Tourism Market Segmentation using Big Data Approach: Where is the Next Non-Stop Destination?

So Young Park
The Pennsylvania State University, sup97@psu.edu

Bing Pan
The Pennsylvania State University, bingpan@psu.edu

Tourism Market Segmentation using Big Data Approach: Where is the Next Non-Stop Destination?

Introduction

Attracting more visitors and managing them effectively are one of major responsibilities of destinations' Destination Marketing Organizations (DMOs). A series of literature on attracting more visitors have used the segmentation method to identify the potential customers and pinpoint the target markets using various criteria (Andereck & Caldwell, 1994; Dolnicar, 2007; Jang, Morrison, & O'Leary, 2002; Müller, & Hamm, 2014; Smith, 1956; Tkaczynski, Rundle-Thiele, & Beaumont, 2009). However, as every person has unique characteristics and interests, categorizing the customers using some demographic or behavioral variables can be a coarse method.

This paper introduces multiple gravity models based on both travelers' interests and actual visitation; it tries to find the regions that are most likely to have a significant increase in visitor volume to the destination if non-stop flight route were to open. The gravity model includes direct flight availability as its key component. The model specifically includes the availability of direct flight as one of the deciding factors for choosing a destination (Tveteras & Roll, 2014), related studies are lacking.

The interest is measured using the web traffic at the destination's Convention and Visitors Bureau (CVB) website. Although interest in the destination at an individual level is hard to observe, interest at a regional level can be quite accurately estimated with the emergence of big data. The web traffic at a CVB can be considered as a proxy for potential interests, as tourists are most likely to search information online prior to visiting a city and a CVB's website are more likely to turn up on the top as the authoritative information source. Hence, it is possible to target markets that have keen interest in the destination and increase the number of visitors via creating direct flight routes.

The paper provides a theoretical framework that based gravity models, the push-pull model of Crompton (1979), and the purchase funnel theory. The theoretical framework is able to calculate each region's magnitude of potential to travel, explain why some regions that have high potential are not travelling as much as expected, and find the most prominent potential markets for the next non-stop route. The first part is done using the gravity model and the latter parts are done using web traffic and direct flight data.

Widely accepted push-pull model of Crompton (1979) provides a theoretical background for the application of the gravity model. Push factors cause the tourists to leave home and pull factors cause them to a choose certain destination. In the gravity model, the push and pull factors for both origin and destination are included. Push and pull factors are the determinants of power in the gravity models: average income, represented by GDP and population, distance, and direct flight.

In marketing, purchase funnel theory explains customers' decision process (Clow, 2013; Jerath, Ma, & Park, 2014), The purchase process consists of four steps: awareness, familiarity, consideration, and purchase. In familiarity and consideration steps, prospective visitors may search online for the destinations with desired features. Web traffic at the CVB's of such destinations represents the behavior of customers at the consideration stage, and can be considered as a proxy for the interest in the area.

At the same time, the availability of direct flight would also come into the decision making process between familiarity and purchase. People tend to pursue faster and more comfortable way to travel (Gronau, 1970) and direct flight offers such option. Hence, some of the destinations could be included or excluded in the destination selection process due to the availability of direct flight.

With this theoretical framework, the paper empirically locates the cities that have high interest in one destination and have great potential to travel but are not visiting it as much as they would if there were direct flights. This study identifies top five potential markets for next non-stop air route.

Literature Review

1. Gravity Models in Social Sciences

Originated from Newton's law of universal gravitation in physics, various disciplines of social science have adopted gravity models. The model has also emerged as one of the popular models to explain the tourism phenomena empirically (Balli, Balli, & Louis, 2016; Durbarry, 2008; Eryigit, Kotil, & Eryigit, 2010; Grosche, Rothlauf, & Heinzl, 2007; Kaplan & Aktas, 2016; Khadaroo & Seetanah, 2008; Morley et al., 2014).

Eryigit et al. (2010) incorporates environmental and financial factors into the classical gravity model to explain the tourist arrivals to Turkey. Beside distance, GDP, and population, the paper included tourism climate index, tourism price index, earthquake, shared border, Iraq War and September 11 terrorist attacks as variables. Employing the gravity model, Balli et al. (2016) also analyzes the bilateral tourism between 34 OECD countries using panel data.

