Journal of Data Analysis and Information Processing, 2017, 5, 49-66
http://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

A Basic Study of the Forecast of Air Transportation Networks Using Different Forecasting Methods

Yuya Takahashi, Rie Osawa, Susumu Shirayama

Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
Email: ytaka1210@gmail.com, rie-u@nakl.t.u-tokyo.ac.jp, sirayama@sys.t.u-tokyo.ac.jp

How to cite this paper: Takahashi, Y., Osawa, R. and Shirayama, S. (2017) A Basic Study of the Forecast of Air Transportation Networks Using Different Forecasting Methods. *Journal of Data Analysis and Information Processing*, **5**, 49-66.
https://doi.org/10.4236/jdaip.2017.52004

Received: March 8, 2017
Accepted: May 12, 2017
Published: May 16, 2017

Abstract

This research applies network structuring theories to the aviation domain and predicts aviation network growth, considering a flight connection between airports as a link between nodes. Our link prediction approach is based on network structure information, and to improve prediction accuracy, it is necessary to estimate the mechanism of aviation network growth. This research critically evaluates the prediction accuracy of two methods: the receiver operating characteristic curve method (ROC) and the logistic regression method. We propose a four-step method to evaluate the relative predictive accuracy among different link prediction methods. A case study of US aviation networks indicated that the ROC method provided better prediction accuracy compared with the logistic regression method. This result suggests that tuning of the prediction distribution and the regression model coefficients can further improve the accuracy of the logistic regression method.

Keywords

Complex Network, Link Prediction, Air Transportation Network

1. Introduction

In recent years, the number of air passengers has been increasing, with the worldwide annual number of passengers up by approximately 34% from 2010 to 2014 [1]. According to this trend, it is assumed that the demand for the air routes will keep increasing and that the aviation network will change in response to the increased the demand.

Our focus in this study is to predict how the aviation network will change in the future to accommodate increased demand. In addition to being of importance for the industrial aspects of aviation, predictive tools for the evolution of

aviation are also critical to improve our understanding of environmental and social impacts. In terms of advantages to industry, accurate and concrete forecasting of flight demands, including passenger fluctuation, will allow airlines to efficiently plan the frequency of flights, select the appropriate size of aircrafts, and optimize flight plans for each airport. Additionally, a forecast of this type may assist aircraft manufacturers in the design of future development plans [2]. The results of network predictions, such as the frequency of flights and the geographic distances traveled, are environmentally relevant since this data can be used to forecast CO_2 emission [3]. Therefore, it will potentially be possible to identify eco-friendly aircrafts and to petition airlines to introduce and prioritize fuel efficient aircrafts. From a social perspective, traffic jams due to traffic concentration can be predicted by analyzing both demand and local characteristics, such as population distribution. This perspective is important to aviation safety because the number of accidents caused by human errors is increasing owing to the traffic jams [4].

Many studies have been conducted to forecast the aviation network based on estimates of the demand for a given air route and to evaluate the impacts caused by change in demand [5] [6] [7]. However, it is challenging to clearly isolate the main factors that affect aviation networks due to the large number of these factors, which include financial considerations (e.g., business conditions), local characteristics (e.g., population distribution, geographic distance, climate, altitude), social factors (e.g., terrorism), and environmental factors (e.g., natural disasters). Currently, most discussions of the variation of the aviation network remain qualitative.

Quantitative studies have, however, been conducted to investigate the characteristics of aviation networks and to consider network changes in terms of the network structure. In these studies, airports are regarded as nodes, and airlines and the number of flights or the number of passengers above a threshold are regarded as links.

Analysis of the global aviation network structure by Guimera *et al*. determined that the network is scale-free and small-world, and that community structure is best explained from the point of view of geopolitical considerations in cities that have airports [8]. Although this study did not consider the variation of aviation networks, it is pioneering in the sense that complex network theory was applied for the first time to aviation networks. In terms of the complex network, the formation of the aviation network has been explored through network models. Sawai and Sato proposed a method to create star networks from random networks in a bottom-up manner, and studied their characteristics [9] [10]. While this method fails to forecast network change based on the current aviation network structure, it provides the knowledge necessary to efficiently reconstruct future aviation networks from the viewpoint of network structure.

Bonnefoy and Hansman predicted the influence of very light jets (VLJs) by considering the overlaps in performance and capability between light jets and VLJs. The authors proposed a method for network structure analysis and a re-

sultant network growth model [11].

Conversely, other works have suggested that the existing prediction methods are unable to sufficiently explain the real-world variation in aviation networks and have proposed other predictive methods based on complex network analysis. Kotegawa *et al*. attempted to predict future aviation networks utilizing three prediction methods and prediction measures based on the network structure [12]. Essentially, the network growth mechanism was understood through investigation of prediction accuracy and attempts to improve the performance of the prediction method. Furthermore, this work indicated that the scale-free network structure is important for aviation transportation efficiency [13]. Additionally, it was determined that when the ratio between the sum of degrees in the actual network structure and the sum of degrees in the same size (node number) of complete graph is high, the actual network more closely approaches the random graph and the more robust the network becomes. In terms of the robustness of utilizing aviation networks as the network structure, Wei *et al*. proposed methods to maximize robustness by adding and cutting links [14].

