BEST: International Journal of Management Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X,
Vol. 6, Issue 3, Mar 2018, 17-26

A CLUSTER ANALYSIS APPROACH TO MARKET SEGMENTATION IN THE AIRLINES INDUSTRY

YUJUAN WU & PAUL D. BERGER

Bentley University, Waltham, Massachusetts, U.S.A.

ABSTRACT

This paper considers a market segmentation of the airlines's industry, using cluster analysis and a set of selected variables, some demographic, some having to do with the responders' flight experiences and some having to do with the responders' airport experiences. Specific airlines/"brands" are not detailed. The data are based on a survey by IBM Watson Analytics, and the sample size exceeds 100,000 customers. We find that we have 6 clusters, each of which has clear distinguishing characteristics. Marketing implications for reaching these clusters are examined.

KEYWORDS: Airlines, Cluster Analysis, Market Segmentation & Survey Data

INTRODUCTION

Consumer preferences have become one of, if not *the* main consideration for marketing departments. Many markets are switching from being a sellers' market to being a buyers' market. For many industries, undifferentiated marketing methods are unable to earn a substantial profit for companies. Often, "one-for-all" marketing strategies are simply unsuccessful in this day and age of targeted marketing. There are situations in which it is impossible, from a practical point-of-view, for companies to target each customer individually; however, there are other situations when one can segment a market down to the individual level. For example, with postal mailing lists or email lists, one can utilize "list segmentation" to profitably do so (e.g., Berger & Magliozzi, 1992) if a sufficient number of variables are available for analysis. Of course, even with the ability to achieve segmentation on an individual level, it is not always cost effective, nor is there always a need for that degree of granulation.

In the airlines's industry, the industry of interest in this paper, it is unlikely that costumers can fruitfully (i.e., profitably) be dealt with on an individual level. There are millions of customers, but there are not corresponding millions of separate marketing strategies. The situation does not "mirror" the situation of a retail store that can draw on individual past-purchase behavior to the same degree, to differentiate among customers on an individual level. Yet, market segmentation can still be a key to success.

This paper uses a cluster analysis approach to market segmentation of customers of the airline industry, using airline-satisfaction survey data. The survey data were provided by IBM Watson Analytics (2014).The study attempts to identify important factors, explain the motivations of customers' choices and eventually give overall business recommendations. We find a set of 6 clusters, with certain variables clearly differentiating the clusters.

LITERATURE REVIEW

Smith (1956) published an article which was the first to introduce the concept of market segmentation. He suggested that the division of a market should be based on customers who shared certain characteristics. Since then,

other authors have built upon Smith's work. Yankelovich (1964) studied industrial market-segmentation based on a practical example. Over 40 years later, Yankelovich and Meer (2006) authored a paper in the same journal on *rediscovering* market segmentation. Reynolds and Jolly (1980) suggested a new view that there are four tests which a market-segmentation should pass—measurability, accessibility, stability, and substantiality. However, there seems to be no complete agreement how to best carry out these tests. Kennedy, Best, and Kahle (1990) proposed a standard process of how to conduct market segmentation. Luo (2003) focused on customer-segmentation analysis based on a very specific product or under a very specific situation. There has also been a lot of literature on how to segment lists of various sources into deciles or individual percentiles, or, indeed, individual consumers (e.g., Lix, Berger & Magiozzi, 1995).

There are many methods that have been proposed and used in the area of market segmentation. The most common forms of consumer market-segmentation are those based on Geographic segmentation, Demographic segmentation, Psychographic segmentation, and Behavioral segmentation. Of course, these "types of segmentation" are simply describing the variables used in making the segmentation. Clearly, many market segmentations use a combination of variables from all of these areas. When the term "market segmentation" is used, however, many immediately think only of psychographics, lifestyles, values, behaviors, and multivariate cluster analysis routines. Market segmentation is a much broader concept, and it pervades the practice of business throughout the world (Thomas, 2016).

This paper uses cluster analysis as the mode of market segmentation. Porter (1998) first defined the term "cluster" as a "Geographic concentration of interconnected companies and institutions in a particular field." Hill and Baumann (2000) utilized complex techniques proposing a method to identify clusters, called "hierarchical cluster analysis" to help people sort this kind of problem statistically. Porter (2003) then employed correlation analysis to identify geographic clusters. However, there are other methods of cluster analysis and no general agreement about which cluster-analysis method is best.