Khadaroo and Seetanah (2008) also adjusts the gravity equation to account for the role of transport infrastructure in the inbound tourists. They added price index, tourism infrastructure, common language, common border, and number of alternative destinations that are in proximity to the gravity framework. Dubarry (2008) includes real effective price of tourism along with language and European dummies into the traditional gravity model to understand the tourism inflow to the United Kingdom. Similarly, Hanafia and Harun (2010) includes consumer price index as a consideration of price sensitivity and also considers economic crisis as possible factor that could affect Malaysia's tourism industry. However, no empirical models have yet included direct flight and forecasted the flight demand.

Morley et al. (2014) provided a theoretical framework for using the gravity model to explain the number of tourism arrivals. The authors ground their theoretical models on consumer theory, assuming that the traveler tries to maximize his utility given their budget. The framework applies the gravity model to explaining tourism inflow and outflow, but neglects the substitution effect between modes of travel, which is a key factor affecting an individual's destination decision making.

2. Market Segmentation

Smith (1956) introduced the market segmentation strategy to tackle the lack of homogeneity among consumer demand. Much tourism literature has essayed to dissect the heterogeneous target markets into few homogeneous segments (Andereck & Caldwell, 1994; Beane, & Ennis, 1987; Dolnicar, 2003; Müller, & Hamm, 2014; Smith, 1956; Tkaczynski, Rundle-Thiele, & Beaumont, 2009).

Most segmentation studies focus on consumers' socio-demographics and behavioral characteristics (Tkaczynski, Rundle-Thiele, & Beaumont, 2009) and are based on visitor surveys. Popular segmentation factors also include geographics, psychographics, motivations, behavioral characteristics, and spending patterns (Cai, Liping, & Mimi Li, 2009; Cha, McCleary, & Uysal, 1995; Kivela, 2017; Woodside, Moore, & Etzel, 1980).

Woodside, Moore, and Etzel (1980) divided the travelers into in state travelers and out of state travelers and Cai, Liping, and Mimi Li (2009) used distance for segmentation. As in the case of motivation, Cha, McCleary, and Uysal (1995) looked at the Japanese overseas travelers and categorized them into sport seekers, novelty seekers, and family/relaxation seekers; Kivela (2017) used gastronomy as the motivation to travel for segmentation; Almeida, Correia, & Pimpão, (2014) also clustered rural tourists into four groups according to motivation, i.e., benefits sought.

In case of behavioral characteristics, Johns and Gyimóthy (2002) takes the case of Bornholm, Denmark and dissects the tourist into active and inactive vacationers; Mudambi, and Baum (1997) focused on nationality to assess expenditure patterns of tourists visiting Turkey; Lehto, O'Leary, and Morrison (2002) analyzes British overseas tourists according to travel philosophies, travel benefits sought, and destination attribute preferences.

However, it was difficult to locate existing literature that used transportation method or focused on regions rather than individuals to segment the target market. Effectiveness of consumer characteristic factors in segmenting the market have been controversial over time and among researchers. Individual characteristics can be quite subjective and are mostly collected via survey, which can be biased and limited. In contrast, transportation measures and region's characteristics such GDP and population are comparatively objective and do not require surveys. Hence, this paper has its value in that it essayed to investigate the tourism markets using the travelling modes and regional socioeconomic characteristics as the cluster factors.

Theoretical Framework

Due to availability of data, this study adopts Charleston, South Carolina, USA as the example. The theoretical framework is demonstrated as the following figures (Figure 1 and 2). The red circle in the middle represents the City of Charleston and grey dots in the orbit are tourist origin Metropolitan Statistical Areas (MSAs) (Figure 1). The size of the grey dots is determined by the gravity model and represents the amount of travel demand. If the MSA has greater GDP and population, is closer to Charleston and has a non-stop route, the residents in the MSA are considered to have higher possibility to travel to Charleston. The black orbits indicate the barriers that exist in reaching Charleston from the MSA's. Each barrier decreases the number of flight passengers from the MSAs and some are completely obstructed even though they have great potential to travel. The barriers can be explained using the purchase funnel theory (Figure 2).

