PAPER

Modeling Inter-Sector Air Traffic Flow and Sector Demand Prediction

Ryosuke MISHIMA[†], *Nonmember and* Kunihiko HIRAISHI[†a)], *Member*

SUMMARY In 2015, the Ministry of Land, Infrastructure and Transportation started to provide information on aircraft flying over Japan, called CARATS Open Data, and to promote research on aviation systems actively. The airspace is divided into sectors, which are used for limiting air traffic to control safely and efficiently. Since the demand for air transportation is increasing, new optimization techniques and efficient control have been required to predict and resolve demand-capacity imbalances in the airspace. In this paper, we aim to construct mathematical models of the inter-sector air traffic flow from CARATS Open Data. In addition, we develop methods to predict future sector demand. Accuracy of the prediction is evaluated by comparison between predicted sector demand and the actual data.

key words: *mathematical modeling, airspace traffic, big data*

1. Introduction

In 2015, the Ministry of Land, Infrastructure and Transportation started to provide information on aircraft flying over Japan, called CARATS Open Data [1], and to promote research on aviation systems actively. The airspace is divided into sectors, which are used for limiting air traffic to control safely and efficiently. Since the demand for air transportation is rapidly increasing, new optimization techniques and efficient control have been required to predict and resolve demand-capacity imbalances in the airspace. In this paper, we aim to develop methods to predict future sector demand. Limiting the number of aircraft in each sector is necessary to safely handled by a human air traffic controller. By the prediction of air traffic demand, controllers can give appropriate delay in departure of aircrafts in order to keep the number of aircraft below an allowable level.

Toward accurate sector demand prediction, we need to have mathematical models that represent airspace traffic flow. There are mainly two approaches to the modeling of air traffic flow. The first approach is based on detailed modeling of individual particle (aircraft), and future trajectory of each particle is estimated by the model. This type of microscopic modeling is called Lagrangian approach (e.g., [2], [3]). Multi-agent simulation is classified as this approach (e.g., [4], [5]), and is used for air-traffic management tools (e.g., [6]). The trajectory-based models predict adequately for short intervals of up to 20 minutes, but the accuracy decreases with the increasing prediction interval [7].

In the second approach, space and time are divided into control regions, and each region has properties such as size, density, and flow rate. This type of macroscopic modeling is called Eulerian approach and is used for estimation of traffic flow between adjacent regions. The cell transmission model is a typical Eulerian model for land road traffic [8], [9]. This modeling approach is extended and applied to air traffic flow [10]–[13]. One of the authors proposed an extended version of cell transmission model that reflects topology of the airspace [14]. As a different Eulerian approach, linear dynamic system models (LDSM) were proposed [15]. In this approach, air traffic flow between adjacent sectors is modeled by the discrete-time linear state equation $x(k+1) = Ax(k) + u(k)$, where $x(k)$ is a nonnegative integer vector each component of which is the sector demand, i.e., the number of aircrafts in the corresponding sector, at time step k, and $u(k)$ is the inputs to the airspace, i.e., the number of aircrafts depart from airports or enter from outside, at time step k. In the original model, time-invariant formulation is used, i.e., the matrix A is fixed. In [16], time-varying formulation is proposed. In the time-varying model, multiple matrices are used for covering the entire prediction period. The matrix A varies according to not only the time but also various situations such as the seasons, the day of the week, weather, and congestion level. To achieve accurate prediction, it is necessary to switch the matrix to adapt the model for the current situation. In [7], a method for selecting an appropriate matrix is proposed. In this method, multiple matrices are prepared beforehand and the best one is selected by the hypothesis testing technique. In addition, prediction based on aggregating multiple models weighted by its posterior probability is proposed. It was reported prediction errors by this LDSM approach do not vary significantly with the length of the prediction period. A summary of existing modeling approach can be found in [17].

In this paper, we show a method to construct the LDSM from CARATS Open data. The method consists of event extraction from flight trajectories to compose matrices in the LDSM and a machine learning technique for selecting representative matrices. We also propose two methods for determining appropriate matrices. The paper is organized as follows. In Sect. 2, concept of the LDSM is explained and how to construct the model from CARATS Open data is presented. In Sect. 3, the idea on switching multiple sector flow models is explained, and how to obtain the multiple models from CARATS data is shown. Next two methods for pre-

Manuscript received November 9, 2021.
Manuscript revised March 8, 2022.
Manuscript publicized April 11, 2022.

