Algorithm

KCA was carried out using a procedure formulated by Hartigan and Wong, and the k-value were determined by a model-based clustering based on parameterized finite Gaussian mixture models. After obtaining the best-fitting clusters for the K-means, then the researchers compared the resulting clusters to understand their reasonable degree of explanation for practical applications.

The R in-built function used here for KCA is called "kmeans" where the input required was the matrix of the 58 countries and 8 variables. The random starting is a parameter in the function. The R package used here for determining the k-value is called "mclust" (Scrucca et al., 2016) where the input required was same as the function "kmeans". The optimal k-value with the largest Bayesian information criterion (BIC) was chosen.

2.2.3 Negative Binomial Regression

The NBR model is used when exploring the factors affecting the inbound tourism. Specifically, the NBR model used in this paper is

$$\hat{y} = exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_6 x_6),$$

where $\widehat{y}$ is the expected number of tourist arrivals, β_0 is a constant, x_1 is GDP, x_2 is GDPpercapita, x_3 is Population, x_4 is Laborforce, x_5 is Labourproductivity, and x_6 is VisitedYes, with coefficient β_1, β_2, β_3, β_4, β_5, and β_6, respectively. x_6 is a dummy variable that indicates whether a country was visited by the president of the Philippines at least once during the period from 2009 to 2018, which is 1 for Yes and 0 for No.

The R package used here is called "MASS" (Venables and Ripley, 2017), where the input required was the vector of the dependent variable and each independent variable. The overdispersion test was also conducted using the R package AER (Kleiber and Zeileis, 2008).

3. RESULTS AND DISCUSSION

3.1 The Pattern of the Number of Tourists

An analysis of the temporal pattern of the different tourism arrival was performed before clustering, which is to give additional information on whether the pattern of the number of tourist arrivals is constant or has a trend for each of the different tourism arrival purposes from 2009 to 2018.

This research did unit root and stationary tests for the time series of each vacation purposes. The results for trend stationary are shown in Table 1 and Table 2.

Table 1. The result of trend stationary KPSS test

Travel Purpose	P-value	KPSS Trend
Holiday	>0.1000	0.1075
Business	0.0871	0.1259
OfficialMission	>0.1000	0.0943
Convention	0.0774	0.1311
VisitingFriends andRelatives	0.0896	0.1245
Incentive	0.0273*	0.1732
HealthMedical	>0.1000	0.1065
Education	>0.1000	0.0958

*significant at 0.05

Table 2. The result of trend stationary DF test

Travel Purpose	P-value	Dickey-Fuller
Holiday	0.0100*	-7.4988
Business	0.4111	1.7511
OfficialMission	0.3291	-2.6387
Convention	0.9768	-0.4572
VisitingFriends andRelatives	0.7498	-1.5343
Incentive	0.6149	-1.8884
HealthMedical	0.9153	-1.0361
Education	0.2865	2.7505

*significant at 0.05

Since most of the p-values for both trend stationary tests are greater than 0.05, we cannot reject the null hypotheses in both tests, which means that the dataset does not have enough observations for most of the travel purposes. While for the holiday purpose of travel, the null hypothesis of the KPSS test is not rejected (p > 0.1000), but the null hypothesis of DF test is rejected (p = 0.011). These results imply that the series is trend stationary for the holiday purpose of travel.

Among the 8 travel purposes, only the pattern of the number of tourist arrivals for holiday purpose has a trend stationarity. The number of tourists for holiday purpose is expected to improve the development of tourism.

3.2 Classification of the Philippine Inbound Tourism Based on K-means

According to the method based on parameterized finite Gaussian mixture models (BIC=2966.047), the optimal k-value is 4. In this paper 1000 random sets of initial center points were chosen for the KCA. The largest Dunn index for the optimal clustering result is around 0.12. The mean percentages from the K-means clustering by travel purpose within each cluster is shown in the Table 3.

