



Research
Novel Methodologies in Air Transportation—Article

Characterizing Flight Delay Profiles with a Tensor Factorization Framework

Mingyuan Zhang^{a,b}, Shenwen Chen^{a,b}, Lijun Sun^c, Wenbo Du^{a,b,*}, Xianbin Cao^{a,b,*}

^aSchool of Electronic and Information Engineering, Beihang University, Beijing 100191, China

^bNational Engineering Laboratory for Comprehensive Transportation Big Data Application Technology, Beijing 100191, China

^cDepartment of Civil Engineering and Applied Mechanics, McGill University, Montreal, QC H3A 0C3, Canada



ARTICLE INFO

Article history:

Received 10 April 2020

Revised 21 July 2020

Accepted 3 August 2020

Available online 19 March 2021

Keywords:

Air traffic management

Flight delay

Latent class model

Tensor decomposition

ABSTRACT

In air traffic and airport management, experience gained from past operations is crucial in designing appropriate strategies when facing a new scenario. Therefore, this paper uses massive spatiotemporal flight data to identify similar traffic and delay patterns, which become critical for gaining a better understanding of the aviation system and relevant decision-making. However, as the datasets imply complex dependence and higher-order interactions between space and time, retrieving significant features and patterns can be very challenging. In this paper, we propose a probabilistic framework for high-dimensional historical flight data. We apply a latent class model and demonstrate the effectiveness of this framework using air traffic data from 224 airports in China during 2014–2017. We find that profiles of each dimension can be clearly divided into various patterns representing different regular operations. To prove the effectiveness of these patterns, we then create an estimation model that provides preliminary judgment on the airport delay level. The outcomes of this study can help airport operators and air traffic managers better understand air traffic and delay patterns according to the experience gained from historical scenarios.

© 2021 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

With the rapid development of the world's civil aviation industry, severe flight delays have remained a major problem and a cause of inconvenience. Such delays not only discourage passengers from considering air transport or choosing the same airline again [1–4], but also force airlines to bear the additional costs of aircraft maintenance and fleet underutilization [5]. Furthermore, flight delays lead to increases in fuel consumption and carbon dioxide emissions, bringing harm to the environment [6,7]. Besides the direct impacts listed above, flight delays have a negative influence on various aspects of the overall economy [8]. In summary, flight delays constitute a severe and widespread problem with significant negative impacts.

Many factors contribute to the complexity and intractability of this problem. These factors are generally summarized as abnormal weather [9,10] and technical reasons including air traffic control [11], insufficient facility capacity, poor scheduling [12], changes

in procedures [13], and limited buffer time [14]. This variety of factors makes it difficult to understand the underlying pattern of flight delays and design appropriate strategies [15]. Recently, data-driven approaches based on historical observations have been shown to be free from previous constraints and to fit the potential dynamic properties [16]. Therefore, one approach to facilitating system cognition and decision-making is to fully utilize and learn from historical data [17]. For example, when faced with bad weather, it is possible to look up past days with similar weather conditions and refer to the actions taken by air traffic controllers on those days. Several previous studies [18–23] have focused on identifying fixed patterns in air traffic management. Liu et al. [19] introduced a semi-supervised learning algorithm that can determine groups of similar days as different patterns. The first step was to measure the distance between hourly weather forecasts, followed by identifying the days with a small total distance. The authors applied this method to conduct two case studies at Newark Liberty International Airport and proved its efficiency. Mukherjee et al. [20] proposed a way to classify patterns based on the impact of severe weather conditions. The authors used a weather index as the input and applied factor analysis to

* Corresponding authors.

E-mail addresses: wenbodu@buaa.edu.cn (W. Du), xbcao@buaa.edu.cn (X. Cao).

identify dominant weather patterns. Then, they clustered days using Ward’s minimum-variance method. Days belonging to the same cluster shared similar weather patterns. Apart from weather patterns, some studies have attempted to identify similar days from other perspectives. Grabbe et al. [21] used a k -means clustering algorithm to identify similar days in terms of ground delay programs. The authors applied an expectation-maximization (EM) algorithm on the start and end times of ground delay programs and scheduled arrival rate data. Other studies have focused on air traffic flow and flight delays to identify similar patterns [24]. Gorripaty et al. [17] measured the principal components in demand and capacity data, but found that there were no natural patterns in demand or capacity after applying clustering analysis to the data. By identifying the periodic patterns of arrival delay for non-stop domestic flights, Abdel-Aty et al. [25] found that some patterns were not detected by statistical methods.

Despite these previous efforts, there are still gaps in understanding flight delay patterns. As mentioned above, an effective method is to find clusters or patterns in spatiotemporal historical data [26,27]. However, due to the high dimensionality of this data, it is difficult to find distinct patterns in Euclidian space, as mentioned in previous studies [28]. Methods of latent component analysis have thus been put forward to reveal the hidden patterns. Instead of directly mining the patterns, methods such as latent distribution analysis [29], latent trait analysis (including item response theory and Rasch models), and grade-of-membership analysis can utilize the signatures derived from the tensor factorization to form a projection subspace with much lower dimensionality, which strengthens the underlying cluster structure of spatiotemporal traffic dynamic patterns [30]. These methods open up new frontiers in the field of transportation science [31], such as urban mobility analysis [32–34], traffic speed prediction [35], missing traffic data completion [36,37], and vessel track recovery [38].

Inspired by the methods described above, the main goal of this paper is to understand potential air traffic and flight delay patterns using massive air traffic data. In the first part of this work, we regard flight records as a multivariate observation sampled from a universal distribution and adopt a Tucker-like latent class model to identify principal patterns of each mode. We then propose an estimation model based on potential patterns to show the efficiency of this framework. The remainder of the paper is organized as follows. In Section 2, we introduce the modeling framework. Section 3 demonstrates a case study in which we apply the model to a dataset of the Chinese aviation system, and Section 4 gives conclusions.

2. Modeling framework

In this section, we introduce the overall framework for modeling multi-way flight data in a probabilistic setting. The aim is to characterize the principal patterns of air traffic and flight delays from different dimensions and their joint interactions. In Section 2.1, we introduce the notation used in our framework. In Section 2.2, the nonnegative Tucker decomposition (NTD) method is presented as a method of understanding the tensor decomposition (TD). Finally, the latent class analysis (LCA) is described in Section 2.3.