[†] The authors are with School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923-1292 Japan.
a) E-mail: hira@jaist.ac.jp

DOI: 10.1587/transfun.2021EAP1147

dicting future sector demand step are proposed. These two method are evaluated by comparison between the outputs of the models and the actual data. Concluding remarks including ideas on further improvement of the proposed approach are described in Sect. 4.

2. Modeling Inter-Sector Air Traffic Flow

2.1 Linear Dynamic System Model

The flight information service for aircrafts is provided in flight information regions (FIR). There is one FIR, Fukuoka FIR, in Japan. An FIR consists of several area control centers (ACC) and each ACC is subdivided into smaller regions called sectors. Sectors are units of the air traffic control. Keeping the amount of air traffic in each sector below its capacity is mandatory in the air traffic control. Detailed information on ACCs and sectors in Fukuoka FIR is available in [18]. The problem of identifying the sectors for given points in the airspace is studied in [19]. This result is used as preprocessing of the model construction.

Figure 1 depicts the concept of LDSM. In this figure, there are three sectors in an ACC. The flow in the model is the number of aircrafts moved between two adjacent sectors in a fixed time interval. The flow from sector i to sector j ($i \neq j$) is denoted by $a_{i,j}$, the number of aircrafts that stay in sector i is denoted by $a_{i,i}$, the flow from outside/airport to sector j is denoted by $u_{in,j}$, and the flow from sector i to outside/airport is denoted by $u_{i,out}$. Let N be the number of sectors in the ACC. Then the sector flow in time interval $[k, k+1)$ is defined by the matrix

$$T(k) = \begin{bmatrix} a_{1,1}(k) & \cdots & a_{1,N}(k) & u_{1,out}(k) \\ \vdots & \ddots & \vdots & \vdots \\ a_{n,1}(k) & \cdots & a_{N,N}(k) & u_{N,out}(k) \\ u_{in,1}(k) & \cdots & u_{in,N}(k) & 0 \end{bmatrix} \quad (1)$$

The traffic demand of sector i at time step k is the sum of the i-th row of $T(j)$, formulated as

$$x_i(k) = \sum_{j=1}^{N} a_{i,j}(k) + u_{i,out}(k) \quad (2)$$

The traffic demand of sector i at time step $k+1$ is obtained by the sum of the i-th column of $T(k)$, formulated as

$$x_i(k+1) = \sum_{j=1}^{N} a_{j,i}(k) + u_{in,i}(k) \quad (3)$$

From (2) and (3), we have

$$\sum_{j=1}^{N} a_{i,j}(k+1) + u_{i,out}(k+1) = \sum_{j=1}^{N} a_{j,i}(k) + u_{in,i}(k) \quad (4)$$

The sector flow model can be written as a standard linear state equation. The state vector $x(k)$ is an N-dimensional nonnegative integer vector $x(k) = [x_1(k), \cdots, x_N(k)]^T$. The

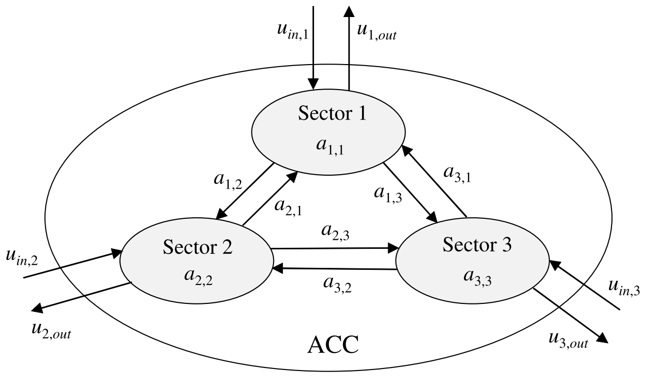

Fig. 1 Sector flow model.

