Cluster Analysis of World's Airports on the Basis of Number of Passengers Handled (Case Study Examining the Impact of Significant Events)

Žambochová Marta [1] | *J. E. Purkyně University, Ústí nad Labem, Czech Republic*

Abstract

Nowadays, the air transportation is one of key means of transport. Unfortunately, there are many factors influencing its popularity and intensity of its use. There are many studies investigating these factors. The paper investigates the possibility of classifying the world's airports in terms of the trend in the number of handled passengers as it is one of the main economic indicators for airports. For this classification we chose cluster analysis. The paper focuses on an important aspect of the process, which chooses the appropriate number of clusters. It turned out that in terms of interpretation of the results, it may not always be the most efficient to set this number at the theoretically best and recommended value. As a result of our analysis, several groups of airports with similar trend of post-event reactions are found. Therefore, this may bring better understanding of the intensity and the range of the impact of particular events on air transportation.[2]

Keywords	JEL code
Cluster analysis, number of clusters, occupancy of airports, Bayesian Information Criterion, Akaike Information Criterion, silhouette coefficient	*C38*

INTRODUCTION

Nowadays, air transport is the fastest growing transport sectors. In order to operate successfully, it is necessary to care not only for its means of transport, i.e. aircraft, but also for the facilities and background – airports and airfields. Assuming an airport to be a normal economic entity, its success is evaluated

[1] Faculty of Social and Economic Studies, Jan Evangelista Purkyne University in Usti nad Labem, Pasteurova 3544/1, 400 96 Ústí nad Labem, Czech Republic. E-mail: marta.zambochova@ujep.cz.

[2] This article is based on contribution from the conference *Robust 2016*.

according to operational and economic indicators. The basic indicators include performance indicators such as the number of aircraft movements, the number of tons of cargo handled, the number of passengers handled, etc. In this paper we deal with the last of these factors – the number of passengers.

The paper (Akamai et al., 2015) focuses on the importance of the amount of passengers for the operation of airports. The paper (Lu et al., 2014) deals with changes in traveller's behaviour during extreme weather conditions such as strong wind. Stability of air traffic at selected airports within a particular time period is reviewed in paper (Grenčíková et al., 2011). In the paper of ours, the stability of air traffic is examined globally. At the same time, the paper searches for factors influencing the possible instability at a certain moment.

This paper focuses on facts influencing the the number of passengers handled at particular airports around the world. The main task of the analysis is a data classification using cluster analysis. Several authors dealt with cluster analysis in the field of aviation before. In paper (Kraft, 2012) authors use cluster analysis to examine the key factors affecting the transport important for settlement of the Czech Republic. The paper is focused on road transport. Similarly, in the paper (Grabbe et al., 2014) cluster analysis is performed when the input variables are particular weather data at given times. Based on this analysis, the authors focus on the impact of weather on air traffic delays. However, in our contribution we used cluster analysis differently. Our main goal is to show the way enabling to find analytically a group of world airports which exhibit the same trend in the number of passengers handled. Based on this or a similar analysis, it would be possible to understand better effects which influence air transport.

1 METHOD OF ANALYSIS

Cluster analysis is based on finding similarities of data objects. It divides sets of objects into several previously unspecified groups (clusters) so that objects within an individual cluster are the most similar and objects from different clusters are the least similar.

Statistical software systems typically include, among other things, both the hierarchical algorithm with the possibility of the result shown in the form of so-called dendrogram, and non-hierarchical iterative *k*-means algorithm. The SPSS statistical system has included the TwoStep method since the version 11.5.

1.1 *K*-means method

The *k*-means method and its variants belong among the most important representatives of *k*-centroid algorithms, which form an important subset of divisive methods. This method is a very popular and widely used iterative clustering process which is suitable for analysis of quantitative data. The principal idea of the algorithm is to divide objects into a predetermined number of clusters so that the sum of distances of component objects from the centre of the cluster is minimal. In other words, its objective is to find minimum of the function:

$$Q = \sum_{\mathbf{x}\in\mathbf{X}} \left\| \mathbf{x} - c(\mathbf{x}) \right\|^2 = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{il} \sum_{j=1}^{d} \left(x_{ij} - c_{lj} \right)^2, \tag{1}$$

where $\mathbf{X}$ represents the set of all analysed objects, n represents the number of objects, d represents the number of dimensions, k is the number of clusters, $\mathbf{x}$ represents an object with coordinates x_{ij}, $c(\mathbf{x})$ is the nearest centroid of the object $\mathbf{x}$, w_{il} is indicator of belonging i-th object to the l-th cluster, c_{lj} is j-th coordinate of the centroid of l-th cluster. Many variations of the basic *k*means procedure are described in literature under different names.

The algorithm is composed of the following steps:

- Step 0: An initial partition of the data file into *k* clusters,
- Step 1: The counting of every cluster's centroid,

- Step 2: The assignment of every object to the closest centroid,
- Step 3: Repeating Steps 1 and 2 until the centroids no longer change.

1.2 TwoStep method

This method uses the BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies) which is explained in detail in (Zhang et al., 1996), or (Zhang et al., 1997). The algorithm creates first a so-called CF-tree, which is progressively filled by incoming data. The advantage of this principle is that it goes through the data file only once. The disadvantage is the sensitivity to the entry data ordering.

The TwoStep clustering procedure consists of the following steps:

- Step 1: Pre-clustering,
- Step 2: Outlier handling (optional),
- Step 3: Clustering.

In the first phase the CF-tree is created and the entering objects are progressively organized. An entry in the leaf node represents a sub-cluster. The non-leaf nodes and their entries are used for entering a new object quickly into a correct leaf node. Each entry is characterized by its CF that consists of the entry's number of objects, mean and variance of all data points belonging to the node. In the second phase the CF-tree is condensed and optimized due to its threshold adjustment. The outliers are eliminated with the help of the proper tree re-designing. In the third phase the impact of entry data order sensitivity is minimized. The leaf nodes of the CF tree are then grouped using an agglomerative hierarchical clustering algorithm. The cluster step takes sub-clusters resulting from the pre-cluster step as input and then groups them into the desired number of clusters.