The results of these studies imply that the structural characteristics and the growth process in the current aviation network affect the future network. For example, according to Bonnefoy and Hansman, the scale-free characteristics and the growth limits of hub airports provide an estimate of the future structure of the aviation network. The network structure described by Sawai and Sato and the robustness optimization described by Wei *et al*. will likely affect the reconstruction of a next-generation aviation network. Furthermore, the methods and measures proposed by Kotegawa *et al*. provide directly useful estimates of the variation of future aviation networks.

However, the work of Bonnefoy and Hansman relied on one network growth model that was not compared to other models and was not validated. Additional difficulties are associated with the previously mentioned studies, for instance the method of Sawai and Sato is somewhat unfeasible, and it is uncertain whether the method of Wei *et al*., which employed a relatively small network consisting of sixteen nodes, can be extended to larger networks. Furthermore, the predictive power of the approach of Kotegawa *et al*. is challenged by low accuracy since multi-year data was not employed.

Based on the above discussion, in the current study, we aim to improve the prediction accuracy of the future aviation network by the method of link prediction coupled with predictive measures calculated from the network structure.

2. Proposed Method

2.1. Outline of Proposed Method

First, link prediction was conducted according to two methods that utilize the measures introduced in Section 2.3. These measures are calculated from the network structure to identify missing links, and to determine which measure achieves the best prediction accuracy and the highest contribution. Next, the growth mechanism was estimated based on the following hypothesis: network

growth depends on the measures that have high contributions. Furthermore, the factors that change the network structure were analyzed.

To compare results, we applied a four-step method that is popular in traffic engineering. This method incorporates population, income, and other statistical data as measures.

2.2. Subject Network

In the current study, we analyzed the annual variation of the aviation network in the US.

The data supplied by the Bureau of Transportation Statistics (BTS) [15], as part of the United States Department of Transportation were employed in the network construction. A sampling of the data is presented in Table 1.

The data set consists of scheduled departures, performed departures, passenger numbers, the origin and destination, including the distance between these locations, and the aircraft type, year, and class. (Although additional data are also available, we utilized these annual data.) The names of airports are according to the 3 letter abbreviations provided by the International Air Transport Association.

Table 1. Sample of the data obtained from the BTS.

Departures_scheduld	Departures_performd	Passengers	Distance	Origin	Destination	Aircraft_type	Year	Class
0	2	0	210	ATK	SCC	556	2014	P
0	1	0	677	DQH	ENA	556	2014	P
0	2	0	59	DQH	SCC	556	2014	P
0	1	0	59	ENA	ANC	556	2014	P
0	1	0	203	FAI	HUS	556	2014	P
0	1	0	277	FVQ	ANC	556	2014	P
0	2	0	329	GAL	ANC	556	2014	P
0	1	0	362	HUS	ANC	556	2014	P
0	1	0	268	OTZ	AIN	556	2014	P
0	1	0	466	PHO	SCC	556	2014	P
0	1	0	269	SCC	AIN	556	2014	P
0	1	0	627	SCC	ANC	556	2014	P
0	3	0	210	SCC	ATK	556	2014	P
0	2	0	59	SCC	DQH	556	2014	P
0	1	0	466	SCC	PHO	556	2014	P
0	1	0	765	STG	ANC	556	2014	P
0	1	15	151	ABI	MAF	676	2014	F
0	1	48	1062	ABQ	CMI	631	2014	F
0	8	290	285	AEX	DFW	676	2014	F

For example, the first record (Table 1; row 2) indicates that the no flights were scheduled but that 2 flights were conducted from Atqasuk airport (ATK) to Deadhorse airport (SCC), and that the distance between ATK and SCC is 210 miles. We can assume these were cargo flights since there were no passengers aboard either flight. Similarly, the last record indicates 0 scheduled flights and 8 flights performed between the Louisiana airport (AEX) and the Dallas/Fort Worth airport, with 290 passengers transported annually.

In total, 436,559 records were obtained for 2014.

To begin, we collected the data for the number of flights performed and the number of passengers between any two given airports per year. We combined data from different aircraft types or class, which are otherwise separated in the records.

Next, we created an adjacent matrix of the airports. The thresholds were defined for the number of flights and the number of passengers. In this study, links are connected if there are two-way flights with a passenger count above the threshold. That is to say, no connection between node pairs (two airports) indicates either that the number of flights is below the threshold or that there are no flights. Predicted links indicate that the number of flights is expected to be above the threshold according to the prediction.

Weighting of the links was not applied in the current work. Therefore, the subject network is a non-directed network, with link prediction applied to this network.

2.3. Prediction Measures

In this study, according to Zhou *et al.* [16], the prediction measures are based on the similarity of the node pairs (two airports) calculated only from the network structure.

The similarity between node x and node y is represented by Score s_{xy}. For example, PA, one of the measures, is calculated as $s_{xy} = k_x k_y$ where k_x is the degree of node x and k_y is the degree of node y. A higher score indicates a greater possibility that a link exists between node x and node y.

In this study, eleven measures were used to analyze the network: JI, PA, CN, SP, Sal, Sør, HPI, HDI, LHN, AA, and RA.

