Title

A Practical Approach to Determining Critical Macroeconomic Factors in Air-traffic Volume based on K-means Clustering and Decision-tree Classification

Abstract

A given region's volume of air passengers and cargo is frequently taken to represent its economic development. This research proposes a practical methodology for investigating the inherent patterns of the relationships between air-traffic volume and macroeconomic development, utilizing data-mining techniques, including K-means clustering and Decision Tree C5.0 classification. Using the case of Taiwan from 2001 to 2014, 32 potential macroeconomic factors ascertained from a literature review were combined with air-traffic volume data to establish a 168-month dataset. After this dataset was grouped into five clusters, decision trees were implemented to determine its critical macroeconomic characteristics. The resulting four critical factors and their thresholds were the Information and Electronics Industrial Production Index (IE Index), at 83.22; National Income Per Capita, at US$3,222; Employed Population, at 10.134 million; and the Japanese Nikkei 225 Stock Average, at 10564.44. Among these, the IE Index was found to be the first critical factor relating to air-traffic volume as well as the only characteristic to distinguish Cluster V – 58 consecutive months from March 2010 to December 2014 inclusive – among others, and the reasonableness of this finding was confirmed via examination of detailed air-traffic statistics. Besides, the effectiveness of the four identified critical factors as predictive variables were validated by comparing forecasted results with actual air traffic volume from 2015 to 2016. Understanding these four critical factors and their relative importance is of great value to policymakers seeking to allocate limited resources optimally and objectively. Therefore, as an effective and efficient means of capturing significant and explainable macroeconomic factors influencing air-traffic volume, the proposed methodology can be applied to strategy formulation, operations management, and investment planning by governments, airports, airlines, and related entities.

Keywords

Air-traffic volume, K-means, decision trees, clustering and classification, macroeconomic factors

0. Introduction

With the advance of economic globalization, the volume of air traffic – comprising the carriage of both passengers and freight – has emerged as closely interrelated with regional economic development [1-4]. A better understanding of the mechanisms that link air-traffic volume to macroeconomic factors would, therefore, equip policymakers to arrive at more informed policy and strategy decisions regarding airports, airlines and related entities. A considerable body of research has estimated or forecasted air-traffic volumes over both the short and long term, and/or has explored the relations between air traffic and urban development, and this has yielded a variety of demand-forecasting models for both air passengers and cargo [5-7]. However, air-traffic volume is potentially influenced by an immense array of factors, most of which are to some extent stochastic, so such models do not always perform satisfactorily due to the limited number of factors each one of them takes into account. Meanwhile, those researchers who have explicitly focused on linkages between air-traffic volume and such macroeconomic factors as employment, trade volume and national income [8, 9] have mainly focused on a handful of suspected factors from their own subfields. In short, the complexity of the issues at hand has rendered it difficult in practice for researchers to consider all the relevant variables. However, the present paper argues that doing so is both necessary and possible.

Specifically, this study aims to uncover the inherent relationship between air-traffic volume and macroeconomic environment, using algorithmic data-mining techniques such as K-means clustering and decision-tree classification to determine the critical macroeconomic factors in this relationship from among an unprecedentedly wide range of potential macroeconomic factors, using the case of Taiwan from 2001 to 2014. After identifying the critical macroeconomic factors, an air-traffic volume forecasting model is developed to validate the effectiveness of the identified critical factors as predictive variables, with the data of Taiwan from 2015 to 2016. The results are expected to be useful in strategy formulation, operations management, and investment-program decisions by governments, airports, airlines, and related entities.

1. Literature Review

In air-traffic operations and management, volume data has always been highly prized by analysts, especially of demand. Accordingly, there is a sizable literature concerning forecasting air-traffic volume by proposing various predictive models, mainly including causal econometric models, time series models, and artificial intelligence models. Grosche et al. [5], for instance, forecast demand using two gravity models that took into account both economic activity and geographical characteristics. However, both ignored air traffic's historical features. Tsui et al. [7], in contrast, employed the Box-Jenkins Seasonal Autoregressive Integrated Moving Average (ARIMA) and ARIMA with additional explanatory variables (ARIMAX) models to estimate air-passenger volume for Hong Kong, based on historical data from 1993 to 2011. With monthly time-series historical data from 1990 to 2010, Scarpel [10] utilized a mixture of local experts' models to forecast air passenger volume at São Paulo International Airport. Their methodologies were capable of producing reasonable predictions, yet not highly accurate, as both ignored some potential impact factors. A more effective approach was Suryani et al.'s [6] dynamic model of the interaction between passenger demand and airport capacity, which took into consideration the cyclic action of the whole air-transportation system and regarded gross domestic product (GDP), population, and inflation as vital determinants of air-passenger volume. Moreover, using big data from search engine queries, Kim and Shin [11] developed a short-term air passenger demand predicting model to take into account the short-term fluctuations in the prediction. For long-term demand, Gelhausen et al. [12] proposed a more versatile model to estimate the effects of Brexit on air passenger volume in German for the years from 2016 to 2018. Air-cargo volume has also been predicted through parallel modeling techniques [13-15]. However, one of the fundamental steps shared by all such models is determining which factors are critical to air-traffic volume, and the question of how best to do this is still a topic of debate.

Many researchers who focus on correlations between air-traffic volume and other factors have come to suspect that particular factors are especially important based on data observation, logical reasoning, or statistical analysis. For instance, Hofer et al. [16] investigated the impact of socio-economic mobility on airfares and passenger volumes, indicating that higher socio-economic mobility is associated with lower airfares and more significant passenger volumes in the U.S. In addition, Dresner et al. [17] suggested that the magnitude of distance-adjusted airfare is one of the most important predictors of passenger demand. More importantly, given that air transport activity is closely related to the economic activity [18-20], recent studies have investigated a number of potential macroeconomic determinants of passenger volumes. For instance, Carmona-Benítez et al. [21] estimated air passenger demand based on economic variables such as the indicator of economic activity, the indicator of economically active population, national consumer price index, and foreign exchange earnings from international arrivals. Dobruszkes et al. [9] employed multiple-regression models to

arrive at the conclusion that GDP, levels of economic decision-making power, and tourism functions were the leading causes of variation in air service among eight independent variables. Hsu and Chao [22] investigated the relationships among commercial revenue, passenger service levels and space allocation in international passenger terminals, and based on their findings, proposed a model for optimizing space allocation for various types of stores. And Fildes et al. [23] selected the growth rate of income, trade, and price as independent variables to forecast the air travel demand from among several potential variables such as GDP, population, employment rates, airfares, and volume of trade. In short, more and more studies noticed the relationship between air-traffic volume and macroeconomic growth (e.g., GDP [9], national consumer price index [21], and employment rate [23]). Still, given how many different possible influences on air-traffic volume prior researchers have proposed [24-26], it is exceedingly difficult to determine which factors are pertinent – let alone critical – to a given region during a given period.

To discover previously unsuspected patterns or correlations in existing data, data mining is being widely applied in many fields, including biochemistry, geographical economics, and banking, among others [27-29]. Two of its basic techniques are clustering and classification. The former [30] involves dividing a dataset into several groups, known as clusters, according to some simple principles that ensure all data in a given cluster shares some specific commonalities. Clustering can be followed by classification [31], in which a classifier is trained to determine or predict which cluster a piece of sample data should be in, based on a set of previous clustering results. By applying these techniques to air-traffic volume analysis, it should be possible to distinguish critical macroeconomic factors scientifically and objectively from among all potential factors.

Clustering analysis is a non-supervisory pattern-recognition method, comprising both distance-based (e.g., K-means) and density-based algorithms (e.g., DBSCAN) among others. The K-means algorithm [32] is popular because it is simple to use. Classification, in contrast, is a supervisory method that creates classification rules based on empirical data; and once such rules are in place, the model can easily decide which class a given sample should belong to. Decision trees [33] constitute a commonly used classification method based on tree-like data structures and can create decision rules that are easily understood and implemented.