state transition matrix $A(k)$ that maps $x(k)$ to $x(k+1)$ with an input $u(k) = [u_{in,1}(k), \cdots, u_{in,N}(k)]^T$ is defined as

$$A(k) := [A_{i,j}] = \begin{bmatrix} a_{1,1}(k)/x_1(k) & \cdots & a_{N,1}(k)/x_N(k) \\ \vdots & \ddots & \vdots \\ a_{1,N}(k)/x_1(k) & \cdots & a_{N,N}(k)/x_N(k) \end{bmatrix} \quad (5)$$

where $A_{i,i} = 1$ and $A_{i,j} = 0$ $(i \neq j)$ if $x_i(k) = 0$. Each component $A_{i,j} = a_{i,j}(k)/x_i(k)$ of $A(k)$ represents the flow rate from sector i to sector j at time k. Then the state equation is given as

$$x(k+1) = A(k)x(k) + u(k) \quad (6)$$

This model is time-varying since the matrix $A(k)$ is given for every time step k. In [7], the matrix is determined by past data. Let H_l denote a hypothesis on the current situation, where l is the index attached to the hypothesis. Based on the past data on air traffic flow, the posterior probability $Pr(H_l \mid X_k)$ that hypothesis H_l is true after sector demand history $X_k = x(k), x(k-1), \cdots, x(0)$ is derived as follows. Let $Pr(x_k|X_{k-1}, H_l)$ be the conditional probability density function of the sector demand vector given sector demand history and hypothesis H_l. This function is derived from the demand prediction model for H_l and the stochastically-given aircraft departure model. The posterior probability of each hypothesis is obtained through Bayes' theorem, and is generated recursively by

$$Pr(H_l|X_k) = \frac{Pr(x_k|X_{k-1}, H_l) \cdot Pr(H_l|X_{k-1})}{\sum_l Pr(x_k|X_{k-1}, H_l) \cdot Pr(H_l|X_{k-1})} \quad (7)$$

with some initial distribution $Pr(H_l|X_0)$ for each H_l.

Then the hypothesis with the largest probability is selected and the matrix is determined by the hypothesis. As an alternative approach, the prediction by aggregating all hypotheses is given by

$$\tilde{x}(k+1) = \sum_{H_l \in \mathcal{H}} Pr(H_l \mid X_k) \cdot \tilde{x}_l(k+1) \quad (8)$$

where $\mathcal{H}$ is the set of all hypotheses, $\tilde{x}_l(k+1)$ is the prediction under hypothesis H_l, and $\tilde{x}(k+1)$ is the aggregated prediction.

2.2 CARATS Open Data

CARATS Open data consists of flight trajectory data of all regular flights in Fukuoka FIR. The sources of the data are radar data and flight plans. For each flight, the following information is recorded at every 10 seconds: (1) time, (2) flight number (unique ID of the flight), (3) latitude, (4) longitude, (5) altitude, (6) aircraft model (e.g., B772, B738, A320). In this paper, the data from 2012 to 2017 is used. The data contains flight trajectory data in one week of every odd month (2012-2016); every month (2017). The total data size is 38GB. As the preprocessing, the sector that contains each point in the airspace is identified and appended to the data.

2.3 Construction of LDSM from Event Log

From the CARATS Open data, the following three kinds of events are extracted (Fig. 2). We first fix the target center. Let $f = 1, \cdots, M$ denote the flight number, where M is the number of aircrafts that appear in the dataset.

- $in(f,i,t)$: At time t, aircraft f moves from outside/airport to sector i of the center.
- $move(f,i,j,t)$: At time t, aircraft f moves from sector i of the center to another sector j of the center.
- $out(f,i,j,t)$: At time t, aircraft f exits from sector i of the center and moves to sector j of an adjacent center or outside/airport ($j = out$). Note that event $out(f,i,j,t)$ ($j \neq out$) corresponds to event $in(f,j,t)$ in the adjacent center containing sector j.

When the time interval $[k, k+1)$ becomes longer, $in(f,i,t)$ and $move(f,i,j,t)$ does not necessarily correspond to $u_{in,i}$ and $a_{i,j}$, respectively, since aircrafts can traverse multiple sectors in the time interval. To handle such a case, we introduce the following binary variables:

- $\underline{u}^{f}_{in,i}(k) = 1$ if aircraft f moves from outside/airport to sector i of the center during time interval $[k, k+1)$; 0 otherwise.
- $\underline{b}^{f}_{i,j}(k) = 1$ if aircraft f moves from outside/airport to sector i of the center during time interval $[k, k+1)$ and it is in sector j ($j \neq i$) of the center at the end of the time interval; 0 otherwise.
- $\underline{a}^{f}_{i,j}(k) = 1$ if aircraft f is in sector i of the center at the beginning of time interval $[k, k+1)$, and is in sector j of the center at the end of the time interval; 0 otherwise.