1) Shortest Path (SP)

The SP measure is defined as the inverse of L_{xy}, where L_{xy} is the vertex distance between node x and node y as shown in Equation (1). When no connection exists, s_{xy} is 0. This measure is created based on the hypothesis that two airports are likely to be connected when there are as few hub airports as possible.

$$s_{xy} = \frac{1}{L_{xy}} \tag{1}$$

2) Common Neighbors (CN)

The CN measure is defined as the number of common nodes between node x and node y as shown in Equation (2). This measure is created based on the hy-

pothesis that the more airports two nodes have in common, the more likely they are to be connected. Here for node x, let $\Gamma(x)$ denote the set of neighbors of x.

$$s_{xy} = \left|\Gamma_x \cap \Gamma_y\right| \tag{2}$$

3) Salton Index (Sal)

The Sal measure is defined as the score obtained by dividing the CN measure with the geometrical mean of the node pair degrees as shown in Equation (3).

$$s_{xy} = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\sqrt{k_x k_y}} \tag{3}$$

4) Jaccard Index (JI)

The JI measure is defined as the score obtained by normalizing the CN measure by the union of the adjacent nodes to the node pair as shown in Equation (4).

$$s_{xy} = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\left|\Gamma_x \cup \Gamma_y\right|} \tag{4}$$

5) Sørensen Index (Sør)

The Sør measure is defined as the score obtained by normalizing the CN measure by the arithmetic mean of the node pair degrees as shown in Equation (5).

$$s_{xy} = \frac{2\left|\Gamma_x \cap \Gamma_y\right|}{k_x + k_y} \tag{5}$$

6) Hub Promoted Index (HPI)

The HPI measure is defined as the score obtained by dividing the CN measure with the lower degree of the node pairs as shown in Equation (6). When the link is adjacent to a hub node, the score tends to be higher because the denominator is determined by the lower degree only. The name of this measure is derived from this attribute.

$$s_{xy} = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\min\left\{k_x, k_y\right\}} \tag{6}$$

7) Hub Depressed Index (HDI)

The HDI measure is similar to the HPI. In contrast the HPI, the higher degree of the node pairs is applied to the denominator as shown in Equation (7). The score tends to be lower when links are adjacent to a hub node.

$$s_{xy} = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\max\left\{k_x, k_y\right\}} \tag{7}$$

8) Leicht-Holme-Newman Index (LHN)

The LHN measure is defined as the score obtained by dividing the CN measure with the product of the node pair degrees as shown in Equation (8). Although this measure appears similar to the Sal, the score tends to be lower when both degrees of the node pair are higher even if all adjacent nodes are common.

$$s_{xy} = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{k_x k_y} \tag{8}$$

9) Preferential Attachment (PA)

The PA measure is defined as the product of node pair degrees as shown in Equation (9). This measure is created based on the hypothesis that the more airports connected to a given airport, the more likely it is to be connected to other airports.

$$s_{xy} = k_x k_y \tag{9}$$

10) Adamic-Adar Index (AA)

The AA measure is defined as the sum of the inverse of the logarithm of the common node degrees. That is to say, the lower degree node of the common adjacent nodes exerts a higher impact on the score as shown in Equation (10). Here, z indicates the common adjacent node of the node pair x and y.

$$s_{xy} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log k_z} \tag{10}$$

11) Resource Allocation Index (RA)

The RA measure was proposed by Zhou *et al.* [16]. This measure is the score obtained by removing the logarithm from the AA as shown in Equation (11). The AA applies the logarithm to prevent the weights from being too small when the degree is high. In the case of the RA, the weights of the nodes with high degree are sufficiently low.

$$s_{xy} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{k_z} \tag{11}$$

2.4. Prediction Methods

1) ROC curve method

The method of Zhou *et al.* [16], which was originally formulated to locate missing links, was applied to the link prediction in this work. First, the prediction measures listed in Section 2.3 were calculated to generate the annual network according to the method outlined in Section 2.2. Next, the ROC (Receiver Operating Characteristic) curve was calculated for each measure, and finally, the point that occupies 95% of the area under the curve (AUC) was considered as a threshold, with node pairs above the threshold predicted to become connected. An example of a network based on this method is shown in **Figure 1**.

Table 2 gives the PA values for the network presented in **Figure 1** in descending order. In the last column of **Table 2**, T indicates that a link exists whereas F indicates no link exists.

The ROC curve was then calculated based on the data presented in **Table 2**. Specifically, the following X and Y parameters were calculated once the threshold was defined.

$$X = \left(\frac{\text{the number of F entries in excess of the threshold}}{\text{the total number of F entries}} \right)$$

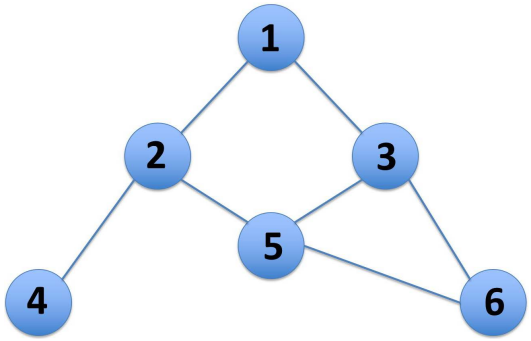

Figure 1. Example of a targeted network.