In general, Cluster analysis, or *clustering*, is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is the main task of exploratory data mining and a common technique for statistical data-analysis. It is used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics (Wikipedia, 2017), and, of course, albeit more recently, in the world of business/management, and specifically, used routinely today in market segmentation.

Based on Wikipedia (2017), cluster analysis was originated in anthropology by Driver & Kroeber in 1932 and, in terms of early use, was introduced to psychology by Zubin in 1938 (Bailey, 1994) and Tryon in 1939 (Tryon, 1939) and used by Cattell (1943) beginning in 1943 for trait-theory classification in personality psychology.

METHODOLOGY

Data Source and the Airline Satisfaction-Survey

The airline satisfaction-survey used in this paper, as well as the data set derived therefrom, was provided by Fraser Anderson, through the IBM Watson Analytics (2014) online website accessed in February, 2017. The survey was carried out during the first three months in 2014. The survey collected data about actual airline performance, as well as how responders felt about airlines's services in general. The dataset consisted of demographic variables, attitudes of customers, and airline information. The respondents constituted a random sample of 129,889 passengers. After eliminating data points

with critical missing values, we were left with 127,151 cases. Based on the survey, we used variables of three types to segment the airline market from the consumers' view.

One type of variable we used to be the demographic information; indeed, it would be unusual to see a market segmentation that did not use demographic variables. Different demographic groups clearly may have different criteria that they consider important to them. For example, many studies have shown that, when it comes to airline issues and airport issues, factors important to business travelers are different than those for non-business travelers (e.g., Wang, Hong & Berger, 2016). Also, students may be more flexible than non-students, perhaps connected to their age and financial situation.

A second type of variable we used to be that of the attributes of the responders' individual flight; a third type of variable used to concerned the responders' airport experience. Responders were asked to connect their answers to their most recent flight and airport experience. Different experiences in these areas could clearly differentiate customer satisfaction and overall attitudes toward airlines (and air*ports*), and suggest placement in a different segment for marketing purposes.

The variables used are listed in Table 1.

Table1: Variables in Our Study

SI. NO	Variables label	Explanation	Coding (if necessary)
1	Age	Age of responder	
2	Age Range		1:<19, 2: 20-29, 3:20-29, 4:40-49, 5:50-59, 6:60-69, 7:70-79, 8: >80
3	Gender	Sex of responder	1: Male, 0: Female
4	Type of Travel	Categorical: Mileage traveler/ Business traveler/ Personal traveler	0: mileage traveler; 1: business traveler; 2: personal traveler.
5	Shopping Amount at Airport	Amount of money spend at airport, $	
6	Eating and Drinking at Airport	Amount of money spend at airport, $。 $	
7	Departure Delay of flight in Minutes		
8	Arrival Delay of flight in Minutes		
9	Flight cancelled	Categorical: Yes/ No	0: No; 1: Yes.
10	Flight time in minutes		

To give a general understanding of our dataset, we provide some descriptive statistics in Table 2. We wonder about the minimum flight time of "8" listed by one respondent, but, while it seems "odd," we had no basis to change it. Perhaps we should have disallowed the data value; we note in the limitations section our perhaps-unwise choice not to explore the issue of outliers. Also, we acknowledge that "Age" and "Age Range 1" are, for the most part, redundant, and it likely would have been wiser not to use both of them.

Table 2: Descriptive Statistics

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Age	129889	70	15	85	46.20	17.321
Age Range1	129889	7	1	8	4.17	1.784
Gender1	129889	1	0	1	.44	.496
Type of Travel1	129889	2	0	2	1.23	.577
Shopping Amount at Airport	129889	879	0	879	26.55	53.081
Eating and Drinking at Airport	129889	895	0	895	68.24	52.210
Departure Delay in Minutes	127544	1592	0	1592	14.98	38.366
Arrival Delay in Minutes	127151	1584	0	1584	15.37	38.762
Flight cancelled1	129889	1	0	1	.02	.135
Flight time in minutes	127151	661	8	669	111.51	71.776
Valid N (listwise)	127151					

Clustering Method

This study adopts the K-means clustering method. As we noted earlier, there are several methods for clustering that are available in statisticalsoftware packages and no general agreement about which is the best. This method is typically easy to interpret and is adequate for a large data set. Shao, Tanner, Thompson, & Cheatham (2007) compared 11 different clustering methods, and concluded that "Overall, it was found that there is no one perfect "one size fits all" algorithm for clustering MD [molecular dynamics] trajectories." They also noted that average-linkage seemed to work best in most of the methods and that hierarchical clustering methods were affected more severely than others by outliers; of course, they acknowledge that they were using the techniques only for MD applications.

