CHALLENGES IN EUROPEAN AEROSPACE
5TH CEAS AIR & SPACE CONFERENCE

CEAS

Clustering of airport cities and cluster dynamic for the air passenger demand forecasting model based on a socio-economic scenario.

Ivan Terekhov
PhD candidate, German Aerospace Center (DLR) Air Transportation Systems
Blohmstr. 18, Hamburg, 21079, Germany
ivan.terekhov@dlr.de
Volker Gollnick (Head of Institute, German Aerospace Center (DLR) Air Transportation systems)

ABSTRACT

This study presents methods of grouping cities into clusters by their socio-economic indicators and tracing changes in the content of cluster within a socio-economic scenario. For cities' grouping, three main clustering approaches have been analyzed: hierarchical, exclusive and probabilistic clustering. Analyzing advantages and disadvantages of these approaches, probabilistic clustering of *normal mixture* has been chosen to separate cities from the air passenger demand (APD) forecasting model. Three parameters, a city's GDP, population and GDP per capita, have been used for clustering. Utilizing these parameters and based on special metrics, separation of cities into 9 clusters has been chosen. Furthermore, this study introduces the "*cluster dynamic*". The cluster dynamic defines how cities are allocated to the various clusters at a given point in time within a socio-economic scenario.

1 INTRODUCTION

The modular environment AIRCAST[1,2] is designed to forecast changes in the air transportation system (ATS) utilizing socio-economic scenarios. An air passenger demand (APD) forecast model of 'origin-destination air travel passenger demand between city-pairs' on a global level called D-CAST[1] is the first layer in a chain of models within AIRCAST[2]. The APD model has two steps: forecasting the topology of the APD network between cities worldwide and calculating demand on existing and new connections. As shown in existing studies[3,4], a partition of elements into groups improves link prediction performance and, thereby, increases the accuracy of the APD topology forecast between cities[5]. Furthermore, studies[6,7] show that the APD has a clear correlation with economic and social indicators. Thus, it is likely that the process of the APD generation is different for different cities. Therefore, these cities could be allocated to a number of groups by their socio-economic indicators, where cities in each group possess similar patterns. Furthermore, within the forecast period it is likely that the placement of cities within particular groups will change as the various city indicators change. Thus, the aim of this paper is it to define qualitative and quantitative features of these groups in the base year (the starting point for forecasting) and the dynamic by which cities change groups.

2 CLUSTERING

In AIRCAST the forecast of future development of the ATS on city level is based on socio-economic scenarios. These scenarios contain indicators including GDP and population of cities and a global average oil price. In AIRCAST the base year is year 2012. For this year, 4,435 cities with at least one airport have been obtained utilizing the ADI[8] database. This number of cities remains fixed for the duration of the scenario. For all cities the GDP[9,10], population[11,12] and geographical coordinates[13,14] have been retrieved from various databases[1]. The problem of allocating cities into groups can be presented as a clustering task. The goal of clustering is to determine a finite set of groups (clusters) to describe a dataset according to similarities among its elements[15,16]. This allows the determination of appropriate methods for

forecasting the APD for each cluster pair and, thereby, increasing the accuracy of the whole APD forecast method.

2.1 Clustering methods

There are various clustering methods. They can be divided into two main groups: hierarchical and partitional. Hierarchical clustering (HC) constructs clusters based on their proximity and forms a hierarchical tree[17]. At first, HC treats each element as a cluster. Then, the two nearest clusters are combined and form a new cluster. This procedure continues until there is only one cluster containing all elements[18]. As a result, there is a hierarchical tree of clusters also known as a dendrogram (Fig.1). Thus, HC builds a system of nested clusters. Using this method, clusters could be retrieved by cutting the dendrogram at different levels. However, HC is appropriate for small sets of data, up to several thousand elements. The method is very sensitive to noise and outliers in the data. Furthermore, HC algorithms are not capable of correcting any previous potential misclassification. Once an object is assigned to a cluster, it will not be considered again[17]. Moreover, HC does not work well in overlapping areas[18] (in these areas, elements from several clusters share the same space).