Table 2. Measure and link existence for each node pair.

Node 1	Node 2	Measure (PA)	Link existence
3	5	9	T
2	5	9	T
2	3	9	F
5	6	6	T
3	6	6	T
...	...	...	...
...	...	...	...
4	6	2	F
4	1	2	F

$$Y = \left(\frac{\text{the number of T entries in excess of the threshold}}{\text{the total number of T entries}} \right)$$

The threshold is defined from the score of the measure, and the (X, Y) coordinates are plotted. The ROC curve is an aggregate data of plotted points. **Figure 2** shows an ROC curve with sixteen thresholds calculated from the data presented in **Table 2**. In this case, the AUC is 0.75.

A perpendicular is drawn down on the X-axis from a certain point on the ROC curve. The area, which is right side of the perpendicular and under the curve, is calculated.

Point [A] in **Figure 2** shows the point that demarks more than 95% of the AUC. If we position the threshold at the 2 and 3 node pair in **Table 2**, [A] is obtained. The PA of the node pair is 9. We predict that node pairs with scores exceeding this value will be connected. Currently, no link exists between nodes 2 and 3 as indicated by the F entry (**Table 2**; row 4). Therefore, the prediction is that these nodes will be connected in the future.

2) Logistic regression and measures method

Kotegawa *et al.* conducted link prediction by logistic regression, which takes into account node degrees, cluster coefficients, weights, and the difference between weights as explanatory variables [12]. We expanded upon this method and conducted the link prediction according to the following four steps.

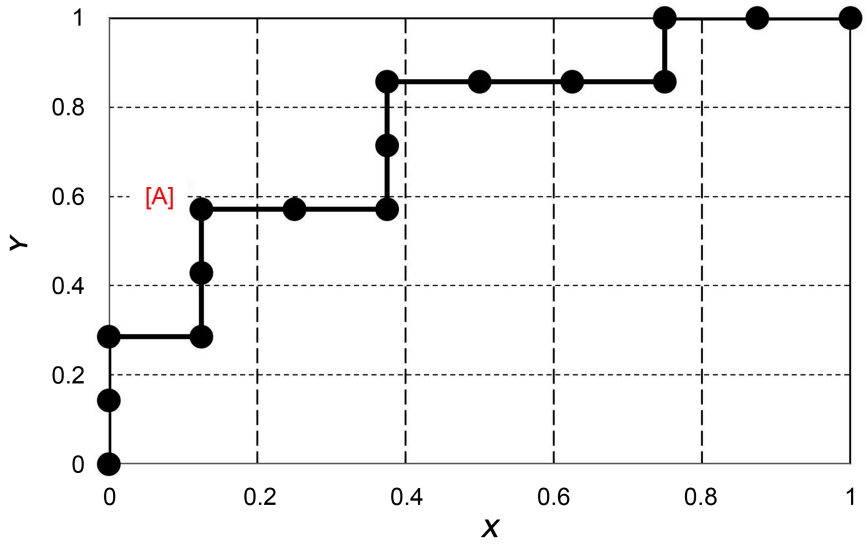

Figure 2. An example of an ROC curve.

a) Detect the node pairs without connection in each year's network.

b) For these node pairs, calculate the values of the eleven prediction measures, which are explanatory variables.

c) Place the link status of these node pairs for the upcoming year in the objective variable y. If the status of the node pair is T, y is equal to 1 and if the status is F, y is equal to 0.

Build the logistic regression model and conduct the link prediction. The following provides additional details to explain step (4). x_i denotes the eleven prediction measures and $\boldsymbol{x}$ is the aggregate of explanatory variables.

$\boldsymbol{x}$ is defined as

$$\boldsymbol{x} = \left(1, x_1, \cdots, x_{11}\right)^t \tag{12}$$

$\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \left(\beta_0, \beta_1, \cdots, \beta_{11}\right) \tag{13}$$

and p gives the probability of the objective variable y being equal to 1, and is defined as

$$p = \frac{\exp\left(\boldsymbol{\beta x}\right)}{1 + \exp\left(\boldsymbol{\beta x}\right)} \tag{14}$$

The above equation is transformed to the linear regression model by a logit transformation as shown below.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{11} x_{11} \tag{15}$$

A logistic regression model is built for each year using data from previous year to predict the link status. For example, for the link prediction from 2009 to 2010, x gives the eleven prediction measures of the node pairs which were not connected in 2008 and y the statuses of the node pairs in 2009. The logistic regression model is then built based on these values, and the 2010 link status is pre-

dicted by applying the eleven measures of the node pairs which were not connected in 2009 to this model.

3) Utilization of the four-step method

The four-step method regards the traffic flow as the movement between zones and predicts the future Origin-Destination (OD) matrix from the current OD matrix based on the two amounts of traffic, which are described as the traffic moving AWAY from a certain zone (generated traffic amount) and the traffic moving INTO a certain zone (attracted traffic amount) [17] [18] [19]. This method is different from two methods mentioned above in that it is not based on the network structure and does not predict link generation. In this study, this method is employed as a reference.