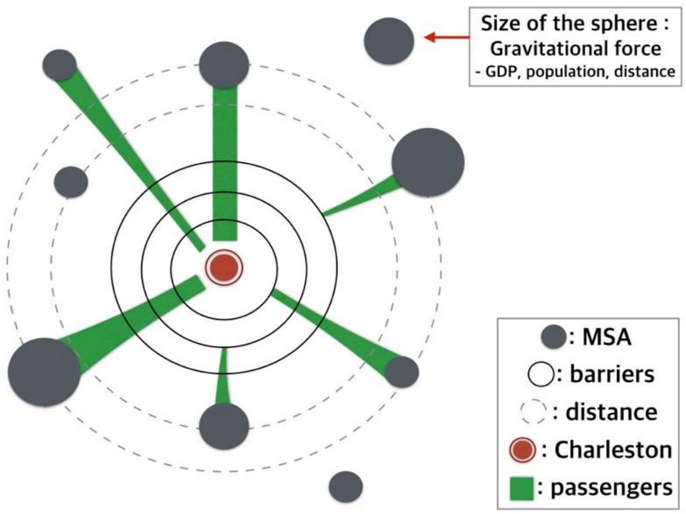

Figure 1. The Origin Markets and the Destination City

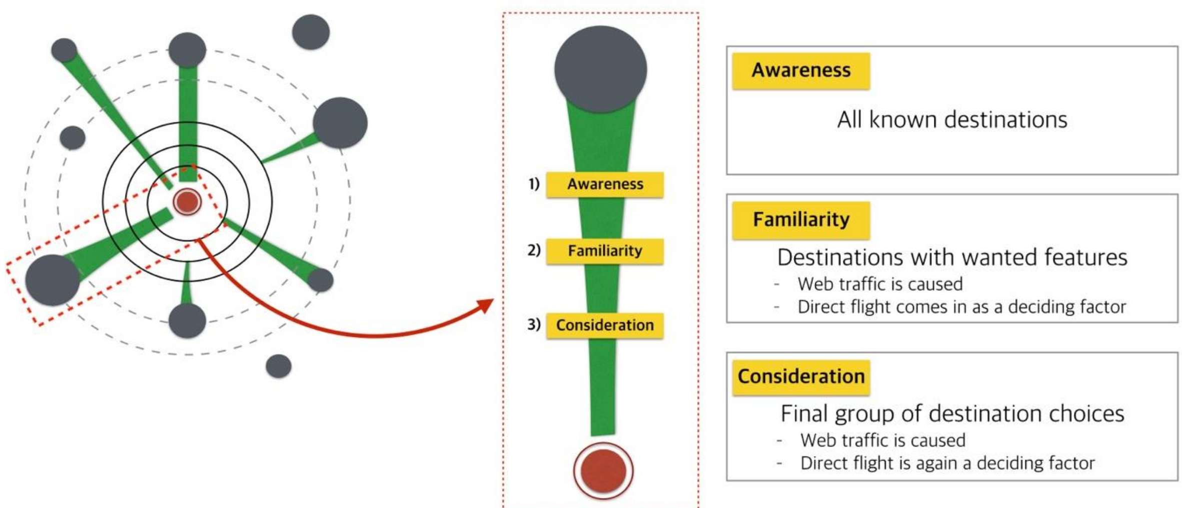

Figure 2. Purchase Funnels and Gravity Models

Purchase funnel model can connect the aggregated gravity model with individual traveler's decision making. Purchase funnel theory basically has four steps: 1. awareness, 2. familiarity, 3. consideration, and 4. purchase (Figure 2) (Clow, 2013; Jerath et. al, 2014).