Table 3. The mean percentages of travel purpose by cluster from the K-means clustering

Cluster	1	2	3	4
Holiday	0.2461	0.3431	0.5077	0.6156
Business	0.0605	0.1806	0.0915	0.0686
OfficialMission	0.0372	0.0098	0.0027	0.0026
Convention	0.1040	0.0439	0.0156	0.0102
VisitingFriends andRelatives	0.0584	0.0565	0.1050	0.0537
Incentive	0.0002	0.0003	0.0003	0.0003
HealthMedical	0.0012	0.0033	0.0053	0.0008
Education	0.1104	0.0310	0.0062	0.0060

Cluster 1, from the K-means clustering, is composed of Cambodia, Laos, Myanmar, Bangladesh, Iran, Nepal, and Nigeria. Cluster 2 is composed of India, Indonesia, Papua New Guinea, Singapore, Sri Lanka, Thailand, Vietnam, Egypt, Greece, Jordan, and Pakistan. Cluster 3 is composed of Japan, Malaysia, Mexico, Netherlands, Peru, Bahrain, Brazil, Colombia, Germany, Hong Kong, Ireland, Italy, Kuwait, South Africa, United Arab Emirates, Austria, Belgium, Norway, United Kingdom, Australia, Canada, New Zealand, USA, and Guam. Cluster 4 is composed of Israel, Spain, Luxembourg, Brunei, China, France, Korea, Russian Federation, Sweden, Argentina, Denmark, Finland, Poland, Portugal, Saudi Arabia, and Switzerland.

From Table 3, the mean percentage of tourist arrivals for educational purpose in cluster 1 is much higher than those in the other clusters. Similarly, the average percentages in cluster 2, cluster 3, and cluster 4 for business, visiting friends and relatives, and holiday purposes, respectively, are much higher than in the other clusters. The main purpose of visiting the Philippine for all the four clusters is holiday, which has the highest percentage in all these four clusters.

3.3 The Results of Negative Binomial Regression

Shown in the Table 4 is the summary of the

NBR for all countries. At a significance level of 0.05, GDP, GDPpercapita, and VisitedYes are positively significant predictors for the number of tourist arrivals. If a country were to increase its GDP by one unit, the logarithm of the number of tourists would be expected to increase by 2.67E-13 unit while holding the other variables in the model constant. For VisitedYes, this is the estimated NBR coefficient comparing Yes to No, given that the other variables in the model are held constant. The logarithm of the number of tourists would be expected to be 1.87 units higher for Yes compared to No, while holding the other variables in the model constant. Similarly, the other variables can be interpreted in the same way.

Table 4. Exploring the factors for all countries

Coefficients	Estimate	z value	Pr(>|z|)
(Intercept)	8.29E+00	87.71	<2e-16*
GDP	2.67E-13	13.09	<2e-16*
GDPpercapita	3.48E-05	16.36	<2e-16*
VisitedYes	1.87E+00	17.03	<2e-16*

*significant at 0.05

From the Table 5, no factors are significant for the number of tourist arrivals. From the table 6, GDP, GDPpercapita, and VisitedYes are significant predictors for the number of tourist arrivals in cluster 2. From the Table 7, Population, Labourproductivity, and VisitedYes are significant predictors for the number of tourist arrivals in cluster 3. From the Table 8, GDP and VisitedYes are significant predictors for the number of tourist arrivals in cluster 4.

Table 5. Exploring the factors for cluster 1

Coefficients	Estimate	z value	Pr(>|z|)
(Intercept)	7.74E+00	86.449	<2e-16*
GDPpercapita	6.40E-05	1.954	0.0506

*significant at 0.05

Table 6. Exploring the factors for cluster 2

Coefficients	Estimate	z value	Pr(>|z|)
(Intercept)	7.19E+00	63.993	2.00E-16*
GDP	9.38E-13	8.043	8.78E-16*
GDPpercapita	3.75E-05	9.191	2.00E-16*
VisitedYes	2.26E+00	16.315	2.00E-16*

*significant at 0.05

Table 7. Exploring the factors for cluster 3

Coefficients	Estimate	z value	Pr(>|z|)
(Intercept)	8.64E+00	47.992	2.00E-16*
Population	4.68E-09	4.294	1.76E-05*
Labour productivity	1.98E-05	10.063	2.00E-16*
VisitedYes	1.34E+00	8.128	4.37E-16*

*significant at 0.05

Table 8. Exploring the factors for cluster 4

Coefficients	Estimate	z value	Pr(>|z|)
(Intercept)	9.36E+00	73.06	2.00E-16*
GDP	1.34E-13	2.81	0.00495*
VisitedYes	2.80E+00	12.15	2.00E-16*

*significant at 0.05

Now the results of the K-means clustering and the NBR can be combined. For example, for tourists from countries in cluster 1 the main reason for visiting the Philippines is for holiday purpose, the education purpose has a much higher mean percentage than in the other clusters, but no factors are significant predictors for the number of tourist arrivals under a significance level of 0.05. The president's visit has a significant impact ($p < 0.05$) on the expected tourist arrivals in cluster 2, cluster 3, and cluster 4.