We show the main procedure for the four-step method below and refer to the original paper where it is described [17].

a) Traffic is sorted according to the origin and the destination to create the OD matrix.

b) The future generated and attracted traffic amounts are predicted through application of the linear regression model.

c) The calculation is repeated using the Frater method until convergence occurs.

Step 2 can be further explained as follows.

Here, G_i and A_i are the generated and attracted traffic amounts in zone i, respectively. For each zone, G_i and A_i are calculated according to the linear regression model shown below.

$$G_i = \beta_0 + \sum_m \beta_m \cdot X_{mi} \tag{16}$$

$$A_i = \gamma_0 + \sum_m \gamma_m \cdot X_{mi} \tag{17}$$

In this paper, the following four explanatory variables, X_{mi}, were employed:

X_{1i}: The employed population in the state to which airport i belongs.

X_{2i}: The per-capita disposable income in the state to which airport i belongs.

X_{3i}: The GDP of the state to which airport i belongs.

X_{4i}: Whether a northeast corridor station exists in the state to which airport i belongs.

X_{4i} is a dummy variable, which is 1 when a northeast corridor station exists in the state to which airport i belongs. The northeast corridor is the railway in the east coast of North America. A majority of east coast states have this type of railway station. We take this variable into consideration because it is assumed that the transfer of the railway affects the utility of airlines.

The OD matrix for future flights was generated by the four-step method, where the number of flights was regarded in terms of traffic flow. Next, we considered airport pairs as pairs that will connect if the number of flights was above the threshold in both directions.

2.5. Evaluation of Prediction Accuracy

When we consider the presence of links as events, the link prediction becomes a

two-classification problem. The F-value is generally used to calculate the accuracy of this type of problem, and therefore we also employ it in the current research.

In case the prediction is positive or negative, and the fact is positive or negative, the prediction result is classified into four groups (TP-true positive, FP-false positive, FN-false negative, and TN-true negative) as shown in Table 3. The sum of TP and FP indicates the total number of elements predicted to be positive, whereas the sum of TN and FN indicates the opposite. Additionally, the sum of TP and FN gives the total number of actual elements that are positive, whereas the sum of FP and TN indicates the opposite.

In this paper, positive and negative values indicate that node pairs are either connected or not connected, respectively. Therefore, TP, FP, FN, and TN describe the following scenarios:

TP: Node pair is predicted to be connected and is actually connected.

FP: Node pair is predicted to be connected but is actually not connected.

FN: Node pair is predicted to remain unconnected but actually becomes connected.

TN: Node pair predicted to remain unconnected and in fact remains unconnected.

Here, we define a as the number of TP, b as the number of FP, c as the number of FN, and d as the number of TN. Additionally, PAG represents Precision, and POD represents Recall. PAG and POD are defined according to Equations (18) and (19), respectively.

$$\mathrm{PAG} = \frac{a}{a+b} \tag{18}$$

$$\mathrm{POD} = \frac{a}{a+c} \tag{19}$$

Furthermore, F as defined in Equation (20) gives the F-value, which indicates the harmonic mean of Precision and Recall.

$$F = \frac{2 \cdot \mathrm{POD} \cdot \mathrm{PAG}}{\mathrm{POD} + \mathrm{PAG}} \tag{20}$$

3. Results and Discussion

3.1. Experimental Method

The network was based on the method described in Section 2.2. Here, links are composed of node pairs (two airports) with more than 3650 flights. An example diagram of the network is presented in Figure 3.

For the case study employed in this work, when the number of flights per year

Table 3. Classification of prediction and fact.

Prediction/Fact	True (1)	False (0)
True (1)	TP	FP
False (0)	FN	TN

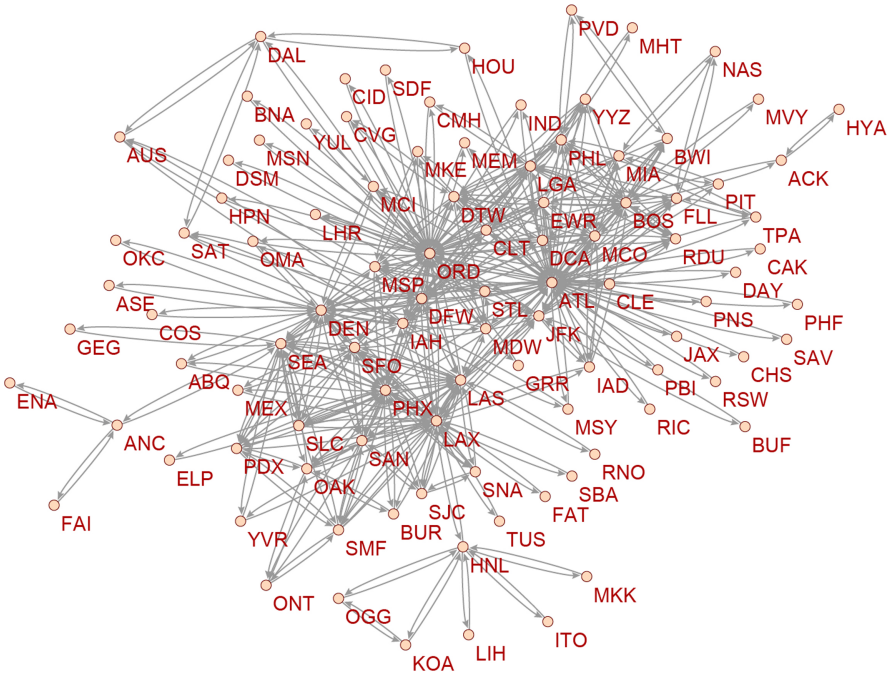

Figure 3. An example of a network.

between two given airports is below 3650, we consider that no connection exists, and a predicted link indicates that the number of flights exceeds the threshold of 3650.