1.3 Criteria for determining the optimal number of clusters

There are many information criteria for determining the optimal number of clusters (Řezanková et al., 2009). Among them, three information criteria are implemented in the SPSS. The first is the Bayesian Information Criterion, (*BIC),* which is used to determine the optimal number of clusters in the TwoStep cluster analysis. For our purpose it is calculated by the following formula:

$$BIC(k) = -2\sum_{i=1}^{k} \lambda_i + w_k \ln(n)\,, \tag{2}$$

where λ_i is the characteristic for the *i*-th cluster determined by the formula:

$$\lambda_i = -n_i \sum_{j=1}^{m_1} \frac{1}{2} \ln\left(s_j^2 + s_{ij}^2\right) + \sum_{j=1}^{m_2} H_{ij}\ , \tag{3}$$

n_i is the number of objects in the *i*-th cluster, m_1 is a number of quantitative continuous variables, m_2 is the number of categorical variables, s_j^2 is the sample variance of the *j*-th continuous variable, s_{ij}^2 is the sample variance of the *j*-th continuous variable in the *i*-th cluster. H_{ij} is the entropy given by the relation:

$$H_{ij} = -\sum_{l=1}^{p_i} \frac{n_{ijl}}{n_i} \ln\left(\frac{n_{ijl}}{n_i}\right), \tag{4}$$

p_j is the number of categories of the *j*-th categorical variables and n_{ijl} is the frequency of the *l*-th category of the *j*-th categorical variables in the *i*-th cluster. Furthermore, w_k is calculated according to the formula:

$$w_k = k\left(2m_1 + \sum_{j=1}^{m_2} p_j - 1\right)\cdot \tag{5}$$

When determining the optional number of clusters, the values of *BIC* are calculated. The estimated initial number of clusters is ruled by the minimum value of *BIC*.

The second criterion is called the Akaike Information Criterion (*AIC*) and is calculated according to the formula:

$$AIC(k) = -2\sum_{i=1}^{k} \lambda_i + 2w_k \ . \tag{6}$$

The optimal number of clusters is determined by the same principle as in the case of *BIC*.

For the evaluation of resulting clusters obtained by divisive methods we use the silhouette coefficient (*SC*), which expresses the silhouette measure of cohesion and separation. The silhouette coefficient for individual *i*-th object from the *h*-th cluster is calculated according to the formula:

$$SC(i) = \frac{\mu_i - \eta_i}{\max\{\mu_i ; \eta_i\}} \ , \tag{7}$$

where n_i is the average distance of the individual *i*-th object with all other objects within the same cluster and:

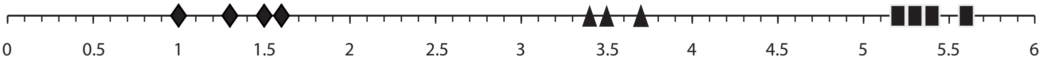
$$\mu_i = \min_{g \neq h} \left(\frac{\sum_{j \in C_g} D_{ij}}{n_g} \right) \ , \tag{8}$$

where C_g is the *g*-th cluster and D_{ij} is the distance between the *i*-th and *j*-th objects.

Using Formula (7), average values for individual clusters are determined as well as the average value for the whole decomposition. The higher the average value is, the more compact the cluster is.

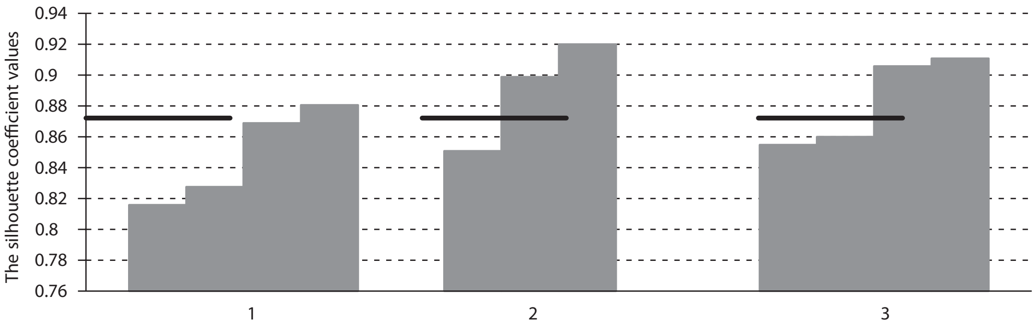
The following three figures show a simple and illustrative example of silhouette coefficient. Figure 1 presents the situation of the eleven objects divided into three clusters. In Figure 2 can be seen graphical representation of both all individual values *SC* (gray bars) and the resulting average *SC* (black dashed line). Figure 3 shows *SC*, which is the output of the system when applying SPSS TwoStep method.

Figure 1 A data file containing 11 objects divided into three clusters



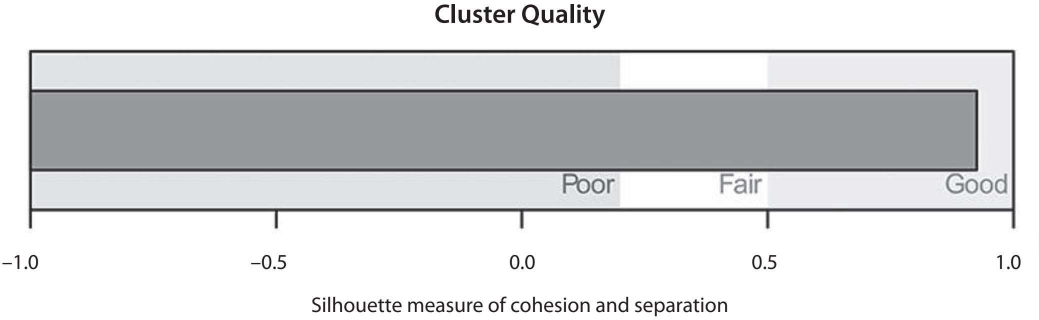
Source: Own construction

Figure 2 Silhouette coefficient for objects in three clusters



Source: Own construction

Figure 3 Silhouette coefficient – the output from SPSS



Source: Own construction

2 CASE STUDY

In our work, we focused on the segmentation of airports using cluster analysis. Each airport stands for an object to be clustered. We analysed data of 838 airports from a total of 977 monitored ones. The data consisted of numbers of passengers who passed through the particular airport per month. Data were collected in the thesis work (Darda, 2014), individual data were obtained partly from the Institute of Civil Aviation (Service technique de l'aviation civile) with headquarter in Paris and from the French Ministry for ecology, sustainable development and energy (Ministère de l'écologie, du Développement durable et de l'Énergie), headquartered in Paris.