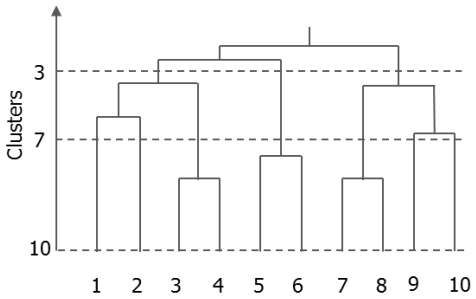

Fig.1. Hierarchical clustering

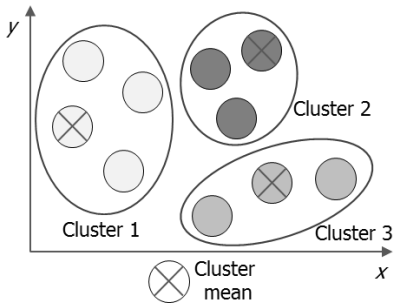

Fig.2. Exclusive clustering

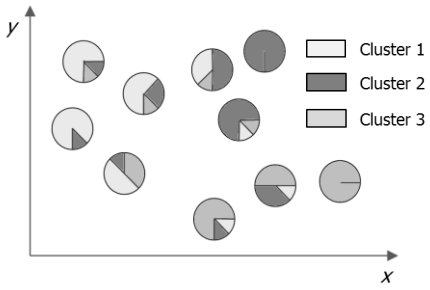

Fig.3. Probabilistic clustering

In partition methods two approaches can be highlighted: exclusive clustering (EC) and probabilistic clustering (PC). For both approaches the number of clusters has to be determined in advance. In an EC approach, elements are only allocated to certain clusters and then can no longer be included in others (hard clustering). One of the most commonly used algorithms in the EC approach is *k-means* algorithm, which is based on a certain number of clusters defined in advance. The main idea of *k-means* is to define means for every cluster (Fig.2). This algorithm picks the randomly chosen number of elements in the initial set equal to the number of clusters defined in advance. These randomly chosen elements are assumed to be cluster means. This is an iterative process. The algorithm performs recalculations of cluster means until a specified criterion is met. The affiliation to clusters is defined for every element in the set by defining the minimum distance between means and elements. The *k-means* algorithm is appropriate for large sets of data, up to hundreds of thousands elements. However, the appropriate number of clusters is unknown. It is necessary to specify number of clusters before one starts the algorithm[18]. The algorithm is sensitive to the selection of the initial partition[19]. In addition, it does not perform well in the case of overlapping areas[18].

With a PC approach, each cluster can be present as a parametric distribution. Thus, the initial set of elements is modeled by a mixture of these distributions. In contrast to *k-means*, where elements are deterministically assigned to one and only one cluster (hard clustering), a PC approach assigns elements to clusters with certain probabilities (soft clustering). The most commonly used algorithm in PC is a *normal mixture* or a *mixture of Gaussians*. The *normal mixture* algorithm is similar to *k-means*. The *normal mixture* uses *expectation-maximization* (EM) algorithm where on *expectation* step (*E-step*)

expected values of the cluster membership for each element is calculated. Here probabilities for all elements are calculated. Then, *maximization* step (*M-step*) recalculates the parameters of each Gaussian to maximize the probabilities found on *E-step*. These steps repeat until convergence. The *normal mixture* algorithm based on a probabilistic approach performs well in overlapping areas[18]. However it is sensitive to the selection of the initial partition[17].

The APD forecast model contains 4,435 cities. For every city numerical attributes are obtained from various databases: GDP and population for 2012 and geographical coordinates. All economic indicators within the study are adjusted to 2005 US dollars[i]. Cities' distributions by population and GDP are presented in Fig.4 and Fig.5 respectively. Based on these distributions, cities' quantiles by population and GDP are presented in Tab.1 and Tab.2 respectively. Cities possess various socio-economic indicators, but they are not separated well, as can be seen in Fig.6. Most cities are concentrated in a small area.