3.2. Experimental Results

1) Prediction results based on ROC curve

Table 4 shows the F-value of each measure calculated using the ROC curve method. In the table, the values listed under the entry "08_09", for example, give the prediction from 2008 to 2009.

In terms of the average F-values, the CN value is highest and the values of PA, AA, and RA increase in the order listed. Additionally, the F-values for the CN measure remain the highest value throughout the years. This indicates that it is possible to predict whether two airports will be connected in the future by investigating whether or not they share many common airports. Furthermore, it is assumed that low degrees of common airports will likely result two airports being connected in the future because the F-values of the AA and RA are high.

In terms of the average Precision, the Precision of the CN measure has the highest score, followed by the values of PA, AA, and RA in the order listed, similar to the previously mentioned trend in the F-value. On the other hand, Recalls for the SP or LHN measures are higher than for the other measures. The average Recall for SP is 0.974. This result means that the vertex distance between most of the airport pairs is 2 for the pairs that will be connected in the future. This statement is equivalent to saying that two airports have common airports with more than 3650 flights. As LHN is defined as the score obtained by dividing the number of common nodes with the product of the node pair degrees, LHN is

Table 4. F-values calculated by the ROC curves method.

Measures /Year	08_09	09_10	10_11	11_12	12_13	13_14	Average
JI	0.01986755	0.019933555	0.013303769	0.013003901	0.018592297	0.026785714	0.018581131
PA	0.044897959	0.039312039	0.030769231	0.027548209	0.062111801	0.055384615	0.043337309
CN	0.080808081	0.080924855	0.064864865	0.045454545	0.075	0.096256684	0.073884839
SP	0.013832853	0.013221154	0.006810443	0.011420414	0.023880597	0.016793893	0.014326559
Sal	0.017424976	0.014522822	0.009960159	0.007168459	0.017094017	0.025940337	0.015351795
Sor	0.01986755	0.019933555	0.013303769	0.013003901	0.018592297	0.026785714	0.018581131
HPI	0.01659751	0.012311902	0.005191434	0.011744966	0.011173184	0.019555556	0.012762425
HDI	0.018120045	0.017094017	0.013559322	0.014577259	0.02173913	0.022764228	0.017975667
LHN	0.014449127	0.013767209	0.00702165	0.011851852	0.02143951	0.017120623	0.014274995
AA	0.055276382	0.052023121	0.033707865	0.023121387	0.056782334	0.055214724	0.046020969
RA	0.047930283	0.046875	0.03	0.022988506	0.038781163	0.062695925	0.041545146

high when the product of the node pair degrees is low. Therefore, the two airports do not contain a hub airport with high degree. That is to say, it indicates that two airports are likely connected when both of the two airports have low degrees.

2) Results from logistic regression method

The accuracy of the logistic regression prediction of TP (*i.e.*, a node pair that is predicted to be connected and is actually connected) is low, the reason for which will be explained below. As an example, Table 5 shows the 2010 aviation network prediction result, which indicates that most of the node pairs (two airports) are predicted to remain unconnected. Additionally, the partial regression coefficient is shown in Table 6. This behavior is also seen in the predictions for other years.

It is known that when the logistic regression model is applied to such imbalanced data, the data suggesting that the node pairs will remain unconnected strongly affects the result. This is likely the main reason for the low accuracy of TP.

The data used in this paper is imbalanced. The number of flights between two airports that exceed the threshold (3650) is much less than the number of flights below the threshold. In this case, generally weighting or other adjustments can be applied. However, in this paper, we do not apply such adjustments because the method of Kotegawa *et al.*, the basis of the current method, did not apply any adjustments.

On the other hand, we can estimate which measures might have impacts on the prediction from the partial regression coefficients shown in Table 6. Although the TP accuracy is low, the TN accuracy (recall-TN refers to the prediction that a node pair is unconnected and remained unconnected) is high. Thus it may be possible to obtain the knowledge of the variation of the aviation network from values of the regression coefficient of each measure in this point. For

Table 5. Prediction in 2009-2010.

Prediction/Fact	True (1)	False (0)
True (1)	0	0
False (0)	12	3338

Table 6. Values of the partial regression coefficient.

JI	PA	CN	SP	Sal	Sor	HPI	HDI	LHN	AA	RA
−0.1220	−0.1732	0.4830	0.3005	−0.3473	−0.2362	0.0571	0.3686	0.2599	0.1927	0.1087

example, the partial regression coefficient of the CN measure is high. We examine this result in the next section.