In the tourism context, first, the consumer is assumed to be aware of myriad of possible destinations with similar traits or attractions; second, in the familiarity step, the consumer has decided the kind of place he/she wants to visit and searches for similar destinations using internet and asking friends, i.e., word-of-mouth. At the second and third stage, the web traffic at CVBs may occur as the consumer searches for the features and activities-to-do at certain destinations. Moreover, the consumer is likely to decide the maximum distance to travel due to time and budget constraint and to encounter the transportation options. Here, existence of direct flight can function as one of the positive factors to keep the destination in one's choice set since it decreases cost coming from the time of travel, uncertainty and discomfort (Nicolau, Màs, 2005).

Individual traveler's decision process (Figure 2) impacted the amount of total visitors to the destination due to those barriers of travel (Figure 1).

Methodology

Based on the proposed theoretical framework, this study fits gravity models with a series of regressions. Though log-normal model is often used, it tends to have serious disadvantages due to its strong assumptions (Flowerdew and Aitkin, 1982; Silva and Tenreyro, 2006). Tourism researchers have incorporated Poisson measure to analyze the tourism demand and travel costs, often stating the method as Count model (Creel & Loomis, 1990; Feather et al., 1994; Hellerstein, 1999; Parsons, 2003; Hellström, 2006). For example, Hellström (2006) employed bivariate Poisson log-normal model to efficiently and accurately look at the household tourism demand. Thus, this study employs generalized linear model (GLM) with link function of the Poisson distribution to analyze Charleston's data. The assumption is that the dependent variable follows the Poisson probability distribution. The estimation is carried out through maximum-likelihood method.

Expected value of flight passengers, $E[V_{ijt}|X]$ follows the Poisson probability distribution and can explained by the matrix of independent variables X which includes population, income, distance, direct flight, and relative price between the transportation modes:

$$\ln(E[V_{ijt}|X]) = \beta_1 P_{it} + \beta_2 P_{jt} + \beta_3 GDP_{it} + \beta_4 GDP_{jt} + \beta_5 D_{ij} + \beta_6 DF_{ijt} + \beta_7 S_{ijt} + \beta_8 D_{ij} * S_{ijt} + \epsilon_{ijt}$$

Where V_{ijt} represents the number of visits from the origin *i* to the destination *j* at time *t*, P_{it} is the population of the origin *i* at time *t*, and P_{jt} is the population of the destination *j* at time *t*. GDP_{it} and GDP_{jt} represent income level of the origin *i* and the destination *j* at time *t* respectively. D_{ij} is the physical distance between the origin *i* to the destination *j*. DF_{ijt} is the dummy variable for the availability of direct flight and S_{ijt} is the relative price of the flying and driving between the origin *i* to the destination *j* at time *t*. Interaction term between D_{ij} and S_{ijt} is included since distance and transportation cost can have high correlation. The study estimates the above model using pseudo-maximum-likelihood method.

The study then compares the fitted values and the actual data to search for the most potential markets where there still is no direct flight. Those with the fitted passenger numbers much greater than the actual ones are considered as the MSA's with greater possibility to increase the flight passengers if there were a non-stop route. Of those with larger gaps, the paper sorts them according to the size of web traffic. Accordingly, MSA's with great potential to travel and with highest interest are found.

Results

With 3 year averaged flight passenger data to and from Charleston between 2012 and 2014, the paper was able to find top five future non-stop flight routes that are expected to gain most passengers by establishing new direct flight routes. Table 1 shows the GLM regression results for the passenger data and the second table is the list of top five MSA's.

Table 1. Regression result of Gravity Model of the Number of Passengers

Variable	**Flight passenger**				
	(1)	(2)	(3)	(4)	(5)
Population	0.0002***	0.0001***	0.0001***	0.0001***	0.0001***
Income	0.0766***	0.0733***	0.0672***	0.0706***	0.0620***
Average flight Distance	-0.0009***	-0.0002***	-0.0002***	-0.0007***	0.0003***
Relative cost		-2.9380***		-5.0074***	-1.6507***
Distance* Relative **cost**		-0.0348***			-0.0303***
Direct Flight			1.6426***	1.4708***	1.3215***
Constant	6.0066***	8.8158***	5.7265***	6.6274***	8.2409***
N	216	216	216	216	216
Pseudo R^2	0.7039	0.7855	0.7911	0.8080	0.8377

legend: * p<.1; ** p<.05; *** p<.01

The Model (1) in Table 1 shows the result of the most basic gravity model, which includes only population, income (GDP) and distance. As population and income were divided by 1,000 to adjust the scale, the coefficients indicate that for every 1000 units of increase in population or income, the number of flight passengers is expected to increase by 0.02% and 7.66% respectively. Distance is negatively correlated with the number of passengers with significance as expected.