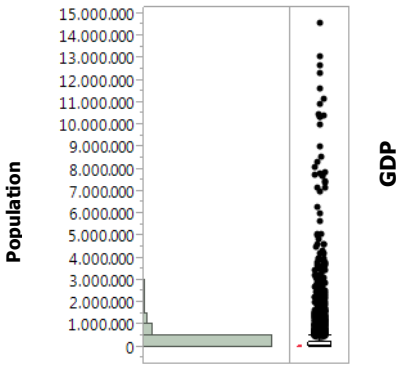

Fig.4. City distribution by population

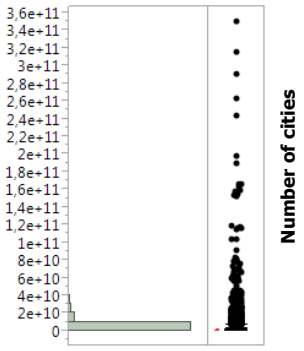

Fig.5. City distribution by GDP

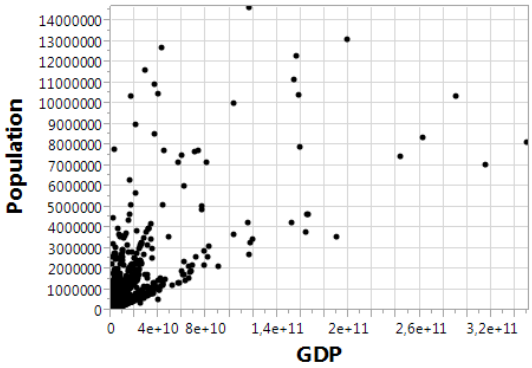

Fig.6. City distribution by population and GDP

Percent of cities	Quantiles	Population	City
100.0%	maximum	14,608,512	Shanghai, China
75.0%	quartile	206,570	Annaba, Algeria
50.0%	median	50,675	Mweka, DR Congo
25.0%	quartile	7,716	Fort Dix, US
0.0%	minimum	2	Portage Creek, US

Tab.1 City quantiles by population

Percent of cities	Quantiles	GDP, billions	City
100.0%	maximum	350	New-York, US
75.0%	quartile	3	Pekanbaru, Indonesia
50.0%	median	0.744	Arcata, US
25.0%	quartile	0.103	Lakselv, Norway
0.0%	minimum	0.00007	Kadhdhoo, Maldives

Tab.2. City quantiles by GDP (indicated here in constant 2005 US dollars)

Thus, cities in this area have quite similar values for GDP and population and lay in overlapping areas. It is difficult then to understand exactly to which group they should be assigned. Despite the simplicity of retrieving clusters, HC algorithms are not capable of correcting potential previous misclassification. Once an object is assigned to a cluster, it will not be considered again. In other words, if a city is assigned at the beginning of the algorithm to one cluster, it will not be taken into account on subsequent clustering steps. EC algorithms are considered to be a form of "hard clustering". They do not work well in the overlapping areas. In other words, if a city is assigned to a cluster, it can no longer be included in others. The PC approach assigns elements to clusters with certain probabilities. It works well when clusters have different sizes and correlation within them. HC and EC algorithms perform well when clusters are well separated, but they fail in overlapping areas[18]. Thus, for clustering cities in the APD model, a PC

[i] In this study, one of the main used socio-economic scenario is the Randers scenario. In this scenario all economic are adjusted to 2005 US dollars. Thus, to unify all economic indicators within the study, they are adjusted to 2005 US dollars.

algorithm of *normal mixture* is used. However, for this type of clustering it is necessary to define the appropriate number of clusters.