3) Results from the four-step method

It is general, the four-step method predicts traffic based on traffic engineering. As mentioned earlier, the four-step method does not use the network structure. The link prediction goal in this work is to predict the number of flights above the threshold. Therefore, we can regard our target as a prediction of the traffic amount, and compare our results with the four-step method prediction.

The distribution of predictions obtained by the four-step method is shown in Table 7. In addition, the coefficients used to determine in the linear regression model of G (generated traffic amounts) and A (attracted traffic amounts) are provided in Table 8. The target is the network from 2013 to 2014.

Table 7 shows that the accuracy of the prediction is low, and Table 8 shows that the coefficients of determination for G and A are low. This result indicates that the explanatory variables used in this paper cannot sufficiently satisfy the linear regression model of G and A.

3.3. Comparison and Consideration of Prediction Results

We first compare the result of the ROC curve method with the logistic regression model method. Although imbalanced data was applied in the ROC curve method, the impact is expected to be small. Specifically, it is noteworthy that Precision is high.

On the other hand, the CN value contributes to the TP in the ROC curve method and contributes to the TN (*i.e.*, a correct prediction that an unconnected node pair will remain unconnected) in the logistic regression model. Based on these results, we examine the prediction of the aviation network from the CN, PA, and AA values, which have high accuracy in the case of the ROC curve, and obtain the following insights:

A node pair (two airports) that will connect in the future (*i.e.*, the number of flights will increase) possesses three main characteristics in accordance with the definition of the CN, PA, and AA.

1) The product of node pair degrees is high.

2) The node pair has many common nodes.

3) The common nodes have low degrees.

Figure 4 shows an example of two airports in such a relationship.

Table 7. Prediction from 2013-2014.

Prediction/Fact	True (1)	False (0)
True (1)	0	99
False (0)	9	2320

Table 8. Coefficient of determination in linear regression.

	Decision Coefficient
G	0.000963
A	0.002016

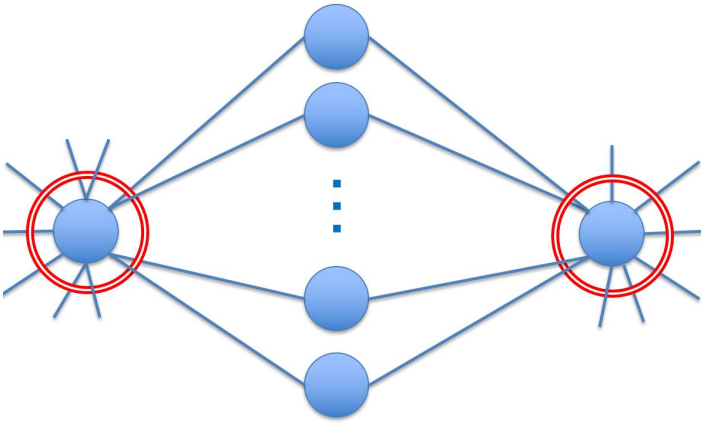

Figure 4. Example of two airports likely to be connected.

Next, we compare the result of the ROC curve method with that of the four-step method.

The four-step method needs statistical measures, such as population or income per airport, to generate the OD matrix and linear regression model. In this paper, we use relatively accessible data, such as the employed population, per-capita disposable income, the GDP, and whether a northeast corridor rail station exists or not in the state where the airport is located. The result may differ depending on the data selected. Therefore, we cannot conclude that the prediction by the four-step method is significantly inferior to the other methods in this study.

However, the ROC curve method can predict links without such data. The result of this study indicates that the ROC curve method has an advantage because it only requires measures from the network structure to predict links.

4. Conclusions

In this study, we highlight that the accurate prediction of future aviation networks is important because of industrial, environmental, and social aspects. Additionally, we considered aviation networks as network structures and tried to forecast their future development using link prediction, with measures based on the network structure.

At first, we defined the prediction measures based on the similarity of the node pairs calculated only from the network structure. Then, we created two methods to utilize those measures. The two methods are the ROC curve and the logistic regression model.

Next, we calculated the measures that achieve the highest prediction accuracy and contribution, and determined the growth mechanism of the aviation networks based on these measures.

As a case study, we applied our prediction method to the aviation networks in the US by creating a network of the number of flights.

In the link prediction for this aviation network, the accuracies of the CN, PA, and AA values were high in the ROC curve method. The CN measure contributed to the TN (unconnected node pair predicted to remain unconnected) in the logistic regression model method. We determined the three characteristics of a node pair (two airports), which increase flights: 1) the product of node pair degrees is high, 2) the node pair has many common nodes, and 3) the common nodes have low degrees.

Furthermore, we determined that the ROC curve method has advantages compared with the result of link prediction based on the four-step method.

We describe what we consider to be the novelty and utility of our work below.

The basis of the link prediction of the ROC curve method is the same as was employed by Zhou *et al*. to located missing links. However, the purpose of that study was to find the missing links, whereas we employed this method to generate predictions and interpret their relevance. The novelty in the current approach is that we demonstrate that it is possible to predict links and to connect this prediction to the growth mechanism of the network.

The link prediction based on the logistic regression model, as was employed in this work, is an expansion of the method of Kotegawa *et al*. While the explanatory variables used by Kotegawa *et al*. were unable to explain the mechanism of network generation, the explanatory variables selected in the current work were able to do so, suggesting that our method is advantageous in this respect.