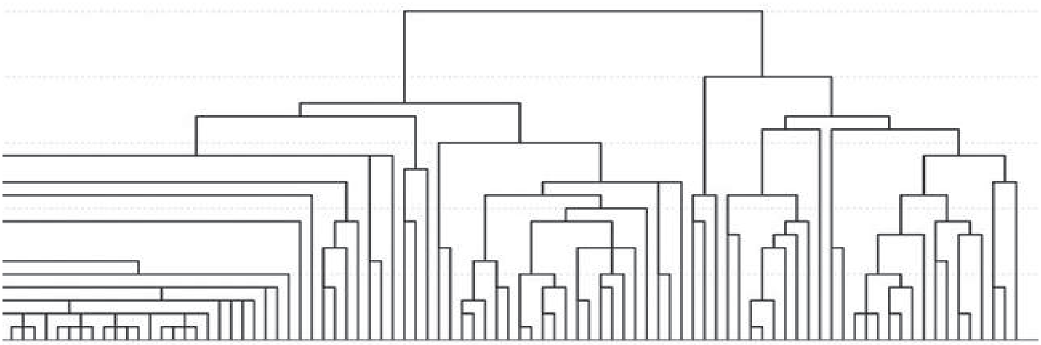
Data were monitored in the period from January 2000 to April 2014. Some airports (mainly Asian) publish data for up to year-end summary, therefore, we restricted our analysis to the period at the end of 2013. World airports, about which we were not able to provide all required information, were not included in the processing. The annual throughput of passengers through each of airports was another factor considered in processing. Airports with the annual throughput lower than 100 000 passengers were excluded. Airports where the statistical data on a monthly basis are published only once per year are also not included in our dataset. This is mainly the case of Asian, particularly Chinese airports where statistics are always published in early April of the following year. Data about several airports were not available since 2000, therefore, we could not incorporate them into the analysis. Complete data about 838 airports were collected from the beginning of 2000 until the end of 2013, thus the input data matrix contains 838 rows and 168 columns.

It should be recalled that the aim of the analysis is first of all to compare the trends in number of handled passengers, so that the absolute values of passengers handled become irrelevant. Therefore, the data for each airport were standardized, subtracting from data on each row corresponding row mean and dividing them by corresponding row standard deviation. We assume these new transformed data to be representatives of quantitative continuous random variables and further as the input for our cluster analysis.

For segmentation, we selected three methods implemented in SPSS. These were the hierarchical method, the TwoStep method and the *k*-means method. The first two methods were used to determine the optimal number of clusters, the third method for the analysis itself.

The final dendrogram was the output of hierarchical clustering (with use Average Linkage method and Euclidean measure). The entire dendrogram was very confusing due to the large number of objects. Its interesting part is shown in Figure 4. Still, it was clearly visible that the suitable number of clusters is two or three.

Figure 4 The interesting part of a dendrogram – the output from SPSS

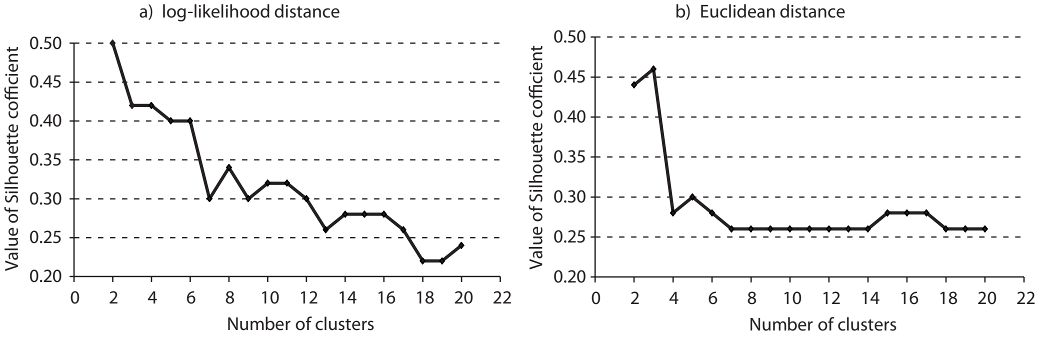


Source: Own construction

Processing with the use of the TwoStep method showed similar results. We used all three indexes implemented in SPSS to monitor the quality of clustering, namely *BIC*, *AIC* and *SC*.

When selecting the log-likelihood distance, the TwoStep method showed three clusters to be the optimum number, for both *BIC* and *AIC* criteria. When the Euclidean distance was selected, the optimal number of clusters turned to be two. Further, we used TwoStep method with both the log-likelihood and Euclidean distance for fixed number k of clusters, $k \in \{2, ..., 20\}$, and calculated corresponding silhouette coefficients SC_k. The values SC_k for each reached distance are plotted in Figure 5.

Figure 5 Graph of silhouette coefficient for TwoStep method



Source: Own construction

It is clear from Figure 5a) and 5b) that the maximum value of *SC* was achieved in case of three and two clusters. The SC value was never lower than 0.2, which means that even in the worst case the quality of clustering was fair. In case of using the Euclidean distance the resulting clusters were highly unbalanced in terms of the number of objects in different clusters. Therefore, for further processing the log-likelihood distance was selected as more appropriate one.