The above literature review indicates that, despite numerous studies having focused on internal and external factors related to air-traffic volume, none has proposed a practical general approach to objectively distinguishing which such factors, among all suspected or potential factors, are critically important. To fill this gap, the current study proposes a methodology for factor selection in air-traffic volume analysis based on data mining.

2. Methodology

As we have seen, it is widely acknowledged that air-traffic volume is to some extent related to the macrosocial, macroeconomic, and political environment. Still, to propose an accurate mathematical model of such volume has proved almost impossible, due to the difficulty of accurately quantifying the impacts of every potential factor. Nevertheless, the present researchers propose that it should be possible to distinguish the *relative* importance of numerous potential factors by using data-mining techniques.

The proposed process for data-mining analysis of air-traffic volume is presented in **Figure 1**. First, the volume data and potentially related macroeconomic statistics are collected and preprocessed to establish a database for data mining. Then, the K-means clustering algorithm is employed to group the data according to their similarities. Next, based on the clustering results, the C5.0 decision-tree

132 classification algorithm is run to determine which factors are the most critical to the relationship
133 between the traffic-volume and macroeconomic data. Lastly, the patterns revealed via clustering and
134 classification serve as a foundation for further analysis.

135 **< Fig. 1. Overview of the process for analysis of air-traffic volume using data-mining techniques >**

136 Given that various attributes or factors are generally recorded in different units, the values in
137 the original dataset's columns were not directly commensurable. To ensure that each attribute was
138 given fair and equal consideration during K-means clustering, the original dataset was normalized
139 before clustering using **Equation 1**, where x_o was the original value of the data, $x_{\min}$ was the minimum
140 value of the column (attribute), $x_{\max}$ its maximum value, and x_n its normalized value. Importantly, this
141 method of normalization eliminates inter-column differences in terms of units of measurement but
142 maintains the relative linear relationships of their values.

$$x_n = \frac{x_o - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

143 After normalization, the K-means algorithm was used to partition the data into k clusters with
144 high intra-cluster and low inter-cluster similarity. The similarity between any two pieces of data was
145 quantified in terms of Euclidean distance [34]. For instance, $\boldsymbol{a}=(x_1, x_2, x_3, \ldots, x_n)$ and $\boldsymbol{b}=(y_1, y_2, y_3, \ldots,$
146 $y_n)$ are two pieces of data in an n-attribute dataset, the distance between which is measured according
147 to **Equation 2**.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

148 As shown in **Equation 3**, the objective of the K-means algorithm is to minimize the total
149 distance between all pieces of data ($\boldsymbol{p}_i$, $i = 1, 2, \ldots, m$, with m being the number of data points) and the
150 centroid center of the cluster ($\boldsymbol{c}_j$, $j=1, 2, \ldots, k$, with k being the number of clusters). After achieving
151 the optimal partitioning, the dataset is divided into k clusters, and the data within any given cluster are
152 more similar to one another than they are to the members of any other cluster.

$$\min D_{total} = \sum_{j=1}^{k} \left[\sum_{p \in C_j} D(p, c_j) \right] \tag{3}$$

153 Based on the clustering results, the decision-tree algorithm is then employed to determine
154 effective rules for deciding whether or not to place a given piece of new data in a given cluster. Among
155 all attributes, the critical ones are distinguished according to information-gain theory [35]: i.e., the
156 attribute that can maximize the information-gain ratio is selected as the most critical one; and then, the
157 attribute that increases that ratio by the second-largest amount; and so on. Finally, tree-like decision
158 rules, including the critical factors and their thresholds, are generated – lending considerably more
159 structure to the air-traffic volume data, and thus enabling analyses of the rationality of the suggested
160 critical factors via both separate evaluation of each cluster, and cross-cluster comparisons. In short,
161 inherent patterns in the data become easier to recognize.

162 **3. The Case of Taiwan from 2001 to 2014**

163 A dominant air hub in the Asia-Pacific region, Taiwan has 17 airports, the dominant one being Taoyuan

International (TIA), with nearly 81% of passengers and more than 94% of cargoes during the 14-year period sampled. As indicated in **Figure 2**, passenger and cargo volumes both show periodical patterns: with the former being slightly higher in July and August every year, partially due to summer vacation, and the latter dropping in January, partially due to the New Year holiday. Moreover, a marked upward trajectory in passenger volumes began in 2009 due to the launch of cross-Strait airline services.

< Figure 2. Air traffic volume, January 2001 through December 2014 >

Based on the above literature review, 32 macroeconomic indicators were identified as potential factors in air-traffic volume; and eight types of traffic-volume statistics generated at TIA were also added to the dataset as a reference for clustering analysis. Specifically, the selected indicators took account of the national economy, stock market, foreign exchange, domestic business environment, import and export trade, industrial development, and human resources. The 32 macroeconomic indicators were further classified into seven groups, as shown in **Figure 3**; and all indicators and statistics are listed within their corresponding categories in **Figure 4**. Therefore, the dataset consisted of 40 columns (i.e., 32 macroeconomic indicators and eight traffic-volume statistics) for 168 months (**Table 1**). The results of the data-normalization procedure described above are briefly listed in **Table 2**.

< Figure 3. Attributes in the dataset for clustering >

< Figure 4. Attribute lists for clustering >

< Table 1. Original dataset for Taiwan air-traffic volume analysis >

< Table 2. Normalized dataset for Taiwan air-traffic volume analysis >

Setting the number of expected clusters (k) was a prerequisite of applying the K-means clustering algorithm to the normalized dataset. Given the size of the data (168 lines), and to ensure that the number of samples in each cluster was higher than 10% of the total number, we set $k = 5$. As shown in Figure 5, the five resulting clusters contained 33, 34, 26, 17, and 58 months, respectively, with the samples within each cluster being temporally uninterrupted: e.g., Cluster V contained 58 consecutive months from March 2010 to December 2014.

< Figure 5. K-means algorithm clustering results >

The use of the decision-tree algorithm yielded four critical macroeconomic factors for distinguishing between clusters, which are presented in **Figure 6-a**. The tree, including these four factors' original thresholds, is visualized in **Figure 6-b**, and a pie chart of both the clustering and classification results is depicted in **Figure 6-c**. This process effectively boosted the data's interpretability by revealing the similarities within, and consistency of, each cluster. For instance, in Cluster IV (17 months from October 2008 to February 2010 inclusive), the Information and Electronics Industrial Production Index (IE Index) was always less than or equal to 83.22; the National Income Per Capita (NIPC), always greater than US$3,222; the Employed Population (EP), always greater than 10.134 million; and the Japanese Nikkei 225 Stock Average (Nikkei 225), always less than or equal to 10564.44.

< Figure 6. Decision-tree algorithm classification results >

As **Figure 6** makes clear, the critical macroeconomic factors for air-traffic volume varied over time. The most notable characteristic of Cluster V, for example, is that the IE Index was greater than 83.22 throughout, unlike in any of the other four clusters. Among the remaining four clusters, Cluster

I was unique in that the NIPC remained less than or equal to US$3,222. Similarly, Cluster II was distinguishable from the remaining two clusters by having an EP of above 10.134 million, while the remaining two clusters were distinguishable from one another in that Cluster III's Nikkei 225 values were less than or equal to the threshold, and Cluster IV's greater than it. In other words, the overall health of the national economy, the size of the working population, stock-market performance, and industrial production successively occupied critical positions vis-à-vis air-traffic volume in Taiwan from 2001 to 2014.

4. Discussion and Evaluation

4.1 Discussion of cluster features

Based on 14 years' worth of Taiwanese air-traffic volume data and the trend curves noted above, a brief summation of all clusters is provided in **Figure 7**. The main legends show the maximum, minimum and average monthly volumes of each cluster, and its variation trends (i.e., whether such volumes increased or decreased per month on average); and the bar legends display the normalized mean values of the corresponding critical factors. Passenger volume generally declined across Clusters I, II and III (i.e., January 2001 through September 2008), though the rate of decline slowed over the same 93-month period. Then, across Clusters IV and V (from October 2008 to December 2014), passenger volume increased. Cargo volume, in contrast, increased in Clusters I, II, and IV and decreased in Clusters III and V. Nevertheless, the monthly average cargo volume in Cluster V was the highest among all clusters.