In the "K-means" algorithm, K is a predetermined number of clusters. The distance between each observation and each cluster mean is examined, and an observation is assigned to the cluster whose mean has the smallest distance from the observation. First, an initial set of means is defined and then subsequent classification is based on the distance between the observation and its center (Dasgupta & Freund, 2009). Then, the cluster mean will be re-calculated and, hence, updated. This step will be repeated again and again until no cluster means change further.

As noted, the K-means clustering method requires us to pre-determine the number of clusters: K. In practice, there are several ways to do. In this paper, we use the Clementine method (Wu, Cheng, & Chen, 2008) to determine the best value of K. First, we set different numbers of clusters, using the K-means cluster method in SPSS, and obtain ANOVA tables. In essence, we perform the clustering method for different values of K, and then pick the K that appears to be the "best of the bests" choice.

By comparing F-statistics in Table 3, it is clear that relatively speaking, there is little difference between the F statistics for the six-clusters result and for the seven- clusters result (as opposed to comparing adjacent results for fewer clusters.) Therefore, we chose the six clusters result. F-statistic results are shown in Table 3 for K = 3 through 8.

Table 3: F-statistic Values when Cluster Numbers (K) = 3-8

	F.(3)	F.(4)	F.(5)	F.(6)	F.(7)	F.(8)
Age	.510*	1.22*	410.75	293.475	251.325	174.123
Age Range1	.400*	.105*	404.056	289.05	247.907	171.038
Gender1	1.963*	1.218*	231.863	159.681	132.653	105.331
Type of Travel1	1.163*	2.228*	75.447	53.977	49.196	35.677
Shopping Amount at Airport	1.44*	1.068*	58836.6	40970.5	33918.8	28863
Eating and Drinking at Airport	3.820*	5.237*[1]	35212.3	23004.9	18984.9	15476.4
Departure Delay in Minutes	89031.1	116912.6	84028.7	54815.3	66929.5	56797.9
Arrival Delay in Minutes	89379.9	116180.8	83339.2	54321.1	65709.9	55565.6
Flight time in minutes	112496.1	74616.8	12.0[2]	39512.0	32886.6	58383.1

ANALYSIS AND DISCUSSION OF RESULTS

In this section, we will interpret our results to support marketingstrategy decisions. Table 4 presents the number of cases/people in each cluster; it is clear that the number of cases in each cluster varies. The largest number of cases is in Cluster 3: 70,253, 55.25% of the total number of cases. The number of cases in Clusters 4 and 6 is next largest, around 20,000; these two are 15.98% and 16.59%, respectively, of the total number of cases. Cluster 1 includes 10,008 cases,

[1] "*" indicates a non-significant F-value $\alpha = .05$. All of the other F-values are significant at $\alpha = .05$.

[2] This value may look to the reader like a typo; while it does appear to be an unusual value, relative to the entire row of values, it is a CORRECT value.

Cluster includes 5,273 cases, and Cluster 5 contains only 201 cases. Having fewer cases in a cluster does not automatically diminish a cluster's importance, since each cluster likely has some vital features, and a small cluster may contain very-high-value, or important-to-identify, customers.

Table 4: Number of Cases in each Cluster

Number of Cases in each Cluster

Cluster	1	10008.000
	2	5273.000
	3	70253.000
	4	20324.000
	5	201.000
	6	21092.000
Valid		127151.000
Missing		2738.000

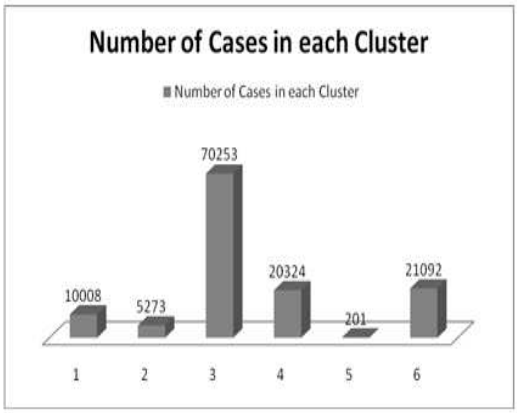

Table 5 is a key table and presents the mean of each variable for each cluster. Figure 1 displays these values in graphical form.