2.1. Notation

We let $x_a = (x_{a1}, \dots, x_{aq})^T$ represent a flight record a , where q represents the number of dimensions—that is, the number of attributes in a single trip record. To characterize the features of flights, each element $x_{a\beta}$ may denote the departure airport of the flight

(x_{a1}), departure time (x_{a2}), departure date (x_{a3}), delay level (x_{a4}), and so forth.

We map the values into discrete values for convenience. We define $x_{a\beta} \in \{1, \dots, w_q\}$ discrete values starting from 1 for attribute β ($\beta = 1, \dots, q$). β denotes the indices of attributes. w_q is the discrete value of the vector of dimension q . Taking the departure airport as an example, $x_{a1} = 1$ represents airport 1 for flight record a . The departure time of a flight is then represented by $x_{a2} \in \{1, \dots, 24\}$, with each value corresponding to a 1 h period in a day. We then present our method using flight record data from four years (2014–2017), with the days ordered by departure date as $x_{a3} \in \{1, \dots, 1461\}$. As the arrival delay is more related to airborne delay, which is difficult to analyze, we consider departure delay rather than arrival delay to investigate the flight delay patterns. Departure delay can better reflect the operation mode of the target airport. In addition, there is a clear definition for the level of departure delays. According to the rules set by the Federal Aviation Administration (FAA) and related research, delayed flights are those that depart more than 15 min later than the scheduled time. Considering practical circumstances, a number of 45 and 90 min later than scheduled are also selected as division thresholds. Thus, we classify the delay of each flight into four levels: ① < 15 min, ② 15–45 min, ③ 45–90 min, and ④ > 90 min. We use $x_{a4} \in \{1, \dots, 4\}$ to represent “on time,” “slight delay,” “moderate delay,” and “serious delay,” respectively.

2.2. Nonnegative Tucker decomposition

It has been shown that TD displays a number of advantages in various contexts, especially when data must be decomposed into a sum of additive components [39]. TD was first proposed by Tucker [40] in 1963. Since then, NTD has been proposed [41,42] to deal with naturally nonnegative and observation data. NTD is a powerful tool for the extraction of nonnegative-parts-based and physically meaningful latent components from high-dimensional tensor data while preserving the natural multilinear structure of the data [24]. In mathematical terms, it decomposes a tensor into a set of matrices and a core tensor.

Given an order tensor χ with K as the tensor dimension. NTD seeks a decomposition of a nonnegative K -way tensor $\chi \in \mathbb{R}_+^{I_1 \times I_2 \times \dots \times I_K}$ (\mathbb{R}_+ is the space of arithmetic number, K is the size of dimensional space, I is one of the orthogonal basis) as mode products of a nonnegative core tensor $\zeta \in \mathbb{R}_+^{J_1 \times J_2 \times \dots \times J_K}$ (J is one of the orthogonal basis) and K nonnegative factor matrices $A^{(K)} \in \mathbb{R}_+^{I_K \times J_K}$:

$$\chi \in \zeta \times {}_1A^{(1)} \times {}_2A^{(2)} \times {}_3A^{(3)} \times \dots \times {}_KA^{(K)} = [\zeta, A^{(1)}, A^{(2)}, \dots, A^{(K)}] \quad (1)$$

where $A^{(1)}, A^{(2)}, A^{(3)}, \dots, A^{(K)}$ are the factor matrices and can be regarded as the principal components of mode K . The entries of ζ show the level of interaction and connections between different components. In this method, both the core tensor ζ and the factor matrices $A^{(K)}$ should be elementwisely nonnegative. Elementwisely, they are given as follows:

$$\chi_{\eta_1, \eta_2, \dots, \eta_K} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \dots \sum_{r_K=1}^{R_K} \zeta_{r_1 r_2 \dots r_K} \partial_{\eta_1 r_1}^{(1)} \partial_{\eta_2 r_2}^{(2)} \dots \partial_{\eta_K r_K}^{(K)} \quad (2)$$

where $\partial_{\eta_K r_K}^{(K)}$ is a factor matrix, R_K is size of factor. r and η are the dimensions of mode K in the factor matrix. The decomposition is then modeled as an optimization problem:

$$\begin{aligned} & \text{Minimize } \frac{1}{2} \|\zeta \times {}_1A^{(1)} \times {}_2A^{(2)} \times {}_3A^{(3)} \times \dots \times {}_KA^{(K)} - \chi\|_F \\ & \text{subject to } \zeta \geq 0, A^{(K)} \geq 0 \end{aligned} \quad (3)$$

where F is matrix norm.

2.3. Latent class analysis

LCA is a statistical method for finding subtypes of related cases (latent classes) from multivariate categorical data [43,44]. The latent class model is written as follows:

$$P_{i_1, i_2, \dots, i_N} = \sum_t p_t \prod_n P_{\partial \mu}^t \quad (4)$$

where P_{i_1, i_2, \dots, i_N} is the probability distribution function, i is dimensional index, N is principal patterns of each mode, T is the number of classes, t is the index of class, p_t is the recruitment that should sum to 1, $P_{\partial \mu}^t$ is the conditional probability, ∂ and μ are the size of probability factors. LCA defines latent classes by the criterion of conditional independence. This means that each variable is statistically independent of every other variable within each latent class. Thus, each element in the probability tensor can be calculated as the sum across all pattern combinations.

$$P_{i_1, i_2, \dots, i_N} = \sum_t p_t \theta_{i_1 t}^{(1)} \times \dots \times \theta_{i_N t}^{(N)} \quad (5)$$

where $\theta^{(N)}$ is the probability vectors characterizing the principal patterns of each mode N .