Summarizing, the choice of either two or three clusters appears to be the best choice while using various indicators. Unfortunately, the resulting clusters were not satisfactorily interpretable in either of these theoretically recommended cases. However, we received interpretatively interesting results using k-means method or TwoStep method when choosing a parameter determining the required number of clusters equal to four and then eight. Either the values of the three indicators of quality or the dendrogram does not condemn this solution. Therefore, we will discuss these cases below.

2.1 Analysis of the results for *k*-means method with Euclidean distance and four clusters

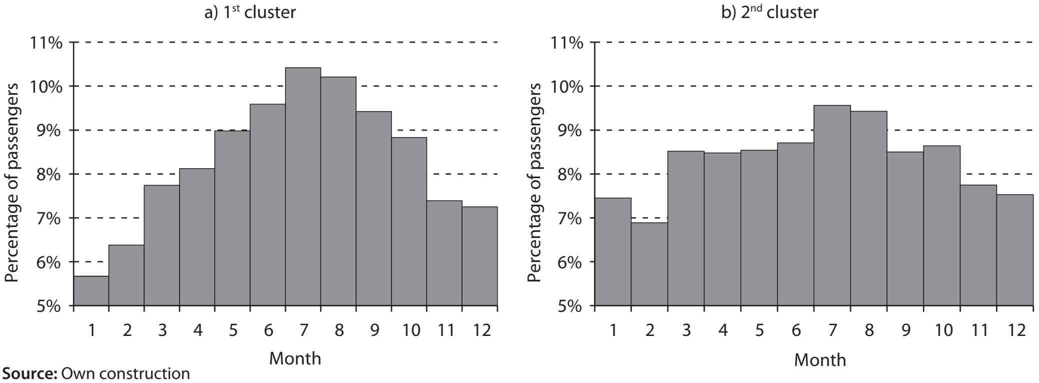
The airports which handled the most passengers during the summer months are included in the first cluster. This cluster contains 33.65% of the monitored airports, and is thus the second largest one. In this category, the number of passengers at the airports increases every month since the beginning of the year until the end of June. The number of passengers reaches the maximum values during July and August. Then again, the number of handled passengers gradually decreases. The end and the beginning of the year have the same or similar values. Therefore, the trend has a recurring character. Many European airports are represented in this category. Many airports from this cluster are also located in North America.

Almost all the airports from cluster one are located in the Northern Hemisphere, where summer culminates in June-September. Demand for air travel increases dramatically in this period, since there is the summer holiday period in many countries. Several airlines operate out of these airports so-called charter and seasonal flights, which carry large numbers of passengers to tourist destinations. This leads to the fact that some airports in the Mediterranean region handle more than 80% of passengers during the summer. During the rest of the year, they handle the remaining 20%. This is demonstrated, for example, on the aggregated group of Greek airports, where 69% of passengers are handled during the summer. Similar indicators can be found also in other Mediterranean airports that experience the greatest rush of passengers during the summer months, such as Spain, France, Italy, Montenegro, Turkey, Egypt and Tunisia. The proportional distribution of handled passengers during the year is shown in Figure 6a).

The second cluster of airports includes 20.88% of the monitored airports. Trends in the number of passengers handled during a year at these airports are similar to those from the first cluster, but with the difference that the increase of passengers in summer is not that dramatic. This means that the traffic and operations in these airports are more balanced during the year. The months of July and August represent the highest percentage of passengers handled, which transcend the boundaries of 9.5%. At the beginning and the end of the year, the percentage is much lower, ranging between 6% and 8%. Proportional distribution of passengers handled during the year is shown in Figure 6b).

In this cluster, airports from any continent are not predominant as in the case of the previous cluster, three continents being more or less equally represented, namely North America, Europe and Asia. As in the first group, the airports included here operate with the most traffic in the summer. We suppose that one of the main reasons why these months prevail again is the summer season in the Northern Hemisphere, that brings increased tourism activity. This is certainly true in the case of Europe and North America, but we do not think that it would be possible to be applied to airports located in Japan.

Figure 6 Proportional distribution of handled passengers during the year which is characteristic for the airports of the clusters

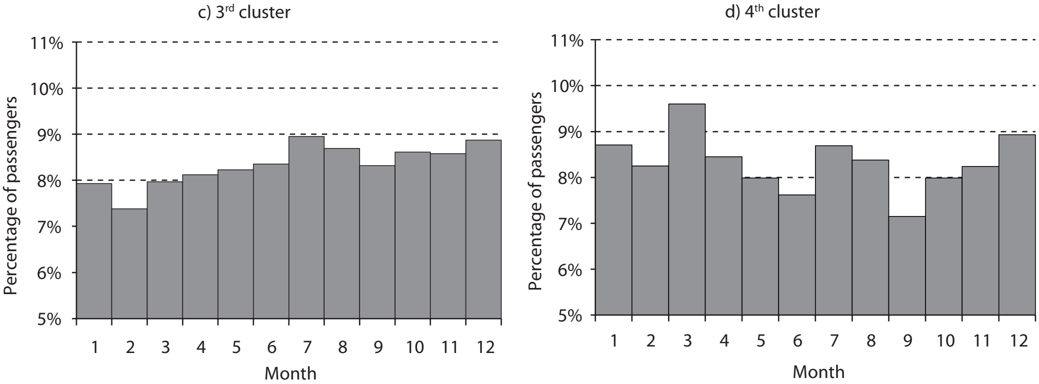


a) 1st cluster

b) 2nd cluster

Source: Own construction

Figure 6 Proportional distribution of handled passengers during the year which is characteristic for the airports of the clusters – continuation



Source: Own construction

Airports with relatively balanced year-round operation represent the third cluster. It is the strongest cluster in the number of airports, precisely 33.77% of the monitored airports. These airports handled 36.35% of passengers of the total number of passengers.

Considering the number of passengers handled per month, these airports are relatively stable during the whole year. Compared with two previous groups, there are not any significant fluctuations in this cluster. In this case, the number of passengers handled per month compared to the total annual number oscillates between 7% and 9%. Therefore, the fluctuation is only about 2%. Proportional distribution of handled passengers during the year is shown in Figure 6c). These airports are located worldwide. We suppose that there exist two main explanations for interpretation of such a distribution.