2.2 Application of normal mixture clustering

The PC of *normal mixture* is chosen to group cities into clusters. As discussed above, the process of APD generation will likely be different for different city clusters. Thus, it is important to define the appropriate number of clusters as well as the number of city parameters for clustering. Utilizing a few parameters could lead to high bias and missed opportunities for cluster insight. Such clustering is not flexible enough to describe the sample well. In contrast, clustering with too many parameters will not be able to fit observed data well, but will be too closely tailored to it. Such models may generalize poorly.[20] AIRCAST socio-economic scenarios contain cities GDP and population. Based on these parameters it is possible to add one more parameter - GDP per capita. This parameter allows a *normal mixture* algorithm to describe clusters with higher precision. Thus, to define the number of city groups with similar socio-economic indicators, clustering in this study is done by utilizing city GDP, population and GDP per capita.
For *normal mixture*, the number of cluster must be set in advance. This is a typical issue for a *normal mixture* clustering approach. It is solved through measurements of standard metrics for different numbers of clusters. In this study two standard metrics are used: the Bayesian information criterion[21] (BIC) and the Akaike information criterion[22] (AIC). Both these metrics are penalized-likelihood information criteria. BIC and AIC choose the model with a particular number of clusters which demonstrates the best penalized log-likelihood. BIC and AIC is a variation of a penalty weight A_n in the information criterion:

$$IC(k) = -2l + A_n p \tag{1}$$

Where k is number of clusters; l is the log-likelihood; p is the number of parameters in the model. For AIC $A_n = 2$, and for BIC $A_n = \ln(n)$; n is sample size. BIC and AIC penalize more for models with additional parameters. The penalty of BIC depends on the sample size and it is usually more "heavy" then AIC. The number of clusters n minimizing BIC and AIC is considered to be the optimal number of clusters for a given set. For clustering, 20 independent restarts of the estimation process with different starting values are used. This avoids the problem of finding a local solution. The maximum number of iterations of the convergence stage of the EM algorithm is 200. The convergence criterion is the difference in the likelihood at which the EM interactions stops and it is equal to 0.00000001. BIC and AIC for 4,435 cities in the base year 2012 of the APD forecasting model is presented in Fig.7. Clustering of these cities is performed based on their GDP, population and GDP per capita.

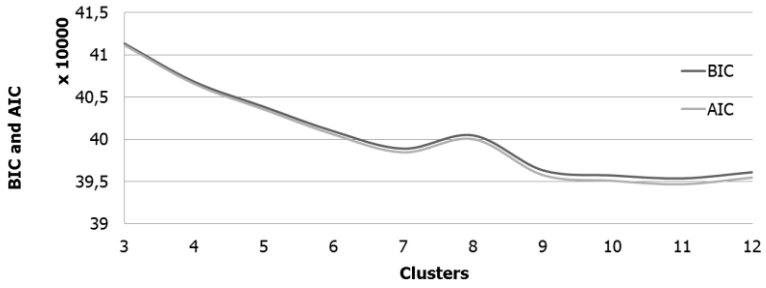

Fig.7. BIC and AIC metric for the different number of clusters for the city set of the ADP forecast model

Based on AIC and BIC in Fig.7, separation into 11 clusters provides the best results. However, some means of these clusters are close to each other. It is difficult to interpret the meaning of these means.

Thus, three separations of the smallest AIC and BIC into 9, 10 and 11 clusters have been considered. Cluster means of these separations are depicted in Fig.8 based on population and GDP per capita. As it can be seen each time, the clustering algorithm detects groups of cities with the largest socio-economic indicators. The main changes in separations are in city groups with populations of less than 1 million. The cluster means for these separations are depicted in Fig.9, Fig.10 and Fig 11 based on their population and GDP per capita.