< Figure 7. Brief summations of all clusters >

Inter-cluster differences embody the inherent correlations between air-traffic volumes and specific macroeconomic factors. In Cluster V, the prominence of the IE Index – identified above as the primary factor distinguishing that cluster – suggests a spike in industrial development during the period in question. Given that both air-passenger and air-cargo volumes were highest in Cluster V, it would be rational to infer a strong connection between the increases in air-traffic volume and industrial development – especially in the information and electronics industries, at least in the recent Taiwanese case. Additionally, Clusters III and IV were associated with both a huge shock to the Nikkei 225 and the lowest overall air-traffic volumes in the sampled 14-year period. Therefore, it is reasonable to associate stock-market slumps with decreases in air-traffic volumes. However, while Cluster I is marked by the NIPC being lower than in any other cluster, there is no noticeable explicit correlation between air-traffic volume and NIPC.

Figure 8 illustrates the geographical distribution of air-traffic volume for each cluster. Looking at the destinations of passengers departing from Taiwan, Asia (**Fig. 8-(a)**) dominated, with more than 80% of the total volume at all times, within which the share of these departures held by Mainland China (**Fig. 8-(b)**) steadily increased once cross-Strait services were launched (i.e., from Clusters III through V). The passenger proportion between Taiwan and America (**Fig. 8-(a)**), in contrast, decreased after the 2008 global economic crisis. Similarly, the proportion of air cargoes imported to Taiwan from other places in Asia increased at a steady rate, while imports from America gradually lost their leading role. Meanwhile, the proportion of Taiwan's exports that went to other Asian countries was more stable than the corresponding proportion of imports; and the proportion of all inbound air cargoes that originated in Mainland China was much higher than the proportion of all such cargoes originating in Taiwan that were sent there. Such geographic variation in air-traffic volumes across clusters was also related to variation in the macroeconomic factors that the present study identified as critical. For instance, the Nikkei 225 experienced a sharp decline in Cluster IV, partially due to the 2008 economic

crisis in the United States, and the proportion of all air-cargo exports from Taiwan that went to America reached its lowest point in Cluster IV.

< Figure 8. Geographical distribution of air-traffic volumes in all clusters >

These explicit and credible links between air-traffic volume and other factors in the case of Taiwan from 2001 to 2014 tend to confirm the effectiveness of the clustering-and-classification approach.

4.2 Evaluation of the identified critical factors

The four macroeconomic factors critical to air-traffic volumes, as revealed by decision-tree analysis aimed at clarifying the characteristics of each cluster, are plotted visually in **Figure 9**. The values of these factors in the sampled 168 months show an obvious crowding effect, i.e., that the data in a given cluster are located in a relatively concentrated area. For instance, the data points in Cluster V are located in the top right corner of **Figure 9-(a)** and at the far-right side of **Figure 9-(b)**, which indicates relatively higher values of IE Index, NIPC, and EP. Oppositely, the data points in Cluster I are located in the bottom left corner of **Figure 9–(a)** and at the far-left side of **Figure 9-(b)**, which corresponds to relatively lower values of IE Index, NIPC, and EP. Additionally, despite the difficulties in separating Cluster IV from Cluster III in **Figure 9-(a)**, it is easy to distinguish between them in **Figure 9-(b)**: data points in Cluster III have greater values of Nikkei 225 than those in Cluster IV. Indeed, the geometrical crowding effect shown in **Figure 9** represents key evidence in support of the clustering-and-classification approach's ability to give structure to data. That is to say, the proposed data-mining approach is capable of capturing the inherent pattern of the given data without artificial intervention.

< Figure 9. Visualization of the clustering results for the four critical macroeconomic factors >

To have a better understanding of the four identified critical macroeconomic factors, the trend curves of them are plotted in **Figure 10**. The IE Index, similar to the cargo volume **(Fig.2-b)**, exhibits a marked annual pattern, with January having the lowest value every year, again partially ascribable to the New Year holiday. The NIPC, on the other hand, shows a different pattern: with people usually earning more income in the first quarter of every year than in the other three quarters. The EP increased steadily from 2001 to 2014, apart from during the 2008-9 global economic crisis. Nevertheless, the overall trend of these three critical factors was upward for the 14-year period as a whole; whereas the Nikkei 225 experienced big ups and downs with no immediately obvious pattern or trajectory.

< Figure 10. Trend curves and thresholds of the four critical factors >

To evaluate the similarity and dissimilarity of all factors, the 40 attributes were grouped into three clusters using K-means according to their trends in variation over time. As shown in **Figure 11**, 24 of the 40 attributes were clustered in T1, because they shared the characteristic of tending to increase markedly as time went by. Another nine of the 40 were placed in T2, due to their tendency to increase over time, but only slightly, and with an obvious shock during the period 2008-10; and the remaining seven were clustered in T3, due to their general decreasing trend. None of the four critical factors clustered in T3, and all but the Nikkei 225 were placed in T1.

< Figure 11. K-means attribute-clustering results >

The present study's methodology identified the IE Index as the most critical macroeconomic factor, and as previously mentioned, this factor was the key distinguishing characteristic of Cluster V. Intuitively, this result appears to be correct, since, electronics cargoes were an important component of all air cargoes (as plotted in **Figure 12**) within the time period represented by Cluster V. In other words, even though the researchers did not input copious details of air-cargo composition, the proposed methodology showed itself capable of capturing this important data relationship.

< Figure 12. Breakdown of cargo volume within Cluster V >

Compared with some established air-traffic volume prediction models [5, 36, 37], the four identified critical macroeconomic factors above are consistent with the predictors of those predictive models in principle, indicating that industrial production index, national income, employment status, and stock market prosperity have been recognized as potential impact factors in air-traffic volume in different models. However, many models [6, 9, 18] seek to quantify the relationship between GDP and air-traffic volume and thus employ GDP as an important predictor for air-traffic demand prediction, whereas GDP is not recognized as one of the critical factors in our sampled dataset. It is worth noting that, through the proposed data-mining approach, the identified critical factors might change from case to case since they are determined by the sampled data instead of artificial causal analysis. Therefore, the results from the proposed methodology are not leading to the conclusion that GDP is not a proper predictor for air-traffic demand prediction, but that the identified four factors are more critical to air-traffic volume in the sample Taiwanese case from 2001 to 2014 when grouping them into 5 clusters. Therefore, not contradictory to the existing predictive models in predictor selection, the proposed methodology is capable of identifying the critical factors objectively according to the data itself instead of artificial causal analysis.

Quantitatively, to validate the effectiveness of the four identified factors as forecasting variables, the ordinary least-square (OLS) linear regression model has been trained based on the data from 2001 to 2014 to predict the monthly air traffic volume of Taiwan Taoyuan International Airport from January 2015 through December 2016. The prediction model consists of two equations with air passenger volume and air cargo volume as dependent variables, respectively, which are specified as follows:

$$y_{pax} = \alpha_0 + \alpha_{\text{IE Index}} x_{\text{IE Index}} + a_{\text{NIPC}} x_{\text{NIPC}} + \alpha_{\text{EP}} x_{\text{EP}} + \alpha_{\text{Nikkei225}} x_{\text{Nikkei225}} \quad \textbf{(4)}$$

$$y_{cargo} = \beta_0 + \beta_{\text{IE Index}} x_{\text{IE Index}} + \beta_{\text{NIPC}} x_{\text{NIPC}} + \beta_{\text{EP}} x_{\text{EP}} + \beta_{\text{Nikkei225}} x_{\text{Nikkei225}} \quad \textbf{(5)}$$

where α, β denote coefficients of the linear regression model. Based on 168 observations in the normalized dataset (Table 2), the regression results of both passenger volume and cargo volume are presented in Table 3. As shown, IE Index, EP, and Nikkei 225 are significant variables for air passenger volume, while IE Index, NIPC, and EP are significant variables for air cargo volume. In both models, the coefficient of IE Index is the highest among the four variables, which is in accordance with the fact that IE Index is identified as the most important critical factor in the proposed data-mining methodology. Moreover, the R-squared is 0.792 for passenger volume and is 0.706 for cargo volume, meaning that the four identified variables are able to explain more than 70% of air-traffic volume variance.