The first explanation is the transitivity of these airports. There are several important airports belonging to this group such as the one in Dubai, Beijing, Hong Kong, Bangkok, Singapore, Istanbul, Shanghai, Seoul and American Charlotte. Most of these airports have lines serving all inhabited continents and most of countries in these continents. In our opinion, the balanced character of their operation lies in a dense network of destinations. For example, the airport in Singapore used to be a key transit point between Australia and the UK until March 2013 as it was used by the Australian airlines QANTAS. QANTAS have selected a new transit airport – Dubai after the termination of cooperation. Nowadays, Dubai became one of the largest transit airports in the world. If we look at the exact statistics on the number of passengers handled at these airports, we find essentially no difference since this line of company QANTAS had a negligible share of passenger transport between Australia and the UK. In other words, it is obvious that a transit airport gain passengers from dozens of lines. Accordingly, the sudden or forward known loss of one or several airlines does not have considerable importance to the fluctuation of handled passengers.

In the case of airports where geographical conditions make it difficult or even impossible to travel by other means of transport are the second group in this category of airports with a balanced operation. Examples which demonstrate this cluster well are airports in India or Brazil, where travelling by train or car from one end of the country to another is very time consuming. Furthermore, in this cluster, there are also airports located in island countries, such as Japan, the Philippines, Indonesia and South-east Asian countries.

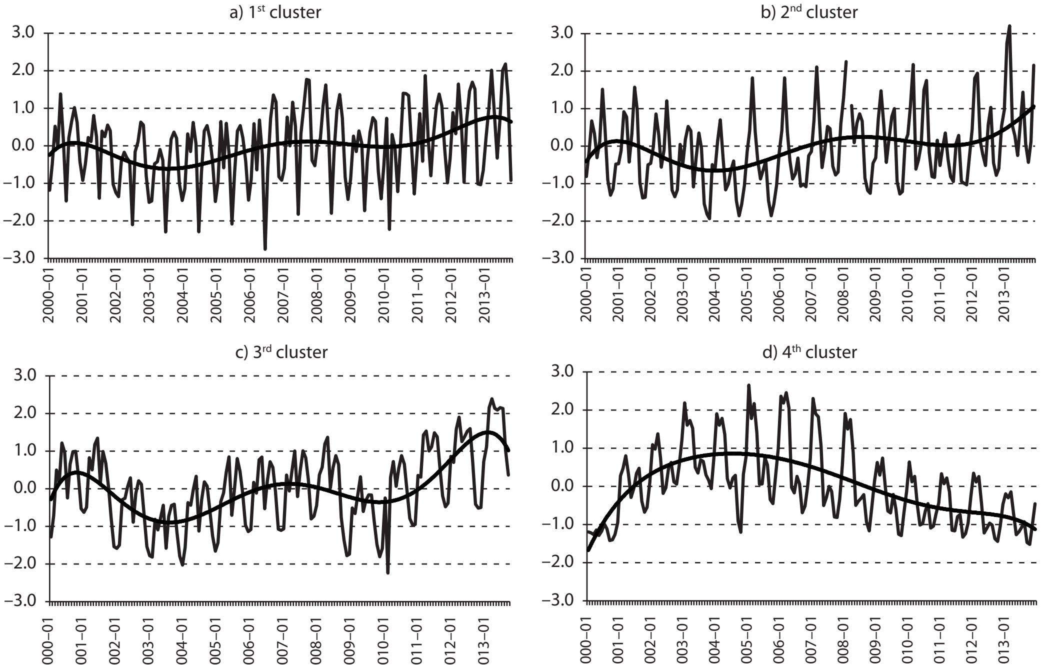
The airports that handle larger numbers of passengers during the months at the beginning and at the end of the calendar year form the last cluster. This is the smallest cluster of airports generated by our analysis. The number of passengers reached only 6.95% of the total number of handled passengers at all examined airports. The greatest number of passengers at these airports occurs in the first quarter of the year. Then the number slightly decreases. It achieves very small numbers in the months of May and June compared with the previous month. Number of passengers increases slightly during the summer months and culminates its growth in the last quarter of the year.

If we disregard at this moment the months of July and August, which are relatively rich in passengers in each category due to the position of most major airports in the Northern Hemisphere and tourism, we find that the months from January to April and October to December provide a large percentage of handled passengers. From 8% to 10% of passengers are handled during these months. Proportional distribution of handled passengers during the year is shown in Figure 6d).

In this group, there are the airports in the Southern Hemisphere, particularly airports in Australia, New Zealand or South America. Trends in the number of passengers handled in this category can be described similarly as we characterized trends of the first and second category. In the Southern Hemisphere, summer culminates in months around the turn of the year. Conversely, there is winter in the Northern Hemisphere. Tourism brings an increased number of passengers into these thermopile areas. A group of airports in the Caribbean Sea and the Gulf of Mexico have the same tendency. The principle of airports functioning in the Southern Hemisphere is nearly symmetrical compared to the Northern Hemisphere.

There is a second significant group of airports which belong to this cluster. These are airports situated very close to the two poles of Earth. These are mainly airports in Scandinavia, Canada and the southern part of South America. We think that very low temperatures and frozen water make it difficult to use land or water transport. Therefore, air transport prevails in these months.

Figure 7 Time series of normalized monthly values of the number of handled passengers in the airports belonging to different clusters



Source: Own construction

SUMMARY

During processing, we wiped out all four time series resulting from averaging the values for all airports of the cluster using polynomials of sixth grade. The courses of these four regression functions were similar (except 4th cluster). It is seen from Figures 7a) to 7d). It can be concluded that the clusters do not differ too much in terms of long term evolution. Substantial difference was demonstrated in terms of seasonality.

In the first group, there are airports with a significant increase in passengers during summer months. The second group of airports shows a similar situation as the first group, but the summer increase in passengers is not as significant. The third group of airports has a balanced number of transported passengers during the whole year. The fourth group consists of airports, where the number of transported passengers is the highest in winter months. The main factor determining this division is therefore seasonal development during the year. The most important characteristics of individual clusters are summarized in the Table 1.