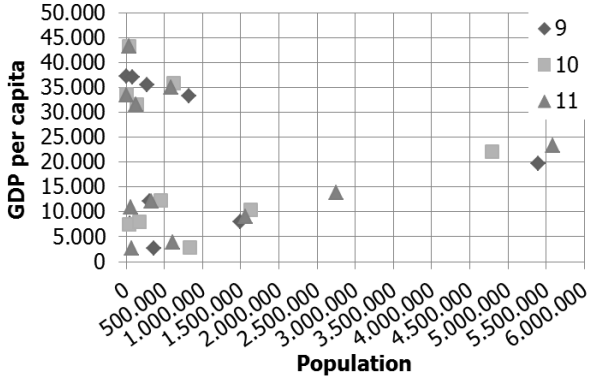

Fig.8. Cluster means of separation into 9, 10 and 11 clusters by population and GDP per capita

For separation into 11 clusters, a few means are in close proximity to each other. These groups of cities have relatively small populations with high GDP and GDP per capita. Furthermore, there are two proximate city groups of small cities with small GDP and GDP per capita (Fig.9). The same proximity groups are for separation into 10 clusters (Fig.10). However, the situation is different for separation into 9 clusters (Fig.11). These cluster means are clearly distinguished from each other and are easily interpreted. The performance of the 9 clusters is good enough, and separation to 10 and 11 clusters do not add much. Thus, despite that it does not demonstrate the best AIC and BIC, for simplicity separation into 9 clusters has been chosen for this study.

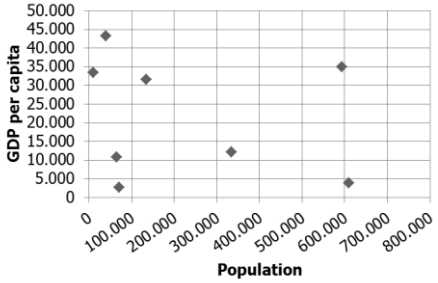

Fig.9. Cluster means for separation into 11 clusters of citieswith less than 1 million inhabitants.

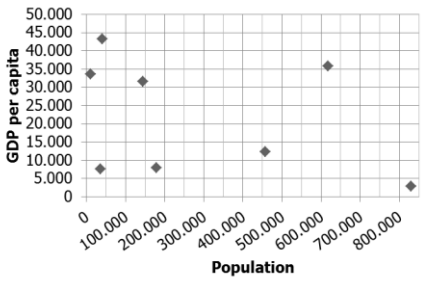

Fig.10. Cluster means for separation into 10 clusters of cities with less than 1 million inhabitants.

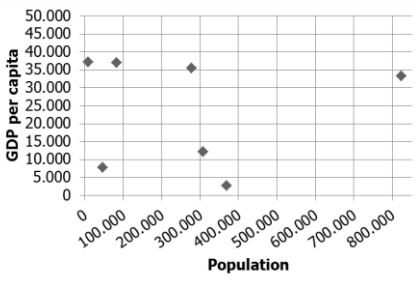

Fig.11. Cluster means for separation into 9 clusters of cities with less than 1 million inhabitants.

All 9 clusters centres are well separated. These clusters cover "small", "middle" and "big" cities by population and "poor", "middle class" and "rich" cities by wealth. Based on these 9 clusters, the APD network in 2012 can be presented as a set of 45 cluster pairs. For the purpose of the study, cluster

names derived from population and per capita GDP of cluster means were adopted. Tab. 3 reflects the number of cities in each cluster, cluster means and cluster names. For the PC of *normal mixture* a complex formula to find probabilities of every city affiliation to each cluster is obtained. Thus, utilizing this formula it is possible to retrieve probabilities for city affiliations to different clusters for developing socio-economic indicators, and, thus, trace how cities are changing their clusters within a given time period.

