Forecasting global air passenger demand network using weighted similarity-based algorithms

Authors Affiliation and Identification:

Ivan Terekhov, PhD candidate, Department System Analysis Air Transport.

Antony Evans, Associate Research Scientist, University of California Santa Cruz, University Affiliated Research Center, Moffett Field, CA 94035.

Volker Gollnick, Head of the German Aerospace Center (DLR) Air Transportation Systems.

Presenter: Ivan Terekhov

Address: German Aerospace Center (DLR), Air Transportation Systems, Blohmstr. 18, Hamburg, 21079, Germany.

Telephone Number: +49 531 295 3848

E-mail address: ivan.terekhov@dlr.de

Abstract

The aim of this study is to define an appropriate approach to forecast the appearance of the air passenger demand between cities worldwide. For air passenger demand link forecasting a weighted similarity-based algorithm is used, with an analysis of nine indices. The weighted resource allocation index demonstrates the best metrics. The accuracy of this method has been determined through a comparison of modeled and known data from three separate years. The known data was used to establish boundaries when applying the similarity-based algorithm. As a result, it was found that a weighted resource allocation index, with defined boundaries, should be utilized for link prediction in the air passenger demand network. Furthermore, it is shown that grouping cities within the air passenger demand network, based on socio-economic indicators, increases the accuracy of the forecast.

Keywords: demand, network, weight, forecast, scenario

Corresponding Authors: Ivan Terekhov

1. Introduction

The modular environment AIRCAST[1,2] aims to forecast future development of the air transportation system (ATS) based on socio-economic scenarios. AIRCAST allows to simulate a range of possible outcomes for the future ATS and assess, for example, the impact of new technology on the number of demand passengers or the size and number of aircraft on particular routes. 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]. In D-CAST, the APD model forecasts the number of passengers as well as changes in the number of connected cities within the forecast period. This paper aims to define an appropriate approach to forecasting the appearance of APD between cities worldwide.

The APD network is a dynamically evolving network in time. This network contains a number of cities (nodes) with links between them. In this study the APD network is considered as an undirected network[1]. The APD network is a weighted network. In other words, each link is characterized by a parameter or a set of parameters. As shown, the APD has interdependences with economic and social indicators[3,4]. Thus, the weight of a link could be considered as a combination of socio-economic indicators between cities in pairs. During the forecasting period, the socio-economic indicators of cities vary. Therefore, the weighting of links is also changing. This variation in weightings over time has an impact on the APD network and, accordingly, the topology of the network will likely change. For example, where the socio-economic indicators of cities (e.g. GDP, population and oil price) show a rapid increase, it is likely that there will appear a number of connected cities with a significant APD where no APD connections previously existed.

There are three main groups of link prediction methods[5] for forecasting connections in the network: similarity-based algorithms, maximum likelihood (ML) and probabilistic models (PM). Similarity-based algorithms are divided into local, global and quasi-local indices[5]. Similarity-based algorithms are the mainstream class of algorithms of link prediction. ML methods and PM are complex and very time consuming. ML is able to handle networks with up to a few thousand nodes in a reasonable time[5]. Furthermore, ML methods do not demonstrate the best accuracy[5]. Mostly, studies consider link prediction in non-weighted networks. Studies on link prediction in weighted networks are mainly conducted utilizing weighted local similarity indices[6,7]. In addition, the APD network is a high clustered network as shown by Ghosh and Terekhov[2]. For highly clustered networks, the common-neighbor-based indices demonstrate relatively good prediction with low complexity[5]. Thus, in this study, only weighted local similarity indices are considered.

The underlying principle of weighted and non-weighted indices of similarity-based algorithms is the same. These algorithms assign a score to each non-existing link in a given network. Then, the links are ranked in descending order according to their score. Links with the highest score should appear in the network. Here, two significant problems arise. In the network one index can perform well and another fail[5]. Thus, the first problem is to define which weighted local similarity index shows the best performance in the APD network. The second problem is to define a criterion for adding new connections to the network with the highest score from the top

of the ranking list. In other words, a boundary condition in the ranking list of non-existing links has to be defined: links from the ranking list between the first link and a boundary link will be added to the network.