The results of this paper can also be compared with other researches. In the paper "The impact of government failure on tourism in the Philippines" (Manuela et al., 2015), the researchers showed that the US FAA downgrade and the EU ban impact tourist receipts negatively. This finding also indicates that political factor has an impact on tourism. In the paper "Role of source-destination proximity in international inbound tourist arrival: empirical evidences from India" (Mishra and Bansal, 2017), the researchers classified the source countries based on the air travel duration to the destination and found that the influencing factors are also different in each classification.

4. CONCLUSIONS

This paper mainly explored the inbound tourism of the Philippines from two aspects. The first one is to sort out the research content of international inbound tourism and the inbound tourism of the Philippines. The second one is to sort out the K-means algorithm and the negative binomial regression model, then combine these two methods to explore the influencing factors of inbound tourism in the Philippines. By investigating the influencing factors of each cluster separately, the following conclusions are drawn.

The number of tourists from countries in cluster 1 (the countries visiting the Philippines with relatively high mean percentage due to educational travel purpose) is not significantly affected by any of the factors considered in this paper. The probable reason could be that the countries in cluster 1 have relatively poor economies, so the general economic indexes are invalid.

Policy-makers should focus on the tourists for holiday purpose to improve the development of tourism, because the tourist arrivals for this reason have an increasing trend. The Department of Tourism may suggest that the president visit more countries because the presidential visits to other countries have a positive impact on the expected tourism arrivals.

Due to time and data limitations, there are still areas for improvement for similar studies. Firstly, six potential factors were considered in this paper, including economic, population, and political factors. But researchers may consider how to introduce more independent variables in future research. Secondly, in the analysis of the temporal pattern, the test results show that the available observations are not enough for most of the travel purposes. Therefore, a larger sample size should be used in future studies. Finally, this study is based on K-means clustering and NBR models, other clustering methods and regression methods may be considered in future studies.

5. ACKNOWLEDGMENTS

The researchers would like to thank Dr. Rechel G. Arcilla, Dr. Jayne Lois San Juan, and Ms. Shirlee R. Ocampo for their suggestions and the Philippine Statistic Authority, the World Bank Open Data, and the Official Gazette for providing necessary data in this paper.

6. REFERENCES

Hitchcock, M., King, V. T., & Parnwell, M. J. (2018). Tourism in South-East Asia: Introduction (pp. 1-31). Routledge.

Kleiber, C., & Zeileis, A. (2008). Applied Econometrics with {R} ({ISBN} 978-0-387-77316-2) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=AER

Maguigad, V. M. (2013). Tourism planning in archipelagic Philippines: A case review. Tourism Management Perspectives, 7, 25-33.

Manuela Jr, W. S., & de Vera, M. J. (2015). The impact of government failure on tourism in the Philippines. Transport Policy, 43, 11-22.

Mishra, S. S., & Bansal, V. (2017). Role of source-destination proximity in international inbound tourist arrival: empirical evidences from India. Asia Pacific Journal of Tourism Research, 22(5), 540-553.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). {mclust} 5: clustering, classification and density estimation using Gaussian finite mixture models [Computer software]. The Comprehensive R Archive Network. Available from https://doi.org/10.32614/RJ-2016-021

The Philippine Star. (2019). 'More Fun For Everyone': A Philippine tourism industry grown by Filipinos for Filipinos. Retrieved from https://www.philstar.com/lifestyle/travel-and-tourism/2019/05/18/1918682/more-fun-everyone-philippine-tourism-industry-grown-filipinos-filipinos

Trapletti, A., & Hornik, K. (2019). tseries: Time series analysis and computational finance (R package version 0.10-47.) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=tseries

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (ISBN 0-387-95457-0) [Computer software]. The Comprehensive R Archive Network. Available from http://www.stats.ox.ac.uk/pub/MASS4

Vera, A. (2019). Tourism an economic growth driver in ASEAN - DOT. Retrieved from https://news.mb.com.ph/2019/05/02/tourism-an-economic-growth-driver-in-asean-dot

Wuertz, D., Setz, T., & Chalabi, Y. (2017). fUnitRoots: Rmetrics - modelling trends and unit roots (R package version 3042.79) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=fUnitRoots

Zhang, W. (2017). 基于改进负二项回归模型的高速公路交通事故起数预测方法研究 [Master dissertation, Changan University]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017869767.nh