Cluster #	Population	GDP, billions	GPD per capita	Number of cities in cluster	Proportion	Size	Wealth
1	8,520	0.3	37,134	1,453	0.32191	Very small	Rich
2	47,010	0.3	7,729	1,055	0.22774	Small	Poor
3	824,546	27	33,219	108	0.02487	Big	Rich
4	307,440	3	12,066	417	0.09684	Middle	Middle
5	5,394,129	77	19,767	76	0.01748	Megacities	
6	82,790	2	37,010	565	0.13312	Small	Rich
7	1,493,549	11	8,032	238	0.05451	Big	Poor
8	278,644	9	35,547	207	0.04738	Middle	Rich
9	369,340	1	2,744	316	0.07615	Middle	Poor

Tab.3. Clusters centers, city distribution among clusters and cluster names. GDP and GDP per capita indicated here in constant 2005 US dollars.

3 CLUSTER DYNAMIC

Over the forecast period, the socio-economic indicators of the cities change. These changes affect the probability of membership of a given city to a certain cluster. This process reveals the changes over time of city distributions within the clusters. Thus, this study introduces the "*cluster dynamic*". The cluster dynamic is a method of calculating the probability that a given element (city) will appear within a given cluster at a given point in time. This method is how the cities are allocated to the various clusters in any given forecast year, based on socio-economic indicators of cities. During the forecast period, cluster centers remain fixed as in the 2012 base year and do not change. In other words, in this study, affiliation calculations are made from a 2012 perspective.

In this study the Randers socio-economic scenario[23] from 2012 to 2050, with time slices every 5 years since 2015, is used. Utilizing the *normal mixture* PC formula from 2012, probabilities of city affiliations are defined from 2015 up to 2050, with 5 year step for 8 time slices (Fig.12).

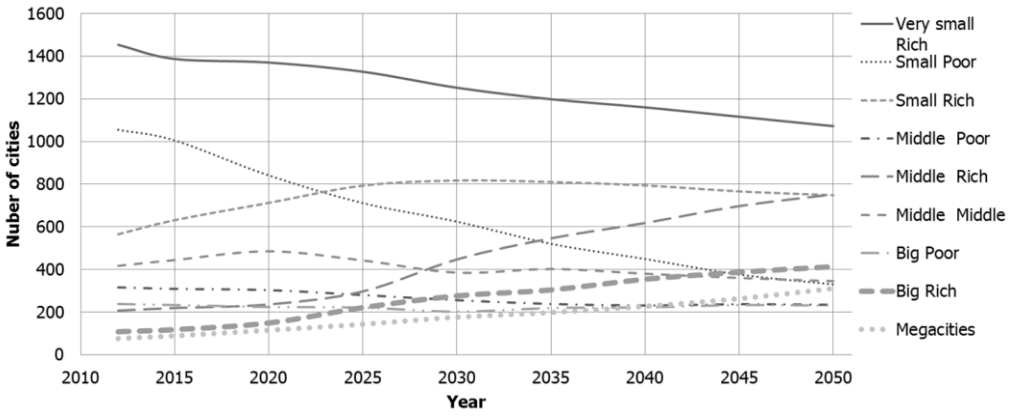

Fig. 12. Cluster dynamic based on Randers scenario

In this study the number of cities is constant within the forecast period and contains 4,435 cities from the base year. Based on the Randers scenario, *"Small Rich"* and *"Very small Rich"* clusters show significant decrease in the number of cities. This is because, in general, the Randers scenario is a positive scenario, where almost all cities demonstrate population and GDP growth. Thus, the cluster dynamic shows cities moving into more "powerful" clusters. As a consequence, *"Middle Middle"*, *"Big Rich"* and *"Megacities"* show significant increases. A transition diagram of cities in clusters between the base year 2012 and the last year of the scenario, 2050, are presented in Fig.13.