< Table 3. Regression results using the normalized dataset from 2001 to 2014 >

By using the regression results in Table 3, the forecasted results are compared with the actual air traffic volume in Figure 13. The average relative error is 13% for passenger volume forecast, with the highest error of 25% and the lowest error of 2%, and 6% for cargo volume forecast, with the highest error of 18% and the lowest error of 1%. Although the forecasted results do not agree perfectly with the actual air traffic volume, the result still shows that the four identified critical factors are effective as predictive variables for air traffic volume forecasting in the case of Taiwan in the near future of the sampled period.

<Figure 13. Forecasted vs. actual air traffic volume, January 2015 through December 2016>

In short, the results obtained from the proposed methodology indicate that it accords well with

the demands of both rationality and practicability for predictive models, meaning that it is suitable for air-traffic volume analysis, and can be effectively utilized to recognize the critical macroeconomic factors affecting such volume from among a wide range of potential factors. Such recognition is fundamental to many forecasting models for air-traffic volume. Moreover, the proposed approach's relative simplicity means that it should be readily adaptable to most regions of the world, adding further to its potential value to policymakers and others responsible for operational strategy.

5. Conclusions

This study proposes a practical approach to identifying the macroeconomic factors critical to air-traffic volumes from among a wide range of potential factors, using the data-mining techniques K-means clustering and decision-tree classification. The results, based on Taiwanese data covering the 168 months of the period 2001-14 that were divided into five clusters and four critical macroeconomic factors from among 32 potential such factors, were evaluated and found to be rational. Moreover, the effectiveness of the identified critical factors for predicting the air traffic volume was validated by comparing the forecasted results and actual air traffic volume from 2015 to 2016. Therefore, as an effective and efficient means of capturing significant and explainable macroeconomic factors influencing air-traffic volume, the proposed approach could be further applied to air-traffic demand estimation, and thus would be of considerable interest to air-transportation strategists and policymakers. Previous studies of air-traffic demand have mainly focused on model selection, which often requires identifying independent variables (impact factors) based on literature reviews. As such, the most important contribution of the methodology proposed in the current study is arguably its ability to select such variables objectively and relatively quickly. In this context, one limitation of this study is the ignorance of causality issues (i.e., the causal relationship between macroeconomic factors and air-traffic volume). However, the result obtained from the proposed methodology is still practically valid based on the papers by and Baker et al. (2015) [19] and Hakim and Merkert (2016) [20] which have validated the causal relationship between air transport and economic growth. After identifying the critical macroeconomic factors as predictive variables, how to quantify the relationship between these factors and air-traffic demand and thus to establish a more accurate demand-estimation model is still a worthwhile endeavor that needs to be tackled in our future research.

References

1. H. Matsumoto, *International urban systems and air passenger and cargo flows: some calculations.* Journal of Air Transport Management, 2004. **10**(4): p. 239-247.
2. J. Khadaroo and B. Seetanah, *The role of transport infrastructure in international tourism development: A gravity model approach.* Tourism Management, 2008. **29**(5): p. 831-840.
3. K. Yamaguchi, *International trade and air cargo: Analysis of US export and air transport policy.* Transportation Research Part E: Logistics and Transportation Review, 2008. **44**(4): p. 653-663.
4. B. Miller and J.-P. Clarke, *The hidden value of air transportation infrastructure.* Technological Forecasting and Social Change, 2007. **74**(1): p. 18-35.
5. T. Grosche, F. Rothlauf, and A. Heinzl, *Gravity models for airline passenger volume estimation.* Journal of Air Transport Management, 2007. **13**(4): p. 175-183.
6. E. Suryani, S.-Y. Chou, and C.-H. Chen, *Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework.* Expert Systems with Applications, 2010. **37**(3): p. 2324-2339.
7. W.H.K. Tsui, H. Ozer Balli, A. Gilbey, and H. Gow, *Forecasting of Hong Kong airport's passenger throughput.* Tourism Management, 2014. **42**: p. 62-76.
8. J.K. Brueckner, *Airline Traffic and Urban Economic Development.* Urban Studies, 2003. **40**(8): p. 1455-1469.
9. F. Dobruszkes, M. Lennert, and G. Van Hamme, *An analysis of the determinants of air traffic volume for European metropolitan areas.* Journal of Transport Geography, 2011. **19**(4): p. 755-762.
10. R.A. Scarpel, *Forecasting air passengers at São Paulo International Airport using a mixture of local experts model.* Journal of Air Transport Management, 2013. **26**: p. 35-39.
11. S. Kim and D.H. Shin, *Forecasting short-term air passenger demand using big data from search engine queries.* Automation in Construction, 2016. **70**: p. 98-108.
12. M.C. Gelhausen, P. Berster, and D. Wilken, *A new direct demand model of long-term forecasting air passengers and air transport movements at German airports.* Journal of Air Transport Management, 2018. **71**: p. 140-152.
13. C.-C. Hwang and G.-C. Shiao, *Analyzing air cargo flows of international routes: an empirical study of Taiwan Taoyuan International Airport.* Journal of Transport Geography, 2011. **19**(4): p. 738-744.
14. A. Regan and R. Garrido, *Modelling freight demand and shipper behaviour: state of the art, future directions*. 2001.
15. B. Graham, *Airport-specific traffic forecasts: a critical perspective.* Journal of Transport Geography, 1999. **7**(4): p. 285-289.
16. C. Hofer, R. Kali, and F. Mendez, *Socio-economic mobility and air passenger demand in the U.S.* Transportation Research Part A: Policy and Practice, 2018. **112**: p. 85-94.
17. M. Dresner, C. Eroglu, C. Hofer, F. Mendez, and K. Tan, *The impact of Gulf carrier competition on U.S. airlines.* Transportation Research Part A: Policy and Practice, 2015. **79**: p. 31-41.
18. V. Profillidis and G. Botzoris, *Air passenger transport and economic activity.* Journal of Air Transport Management, 2015. **49**: p. 23-27.
19. D. Baker, R. Merkert, and M. Kamruzzaman, *Regional aviation and economic growth:*