In our research, we apply latent class models, which assume that each observation is generated from a mixture of underlying classes, and each class is associated with a unique probability distribution. Thus, the joint distribution is regarded as a mixture of product-multinomial and the probability of observing a flight x_a . With the notations in Section 2.1, all the flight records x can be summarized into an m -way tensor with dimension $\phi = w_1 \times w_2 \times \dots \times w_m$ and each cell (v_1, v_2, \dots, v_m) (where v represents a dimension) is the counting of flight numbers $\sum \delta(x_{a1} = v_1, \dots, x_{am} = v_m)$. δ is a two-valued indicator function. $\delta = 1$ if it is true and 0 otherwise. To better understand the interior connections of the dataset, we put these collected flight records into a probabilistic tensor, each cell of which represents the probability of a flight belonging to that particular cell. The probabilistic tensor with each cell is the probability of a flight belonging to a particular cell $p_c(x_{a1} = v_1, \dots, x_{am} = v_m)$ (where p_c is the probability of a particular cell). The probability of observing a flight x_a (i.e., the probability mass function) can be reformulated by the Tucker decomposition in a similar way.

$$P(x_a | \theta) = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \dots \sum_{r_N=1}^{R_N} \pi_{r_1 r_2 \dots r_N} \prod_{n=1}^N \theta_{x_a r_n}^{(N)} \quad (6)$$

The core tensor π captures the interactions across different modes. $\theta_{x_a r_n}^{(N)}$ are the probability vectors characterizing the principal patterns of each mode N . Each element in the probability tensor can be calculated as the sum across all pattern combinations.

$$P(x_{a_1} = v_1, \dots, x_{a_m} = v_m | \theta) = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \dots \sum_{r_N=1}^{R_N} \pi_{r_1 r_2 \dots r_N} \theta_{i_1 r_1}^{(1)} \theta_{i_2 r_2}^{(2)} \dots \theta_{i_N r_N}^{(N)} \quad (7)$$

$\theta_{i_n r_n}^{(N)}$ is the probability vectors characterizing the main patterns of each mode. This model, which is equivalent to a nonnegative Tucker (NNT) decomposition [45,46], can identify universal patterns of different dimensions and reveal the interactions through the core tensor. The EM algorithm is applied to efficiently derive the model [47].

3. Case study

3.1. Flight operation data

The flight operation data is used as a proxy for the collective features of flights in this analysis. The dataset analyzed in this paper was provided by the Civil Aviation Administration of China (CAAC). Given that the purpose of our study is to provide decision-making support for air traffic and airport management, flight delay conditions are of particular interest. Moreover, air traffic flow is the basis of delay conditions. Therefore, it is natural to choose air traffic flow and flight delays as the main objects of our study. Departure delays are chosen, as they can better reflect the level of congestion of a departure airport and airspace. The categorical values of flight data are shown in Table 1. The database contains 13 492 326 domestic flights. All flights connect 224 airports, among which Beijing Capital International Airport has the most flights, accounting for 6.3% of flights in all airports. The departure date of the flights is from 1 January 2014 to 31 December 2017, and the departure time could be in any hour within a day. As can be seen, all the flights can be divided into four groups according to the departure delay time. The percentage of flights in the four groups are 37%, 38%, 14%, and 11%, respectively. The average departure delay of these flights is 31.08 min. Canceled flights are ignored in this study. 9 February 2016 was the day with the greatest number of flights, which was 12 419. 1 January 2014 was the day with the least number of flights, which was 7009. The busiest time period for airports is from 8:00 to 9:00, with 6% of all flights taking off during this time period.

3.2. Factorization analysis

In our research, we assume that flight records are multivariate variables sampled from a universal distribution. The 13 492 326 flights are aggregated to the same tensor. Each observation contains four variables, including departure airport, order of the day during the four years, time of day, and level of delay. The total combinations are $224 \times 1461 \times 24 \times 4$. Here, we use a small 3 (departure airport, A) \times 4 (order of day in four years, D) \times 5 (time of day, H) \times 4 (level of delay, L) core tensor π to capture the interactions across different modes. Although a bigger core tensor could contain more information and reflect comprehensive relationships among different modes, this small core tensor can facilitate the interpretation of results. Moreover, existing research shows that the results are basically consistent across different sizes of core tensors [48]. In the following paragraph, we present the main results based on the core tensor size $[3 \times 4 \times 5 \times 4]$ as an example. The hour, day, airport, and delay-level factor matrix allows us to interpret the pattern due to its distribution profile.

Fig. 1(a) depicts hour profiles of the five patterns. The component pattern H_1 , covering 18.7% of all flights, reveals a gradual rise from 11:00 and reaches its peak at 24:00, followed by a decrease until the early morning. Pattern H_2 shows similarities to a Gaussian distribution with a peak at 17:00, and covers about 24.4% of all flights. Compared with patterns H_1 and H_2 , patterns H_3 and H_5

Table 1
Categorical values of flight data.

Attribute	Category	Description
Departure airport	224	Origin airport
Order of day	1461	Ordered by date during four years
Time of day (h)	24	0:00–1:00, 1:00–2:00, ..., 23:00–24:00
Level of delay	4	① < 15 min, ② 15–45 min, ③ 45–90 min, ④ > 90 min

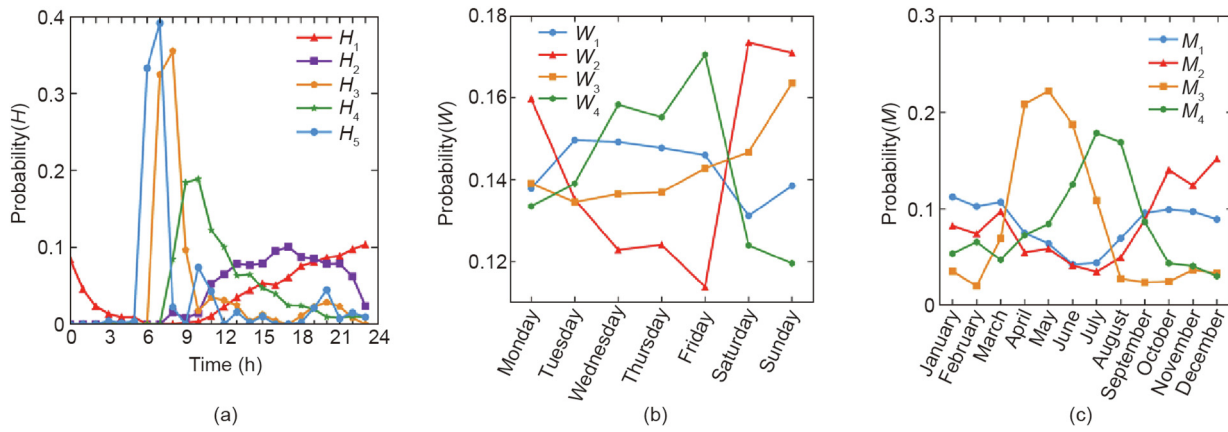


Fig. 1. Principal patterns in different modes. (a) The Y-axis is the probability of each pattern (column) in the hour factor matrix Probability(H); (b) the Y-axis is the probability of day of the week along with the day factor matrix Probability(W); (c) the Y-axis is the probability along with the day factor matrix Probability(M).