Model (2) includes the relative cost and its interaction term with the distance. Relative cost was computed by dividing average flight ticket cost by average driving distance times 2.184. $2.184 is the regular gasoline prices (dollars per gallon) as of November 14, 2016 (U.S. Energy Information Administration). When the relative cost is high, the flight cost is higher compared to the driving cost, and decreases the number of flight passengers significantly.

Model (3) to (5) gradually increases the number of variables included in the regression to show the robustness of the model by including the existence of non-stop flight dummy. The R-squared values increase with the inclusion of the direct flight variable. All other variables' coefficients stay the same except for the distance in Model (5) due to the interaction term. In Model (5), the existence of direct flight increases the number of passengers by over 130%.

Table 2. Top Five Potential MSAs for Direct Flights

Rank	**Name**
1	Los Angeles-Riverside-Orange County, CA
2	Santa Barbara-Santa Maria-Lompoc, CA
3	Albany-Schenectady-Troy, NY
4	Savannah, GA
5	Burlington, VT

Conclusion and Discussion

In summary, this paper has developed a theoretical framework that encompasses the purchase funnel theory and the gravity model, which helped identifying the prominent markets for the next direct flight. In this way, direct flight information and web traffic data were included in the analysis and it was possible to increase the explaining power of visitor volumes and the accuracy of forecasting the future flight passengers. With the flight passenger data and web traffic flow between Charleston and 216 MSA's, this paper empirically showed the impact of direct flight route availability and found the potential cities from which number of flight passengers can be increased if there were direct flights.

From the regression results, the paper was able to validate that the availability of non-stop route clearly impacts the number of visitors via air. However, there is also a high possibility that greater number of passengers could have been a trigger for the creation of non-stop flight route. Further study is needed to prove the causality relationship from the direct flight to the number of passengers.

References

Almeida, A. M. M., Correia, A., & Pimpão, A. (2014). Segmentation by benefits sought: the case of rural tourism in Madeira. *Current Issues in Tourism*, *17*(9), 813-831.

Anderson, J. E., & Kraus, M. (1981). Quality of service and the demand for air travel. The Review of Economics and Statistics, 533-540.

Andereck, K. L., & Caldwell, L. L. (1994). Variable selection in tourism market segmentation models. Journal of Travel Research, 33(2), 40-46.

Antrás, P., & Rossi-Hansberg, E. (2008). Organizations and trade (No. w14262). National Bureau of Economic Research.

Beane, T. P., & Ennis, D. M. (1987). Market segmentation: a review. European Journal of Marketing, 21(5), 20-42.

Balli, F., Balli, H. O., & Louis, R. J. (2016). The impacts of immigrants and institutions on bilateral tourism flows. Tourism Management, 52, 221-229.

Cai, Liping A., and Mimi Li. "Distance- segmented rural tourists." Journal of Travel & Tourism Marketing 26, no. 8 (2009): 751-761.

Cha, S., McCleary, K. W., & Uysal, M. (1995). Travel motivations of Japanese overseas travelers: A factor-cluster segmentation approach. Journal of travel research, 34(1), 33-39.

Clow, D. (2013, April). MOOCs and the funnel of participation. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (pp. 185-189). ACM.

Creel, M. D., & Loomis, J. B. (1990). Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. American journal of agricultural economics, 72(2), 434-441.

Crompton, J. L. (1979). Motivations for pleasure vacation. Annals of tourism research, 6(4), 408-424.

Dolnicar, S. (2003). Tracking data-driven market segments. Tourism Analysis, 8(2), 227-232.

Dolnicar, S. (2007). 8 Market Segmentation in Tourism. Tourism Management: Analysis, Behaviour, and Strategy, 129.

Durbarry, R. (2008). Tourism taxes: Implications for tourism demand in the UK. Review of Development Economics, 12(1), 21-36.