Table 5: Mean of Each Variable for each Cluster

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
Age	45.06	45.73	45.21	50.34	47.83	45.92
Age Range1	4.05	4.12	4.07	4.59	4.33	4.14
Gender1	.31	.45	.44	.48	.46	.44
Type of Travel1	1.23	1.21	1.22	1.29	1.18	1.22
Shopping Amount at Airport	168.90	20.40	12.16	16.58	28.46	18.16
Departure Delay in Minutes	10.34	139.88	8.40	8.97	460.66	9.03
Arrival Delay in Minutes	10.68	142.35	8.57	9.16	463.14	10.21
Flight time in minutes	99.35	105.96	82.53	89.60	118.79	236.24
Eating and Drinking at Airport	71.20	63.89	46.61	148.51	62.55	62.60

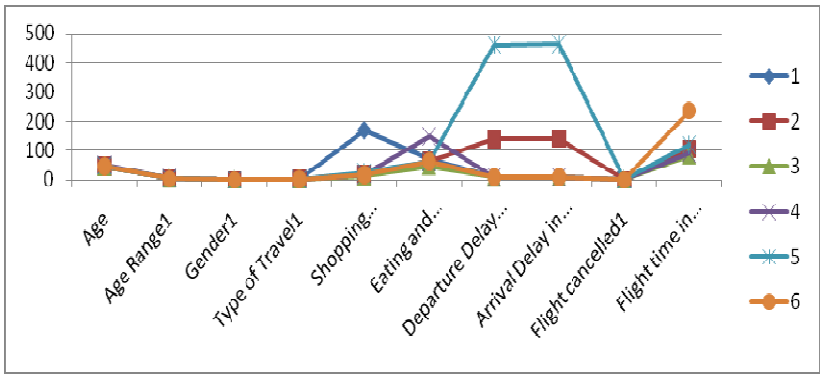

Figure 1: Graphical Representation of Cluster Means for each Variable

Type of traveler needs further elaboration since the mean is not a useful representation of the (more than two categories) nominal variable. Figure 2 displays the frequency distribution for this variable for each cluster.

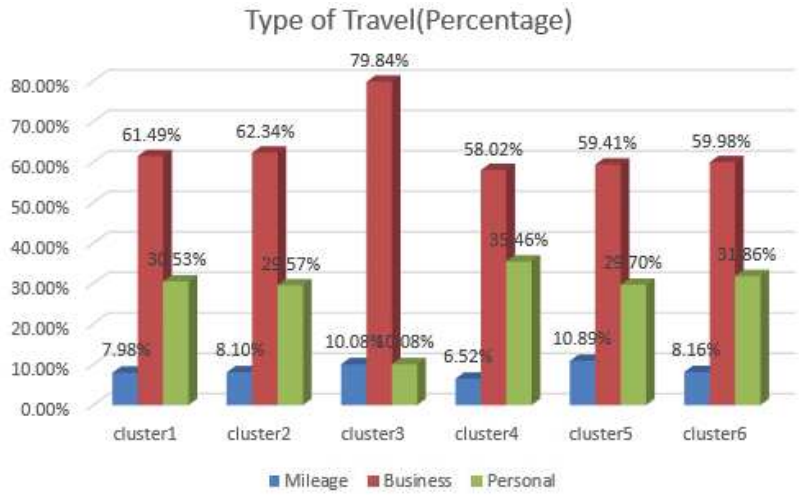

Figure 2: Frequency Distribution for Type of Traveler

We can see from Table 5 that the two age measures do not vary substantially; the average age ranges from 45 to 51 for every cluster. So, we will not focus heavily on the age or age-category variables. Among the six clusters, however, we can identify behavioural differences between customers that, in a sense, "define" a cluster.

Cluster 1 is distinguished from the other clusters primarily by the amount spent shopping at the airport (excluding eating and drinking). The mean amount is about $169, and this cluster contains people who spend far more than people of any other cluster; in fact, the next highest cluster average is under $30. It is also true that cluster 1 contains a somewhat lower percentage of males (31%) than other clusters (all the other percents being between 44% and 48%). None of the other variables are at the top or bottom for this cluster. We can name this cluster, "Shoppers." Members of this cluster could, perhaps, be best targeted-marketed through a bit-more-female-oriented set of advertisements at airport shops.