In addition, as shown by Zheleva et al[8], the combination of network structure, node attributes, and node community features improve link prediction performance. In the APD network, the network structure and node attributes are known. For node communities, cities are distributed to groups by proximity of their socio-economic indicators. For example, cities with large GDP and population are united to the *big-rich* group and cities with large population and small GDP are united to the *big-poor* group. Since cities in general possess different socio-economic indicators in these groups (clusters)[1], the process of link appearance in each cluster pair of the APD network could be different. Thus, similarity-based algorithm which shows the best performance in one cluster is probably different in another cluster. For example, different weighted similarity algorithms could perform better between *big-rich* cities and *small-poor* cities, than between *megacities* and *middle-rich* cities. Furthermore, it is likely that every cluster pair has its own boundary. In this paper, the performance of similarity-based algorithms for each cluster pair will be analyzed. The boundary for each cluster pair will be defined utilizing the algorithm with the best performance.

Two standard metrics are used to identify the appropriate index for each cluster pair: the area under the receiver operating curve[9] (AUC) and precision[10]. In this study, these metrics have been applied to the APD topology for 2009. For boundary identification, a set of forecasts of the APD network has been made: from 2009 to 2010, from 2010 to 2011 and from 2011 to 2012.

For 2009, 2010, 2011 and 2012 origin-destination city pairs worldwide (topology) have been obtained from Sabre Airport Data Intelligence[11] (ADI) database. For link weighting calculation, GDP[12,13], population[14,15] and geographical coordinates[16,17] of the cities have been obtained from various databases[1]. For the average air fare between cities a simple air fare model[2] has been adopted.

	Cluster mean				
Cluster	**Population**	**GDP, $**	**GDP p/c, $**	**Size**	**Wealth**
1	8,519	3.07E+08	37,134	Very small	Rich
2	47,009	3.79E+08	7,728	Small	Poor
3	824,546	2.71E+10	33,219	Big	Rich
4	307,440	3.74E+09	12,066	Middle	Middle
5	5,394,129	7.74E+10	19,767	Megacities	
6	82,789	2.97E+09	37,009	Small	Rich
7	1,493,548	1.16E+10	8,032	Big	Poor
8	278,644	9.73E+09	35,546	Middle	Rich
9	369,339	1.1E+09	2,743	Middle	Poor

Tab. 1: Clusters centers and cities distribution among clusters in 2012. GDP and GDP per capita indicated here in constant 2005 US dollars.

2. Definition of the weighted similarity-based algorithm for the APD network

The initial set of 4,435 cities obtained from the ADI data base has been divided into 9 clusters, based on their socio-economic indicators[1] in 2012: GDP, city population and GDP per capita. All economic indicators within the study are adjusted to 2005 US dollars. Tab. 1 reflects the number of cities in each cluster and cluster means (cluster centers). For the purposes of the study, short hand cluster names, derived from cluster means (population, GDP and per capita GDP), have been adopted (i.e., *very small rich* cities, *small poor* cities, etc.).

In the APD network every cluster is defined as a set of cities and weighted connections. These connections link cities in one cluster with cities in other clusters and link cities within a cluster. Weights in this study are considered as a combination of average air fare[2], distance between cities and main socio-economic indicators such as city GDP and city population. The weight on the connection between cities *x* and *y* is presented as follow:

$$w_{xy} = (GDP_x * GDP_y)^{\alpha} * \left(Pop_x * Pop_y\right)^{\beta} * \left(d_{xy}\right)^{\gamma} * \left(AF_{xy}\right)^{\delta} * \varepsilon + \theta \qquad (1)$$

Where $GDP_{x,y}$ is a gross domestic product of city *x* and *y*; $Pop_{x,y}$ is a population of city *x* and *y*; d_{xy} is a distance between city *x* and *y*; AF_{xy} is an average air fare between city x and y; $\alpha, \beta, \gamma, \delta$ are elasticities of GDP, population distance and average air fare respectively; ε is a dummy variable; θ is a free parameter. In this study it is assumed that $\alpha = 1$, $\beta = 1$, $\gamma = -1$, $\delta = -1$, $\varepsilon = 1$ and $\theta = 0$. Thus, the equation (1) turns to a variation of the Newton's gravity model and the weight could be interpreted as an abstract attractive force between cities. Furthermore, the gravity model has been used in number of studies[4,18] to predict APD between city pairs.