A comparison of the link prediction results by the ROC curve and the four-step method indicates that the ROC curve method is better since it only requires measures from the network structure to predict the links and it achieves a certain level of accuracy.

In the future, we plan on further exploring the following aspects:

- To narrow down the number of predictions to achieve improved Precision in the ROC curve method.
- To create a way to evaluate the accuracy for imbalanced data, in which the number of positives and the number of negatives are significantly different.
- To extend the models to predict the disappearance of links.

References

[1] ICAO (2015) The World of Air Transport in 2015. http://www.icao.int/annual-report-2015/Pages/the-world-of-air-transport-in-2015.a

spx

[2] Liu, H., Tian, Y., Gao, Y., Bai, J. and Zheng, J. (2015) System of Systems Oriented Flight Vehicle Conceptual Design: Perspectives and Progresses. *Chinese Journal of Aeronautics*, **28**, 617-635.

[3] Sarkar, A.N. (2012) Evolving Green Aviation Transport System: A Holistic Approach to Sustainable Green Market Development. *American Journal of Climate Change*, **1**, 164-180. https://doi.org/10.4236/ajcc.2012.13014

[4] Moon, W.C., Yoo, K.E. and Choi, Y.C. (2011) Air Traffic Volume and Air Traffic Control Human Errors. *Journal of Transportation Technologies*, **1**, 47-53. https://doi.org/10.4236/jtts.2011.13007

[5] De Neufville, R., Odoni, A., Belobaba, P. and Reynolds, T. (2013) Airport Systems: Planning, Design and Management. McGraw-Hill, New York.

[6] Scheelhaase, J.D., Dahlmann, K., Jung, M., Keimel, H., Nieße, H., Sausen, R., Schaefer, M. and Wolters, F. (2016) How to Best Address Aviation's Full Climate Impact from an Economic Policy Point of View?—Main Results from AviClim Research Project. *Transportation Research Part D*, **45**, 112-125.

[7] Malavoltia, E. and Podesta, M. (2015) Strategic Reactions of Airlines to the European Trading Scheme. *Transportation Research Procedia*, **8**, 103-113.

[8] Guimera, R., Mossa, S., Turtschi, A. and Amaral, L.A.N. (2005) The Worldwide Air Transportation Network: Anomalous Centrality, Community Structure, and Cities' Global Roles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7794-7799. https://doi.org/10.1073/pnas.0407994102

[9] Sawai, H. (2012) Reorganizing a New Generation Airline Network Based on an Ant-Colony Optimization-Inspired Small-World Network. *Proceedings of WCCI* 2012 *IEEE Congress on Evolutionary Computation*, Brisbane, 10-15 June 2012, 1-8. https://doi.org/10.1109/CEC.2012.6256125

[10] Sawai, H. and Sato, A. (2016) Towards High-Performance Global Air Transportation Networks Using Socioeconomic-Environmental Data. *Proceedings of WCCI* 2016 *IEEE Congress on Evolutionary Computation*, Vancouver, 24-29 July 2016, 2888-2895. https://doi.org/10.1109/CEC.2016.7744154

[11] Bonnefoy, P. and Hansman, R. (2007) Potential Impacts of Very Light Jets in the National Airspace System. *Journal of Aircraft*, **44**, 1318-1326. https://doi.org/10.2514/1.26956

[12] Kotegawa, T., DeLaurentis, D. and Sengstacken, A. (2010) Development of Network Restructuring Models for Improved Air Traffic Forecasts. *Transportation Research Part C*, **18**, 937-949.

[13] Kotegawa, T., Fry, D., DeLaurentis, D. and Puchaty, E. (2014) Impact of Service Network Topology on Air Transportation Efficiency. *Transportation Research Part C*, **40**, 231-250.

[14] Wei, P., Chen, L. and Sun, D. (2014) Algebraic Connectivity Maximization of an Air Transportation Network: The Flight Routes' Addition/Deletion Problem. *Transportation Research Part E*, **61**, 13-27.

[15] Bureau of Transportation Statistics. http://www.rita.dot.gov/bts/

[16] Zhou, T., Lu, L. and Zhang, Y.C. (2009) Predicting Missing Links via Local Information. *The European Physical Journal B*, **71**, 623-630. https://doi.org/10.1140/epjb/e2009-00335-8

[17] Fuse, M., Nakajima, K. and Yagita, H. (2007) Outflow of Resources from Japan Focusing on End-of-Life Vehicles. *Materials Transactions*, **48**, 2436-2444. https://doi.org/10.2320/matertrans.MAW200712

[18] De Jong, G., Gunn, H. and Walker, W. (2004) National and International Freight Transport Models: An Overview and Ideas for Future Development. *Transport Reviews*, **24**, 103-124. https://doi.org/10.1080/0144164032000080494

[19] Chow, J.Y.J., Yang, C.H. and Regan, A.C. (2010) State-of-the-Art of Freight Forecast Modeling: Lessons Learned and the Road Ahead. *Transportation*, **37**, 1011-1030. https://doi.org/10.1007/s11116-010-9281-1