Table 1 Characterization of the clusters

	1st cluster	**2nd cluster**	**3rd cluster**	**4th cluster**
Number of airports	282	175	283	98
Cluster proportion (%)	33.65	20.88	33.77	11.69
Proportion of transported passengers (%)	44.68	12.02	36.35	6.95
Prevailing geographic location	Europe, North America, Japan	Japan, North America, Europe	Asia, Africa, Australia	Mexico, Scandinavia, SE Asia, New Zealand

Source: Author's calculations

2.2 Analysis of the results for *k*-means method with Euclidean distance and eight clusters

We also received interpretatively interesting results when we used the *k*-means method and selected the parameter determining the required number of clusters equal to eight. In all 8 clusters the significant impact of the world economic crisis 2008 was obvious.

There were 49 airports in the first cluster. Of these, 35 were from Asia, and more than half of them were from South Korea (the largest of representatives was the Gimpo Airport), also from Thai (the largest representative was the Phitsanulok airport), but also from Japan (Kansai airport). Among major airports from other continents, there were for instance the New Orleans airport from the USA, or Swedish Växjö.

All airports in this cluster suffered a significant decrease in the number of checked passengers after September 11, 2001. There is another large decline at the end of 2003. This decline began to mitigate until the end of 2005. It can be deduced from knowledge of world events that airport operations were in part influenced by the terrorist attacks of September 11, 2001, but even more by period of SARS epidemic, which erupted at the end of 2002. The impact of the SARS epidemic on aviation is described in Loh and Elaine (2006), see Figure 8a.

There were 29 airports forming the second cluster. Of these, 18 came from Europe, especially from the tourist centres. The Spanish airports Grand Canaria and Tenerife South, the Austrian airports in Innsbruck and Salzburg and Italian Turin, belong among the largest of them. Since airports in this cluster are characterized by their distinct seasonality, they experience a strong increasing trend of handled passengers in the observed period. Events of September 11, 2001 are slightly noticeable (Figure 8b).

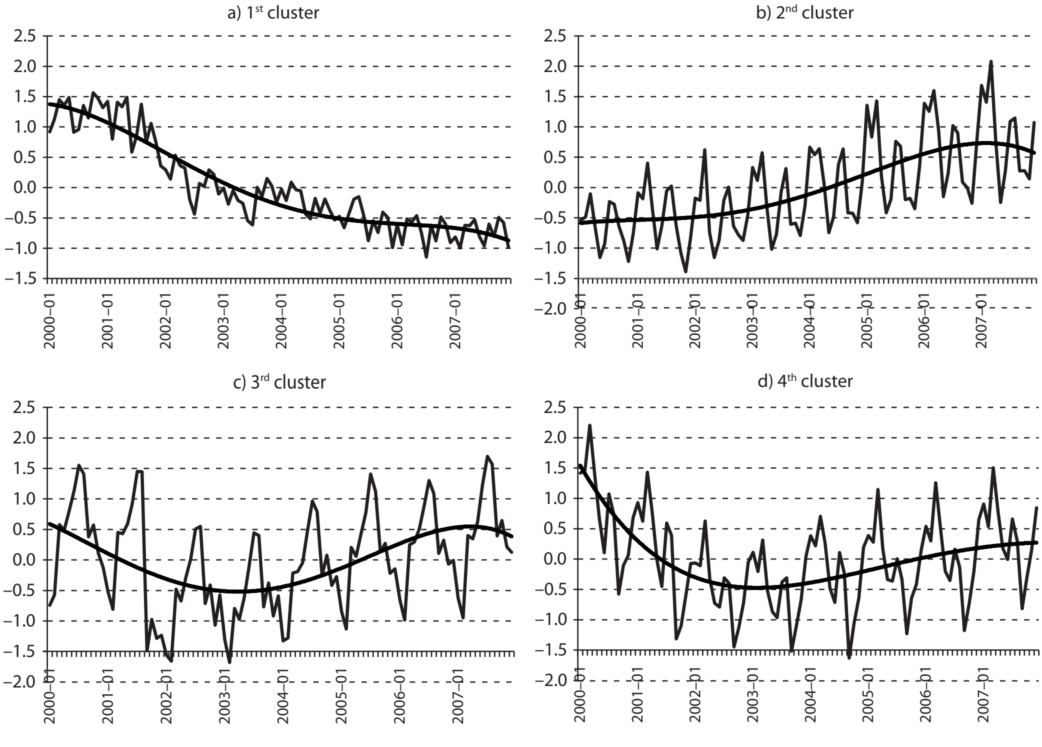
There were 98 airports included in the third cluster. The vast majority of them come from the US. Chicago, Los Angeles and Dallas-Fort Worth are the most important representatives. In this cluster, there are also 12 European airports. The airport Paris Orly and the airport in Brussels, but also Nordic airports, are among the most important ones.

The very strong decline since September 2001 at airports in this cluster is obvious in Figure 8c). The number of handled passengers decreased rapidly compared to the year 2000. After the slump, the same number as in 2000 was not achieved until 2007. Many publications deal with the influence of terrorist attacks of 11 September 2001, namely publication of airlines as (IATA) and scientific papers, such as (Dempsey 2003; Chen, C.-C et al., 2009; Cui and Li, 2015).

The fourth cluster consisted of 29 airports, 15 of them is located in Mexico (Acapulco is the largest representative). This cluster includes e.g. the US Miami airport or the Puerto airport in the Dominican Republic. In this cluster, the decrease in the number of checked passengers since September 2001 was not that dramatic compared with the previous one. However, the decline at the end of 2003 is more significant. Unlike airports from the second cluster, these airports failed to restore the status of early 2000, yet by this time (see Figure 8d).