The average air fare in turn is presented by Ghosh and Terekhov[2] as a simple model of oil price and distance based on historical data:

$$AF_{xy} = (OP * 2 * 10^{-4} + 0.0653) * d_{xy} + 140 \qquad (2)$$

Where OP is an average oil price in a given year. Based on equation (2) and assumptions in equation (1) the weight between cities *x* and *y* could be presented as:

$$w_{xy} = \frac{GDP_x * GDP_y * Pop_x * Pop_y}{(OP * 2 * 10^{-4} + 0.0653) * d_{xy}^2 + 140 * d_{xy}} \qquad (3)$$

Within this study nine indices of similarity-based algorithms have been analyzed. Based on study of Lü and Zhou[5] the weighted common neighbors (WCN), weighted Adamic-Adar index (WAA) and weighted resource allocation index (WRA) have been applied to the APD network. Also similarity indices for unweighted networks have been adapted for weighted networks utilizing the proposed simple method by Murata and Moriyasu.[6] These indexes are the weighted Salton index (WSA), weighted Sorensen index (WSO), weighted hub promoted index (WHPI), weighted hub depressed index (WHDI), weighted Leicht-Holme-Newman index (WLHN) and weighed preferential attachment index (WPA). These similarity indexes are presented in Tab.2.

Two standard metrics *AUC*[5,7] and *precision*[5,7] have been used to determine the accuracy of each index. Initially, for an undirected weighted network, all existing and non-existing links are known. From this set of existing links a group of links – the probe set – is excluded. The remaining existing links are the testing set. The score of each index in the network formed by the testing set is calculated for all non-existing links and the probe set. *AUC* shows the probability that a randomly chosen link from the probe set has a higher score than a randomly chosen link from the set of non-existing links. By Lü and Zhou[5] *AUC* is as follows:

$$AUC = \frac{n' + 0{,}5 * n''}{n} \tag{4}$$

Where n' shows how many times links from the probe set have a higher score then randomly chosen links from the non-existing links set. n'' denotes how many times links from the probe set have the same score as randomly chosen links from the non-existing links set. And n is a number of independent comparisons. For the *precision* metric, the set of probe links and non-existing links is ordered in descending order according to their scores. From this list the top-L links are selected as the predicted once. Among these links, there are L_r links are right (links from the probe set). The *precision* is a ratio of L_r to L. Thus, higher *precision* means higher prediction accuracy[5]. Both metrics are numbers between 0 and 1. The closer the metric is to 1, the better the performance of the index in a given network.

Weighted common neighbors (WCN)	$s_{xy}^{WCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x,z) + w(z,y)$	(5)
Weighted Adamic-Adar (WAA)	$s_{xy}^{WAA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\log(1 + s(z))}$	(6)
Weighted Recourse Allocation (WRA)	$s_{xy}^{WRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\mathrm{s(z)}}$	(7)
Weighted Salton index (WSA)	$s_{xy}^{WSA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\sqrt{s(x) * s(y)}}$	(8)
Weighted Sorensen index (WSO)	$s_{xy}^{WSO} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{2(\, w(x,z) + w(z,x)\,)}{\mathrm{s(x) + s(y)}}$	(9)
Weighted hub promoted index (WHPI)	$s_{xy}^{WHPI} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\min\{\mathrm{s(x)}, \mathrm{s(y)}\}}$	(10)
Weighted hub depressed index (WHDP)	$s_{xy}^{WHDI} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\max\{\mathrm{s(x)}, \mathrm{s(y)}\}}$	(11)
Weighted Leicht-Holme-Newman index (WLHN)	$s_{xy}^{\mathrm{WLHN}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{s(x) * s(y)}$	(12)
Weighed preferential attachment index (WPA)	$s_{xy}^{WPA} = s(x) * s(y)$	(13)

Tab.2 Weighted similarity-based algorithm indexes

In this study for *AUC* and *precision* calculations, the APD network of 2009 has been utilized. For this year, 3,919 cities have been obtained. These cities are allocated to 9 clusters according to their socio-economic indicators, based on cluster means of 2012. It is assumed that cluster means remain fixed as in the 2012 (base year) and do not change. In other words, clustering of the cities

in 2009 has been made from the perspective of clustering in 2012. Based on city clusters, 471,824 real connections in 2009 are distributed between 45 cluster pairs. Non-existing links have been obtained for each cluster pair. The total number of non-existing links in the APD network of 2009 is 7,205,497. For the calculation of the two metrics, sets of existing and non-existing links have been used.

Based on existing studies[5,7] the network has been divided into two sets: testing and probe in proportions 90% and 10%, respectively. Each *AUC* and *precision* value has been obtained by averaging of 10 realizations with independent random separations of random and probe sets. Metrics for the whole network and each cluster pair for different indexes have been calculated as well as their standard deviations. Results are presented in Tab.3 for the whole network and the average of 45 cluster pairs' metric values. *AUC* and *precision* are used to determine the accuracy of each index for the whole network and for clusters. The index with the best metrics values will be chosen for the topology forecast in the APD network. The closer the metric is to 1, the better the performance of the index in a given network.