Cluster 2 is somewhat distinguished by departure delay time, with average 140 minutes and arrival delay time, with average 142 minutes (these two values are, logically, highly positively correlated.) These are the 2nd largest delay values of any cluster, exceeded only by those of **Cluster 5** which has delays about 3 times as large - clearly very unusual average delay times, truly *outliers*, reinforced by the fact that Cluster 5 has only 201 people or less than.2% of the total sample. For Cluster 2 and Cluster 5, some thought should be given to targeted apologetic messages, perhaps offering coupons to members of these clusters to "compensate" them for the delays. None of the other variables are "stand-outs" for these clusters. We can name Cluster 2, "The delayed." We can name Cluster 5, "The unfortunates!!!!" - while rejoicing that there are so few people in this cluster!!

Cluster 3 is comprised of more than half of the customer base. The mean delays (both departure and arrivals) are the lowest of any cluster, and the amount of shopping and eating/drinking dollars spent is also the smallest. Also, we can see from Figure 2 that the percent of business travelers is noticeably higher for this cluster, with a corresponding drop in the percentage of personal travelers. Additionally, their flight times are also the shortest. We can name this cluster, "The business traveler" (who is, perhaps heavily involved in the business aspects of the trip and views him/herself as having less time for non-business activities, such as shopping.) More mass marketing, rather than more specialized,

targeted marketing, may be the way to routinely reach this group. It is suggested that shops and restaurants aspects of the airport/airlines should be de-emphasized, as well as, perhaps, highlighting any attributes that attract the business traveler, such as wifi, computer/terminal availability, and another communication vehicle

Cluster 4 is very distinguishable from the other clusters by their mean of $149 spent for eating and drinking at the airport. This $149 is *more than double the mean of the next highest cluster value*. We can also note that their average age is the highest of the clusters, although only slightly higher than the mean for other clusters. We can name this cluster as "Foodies and Drinkers." Advertisements at airport restaurants and bars seem appropriate to target this cluster.

Cluster 6 is distinguished primarily by its flight times. The mean is 236 minutes, more than double the average flight time of any other cluster. We can name this cluster "The long-distance travelers." Ads on airplane's screens may be good choices to target these travelers.

From the above analysis, it is clear that each of the six clusters has one or more unique characters that differentiate them from other clusters. Marketing managers can reach out to a certain type of customer based on their cluster membership - for some clusters and people, their own demographics, for others their airport behavior, and for yet others, their flight time results.

CONCLUSIONS

Table 6 summarizes the salient features of the 6 clusters.

Table 6: Cluster Summary

SI. NO	Segment Name	# Customers	Description
1	The Shoppers	10008	Spends the most on shopping at airport, with average around $169. Also, is more predominantly female than any other cluster.
2	The Delayed	5273	Notably large, but not outrageous, amount of delay time, around 140 minutes each for departure and arrival.
3	The Business Travelers	70253	Spends the least on both shopping and eating. Shortest average delay times and flight times.
4	The Foodies and Drinkers	20324	A little bit older than those of the other clusters, and overwhelmingly spends the most on food and drinks.
5	The Unfortunates!!	201	Enormous delay times. Ranked second high on shopping Amount/
6	The Long-Distance Traveler	21092	Long flight time, with average about 4 hours (236 minutes).

The cluster analysis based our survey data explains a lot about potential segmentation. The results of the cluster analysis can help marketing departments to understand their market segmentation and further support product positioning.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

It is common that survey data have missing values. Responders either do not have the patience to finish the survey or, are unwilling to provide certain information. This does not appear to be a serious problem here, for the number of missing values is, percentage-wise, relatively low. Nevertheless, there are different ways to handle missing values - for example - deleting the case entirely, using the mean to fill in for the missing data point, and several more sophisticated methods. In this study, we simply deleted incomplete cases. It is possible that a different treatment of missing values would lead to some modification of our results.

We used one specific method of cluster analysis, the K-means clustering method. It is possible that the use of a different clustering algorithm would alter our results somewhat. Future work might consider this issue.

The K-means clustering method, as is the case to different degrees for all clustering methods, is sensitive to outliers. We did not address the issue of outlier values. It might be useful if future research in this field gave serious attention to the elimination of outliers. We can see from Table 2 that three of our metric variables (i.e., not counting the two-category nominal variables of Gender1 and Flight Cancelled 1, for which is it not uncommon for a standard deviation to exceed a mean) have a standard deviation that exceeds the mean. This often indicates that outliers are present. It is possible that the elimination of outliers might moderately alter our results. While there are many ways to define an outlier, any common method is very likely to be superior to not addressing the topic at all.