cointegration and causality analysis in Australia. Journal of Transport Geography, 2015. **43**: p. 140-150.
20. M.M. Hakim and R. Merkert, *The causal relationship between air transport and economic growth: Empirical evidence from South Asia.* Journal of Transport Geography, 2016. **56**: p. 120-127.
21. R.B. Carmona-Benítez, M.R. Nieto, and D. Miranda, *An Econometric Dynamic Model to estimate passenger demand for air transport industry.* Transportation Research Procedia, 2017. **25**: p. 17-29.
22. C.-I. Hsu and C.-C. Chao, *Space allocation for commercial activities at international passenger terminals.* Transportation Research Part E: Logistics and Transportation Review, 2005. **41**(1): p. 29-51.
23. R. Fildes, Y. Wei, and S. Ismail, *Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures.* International Journal of Forecasting, 2011. **27**(3): p. 902-922.
24. S.L. Brown and W.S. Watkins, *The Demand for Air Travel: A Regression Study of Time-series and Cross-sectional Data in the U.S. Domestic Market*. 1970: National Research Council . Highway Research Board.
25. E. Erraitab, *An Econometric Analysis Of Air Travel Demand: The Moroccan Case.* European Scientific Journal, 2016. **12**(7).
26. J.D. Jorge-Calderón, *A demand model for scheduled airline services on international European routes.* Journal of Air Transport Management, 1997. **3**(1): p. 23-35.
27. L. Guoxiang and Q. Zhiheng. *Data Mining Applications in Marketing Strategy*. in *2013 Third International Conference on Intelligent System Design and Engineering Applications*. 2013.
28. S. Tsumoto and S. Hirano, *Risk mining in medicine: Application of data mining to medical risk management.* Fundamenta Informaticae, 2010. **98**(1): p. 107-121.
29. J. Mennis and D. Guo, *Spatial data mining and geographic knowledge discovery—An introduction.* Computers, Environment, Urban Systems, 2009. **33**(6): p. 403-408.
30. A.K. Jain and R.C. Dubes, *Algorithms for clustering data.* 1988.
31. P. Breheny, *Classification and regression trees.* 1984.
32. A.K. Jain, *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
33. R. Pandya and J. Pandya, *C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning.* International Journal of Computer Applications, 2015. **117**(16): p. 18-21.
34. J.A. Hartigan and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C, 1979. **28**(1): p. 100-108.
35. E. Harris. *Information Gain Versus Gain Ratio: A Study of Split Method Biases*. in *ISAIM*. 2002.
36. J.T. Fite, G. Don Taylor, J.S. Usher, J.R. English, and J.N. Roberts, *Forecasting freight demand using economic indices.* International Journal of Physical Distribution & Logistics Management, 2002. **32**(4): p. 299-308.
37. Z. Wadud, *The asymmetric effects of income and fuel price on air transport demand.* Transportation Research Part A: Policy and Practice, 2014. **65**: p. 92-102.

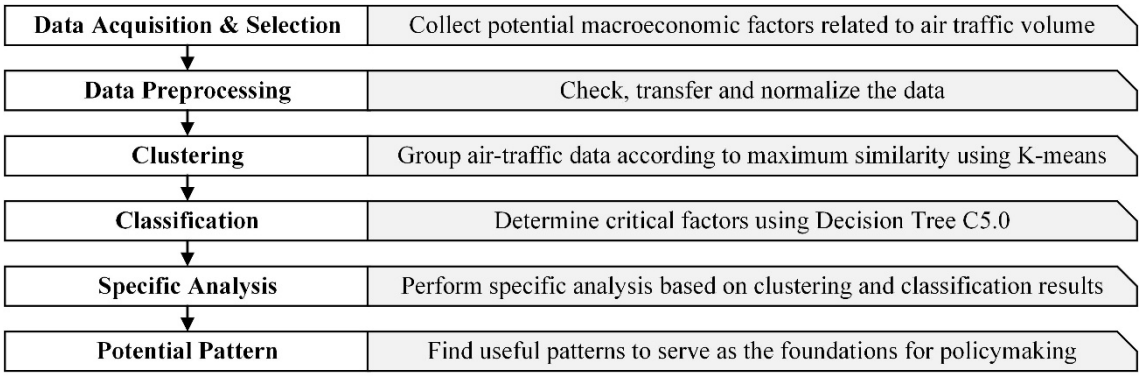
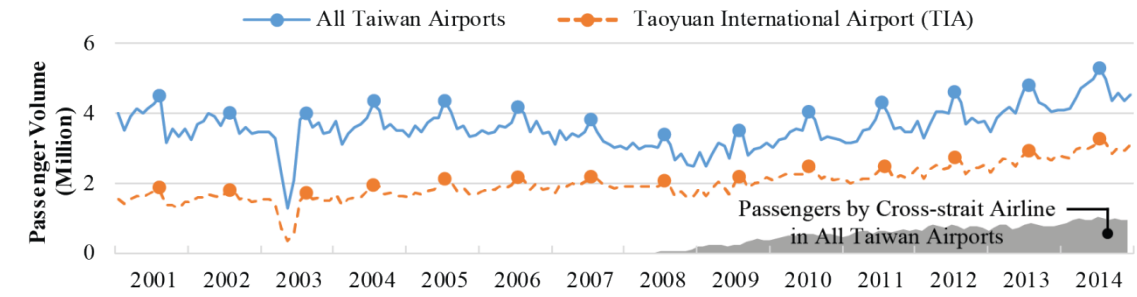
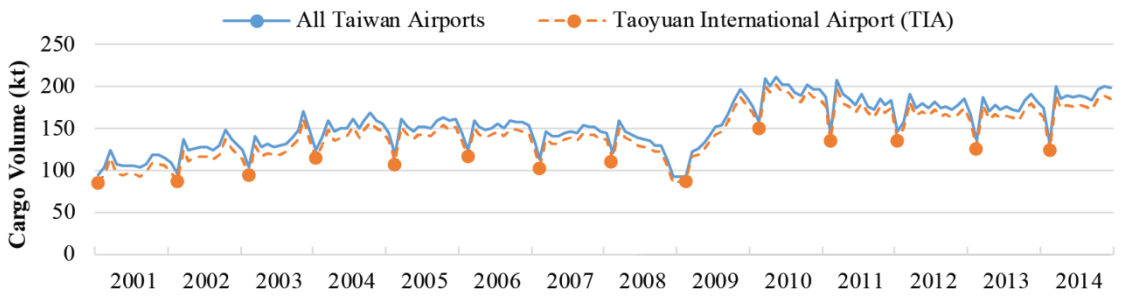


Figure 1. Overview of the process for analysis of air-traffic volume using data-mining techniques