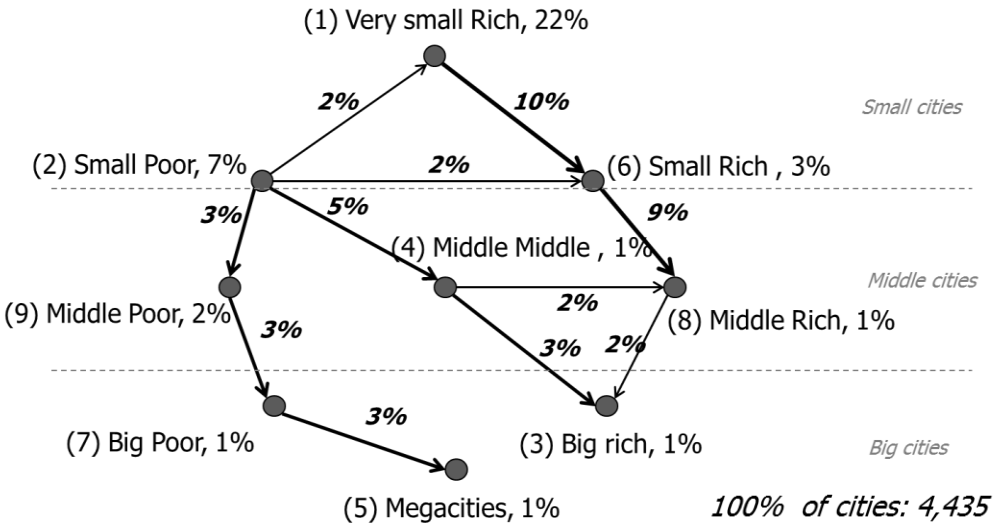

Fig.13. Transition diagram between the base year 2012 and the last year of the scenario 2050

The diagram shows 9 clusters on three levels by city population: small, middle and big. Cities that remain in clusters and change the clusters are indicated in percentage of the total number of cities. Arrows demonstrate to which clusters cities are moving. The diagram shows transitions between clusters with more than 1% of cities. Based on this diagram, tendencies can be seen showing the moving of cities between clusters. It is possible to trace three main paths: from cluster *"Small Poor"* to *"Megacities"*, from "*Small Poor*" to *"Big Rich"* and from *"Very small Rich"* to "Big Rich*"*. Cities in the *"Small Poor"* cluster show the highest diversity. These cities move to 4 different clusters. All cities have tendencies to move to the end point clusters of *"Megacities"* and *"Big Rich"*. Thereby, it is shown that there is a clear correlation with the Randers scenario, which has a positive tendency.
Since the process of APD generation is different in different clusters, the content of clusters have a significant influence on the accuracy of the APD forecasting model. The introduction of the cluster dynamic provides a comprehensive approach to trace these changes within and between clusters, based on a given socio-economic scenario.

4 CONCLUSION

This study presents qualitative and quantitative features of different groups of cities in the base year and the dynamic of cities moving between groups. This study presents methods of grouping cities into clusters by their socio-economic indicators and tracing changes in the content of cluster within a socio-economic scenario. For city grouping, three main clustering approaches have been considered: hierarchical, exclusive and probabilistic clustering. By analyzing the advantages and disadvantages of

these approaches, probabilistic clustering of *normal mixture* has been chosen to separate cities from the APD forecasting model. This clustering approach performs better than others in overlapping areas. This is essential in the case of the APD forecast model cities. Clustering is based on socio-economic indicators of cities that including city GDP and population figures. Thus, the three parameters of city GDP, population and GDP per capita have been defined to fit observed data. Utilizing these parameters and special metrics AIC and BIC, separation into 9 clusters has been chose. Notwithstanding that the separation does not demonstrate the best AIC and BIC, cluster means are distinguished well from each other and the number of city parameters for clustering are easily interpreted. Furthermore, this study introduces the "*cluster dynamic*". This method demonstrates how the cities are allocated to the various clusters at a given point in time within a socio-economic scenario.
Clustering is an important part of the APD forecasting model. The results of this study have significant impact on the accuracies of link predictions in the APD network[5]. Moreover, clustering results can help to understand the air passenger demand generation within and between different groups of cities in further studies. The next step in the study is to apply the cluster dynamic method to various scenarios within AIRCAST and, based on this cluster model, trace the changes in the APD on city level worldwide.