are more concentrated. They increase significantly in the morning and display a sudden drop by noon. Pattern H_4 reaches its peak at 11:00 and then decreases continuously over the rest of the day. The proportions of H_3 , H_4 , and H_5 are 22.9%, 22.2%, and 11.8%, respectively. The day factor matrix has 1461 rows and four columns, which depict the corresponding relationship between original days and day patterns. We analyze the day patterns across different months of the year and days of the week. To do so, we sum up the daily traffic probability along different months and days of the week to identify how the traffic flow interacts with the dates. Patterns across days of the week are shown in Fig. 1(b). W is the weekly pattern and M is the monthly pattern. Most traffic observations of pattern W_1 are concentrated on weekdays, while the peaks of pattern W_3 are on the two-day weekend. Interestingly, we found that patterns W_2 and W_4 also show opposite trends. W_2 is mainly concentrated on Monday, Saturday, and Sunday, while pattern W_4 is concentrated on Wednesday, Thursday, and Friday. Fig. 1(c) shows patterns of different months. We can observe clear seasonal diversity in that no pattern is equally distributed across the seasons. Pattern M_1 is mainly concentrated in autumn and winter, while pattern M_2 is concentrated in winter and spring. Pattern M_3 is concentrated in spring, and pattern M_4 is mainly distributed in summer, which is the travel peak of the aviation system.

Even though the estimation process does not take any spatial location information into account, we find that these components can be identified by the geographic locations of airports. As the traffic hub of China, Beijing Capital International Airport is obviously prominent in patterns A_1 and A_2 . Furthermore, pattern A_1 is mainly distributed in the southeast, while pattern A_2 is mainly distributed in the southwest. Pattern A_3 is mostly composed of airports in the midwest.

Fig. 2 shows a composition of the level of delay for each pattern (column) in the delay-level factor matrix. $L_1, L_2, L_3,$ and L_4 represent flights of different levels, varying from “on time” to “serious delay” (as mentioned in Section 2.1). As can be seen, the delay-level patterns remain almost the same compared with the original tensor, with flights of different delay levels being identified. Fig. 2 shows the composition of each delay-level pattern, which covers 40.7%, 10.5%, 29.1%, and 19.7% of total traffic flow, respectively. It is noted that delay is a manifestation of air traffic congestion, and delay level thus closely interacts with traffic flow characteristics in spatiotemporal factors. To conduct further analysis, we adopt conditional probability across the delay-level mode and other modes to investigate the interactions. We calculate the conditional probability distribution Probability($L|H$) by Bayes’ theorem. Given the hour patterns, the conditional distribution of delay-level patterns is

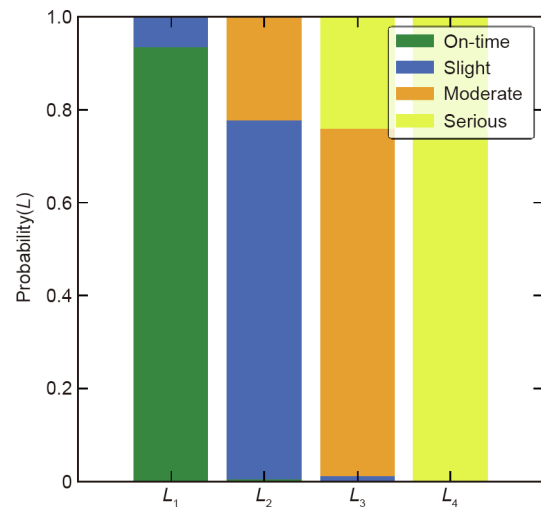


Fig. 2. Delay-level composition probability for each pattern (column) in the delay factor matrix.

$$\text{Probability}(L|H) = \begin{bmatrix} & L_1 & L_2 & L_3 & L_4 \\ H_1 & 0.2857 & 0.4294 & 0.0997 & 0.1852 \\ H_2 & 0.3894 & 0.0340 & 0.2771 & 0.2990 \\ H_3 & 0.4049 & 0.0000 & 0.4270 & 0.1681 \\ H_4 & 0.4017 & 0.0585 & 0.3197 & 0.2201 \\ H_5 & 0.6538 & 0.0290 & 0.2977 & 0.0196 \end{bmatrix} \quad (8)$$

Probability($L|H$) shows how the delay-level patterns interact with the hour patterns. As can be seen, pattern L_1 , denoting the on-time component, strongly interacts with all the hour patterns, which indicates that on-time flights occupy the mainstream at any time of day. To characterize the rules of delay, we now discuss the delay-level patterns excluding L_1 . H_1 , denoting the traffic pattern of an increasing trend throughout the entire day, is associated with delay pattern L_2 , which corresponds to slightly delayed flights. H_2 is mainly covered by the hour patterns L_3 and L_4 , denoting that the high-level delays are highly correlated to the traffic in the afternoon. $H_3, H_4,$ and H_5 are mainly covered by L_3 and L_4 . This can be explained by the fact that flights during the travel peak in the morning are likely to be seriously delayed due to the heavy air traffic in most airports.

$$\text{Probability}(L|M) = \text{Probability}(L|W) = \begin{bmatrix} & L_1 & L_2 & L_3 & L_4 \\ M_1/W_1 & 0.3783 & 0.3433 & 0.2638 & 0.0146 \\ M_2/W_2 & 0.5889 & 0.3195 & 0.0780 & 0.0135 \\ M_3/W_3 & 0.4584 & 0.2777 & 0.1735 & 0.0904 \\ M_4/W_4 & 0.2223 & 0.2035 & 0.2560 & 0.3181 \end{bmatrix} \quad (9)$$