(a) Passenger volume



(b) Cargo volume

Figure 2. Air-traffic volume, January 2001 through December 2014

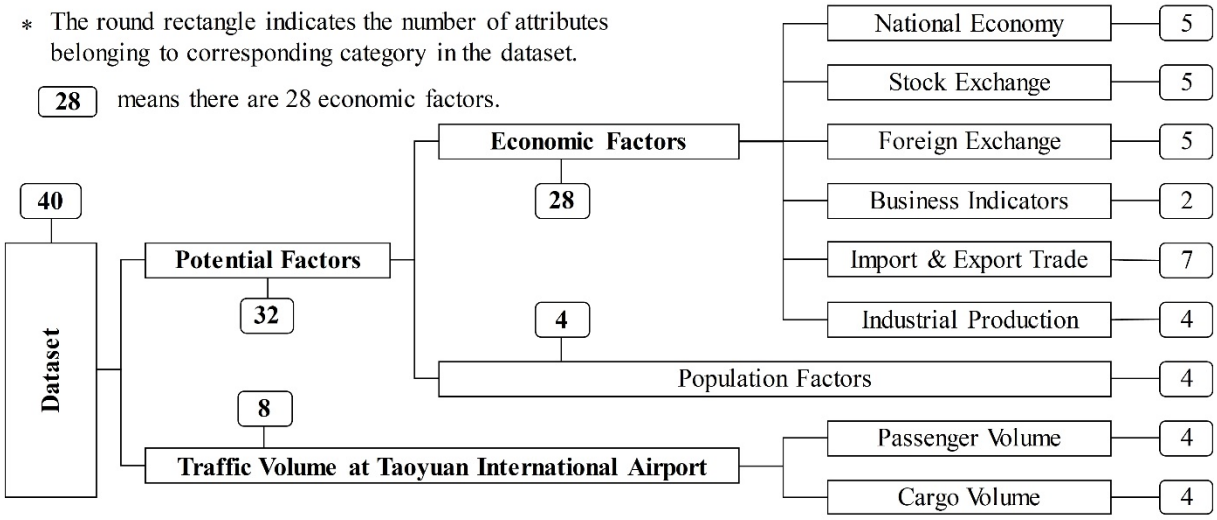


Figure 3. Attributes in the dataset for clustering

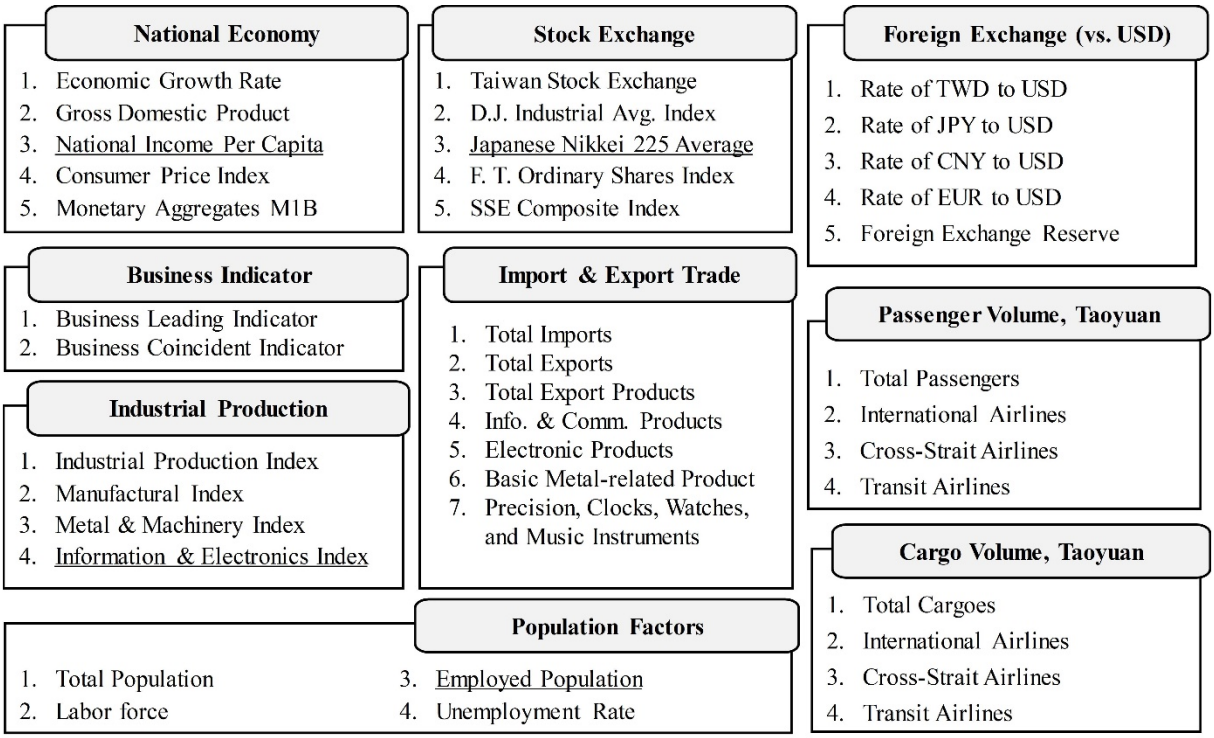


Figure 4. Attributes lists for clustering

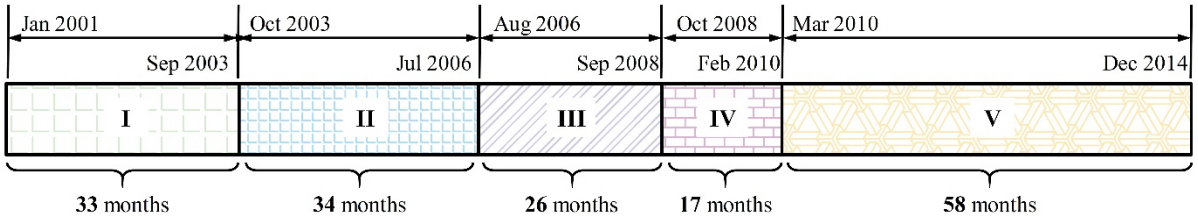
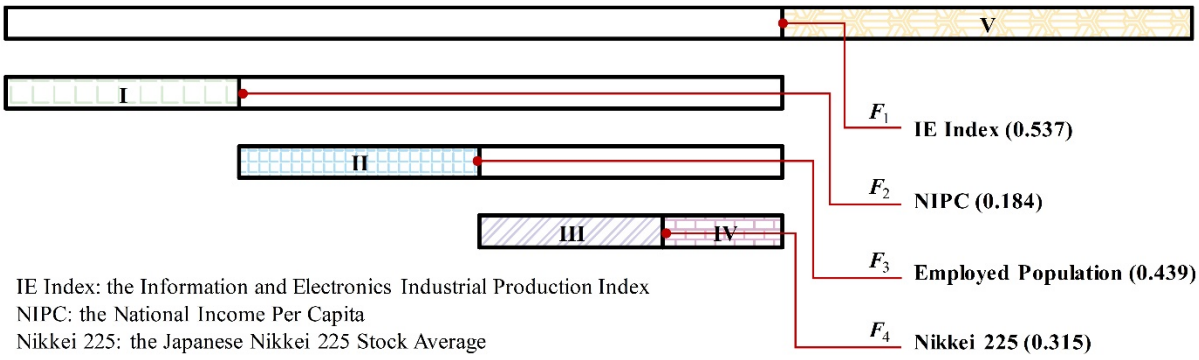
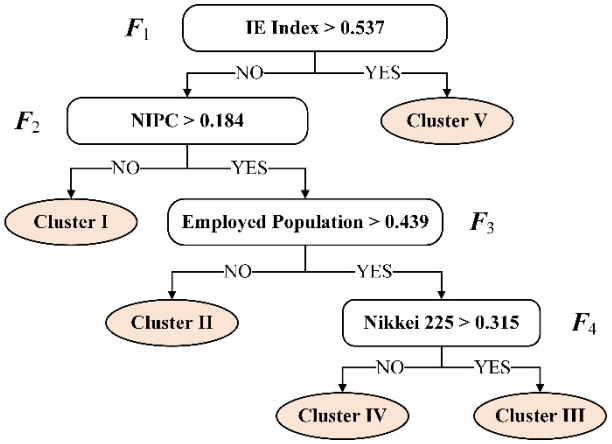


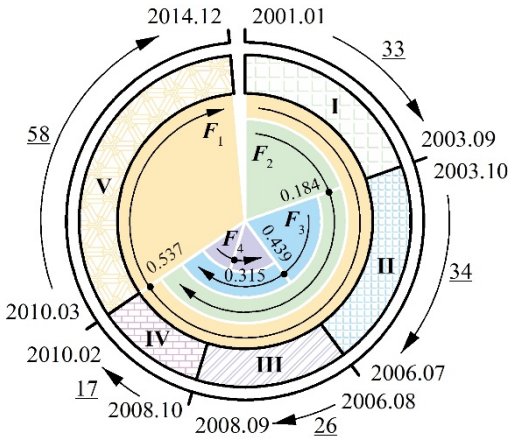
Figure 5. K-means algorithm clustering results