The fifth cluster included 121 airports, of which 104 were located in North America, primarily in the USA. Among the most important these were airports in Atlanta, Denver, Houston and Las Vegas, Pheonix. Airports in Mexico City or Canadian Montreal were another important representatives of this cluster. It is very strong decrease in the number of passengers handled after September 2001 which is characteristic

Figure 8 Time series of normalized monthly values of the number of handled passengers at the airports belonging to different clusters



a) 1st cluster

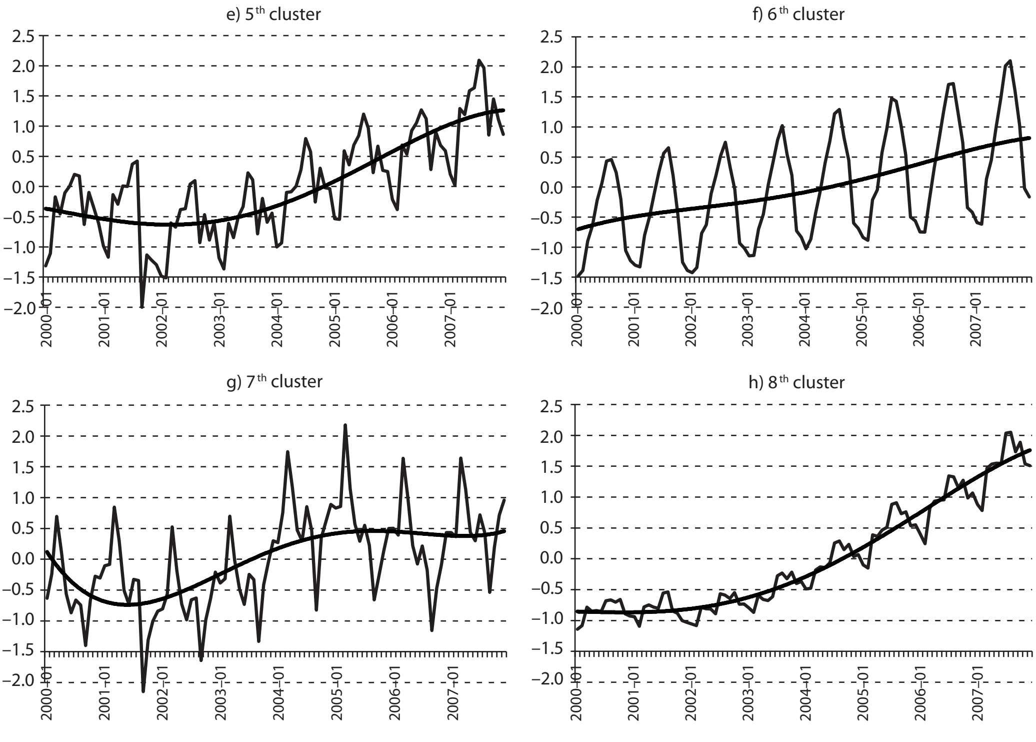
b) 2nd cluster

c) 3rd cluster

d) 4th cluster

Source: Own construction

Figure 8 Time series of normalized monthly values of the number of handled passengers at the airports belonging to different clusters – continuation



Source: Own construction

for airports in this cluster. Thenceforth, a growing trend is evident. As seen from Figure 8e), the current condition is at a higher level than at the beginning of the observed period.

There were 213 airports forming the sixth cluster. The vast majority of them are European. A slightly increasing trend of handled passengers during the whole period is evident for all these airports. In addition, strong seasonal behaviour is manifested here. There is above-average amount of passengers checked during holidays and vice versa strongly below-average amount around the turn of the year (Figure 8f).

In the seventh cluster, there were 26 airports included. Most of them are located in North America. Tampa and Fort Lauderdale in the USA and Mexican Cancun are the most important representatives. The growing trend throughout the incriminated period with a noticeable downturn after September 2001 is characteristic for representatives of this cluster (see Figure 8g).

The eighth cluster consisted of 268 airports. Almost half of them are located in Europe. Most Euro pean airports are from Spain (Madrid Barajs is the largest one) and Italy (Milan Linate). Strongly represented are also Australian airports (Sydney and Melbourne) as well as airports in Asia (Hong Kong, Beijing, Mumbai). Some of Canadian and Mexican airports are included here, too. This cluster also contains African and South American airports and airports in the Middle East. Significantly increasing trend of handled passengers is typical for airports included in this cluster. This is clearly seen in Figure 8h. he description of the cluster is in accordance with the description of development, for example air traffic in China, which is described in the paper (Wang J., Mo and Wang F., 2014).

Table 2 Characterization of the clusters

	1st cl.	2nd cl.	3rd cl.	4th cl.	5th cl.	6th cl.	7th cl.	8th cl.
Number of airports	49	30	98	33	121	213	26	268
Cluster proportion (%)	5.85	3.58	11.69	3.94	14.44	25.42	3.10	31.98
Proportion of transported passengers (%)	2.66	1.66	24.09	1.65	22.89	28.51	2.85	15.68
Prevailing geographic location	South Asia, North America	Europe	USA	Mexico	North America	Europe	North America	Europe, North America

Source: Author's calculations

Table 3 Cross numbers of airports belonging to the individual clusters

	1st cl.	2nd cl.	3rd cl.	4th cl.	5th cl.	6th cl.	7th cl.	8th cl.	Suma
1st cl.	0	1	41	0	118	122	0	0	282
2nd cl.	47	4	19	0	0	91	14	0	175
3rd cl.	0	21	7	0	3	0	0	252	283
4th cl.	2	4	31	33	0	0	12	16	98
Suma	49	30	98	33	121	213	26	268	838

Source: Author's calculations

SUMMARY

Result of our clustering indicates that various global disasters are important factors affecting the number of handled passengers. Within individual clusters it is possible to distinguish airports, which has not been much affected by these disasters, from the airports on which these disasters had either short-term or even long-term impact. The most important characteristics of individual clusters are summarized in Table 2. Cross numbers of airports belonging to the individual clusters are shown in Table 3.