We used secondary data provided on a website and with a 3- month survey-period. It is possible that for a different time period, results could change. Future studies might wish to consider a longer time-period over which to collect data.

Our variable selection was dictated by the survey. There may be other variables that allow for a superior clustering of the customer base. Of course, some of these variables may be difficult to obtain (e.g., an income of the responder). However, variables such as how frequently the responder flies, whether he/she is traveling alone or with one or more companions, and other possible variables, may be placed on the survey without adding materially to responder discomfort (such as would often be the case if one's income was requested), and thus, not add much to the number of cases with missing data.

REFERENCES

1. Bailey, K. (1994), "Numerical taxonomy and cluster analysis," *Typologies and Taxonomies*, p. 34.ISBN 9780803952591.
2. Berger, P., and Magliozzi, T. (1992), "The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis," *Journal of Direct Marketing*, 6(1), pp. 13-22.
3. Cattell, R. (1943), "The description of personality: basic traits resolved into clusters," *Journal of Abnormal and Social Psychology*, 38(4), pp. 476–506. doi:10.1037/h0054116.
4. Dasgupta, S. and Freund, Y. (2009), "Random trees for vector quantization," *IEEE Transactions on Information Theory*, 55, pp. 3229-3242.
5. Hill, G., and Baumann, R. (2000), "Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge," *Diploma thesis, Vienna University of Economics and Business Administration, Austria.*
6. IBM Watson Analytics – 2014 **https://www.ibm.com/communities/analytics/watson-analytics-blog/sample-data-airline-survey/** , accessed Feb. 2017
7. Kennedy, P., Best, R., and Kahle, L. (1988), "An Alternative Method for Measuring Value-Based Segmentation and Advertisement Positioning," Current Issues and Research in Advertising, 11(1-2), pp. 139-155.

8. Lix, T., Berger, P., and Magliozzi, T. (1995), "New customer acquisition: prospecting models and the use of commercially available external data," *Journal of Direct Marketing*, 9(4), pp. 8-18. DOI: 10.1002/dir.4000090403
9. Luo, M. (2003), "Logistics barriers for multinational corporations doing business in China," MIT PhD thesis, http://hdl.handle.net/1721.1/28509
10. Porter, M. (1998), "Clusters and the New Economics of Competition," *Harvard Business Review*, November/December, p. 77-95.
11. Porter, M. (2003), "The Economic Performance of Regions," Regional Studies, 37(6-7), pp. 549-578. https://doi.org/10.1080/0034340032000108688
12. Reynolds, T. and Jolly, J. (1980), "Measuring Personal Values: An Evaluation of Alternative Methods," *Journal of Marketing Research, 17(4), pp. 531-536.* DOI: 10.2307/3150506
13. Shao, J., Tanner, S., Thompson, N., and Cheatham, T. (2007), "Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms," *Journal of Chemical Theory and Computation*, 3(6), pp. 2312-2334.
14. Smith, W. (1956), "Product differentiation and market segmentation as alternative marketing strategies," *Journal of Marketing*, 21(1), pp. 3-8.
15. Thomas, J. (2016), "Market segmentation," *Decision Analyst*, https://www.decisionanalyst.com/whitepapers/marketsegmentation/, accessed February, 2018.
16. Tryon, R. (1939), "Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality, Edwards Brothers.
17. Wang, X, Hong, M. & Berger, P. (2016), Determining Key Factors in Consumer Evaluation of an Airport, *Journal of Marketing Management*, 4(1), pp. 19-30. DOI: 10.15640/jmm.v4n1a3
18. Wikipedia (2017), "Cluster analysis," https://en.wikipedia.org/wiki/Cluster_analysis, accessed, February, 2018.
19. Wu, S., Cheng, K, and Chen, F. (2008), "Application of cluster analysis in telecommunication customers segmentation" 技术经济与管理研, 1004 292X(2008)01 00010 03.
20. Yankelovich, D. (1964), "New Criteria for Market Segmentation," *Harvard Business Review*, March/April, pp. 83.
21. Yankelovich, D. and Meer, D, (2006), "Rediscovering market segmentation," *Harvard Business Review*, February, pp. 143-175.