Metrics		**WCN**	**WAA**	**WRA**	**WSA**	**WSO**	**WHPI**	**WHDI**	**WLHN**	**WPA**
AUC	The whole network	0.73058	0.76106	0.661594	0.778773	**0.852472**	0.642671	0.822776	0.449384	0.656475
	Standard deviation	0.000991	0.000604	0.000763	0.001237	0.001444	0.001251	0.001671	0.001628	0.000586
	Cluster average	0.843271	0.948161	**0.963566**	0.859771	0.879698	0.725055	0.879018	0.63934	0.823745
	Standard deviation	0.020309	0.003291	0.002676	0.017402	0.007919	0.007277	0.008145	0.01512	0.020787
Precision	The whole network	0.790213	0.847255	**0.912041**	0.824091	0.818912	0.484531	0.788109	0.183	0.662276
	Standard deviation	0.003507	0.003498	0.002192	0.004744	0.003834	0.005788	0.003807	0.004153	2.15E-03
	Cluster average	0.910552	0.988456	**0.99109**	0.92102	0.866016	0.481566	0.924206	0.57902	0.885594
	Standard deviation	0.008343	0.003519	0.002261	0.006745	0.00853	0.012802	0.008149	0.072423	0.010541

Tab.3 AUC and precision values and their standard deviations for the whole APD network and average values for cluster pairs of 2009

The data in Tab.3 demonstrates that only one index - weighted hub promoted index (WHPI) has a higher *precision* value in the whole network than the cluster average. However, this value is low compared to other indices. All other indices show higher *AUC* and *precision* numbers in clusters than in the whole network. This proves the necessity of separating cities into groups according to their socio-economic indicators, so as to improve the link forecasting performance. The best *AUC* number for the whole network is WSO. But this number is smaller than *AUC* for WRA in clusters. The WRA index shows the best *AUC* and *precision* results in clusters pairs. This is expected, since WRA gives a higher score to a non-existing connection between two nodes if these nodes have many common neighbors with large weights. It is important to note, that the WRA index has the best performance of *AUC* and *precision* in each cluster pair. This disproves the assumption that cluster pairs in the APD network have different similarity indices demonstrating the best performance.

Based on the aforementioned analysis, the weighted resource allocation (WRA) index is chosen for the topology forecast in the APD network. The score for each non-existing link in

each cluster pair will be calculated utilizing the WRA index. Next, it is necessary to validate the method based on historical data.

3. Model validation

For the validation the APD topology of four years from 2009 to 2012 has been utilized. Data for these years from the ADI database (ADP networks of 2009, 2010, 2011 and 2012) have been retrieved. Socio-economic data and geographical coordinates for cities from the same databases as for 2012 have been obtained. The conditions required for the appearance of new cities in the APD network are not clear and hard to predict[5]. Thus, for the analysis, sets of cities from four networks have been reviewed. Cities which are presented in all four networks have been allocated to the set of common cities. Thereby, there is a constant set of common cities for all 4 networks. In Tab.4 topological characteristics of four networks with original and common cities are presented.

Year	Original number of cities	Number of common cities	Original number of connections	Number of connections with common cities	Number of non-existing connections with common cities
2009	3,930	3,930	487,442	486,857	7,233,628
2010	3,933		506,619	504,895	7,215,590
2011	4,003		524,021	521,219	7,199,266
2012	4,435		531,360	527,308	7,193,177

Tab.4 Topological characteristics of four APD networks with original and common cities

Three analyses based on modified networks with common cities to define accuracies have been made. Within the analyses new connections are calculated utilizing the WRA index and compared with the real data. These connections are calculated for topologies of: 2010 from 2009, 2011 from 2010 and 2012 from 2011. For 2010, 2011 and 2012 the amount of new real added connections is known and shown in Fig.1. New connections in cluster pairs in the figure are sorted in descending order based on the amount of new connections in 2012. The number of new connections is almost the same for most cluster pairs. However, there is not enough data to analyze probable tendencies.

For all three analyses new calculated connections have been compared with new real added connections. For example, new calculated connections in 2010 from 2009 have been compared with the real topology of 2010. The analysis procedure is as follows: socio-economic indicators and cluster accessory of 2010 are assigned to cities in 2009. Thus, the APD network 2009 turns to an incomplete network of the APD network 2010. The score for all non-existing connections in every cluster pair of the 2009 network is calculated using the WRA index. Connections are ordered in descending order by their score. Then the calculated data is compared to real data. Thus, it is possible to define the accuracy of the proposed forecast method. The accuracy is defined as a ratio between the amount of real new added connections in 2010 and the number of real new connections in the ordered list. This number is between 0 and 1. The forecast method has a higher accuracy the closer the ratio is to 1. Accuracies for each cluster pair for years 2010,

2011 and 2012 are presented in Fig.2. Cluster pairs are ordered in descending order by accuracy in 2012.