(a) Critical factors discerned by the decision-tree classification algorithm

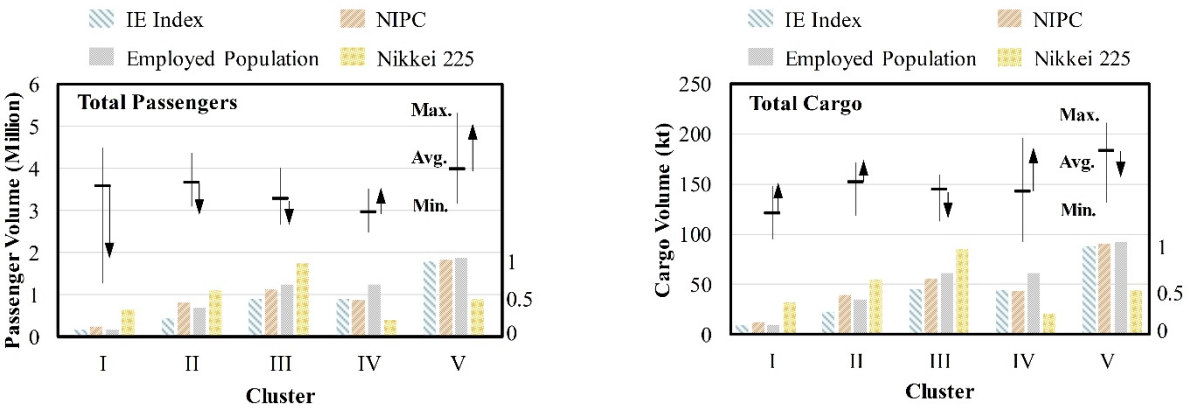


(b) Visualization of decision-tree results



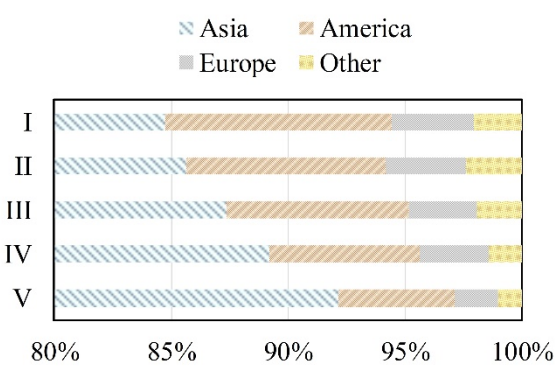
(c) Pie chart of classification results

Figure 6. Decision-tree algorithm classification results

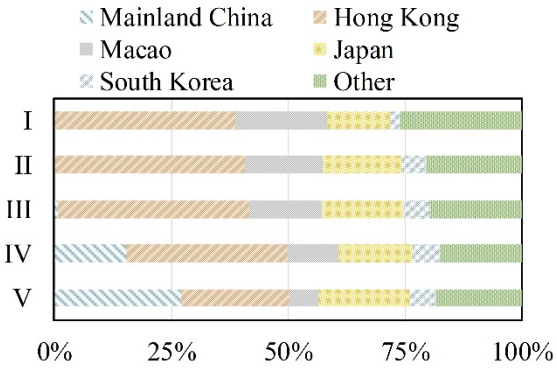


(a) Total Passengers (b) Total Cargo

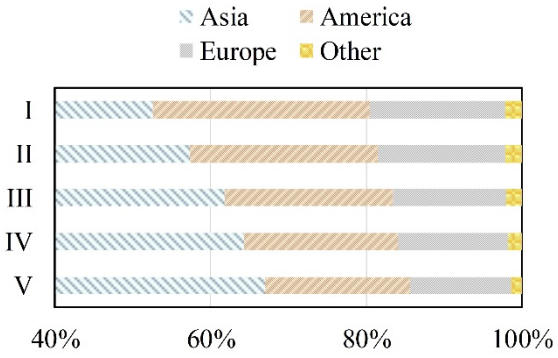
Figure 7. Brief summations of all clusters



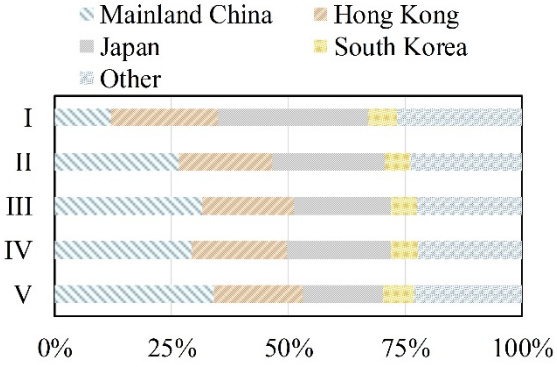
(a) Breakdown of worldwide destinations



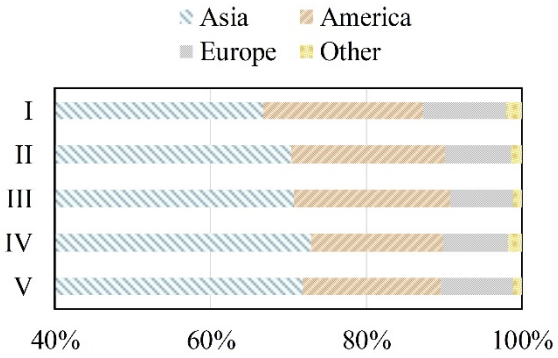
(b) Breakdown of destinations in Asia



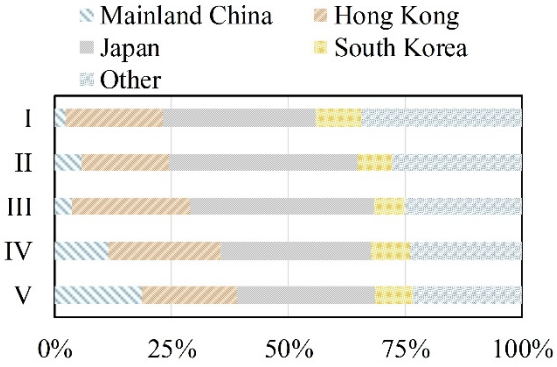
(c) Imports, worldwide



(d) Imports from Asia

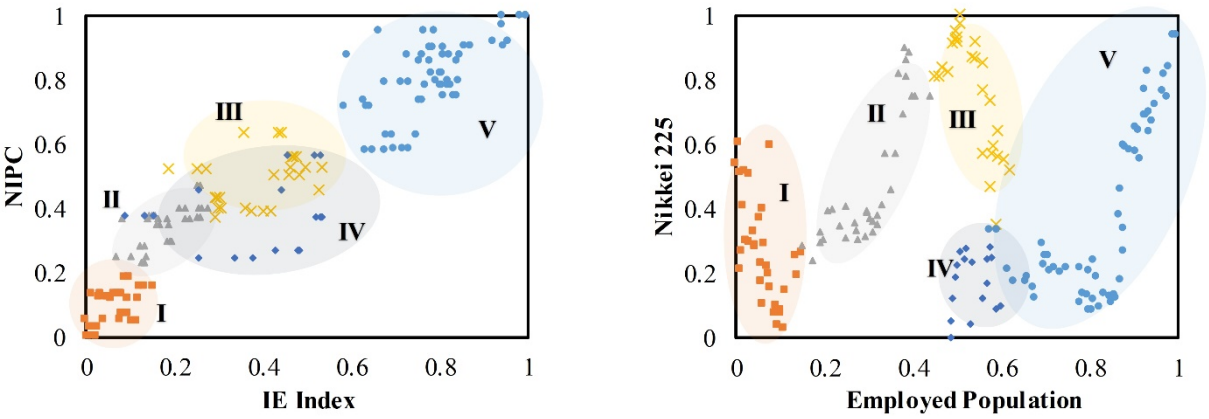


(e) Exports, worldwide



(f) Exports to Asia

Fig.8 Geographical distribution of air traffic volume in all clusters



(a) IE Index vs. NIPC (b) Employed Population vs. Nikkei 225

Figure 9. Visualization of clustering results for the critical macroeconomic factors