Eryigit, M., Kotil, E., & Eryigit, R. (2010). Factors affecting international tourism flows to Turkey: a gravity model approach. Tourism Economics, 16(3), 585-595.

Flowerdew, R., & Aitkin, M. (1982). A method of fitting the gravity model based on the Poisson distribution. Journal of regional science, 22(2), 191-202.

Gasoline and Diesel Fuel Update. (2016, November). Retrieved from U.S. Energy Information Administration website http://www.eia.gov/petroleum/gasdiesel/

Gronau, R. (1970). The value of time in passenger transportation: The demand for air travel. NBER Books.

Grosche, T., Rothlauf, F., & Heinzl, A. (2007). Gravity models for airline passenger volume estimation. Journal of Air Transport Management, 13(4), 175-183.

Hanafiah, M. H. M., & Harun, M. F. M. (2010). Tourism Demand in Malaysia: A cross-sectional pool time-series analysis. International Journal of trade, economics and Finance, 1(1), 80.

Hellerstein, D. M. (1991). Using count data models in travel cost analysis with aggregate data. American journal of agricultural economics, 73(3), 860-866.

Hellström, J. (2006). A bivariate count data model for household tourism demand. Journal of Applied Econometrics, 21(2), 213-226.

Jang, S. C., Morrison, A. M., & T O'Leary, J. (2002). Benefit segmentation of Japanese pleasure travelers to the USA and Canada: selecting target markets based on the profitability and risk of individual market segments. Tourism Management, 23(4), 367-378.

Jerath, K., Ma, L., & Park, Y. H. (2014). Consumer click behavior at a search engine: The role of keyword popularity. Journal of Marketing Research, 51(4), 480-486.

Johns, N., & Gyimóthy, S. (2002). Market segmentation and the prediction of tourist behavior: The case of Bornholm, Denmark. Journal of Travel Research, 40(3), 316-327.

Kaplan, F., & Aktas, A. R. (2016). The Turkey Tourism Demand: A Gravity Model. Available at SSRN 2785797.

Khadaroo, J., & Seetanah, B. (2008). The role of transport infrastructure in international tourism development: A gravity model approach. Tourism management, 29(5), 831-840.

Kivela, J. (2017). Gastronomy Tourism: Croatia, a Land of Wine and Plenty, or Beyond Pizza and Grill!. In Evolution of Destination Planning and Strategy (pp. 265-278).

Lehto, X. Y., O'Leary, J. T., & Morrison, A. M. (2002). Do psychographics influence vacation destination choices? A comparison of British travellers to North America, Asia and Oceania. Journal of Vacation Marketing, 8(2), 109-125.

Mudambi, R., & Baum, T. (1997). Strategic segmentation: an empirical analysis of tourist expenditure in Turkey. Journal of Travel Research, 36(1), 29-34.

Müller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis–A methodological approach. Food Quality and Preference, 34, 70-78.

Newton, I. (1966). Philosophiae naturalis principia mathematica, London 1687. English translation: Sir Isaac Newton?s Mathematical Principles, Univ. of Calif. Press, Berkeley.

Nicolau, J. L., & Màs, F. J. (2006). The influence of distance and prices on the choice of tourist destinations: The moderating role of motivations. Tourism Management, 27(5), 982-996.

Parsons, G. R. (2003). The travel cost model. In A primer on nonmarket valuation (pp. 269-329). Springer Netherlands.

Silva, J. S., & Tenreyro, S. (2006). The log of gravity. The Review of Economics and statistics, 88(4), 641-658.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of marketing, 21(1), 3-8.

Tkaczynski, A., Rundle-Thiele, S., & Beaumont, N. (2009). Destination segmentation: A recommended two-step approach. Journal of Travel Research.

Tveteras, S., & Roll, K. H. (2014). Non-stop flights and tourist arrivals. Tourism Economics, 20(1), 5-20.

Woodside, A. G., Moore, E. M., & Etzel, M. J. (1980). Vacation travel behavior and perceived benefits of home states residents. Business and Economic review, 26(5), 28-34.