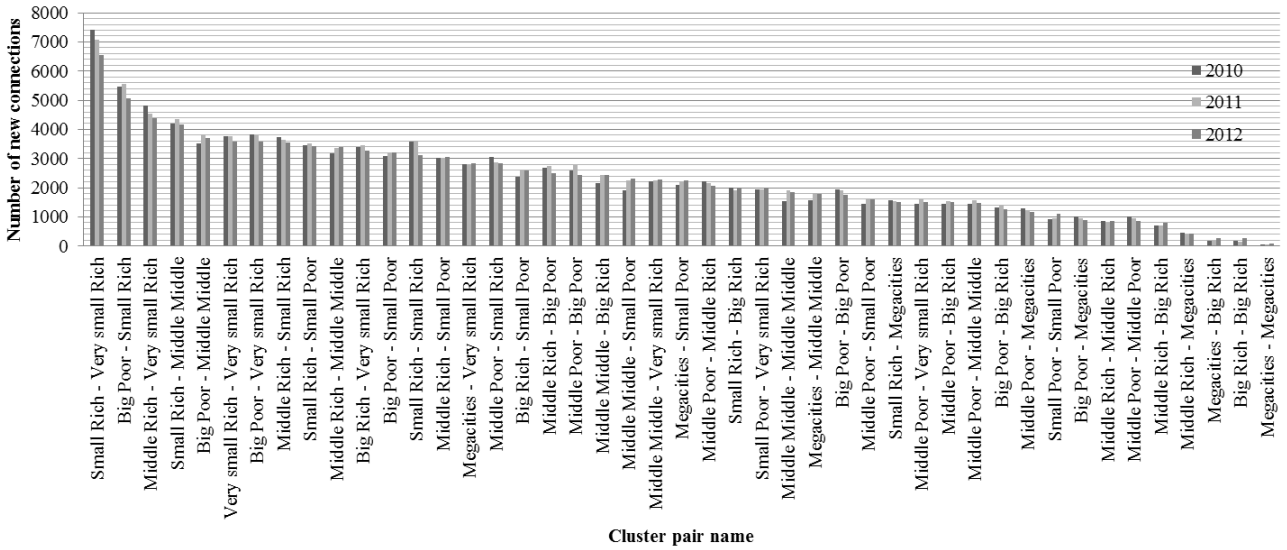

Fig.1 Number of new real connections in 2010, 2011 and 2012

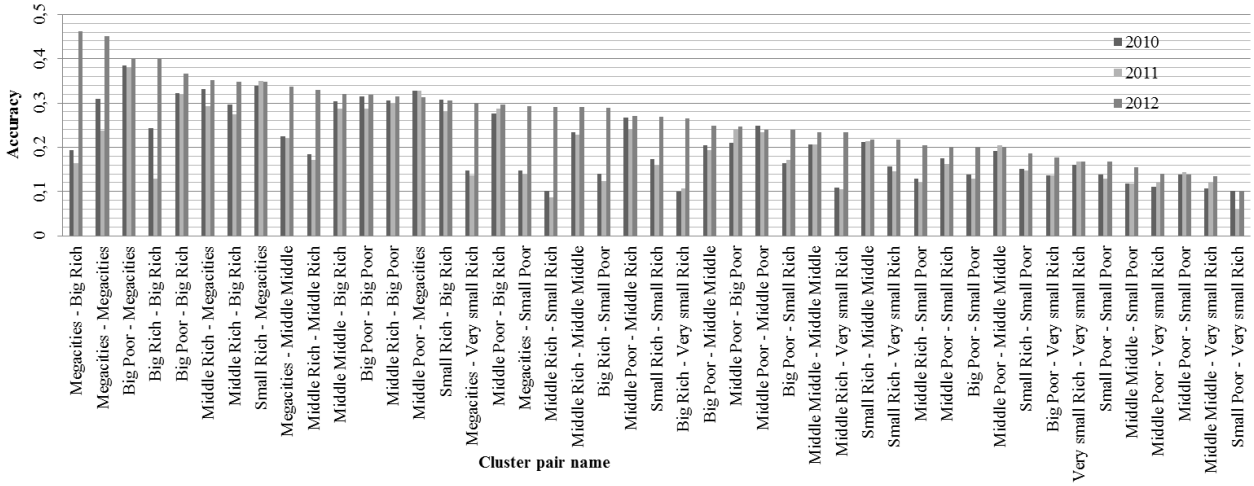

Fig.2 Accuracies by cluster pairs

In this study, all accuracies are below 0.5, meaning that there is more than 50% error in the prediction. However, the accuracy of 0.266095 for the 2012 APD network is higher than, for example, in T. Murata and S. Moriyasu[4] study of link prediction in a weighted network of Question-Answering Bulletin Boards. Furthermore, in this study accuracies are different in different years. For each cluster pair, the accuracies for 2010 and 2011 are almost always lower than for 2012. This is probably related to the economic crisis of 2008, and 2010 and 2011 likely show the tail of this crisis when the world economy was not fully recovered. The processes of the APD generation in 2010 and 2011 are probably different from the proposed "attractiveness" approach in equation (3). Furthermore, there are different accuracy numbers for cluster pairs. Relationships between cities with "strong" socio-economic indicators in some cluster pairs (for

example *small-rich – megacities*) are described better using equation (3) than for "weak" cluster pairs (for example *big-poor – small-poor*). Therefore, the accuracy is higher for cities with relatively high socio-economic indicators. In addition, it should be noted that cluster pairs are not equal in terms of number of passengers. For the forecast model it is important to have a high accuracy for connections with a high APD. Tab.5 presents accuracies for accumulative number of passengers by cluster pairs in Fig.2 for 2010, 2011 and 2012. Numbers in brackets indicate accumulated numbers of cluster pairs corresponding to a given accumulated percentage of passengers.

	50% passengers	90% passengers	100% passengers
2010	0.324165 (10)	0.240846 (32)	0.206051 (45)
2011	0.296976 (13)	0.222804 (34)	0.195393 (45)
2012	0.396876 (7)	0.327051 (25)	0.266095 (45)

Tab.5 Average accuracies for years 2010, 2011 and 2012 for a given percentage of passengers. Numbers in brackets indicate how many cluster pairs generate a given percentage of passengers.