---

[1] Terekhov, I., Ghosh, R., Gollnick, V. "A concept of forecasting origin-destination air passenger demand between global city pairs using future socio-economic development scenarios", *53rd AIAA Aerospace Sciences Meeting*, Kissimmee, Florida, USA, 2015.

[2] Ghosh, R., Terekhov, I., "Future Passenger Air Traffic Modelling: Trend Analysis of the Global Passenger Air Travel Demand Network", *53rd AIAA Aerospace Science Meeting*, Kissimmee, Florida, 2015.

[3] Zheleva, E., Golbeck, J., Kuter, U., "Using Friendship Ties and Family Circles for Link Prediction", *Advances in Social Network Mining and Analysis Lecture Notes in Computer Science*, Vol. 5498, 2012, pp. 97-113.

[4] Lü, L., Zhou, T., "Link prediction in complex networks: A survey", *Physica A*, Vol. 390, 2011, pp. 1150-1170.

[5] Terekhov, I., Evans, A., Gollnick, V. "Forecasting global air passenger demand network using weighted similarity-based algorithms", *the 19th ATRS World Conference*, Singapore, 2015.

[6] Boeing, Current Market Outlook 2013-2032, USA, 2013, http://www.boeing.com/assets/pdf/commercial/cmo/pdf/Boeing _Current_Market_Outlook_2013.pdf [cited 19.11.2014].

[7] Dray L., Evans A.D., Reynolds T., Schäfer A., 2010. "Mitigation of Aviation Emissions of Carbon Dioxide: Analysis for Europe", *Transportation Research Record*, 2177, pp. 17-26.

[8] Sabre Airline Solutions, Aviation Data Intelligence, http://www.sabreairlinesolutions.com /home/software_solutions/airports/, [cited 19.11.2014].

[9] UN, National Accounts Main Aggregates Database, http://unstats.un.org/unsd/snaama/dnllist.asp, [cited 19.11.2014].

[10] The World Bank, World Bank Open Data, http://data.worldbank.org/indicator/NY.GDP.MKTP.CD, [cited 19.11.2014].

[11] UN, World population Prospects: The 2012 Revision, http://esa.un.org/unpd/wpp/Excel-Data/population.htm, cited 19.11.2014].

[12] MaxMind, Free World Cities Database, https://www.maxmind.com/en/worldcities, [cited 19.11.2014].

[13] Our Airports, http://ourairports.com/data/, [cited 19.11.2014].

[14] OpenFlights, Airport database, http://openflights.org/data.html, [cited 19.11.2014].

[15] Hruschka, R.E., Freitas, A.A., de Carvalho A.L., "A Survey of Evolutionary Algorithms for Clustering", *IEEE Transactions on Systems, Man, and Cybernetics—part C: Applications and Reviews*, vol. 39, no. 2, March 2009

[16] Berkhin, P. "A survey of clustering data mining techniques." *Grouping multidimensional data. Springer Berlin Heidelberg*, 25-71, 2006.

[17] Xu, R., Wunsch, D, "Survey of clustering algorithms", *Neural Networks, IEEE Transactions on, 16(3)*, 645-678, 2005.

[18] SAS Institute Inc., "JMP® 11 Multivariate Methods", *Cary, NC: SAS Institute Inc.*, 2014.

[19] Jain, A.K., Murty, M.N., Flynn, P.J., "Data clustering: a review", *ACM computing surveys (CSUR)*, 31(3), 264-323, 1999.

[20] Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria. *The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University*.

[21] Schwarz, G. (1978). "Estimating the dimension of a model". *The annals of statistics*, *6*(2), 461-464.

[22] Akaike, H. (1974). "A new look at the statistical model identification". *Automatic Control, IEEE Transactions on*, 19(6), 716-723.

[23] Randers, J., "2052: A global Market Forecast for the Next Forty Years", *Chelsea Green Publishing*, White River Junction, Vermont, 2012.