Based on the factor matrices of the day pattern and delay-level pattern, we can find that on-time and slight delay patterns occur the most in M_1/W_1 , M_2/W_2 , and M_3/W_3 ; however, pattern M_4/W_4 is very unusual. Along with the week and month dimensions, M_4/W_4 indicates that the traffic is mostly concentrated on weekdays and in the summer, and tends to interact with serious delays.

$$\text{Probability}(L|A) = \begin{bmatrix} & L_1 & L_2 & L_3 & L_4 \\ A_1 & 0.2369 & 0.2015 & 0.2830 & 0.2783 \\ A_2 & 0.3081 & 0.0417 & 0.4394 & 0.2107 \\ A_3 & 0.6402 & 0.0950 & 0.1434 & 0.1214 \end{bmatrix} \quad (10)$$

As shown in Probability($L|A$) airport patterns are also correlated with the delay-level patterns. The obvious point is that pattern A_3 is mainly covered by L_1 , which may denote that air traffic from the airports in the midwest have fewer delays. Compared with A_3 , A_1 and A_2 are more likely to be seriously delayed. This can be explained by the midwest airports' relatively lighter traffic flow and adequate airspace resources. As shown in the analysis above, delay is heavily influenced by time and space. To further explore the interactions among the patterns, we present the disaggregated distribution along the temporal and spatial dimensions. As can be seen in Fig. 3, the value of (L_2, H_1) is greater than that of the other cells in A_1 , indicating that when afternoon or evening flights take off from the southeast airports, they are accompanied by slight delays. We also observe a considerable amount of traffic within the A_2 pattern in (L_3, H_3), while such traffic rarely appears in A_1 and A_3 . This can be explained by the fact that flights from southwest airports during the morning peak often have moderate delays. As mentioned before, due to their excess capacity, the midwest airports (A_3) have a low probability of flight delays. This phenomenon is consistent with cells (L_1, H_2-H_4) in A_3 .

The numerical examples show that the factorization allows us to explain the complex dependence and higher-order interactions based on latent factors. The core tensor π characterizes the interactions between different modes in a very efficient and informative manner. This framework helps us to comprehend and interpret the latent interactions and complex dependence among the patterns in large datasets, enhancing our understanding of air traffic management.

On this basis, another significant problem is generated. Can the delay level be estimated if we only consider the information about latent patterns extracted from the time and space information? Although previous studies have demonstrated that flight delays can be attributed to many complex factors [49], this issue may be of great significance. Firstly, potential patterns based on historical information are generated by the collective effect of various delay causes. For example, extreme weather occurs frequently in summer, and we find that summer interacts strongly

with the serious delay pattern in this paper. Secondly, as spatiotemporal information does not require us to understand operational characteristics in great detail, the departure time and departure airport alone can help us make a preliminary assessment of delays beforehand. Thirdly, the highly dynamic nature and complexity of aviation systems lead people to believe that delays cannot be estimated through basic information. Thus, the effectiveness of the potential patterns produced by our framework is also demonstrated if accurate estimation is achieved. For further exploration, the random forest (RF) algorithms employed here to construct an estimation model, due to its adaptability in regard to this problem [50]. The advantages of RF include the following: ① It generates an internal unbiased estimate of the generalization error as the forest building progresses; ② it runs efficiently on large databases; and ③ it has the ability to model interactions among variables [51]. Specifically, an RF is an ensemble tree $\{\Gamma_1(\mathbf{X}), \dots, \Gamma_B(\mathbf{X})\}$, where B is the number of trees and $\mathbf{X} = (x_1, \dots, x_\rho)$ is a ρ -dimensional vector of descriptors. The ensemble produces B outputs $\hat{y}_1(x) = \Gamma_1(\mathbf{X}), \dots, \hat{y}_B(x) = \Gamma_B(\mathbf{X})$, where $\hat{y}_b(x)$ is the estimation of the b th RF tree ($b = 1, \dots, B$). The outputs of all trees are aggregated to produce one final estimation, $\hat{y}_{\text{rf}}^B(x)$. For classification problems, $\hat{y}_{\text{rf}}^B(x)$ is the class estimated by the majority of trees. The training procedure is as follows:

(1) **Prepare the training data.** We prepare ρ -dimensional the training data samples along with their class label.

(2) **Select the parameters.** E : maximum depth of each tree; C : minimum number of samples required to split an internal node; V : the variables at each split; M_L : the minimum number of samples required to be at a leaf node.

(3) **Grow a classification tree.** For $b = 1$ to B , draw a bootstrap sample Z^* of size (S_d) from the training data. Use about two-thirds of the original training samples to grow a classification tree. Leave the remaining one-third of the samples as so-called out-of-bag (OOB) samples.

(4) **Grow trees.** For each bootstrap sample, grow a tree $\Gamma_b(\mathbf{X})$ with the following modifications. At each node, pick the best variable/split-point and split the node into two nodes until the sample number of this node is less than C . The tree is grown to the maximum size E and not pruned back. Repeat the steps above until B such trees are grown.

(5) **Results output.** Estimate new data by aggregating the estimations of the trees with the majority vote of all analogous trees in the forest. Output $\hat{y}_{\text{rf}}^B(x) = \text{majority vote } \{\hat{y}_{b=1}^B(x)\}_{b=1}^B$.

To evaluate the performance of the estimation model, four indices are employed, referring to Eqs. (11)–(14). $F1_{\text{macro}}$ is the macro-averaged score and weights all the classes ($1, \dots, u$) equally, regardless of how many samples S_u belong to a given class u . $F1_{\text{micro}}$ is the micro-averaged score and weights all the samples equally, thus favoring the performance of common classes. Weighted score finds the average in each label, weighted by the number of true instances for each class. Accuracy_{OOB} is the mean accuracy on each training sample Z^* , using only the trees that do not have Z^* in their bootstrap sample [52].