The model validation on historical data shows higher accuracy compare to existing studies. Furthermore, the accuracy is even higher in clusters with large number of passengers. Thus, at this stage of the study, this accuracy seems to be sufficient. However, the accuracy probably could be enhanced by defining appropriate coefficients in equation (1). Next, it is necessary to analyze the WRA index boundary criteria in the ordered lists of non-existing connections for each cluster pair.

4. Boundaries

For the boundary analysis, the APD network of 2012 is considered based on the assumption that socio-economic indicators of APD networks 2010 and 2011 have been influenced by the economic crisis of 2008. Based on the aforementioned analysis there are two ways to define boundaries: either using the number of new added connections in each cluster pair or the boundary scores for each cluster pair. In other words, for the first method, a fixed number of connections will be added to the network from the ordered list of non-exiting connections, which is in descending order according to their score. In the second method, all connections where the score exceeds the boundary score in the ordered list will be added to the APD network. The number of new added connections based on 2012 data are presented in Fig.3. Boundary scores for each cluster pair based on 2012 analysis are shown in Fig.4.

For example, the APD network topologies of a cluster pair in year *y* (Fig.5) and the next year *y+1* (Fig.6) are known. Socio-economic indicators of cities from year *y+1* are assigned to the same cities in year *y*. Utilizing the WRA index, scores for all non-existing connections are calculated (Tab.6). Connections are ordered in descending order by their score (Tab.7). The accuracy of the method can be defined using new real added connections to network in year *y+1*. The number of forecasted links from the top of the list is equal to the number of new real added connections. The accuracy is defined as the ratio of relevant connections in the list of non-existing connections to number of new real added connections. There are two types of criteria of adding connections to the ADP network. The first criterion is a fixed amount of connections.

This amount of connections added every year is equal to the number of new real added connections from year *y+1*. The second criterion is the boundary score. Each connection with a score higher than the boundary score in the year *y+1* is added to the network.