3 DISCUSSION

It is obvious that from the interpretative point of view there is no optimal number of clusters. The resulting interpretation of division into 8 clusters led us to the idea that the other key accident may occur in some clustering results. In literature, the impact of the eruption of the Eyjafjallajökull volcano in Iceland on air transport (April–May 2010) is often quoted. The volcanic cloud created after the explosion, gradually closed European airports between 15 and 23 April as the cloud progressed through Europe. The paper (O'Regan, 2011) describes serious consequences for the operation of air transport in Europe after the eruption of Eyjafjallajokull.

Unfortunately, in this case, the classification was not very successful. Cluster which could be identified as groups of airports affected by this event appeared only when the value of the parameter determining the number of clusters was set at 15. However, in such a large number, clusters cannot be unequivocally interpreted.

What may justify this failure in identification using cluster analysis? The time period affected by the event was very short. Restrictions in aviation lasted only a month. The trend of development of numbers of passengers in neighbouring periods overrode this difference. Therefore, airport affected by the volcanic eruption were not separated well into a separate cluster.

CONCLUSION

We show that cluster analysis can be used to classify airports on the basis of the number of handled passengers per each individual month. It turned out that this classification has quite interesting interpretation. In one case, we managed to classify the world's airports in terms of seasonal development in the number of handled passengers. In a more detailed division we managed to classify the world's airports in terms of their reactions on world events which had an impact on air traffic. It turned out that an important factor resulting in event classification is the sufficient length of the period during which the consequences of this event persisted at particular airports. Conversely, it proved that the regional arrangement is not important for classification in the first place. Moreover, it turned out that if the cluster analysis was used for closer examination of the data structure from different perspectives, it is not good to restrict itself just to division into "ideal" number of clusters.

A side clustering of the airports here requires another type of analyses of available data, especially of trends and changes in them, and the reasons behind eventual changes. Unfortunately, these problems are behind the scope of this paper. On the other hand, our preliminary results based on so called change point analysis as described, e.g. in (Antoch et al., 2007; Antoch et al., 2004; Antoch et al., 2008; Antoch et al., 1997; Antoch and Jarušková, 2017), show very promising results. It appears that if we take into account the fact that these types of data are obtained sequentially in time, the cluster analysis with time series analysis and change point analysis can lead to more profound explanation of studied problems. We will cover this approach elsewhere.

ACKNOWLEDGMENT

This article was supported by Jan Evangelista Purkyně University in Ústí nad Labem.

References

AKAMAVI, R. K., MOHAMED, E., PELLMANN, K. et al. Key determinants of passenger loyalty in the low-cost airline business. *Tourism management*, 2015, 46, pp. 528–545.

ANTOCH, J., GREGOIRE, G., HUŠKOVÁ, M. Tests for continuity of regression functions. *Journal of Statistical Planning and Inference*, 2007, 137(3), pp. 753–777.

ANTOCH, J., GREGOIRE, G., JARUŠKOVÁ, D. Detection of structural changes in generalized linear models. *Statistics & Probability Letters*, 2004, 69(3), pp. 315–332.

ANTOCH, J., HUŠKOVÁ, M., JANIC, A., LEDWINA, T. Data driven rank test for the change point problem. *Metrika*, 2008, 68(1), pp. 1–15.

ANTOCH, J., HUŠKOVÁ, M., PRÁŠKOVÁ, Z. Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*, 1997, 60(2), pp. 291–310.

ANTOCH, J. AND JARUŠKOVÁ, D. Detection of breaks in capital structure. A case study [online]. *Statistika: Statistics and Economy Journal*, 2017, 97(1), pp. 32–43.

CHEN, C.-C., CHEN, J., LIN, P.-C. Identification of significant threats and errors affecting aviation safety in Taiwan using the analytical hierarchy process. *Journal of Air Transport Management*, 2009, 15(5), pp. 261–263.

CUI, Q. AND LI, Y. The change trend and influencing factors of civil aviation safety efficiency: The case of Chinese airline companies. *Safety Science*, 2015, 75, pp. 56–63.

DARDA, P. *The economic importance of passengers for airports.* Master's thesis (in Czech), Univerzita J. E. Purkyně, ESF: Ústí nad Labem, 2014.

DEMPSEY, P., S. Aviation security: The role of law in the war against terrorism. *Columbia Journal of Transnational law*, 2003, 41(3), pp. 649–733.

GRABBE, S., SRIDHAR, B., MUKHERJEE, A. Clustering days and hours with simile airport traffic and weather conditions. *Journal of Aerospace Information Systems*, 2014, 11(11), pp. 751–763.

GRENČÍKOVA, J., KRIŽAN, F., TOLMÁČI, L. Stability and actuality of aviation networks in Bratislava and Prague. *Moravian Geographical Reports*, 2011, 19(1), pp. 17–31.

IATA, *The Impact of September 11 2001 on Aviation* [online]. 2014. [cit. 04.03.14]. <http://www.iata.org/pressroom/documents/impact-9-11-aviation.pdf>.

KRAFT, S. A transport classification of settlement centres in the Czech Republic using cluster analysis. *Moravian Geographical Reports*, 2012, 20(3), pp. 38–49.

LOH, E. The impact of SARS on the performance and risk profile of airline stock. *International Journal of Transport Economics*, 2006, 33(3), pp. 401–422.

LU, Q., CH., ZHANG, J., PENG, Z., R., et al. Inter-city travel behaviour adaptation to extreme weather events. *Journal of Transport Geography*, 2014, 41, pp. 148–153.

O'REGAN, M. On the edge of chaos: European aviation and disrupted mobilities. *Mobilities*, 2011, 6(1), pp. 21–30.

ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. Clusters number determination and statistical software packages. *DEXA 2008: 19th International Conference on Database and Expert Systems Applications*, 2008, pp. 549–553.

WANG, J., MO, H., WANG, F., Evolution of air transport network of China 1930–2012. *Journal of Transport Geography*, 2014, 40 (October), pp. 145–158.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, 1996, 25(2), pp. 103–114.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: A new data clustering algorithms and its applications. *Journal of Data Mining and Knowledge Discovery*, 1997, 1(2), pp. 141–182.