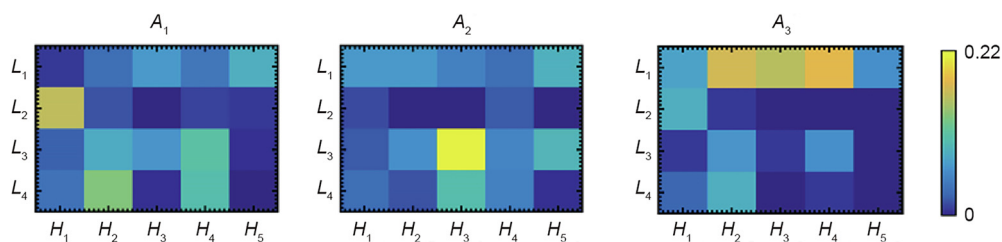


Fig. 3. Interaction of delay level and hour for different airport patterns.

$$F1_{macro} = \frac{2 \times P_{macro} \times R_{macro}}{P_{macro} + R_{macro}} \left(P_{macro} = \frac{1}{u} \sum_{u=1}^u Precision_u, \right. \\ \left. R_{macro} = \frac{1}{u} \sum_{u=1}^u Recall_u \right) \quad (11)$$

where P_{macro} is the macro-averaged precision, R_{macro} is the macro-averaged recall, $Precision_u$ is the number of true positives divided by the total number of elements labelled as belonging to the positive class for a class u , and $Recall_u$ is defined as the number of true positives divided by the total number of elements that actually belong to the positive class for a class u .

$$F1_{micro} = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}} \left(P_{micro} = \frac{\sum_1^u TP_u}{\sum_1^u TP_u + \sum_1^u FP_u}, \right. \\ \left. R_{micro} = \frac{\sum_1^u TP_u}{\sum_1^u TP_u + \sum_1^u FN_u} \right) \quad (12)$$

where P_{micro} is the micro-averaged precision, R_{micro} is the micro-averaged recall, TP_u is the true positive rate of class u , FP_u is the false positive rate of class u , and FN_u is the false negative rate of class u .

$$Weighted\ score = \frac{1}{S} \sum_1^u \frac{2 \times Precision_u \times Recall_u}{Precision_u + Recall_u} \times S_u \quad (13)$$

where S_u is number of samples of class u , S is number of samples.

$$Accuracy_{OOB} = \frac{TP_{OOB} + TN_{OOB}}{TP_{OOB} + FP_{OOB} + TN_{OOB} + FN_{OOB}} \quad (14)$$

where TP_{OOB} is the true positive rate of OOB samples, FP_{OOB} is the false positive rate of OOB samples, FN_{OOB} is the false negative rate of OOB samples, and TN_{OOB} is the true negative rate of OOB samples.

For this study, the data of latent spatiotemporal patterns are used. The classification problem involves the identification of four delay levels (on time, slight delay, moderate delay, and serious delay). To estimate how accurately a model will perform in practice, this paper adopts a five-fold cross-validation strategy. One round of cross-validation involves partitioning all the records into five complementary subsets, performing the training process on four subsets, and then validating the analysis on another testing set. Next, the validation results are averaged over the five rounds to give an estimation of the model's performance. A total of 13 492 326 records are used. As mentioned above, the probability values of the flights in each mode (latent spatiotemporal patterns)

are selected as the features; that is, $\{A_1-A_4, H_2-H_5, W_1-W_4, M_1-M_4\}$. Each feature ranges from 0 to 1.

The number of trees is the most important variable. It should be large enough to let the generalization error of the RF converge. In Fig. 4, we find that $Accuracy_{OOB}$ changes from 53.1% to 53.4% when B increases from 100 to 150, and increases slightly when B is greater than 150. This indicates that the RF classifier is almost insensitive to the increase of B when it is greater than 150. To obtain a better parameters set, a grid search method is used in this paper. Several combinations of other parameters are tried to determine the optimal values for this model.

The overall performance and the accuracy of individual classes are given in Fig. 5. As prior studies have noted, delays are mostly determined by dynamic operational factors. It is not possible to give a judgment much earlier than the actual occurrence of the delay [53,54]. Our model only considers the latent patterns revealed through basic flight information, which may not be viewed optimistically. In this case, the RFs achieved more than 50% in all performance indices and achieved accuracies of 60.0%, 46.0%, 44.0%, and 65.0%, respectively, for the individual classes, which may be a very positive sign. It signifies that the probability of different levels of flight delay can be estimated preliminarily based only on the time and airport information. Furthermore, this application supports the effectiveness of the potential patterns. We can also see that the “on time” pattern and “serious delay” pattern are classified more accurately due to their distinctive features. Even if the algorithm has more difficulty in classifying the “slight

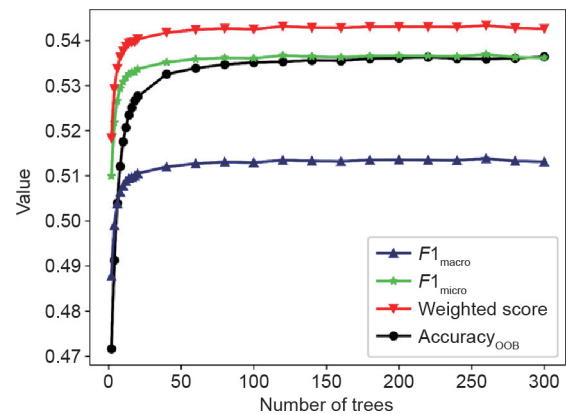


Fig. 4. Model performance with different numbers of trees.