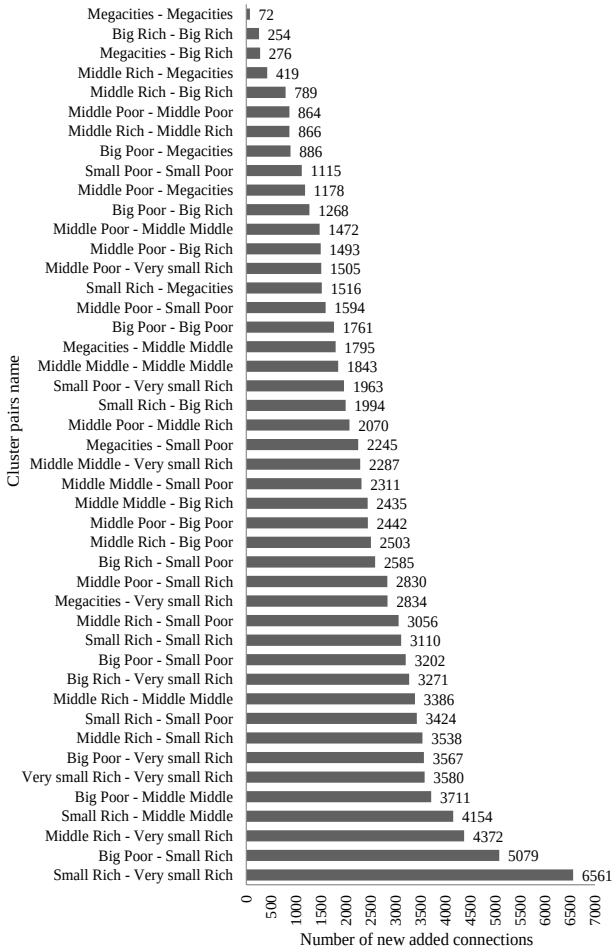

Fig 3. Number of new added connections for every cluster pair for the ADP network of 2012

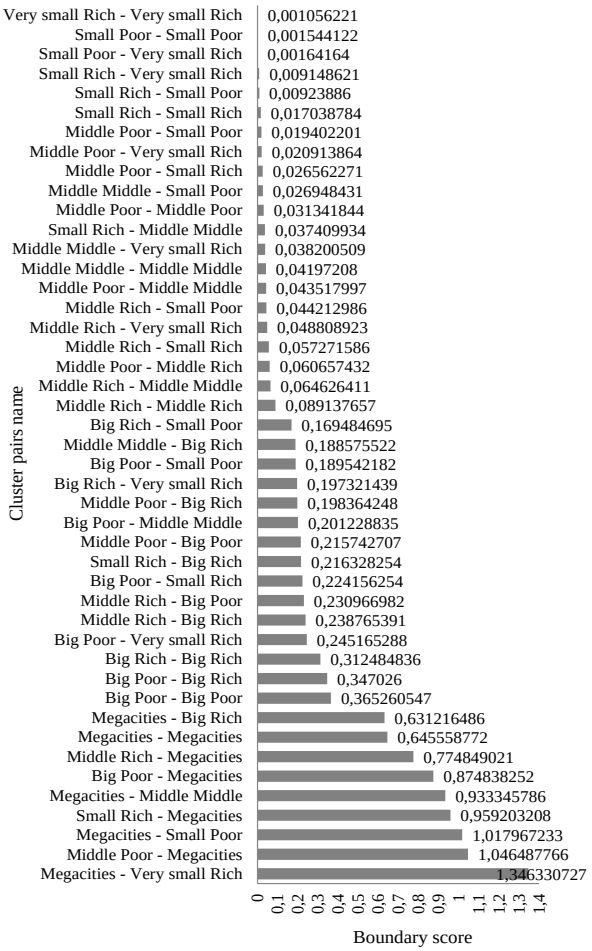

Fig 4. Boundary scores for every cluster pair for the ADP network of 2012

In this study the possibility of link elimination has not yet been considered. It is assumed that if an APD connection has appeared in the network, it will remain in the network throughout the period of the forecast. Thus, in both approaches a situation could arise where all cities within a cluster pair are connected to each other. This is more likely to occur when applying the first method, that of using the number of new added connections in each cluster pair. For example, cluster pair *middle-rich – small-rich* in 2012 has 207 and 565 cities respectively. This cluster pair in 2012 has 27,628 connections including 3,538 new connections added from 2011 (shown in Fig.3). The number of non-existing connections is 89,327. If it is assumed that the number of added connections will remain fixed, all cities in this cluster pair will be connected to each other within ~25 years. For second method, applying boundary scores, the year when all cities are interconnected in the cluster is hard to predict. This will depend on various factors such as network configuration, city clustering, socio-economic scenario, etc. Nevertheless, at this stage of

the study, it seems reasonable to use the second method of boundary definition – boundary scores.

It is important to note that each cluster pair has different boundaries either for the fixed number of connections method or the boundary score method. This proves the assumption that each cluster pair has its own boundaries. However, this adding process requires an additional study.