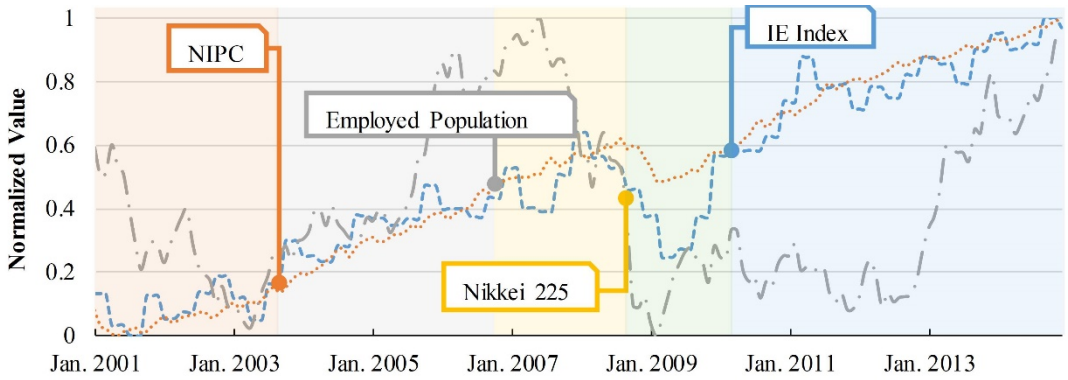


Figure 10. Trend curves and thresholds of four critical factors

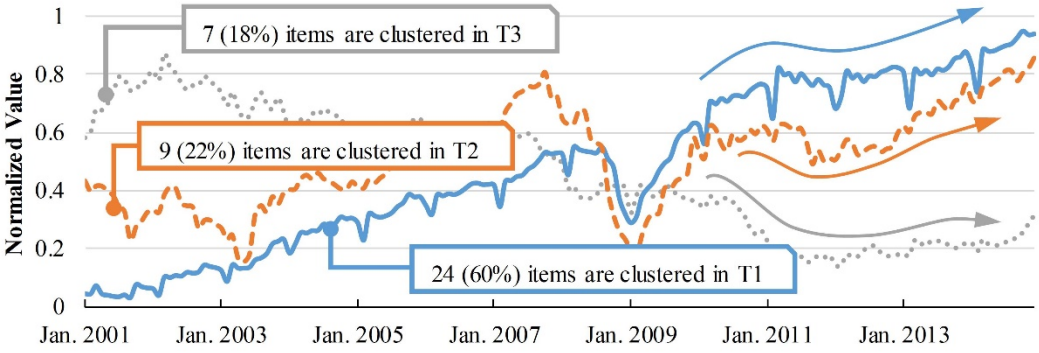
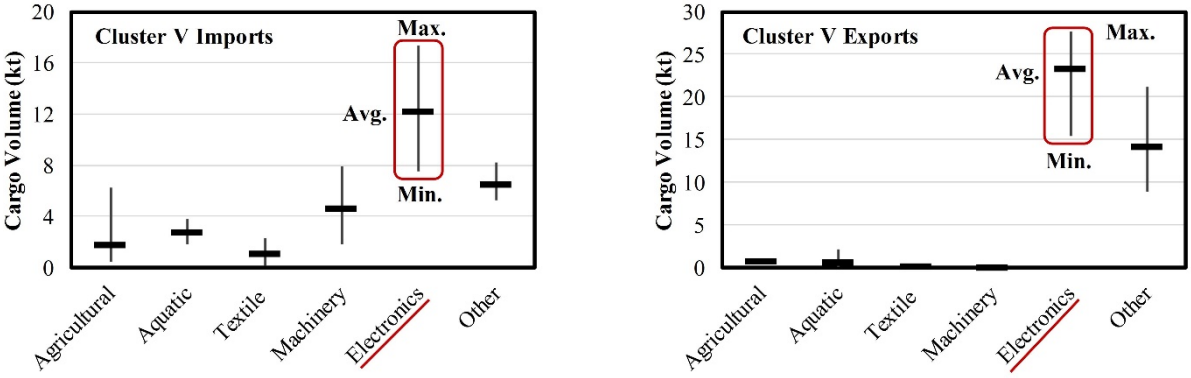
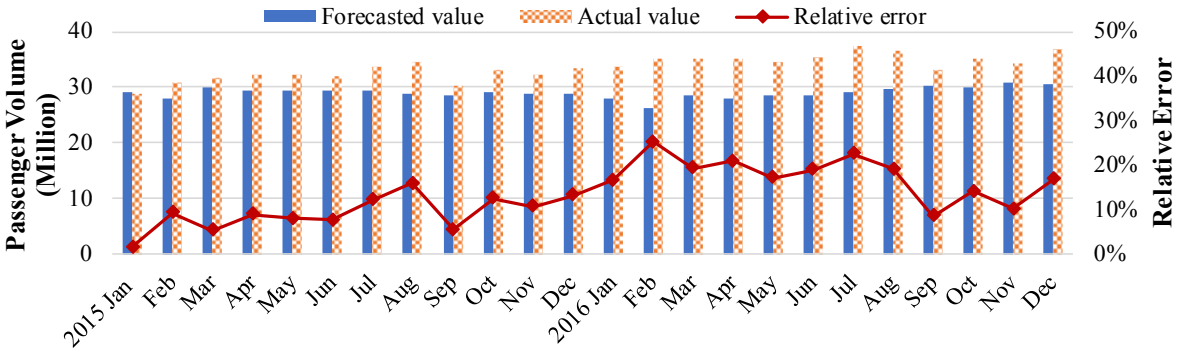


Figure 11. K-means attribute-clustering results

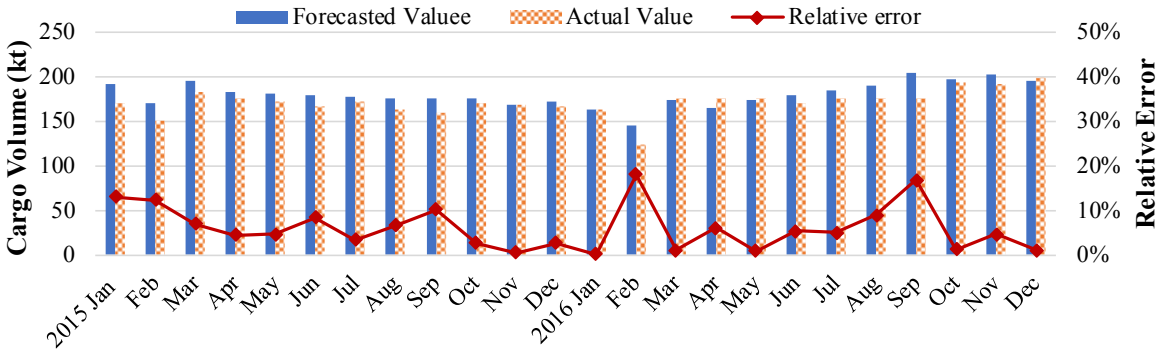


(a) Imports (b) Exports

Figure 12. Breakdown of cargo volume within Cluster V



(a) Passenger volume



(b) Cargo volume

Figure 13. Forecasted vs. actual air traffic volume, January 2015 through December 2016

Table 1. Original dataset for Taiwan air-traffic volume analysis

Attribute	1	2	3	…	40
	Economic Growth Rate	Gross Domestic Product	National Income Per Capita	…	Transit Cargo Volume, Taoyuan
Unit	%	US$	US$	…	kt
1	1.59	2533134	3109	…	19980
2	1.59	2533134	3109	…	19585
3	1.59	2533134	3109	…	22786
…	…	…	…	…	…
168	3.47	4228272	4875	…	90751

Table 2. Normalized dataset for Taiwan air-traffic volume analysis

Attribute	1	2	3	…	40
	Economic Growth Rate	Gross Domestic Product	National Income Per Capita	…	Transit Cargo Volume, Taoyuan
1	0.46	0.04	0.13	…	0.03
2	0.46	0.04	0.13	…	0.03
3	0.46	0.04	0.13	…	0.06
…	…	…	…	…	…
168	0.55	1.00	0.97	…	0.84

Table 3. Regression results using normalized dataset from 2001 to 2014

Variable	*DV: passenger volume*			*DV: Cargo volume*		
	coeff.(α)	Lower 95%	Upper 95%	coeff.(β)	Lower 95%	Upper 95%
Intercept	0.316***	0.288	0.343	0.257***	0.209	0.305
IE Index	0.267***	0.133	0.401	0.993***	0.763	1.224
NIPC	0.029	-0.131	0.190	0.367**	0.091	0.644
EP	0.187*	0.019	0.356	-0.696***	-0.986	-0.406
Nikkei 225	0.105***	0.057	0.152	0.005	-0.076	0.087
Observations		168			168	
F		155.09***			97.77***	
R-squared		0.792			0.706	

*Significance at p <0.05.

** Significance at p <0.01.

*** Significance at p <0.001.