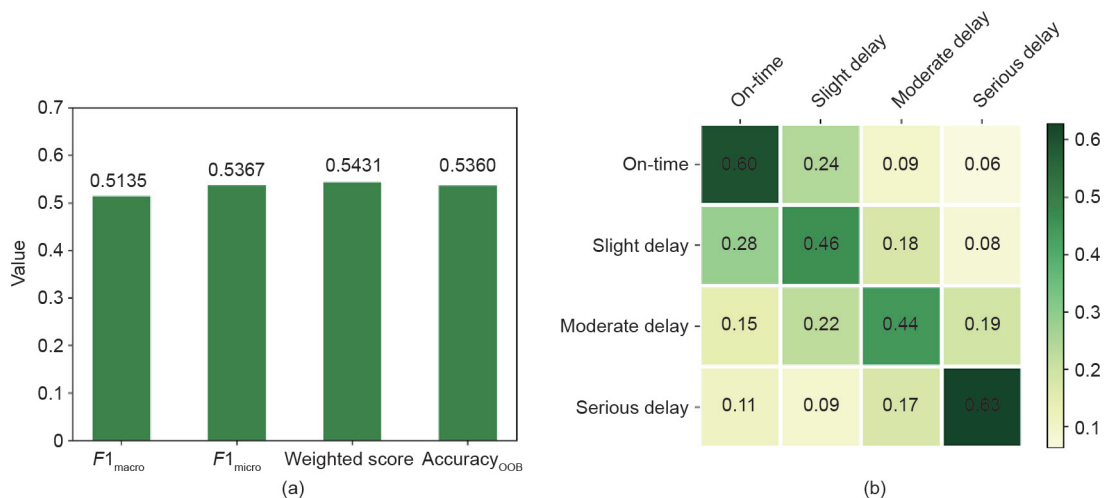


Fig. 5. Classification results. (a) Overall performance through different indices; (b) the confusion matrix.

delay” and “moderate delay” classes, the deep color near the diagonal of the confusion matrix indicates that the wrong estimations have a small bias.

4. Conclusions

In this paper, we develop a probabilistic factorization framework that transforms massive flight record data into a spatiotemporal tensor. Our purpose is to investigate the spatiotemporal dynamic patterns of air traffic and flight delays. We assume that each flight observation is a sample generated from a universal joint distribution. Then, we formulate this joint task using nonnegative tensor factorization, which has been shown to be a useful analytical tool for high-dimensional massive data. The results show that clear patterns are identified in different modes. The interactions of different modes are also characterized in the core tensor, which interprets the relationship between delays and spatiotemporal patterns. The detailed analysis clearly shows that serious delays are strongly correlated to afternoon, weekday, and summer traffic. Flights from midwest airports have a low probability of delay at any time of day.

This framework can shed light on the modeling of flight delays by providing an understanding of flights in both the spatial and temporal dimensions. Moreover, the potential patterns have been shown to display a certain effect on delay estimation. As the integration of space and time information, potential patterns can give positive estimation results regarding the delay level. In the context of highly dynamic environments and the complexity of delays, this output gives us a new comprehension of how air traffic and flight delays interact with time and space. This framework provides a deep understanding of massive aviation data through the latent class model and probabilistic factorization method. The outcomes of this study can help airport operators and air traffic managers to better prepare air traffic and airport arrangements based on the experience gained from historical scenarios. Further analysis can be performed to tackle the challenge of understanding interactions involving more factors, such as weather and air route attributes.

Acknowledgements

The authors would like to thank Yu Zhang from the University of South Florida and Shenqi Zhang from Carnegie Mellon University for technical support and suggestions, which helped a lot to improve the quality of this work. This paper is supported by the National Key Research and Development Program of China (2019YFF0301400) and the National Natural Science Foundation of China (61671031, 61722102, and 61961146005).

Compliance with ethics guidelines

Mingyuan Zhang, Shenwen Chen, Lijun Sun, Wenbo Du, and Xianbin Cao declare that they have no conflict of interest or financial conflicts to disclose.