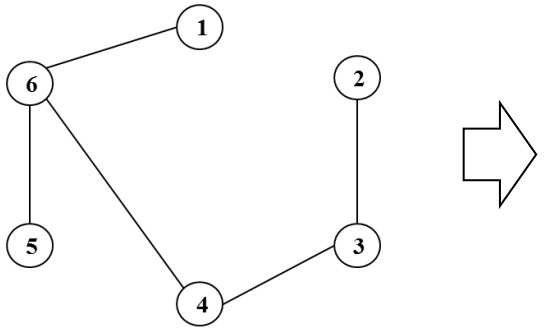

Fig.5 The APD network topology of a cluster pair in a year *y*

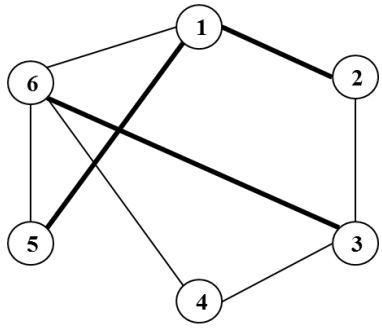

Fig.6 The APD network topology of the cluster pair in a year *y+1*. Thick lines depict new real added connections to the network of year *y*

Existing connections	Non-existing connections	Score of non-existing connections
1-6	**1-2**	S12
2-3	1-3	S13
3-4	1-4	S14
4-6	**1-5**	S15
5-6	2-4	S24
	2-5	S25
	2-6	S26
	3-5	S35
	3-6	S36
	4-5	S45

Tab.6 Existing connection and all non-existing links in the APD network in year *y*. The score for each non-existing link is calculated. New real added connections in the APD network in year *y+1* are marked in bold.

Non-existing connections	Score of non-existing connections	Real new added connections	
3-6	S36	1-2	Fixed amount of connections
2-5	S25	1-5	
1-2	S12	3-6	
2-6	S26		
4-5	S45		
1-3	S13		
2-4	S24		
1-3	S13		
1-5	S15		
3-5	S35		

Boundary score

Tab.7 Non-existing connections are ordered in descending order based on their score. Two types of boundaries based on the number of new real added connections are presented: fixed amount of connections and the boundary score. The forecast predicts two actual connections out of three. Thus, the accuracy in this case is 0.6666.

5. Conclusion

This paper presents the study of topology forecast in the APD network utilizing a socio-economic scenario. The study shows that the weighted resource allocation (WRA) index demonstrates the best performance. *AUC* and *precision* metrics are higher for cluster pairs than for the whole APD network. This proves the necessity of separating cities into groups by their socio-economic indicators to improve the link forecasting performance. Thus, the WRA index is

used to calculate scores for all non-existing links in each cluster pair. This disproves the assumption that cluster pairs in the APD network have different similarity indices demonstrating the best performance. For existing years the modeling is applied and results are compared with real data. The accuracy of the similarity-based algorithm for the APD network is higher than in related studies. The study shows two methods of adding new connections from the ordered score list of non-existing connections. The first method is to add a fixed number of connections based on the historical analyses (Fig.1, Fig.3). The other method is to use as the boundary a score number from the ordered list (Fig.4). Both methods prove the assumption that each cluster pair has its own boundary. It seems reasonable to use the second approach with the boundary score. However, this adding process will require further study.

It is believed that accuracy could be enhanced by defining appropriate coefficients in equation (1). It is likely that every cluster pair could have its own coefficients. This assumption needs an additional study. Furthermore, in future studies, the possibility of link elimination from the APD network should be considered. It seems that an approach of link elimination is similar to that of link addition. In addition, a whole network forecast based on a socio-economic scenario should be made. This will allow assessing the two methods of adding links into the network and thus determine the most appropriate method.

---

[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] 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].

[4] 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.

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

[6] Murata T., Moriyasu S., "Link prediction of social networks based on weighted proximity measures", *IEEE/WIC/ACM International conference on Web Intelligence*, Fremont, California, 2007

[7] Lü L., Zhou T., "Link prediction in weighted networks: The role of weak ties", EPL, *A letters Journal Exploring the Frontiers of Physics*, 89 (2012) 18001, January 2010.

[8] 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.

[9] Hanely J.A., McNeil B.J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, 143: 29-39, April 1982.

[10] Herlocker J.L., Konstann J.A., Terveen L.G., Riedl J.T., "Evaluating Collaborative Filtering Recommender Systems", *ACM Transactions on Information Systems*, Vol. 22, No. 1, pp. 5–53, January 2004.

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

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

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

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

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

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

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

[18] Grosche, T., Rothlauf, F., Heinzl, A., "Gravity models for airline passenger volume estimation", *Journal of Air Transport Management*, Vol. 13, 2007, pp. 175-183.