References

- [1] Folkes VS, Koletsky S, Graham JL. A field study of causal inferences and consumer reaction: the view from the airport. *J Consum Res* 1987;13(4):534–9.
- [2] Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. *Transp Res Part E Logist Trans Rev* 2012;48(2):460–9.
- [3] Ferrer JC, e Oliveira PR, Parasuraman A. The behavioral consequences of repeated flight delays. *J Air Transp Manage* 2012;20(3):35–8.
- [4] Vlachos I, Lin Z. Drivers of airline loyalty: evidence from the business travelers in China. *Transp Res Part E Logist Trans Rev* 2014;71:1–17.
- [5] Cook AJ, Tanner G. European airline delay cost reference values. Report. London: University of Westminster; 2011 Mar.
- [6] Pejovic T, Noland RB, Williams V, Toumi R. A tentative analysis of the impacts of an airport closure. *J Air Transp Manage* 2009;15(5):241–8.
- [7] Ryerson MS, Hansen M, Bonn J. Time to burn: flight delay, terminal efficiency, and fuel consumption in the National Airspace System. *Transp Res Part A Policy Pract* 2014;69:286–98.
- [8] Ball M, Barnhart C, Dresner M, Hansen M, Neels K, Odoni A, et al. Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States. Report. National Center of Excellence for Aviation Operations Research; 2010.
- [9] Abdelghany KF, Shah SS, Raina S, Abdelghany AF. A model for projecting flight delays during irregular operation conditions. *J Air Transp Manage* 2004;10(6):385–94.
- [10] Robinson PJ. The influence of weather on flight operations at the Atlanta Hartsfield International Airport. *Weather Forecast* 1989;4(4):461–8.
- [11] Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. *J Air Transp Manage* 1999;5(3):113–34.
- [12] Wong JT, Li SL, Gillingwater D. An optimization model for assessing flight technical delay. *Transp Plann Technol* 2002;25(2):121–53.
- [13] Mueller ER, Chatterji G. Analysis of aircraft arrival and departure delay characteristics. In: Proceedings of the AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum; 2002 Oct 1–3; Los Angeles, CA, USA; 2002.
- [14] Wu CL. Inherent delays and operational reliability of airline schedules. *J Air Transp Manage* 2005;11(4):273–82.
- [15] Schaefer L, Millner D. Flight delay propagation analysis with the detailed policy assessment tool. In: Proceedings of the 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace; 2001 Oct 7–10; Tucson, AZ, USA; 2001.
- [16] Han Y, Moutarde F. Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. *Int J Intell Transp Syst Res* 2016;14:36–49.
- [17] Gorripaty S, Liu Y, Hansen M, Pozdnukhov A. Identifying similar days for air traffic management. *J Air Transp Manage* 2017;65:144–55.
- [18] Hoffman B, Krozel J, Penny S, Roy A, Roth K. A cluster analysis to classify days in the National Airspace System. In: Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit; 2003 Aug 11–14; Austin, TX, USA; 2003.
- [19] Liu Y, Seelhorst M, Pozdnukhov A, Hansen M, Ball MO. Assessing terminal weather forecast similarity for strategic air traffic management. In: Proceedings of the 6th International Conference on Research in Air Transportation; 2014 May 26–30; Istanbul, Turkey; 2014.
- [20] Mukherjee A, Grabbe SR, Sridhar B. Classification of days using weather impacted traffic in the National Airspace System. In: Proceedings of the 2013 Aviation Technology, Integration, and Operations Conference; 2013 Aug 12–14; Los Angeles, CA, USA; 2013.
- [21] Grabbe SR, Sridhar B, Mukherjee A. Clustering days with similar airport weather conditions. In: Proceedings of the 14th AIAA Aviation Technology, Integration, and Operations Conference; 2014 Jun 16–20; Atlanta, GA, USA; 2014.
- [22] Bloem M, Bambos N. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. In: Proceedings of the 14th AIAA Aviation Technology, Integration, and Operations Conference; 2014 Jun 16–20; Atlanta, GA, USA; 2014.
- [23] Sternberg A, Carvalho D, Murta L, Soares J, Ogasawara E. An analysis of Brazilian flight delays based on frequent patterns. *Transp Res Part E Logist Trans Rev* 2016;95:282–98.
- [24] Zhou G, Cichocki A, Zhao Q, Xie S. Efficient nonnegative Tucker decompositions: algorithms and uniqueness. *IEEE Trans Image Process* 2015;24(12):4990–5003.
- [25] Abdel-Aty M, Lee C, Bai Y, Li X, Michalak M. Detecting periodic patterns of arrival delay. *J Air Transp Manage* 2007;13(6):355–61.
- [26] Zhou X, List GF. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. *Transp Sci* 2010;44(2):254–73.
- [27] Huang J, Levinson D, Wang J, Zhou J, Wang ZJ. Tracking job and housing dynamics with smartcard data. *Proc Natl Acad Sci USA* 2018;115(50):12710–5.
- [28] Du WB, Zhang MY, Zhang Y, Cao XB, Zhang J. Delay causality network in air transport systems. *Transp Res Part E Logist Trans Rev* 2018;118:466–76.
- [29] Mislevy RJ. Estimating latent distributions. *Psychometrika* 1984;49(3):359–81.
- [30] Woodbury MA, Manton KG. Grade of membership analysis of depression-related psychiatric disorders. *Sociol Methods Res* 1989;18(1):126–63.
- [31] Sun L, Axhausen KW, Lee DH, Huang X. Understanding metropolitan patterns of daily encounters. *Proc Natl Acad Sci USA* 2013;110(34):13774–9.
- [32] Zhang F, Wilkie D, Zheng Y, Xie X. Sensing the pulse of urban refueling behavior. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2013 Sep 8–12; Zurich, Switzerland; 2013. p. 13–22.
- [33] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2012 Aug 12–16; Beijing, China; 2012. p. 186–94.
- [34] Ma X, Wu YJ, Wang Y, Chen F, Liu J. Mining smart card data for transit riders' travel patterns. *Transp Res Part C Emerg Technol* 2013;36:1–12.
- [35] Dauwels J, Aslam A, Asif MT, Zhao X, Vie NM, Cichocki A, et al. Predicting traffic speed in urban transportation subnetworks for multiple horizons. In:

- Proceedings of the 13th International Conference on Control Automation Robotics & Vision; 2014 Dec 10–12; Singapore; 2014. p. 547–52.
- [36] Ran B, Tan H, Wu Y, Jin PJ. Tensor based missing traffic data completion with spatial-temporal correlation. *Physica A Stat Mech Its Appl* 2016;446:54–63.
- [37] Asif MT, Mitrovic N, Dauwels J, Jaillet P. Matrix and tensor based methods for missing data estimation in large traffic networks. *IEEE Trans Intell Transp Syst* 2016;17(7):1816–25.
- [38] Liu C, Chen X. Vessel track recovery with incomplete AIS data using tensor CANDECOM/PARAFAC decomposition. *J Navig* 2014;67(1):83–99.
- [39] Gaussier E, Goutte C. Relation between PLSA and NMF and implications. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2005 Aug 15–19; Salvador, Brazil; 2005. p. 601–2.
- [40] Tucker LR. Implications of factor analysis of three-way matrices for measurement of change. In: Harris CW, editor. Problems in measuring change. Madison: University of Wisconsin Press; 1963. p. 122–37.
- [41] De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *SIAM J Matrix Anal Appl* 2000;21(4):1253–78.
- [42] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;401(6755):788–91.
- [43] Hagenaaers JA, McCutcheon AL. Applied latent class analysis. Cambridge: Cambridge University Press; 2002.
- [44] McCutcheon AL. Latent class analysis. Newbury Park: Sage Publications, Inc.; 1987.
- [45] Peng W, Li T. On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis. *Appl Intell* 2011;35(2):285–95.
- [46] Ding C, Li T, Peng W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput Stat Data Anal* 2008;52(8):3913–27.
- [47] McLachlan GJ, Krishnan T. The EM algorithm and extensions. 2nd ed. Hoboken: John Wiley & Sons; 2008.
- [48] Sun L, Axhausen KW. Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transp Res Part B Methodol* 2016;91:511–24.
- [49] Hao L, Hansen M, Zhang Y, Post J. New York, New York: two ways of estimating the delay impact of New York airports. *Transp Res Part E Logist Trans Rev* 2014;70:245–60.
- [50] Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- [51] Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. *Ecology* 2007;88(11):2783–92.
- [52] Gislason PO, Benediktsson JA, Sveinsson JR. Random forests for land cover classification. *Pattern Recognit Lett* 2006;27(4):294–300.
- [53] Rebollo JJ, Balakrishnan H. Characterization and prediction of air traffic delays. *Transp Res Part C Emerg Technol* 2014;44:231–41.
- [54] Sternberg A, Soares J, Carvalho D, Ogasawara E. A review on flight delay prediction. 2017. arXiv:1703.06118.