Using these binary variables, we can reformulate the state equation. Firstly, $a_{i,j}(k)$ is defined as

$$a_{i,j}(k) = \sum_{f=1}^{M} \underline{a}^{f}_{i,j}(k) \tag{9}$$

Next we define $b_{i.j}(k)$ as

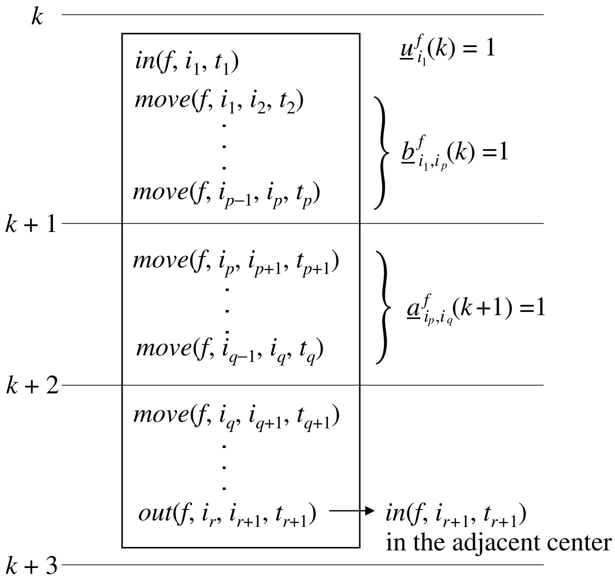
k
$in(f, i_1, t_1)$
$\underline{u}^f_{i_1}(k) = 1$
$move(f, i_1, i_2, t_2)$
$move(f, i_{p-1}, i_p, t_p)$
$\underline{b}^f_{i_1,i_p}(k) = 1$
$k + 1$
$move(f, i_p, i_{p+1}, t_{p+1})$
$move(f, i_{q-1}, i_q, t_q)$
$\underline{a}^f_{i_p,i_q}(k+1) = 1$
$k + 2$
$move(f, i_q, i_{q+1}, t_{q+1})$
$out(f, i_r, i_{r+1}, t_{r+1})$
$in(f, i_{r+1}, t_{r+1})$
in the adjacent center
$k + 3$

Fig. 2 Event log of a flight.

$$b_{i,j}(k) = \sum_{f=1}^{M} \underline{b}^{f}_{i,j}(k) \tag{10}$$

Let $u_{in,i}(k)$ be defined as

$$u_{in,i}(k) = \sum_{f=1}^{M} \underline{u}^{f}_{in,i}(k) \tag{11}$$

We give matrix $A(k)$ by (5). In addition, we define a matrix $B(k)$ as

$$B(k) := [B_{i,j}] = \begin{bmatrix} b_{1,1}(k)/u_{in,1}(k) & \cdots & b_{N,1}(k)/u_{in,N}(k) \\ \vdots & \ddots & \vdots \\ b_{1,N}(k)/u_{in,1}(k) & \cdots & b_{N,N}(k)/u_{in,N}(k) \end{bmatrix} \tag{12}$$

where $B_{i,i} = 1$ and $B_{i,j} = 0$ ($i \neq j$) if $u_{in,i}(k) = 0$. Then the state equation is

$$x(k+1) = A(k)x(k) + B(k)u(k) \tag{13}$$

3. Sector Demand Prediction

3.1 LDSM with Representative Matrices

From the flight trajectory data, we obtain a large number of matrices A and B that reflect various situations. Some of them are very similar and some of them are different. Toward construction of sector demand prediction models, we propose a fully discretized approach. We first collect a set of matrices A and a set of matrices B from the data, and then pick up a fixed number of representative matrices from them. We use the K-means clustering for the selection of matrices.

The objective of the clustering is to obtain representative matrices uniformly distributed in the set of all matrices. By applying the K-means clustering to the sets of collected matrices, we obtain K clusters of matrices A and K clusters of matrices B. Next we chose the center matrix from each cluster as a representative matrix. Let $\mathcal{M}_{d,K}^{A}$ and $\mathcal{M}_{d,K}^{B}$ denote the sets of center matrices for matrix A and matrix B, respectively, given the time interval d and the number of clusters K. In the experiments, we use sklearn.cluster.Kmeans [20] as the clustering algorithm. The center matrices are also computed by this tool.

For the LDSM with representative matrices, we will study the following issues:

1. *Evaluation of prediction errors*: Since the LDSM uses a limited number of matrices, there may exists prediction errors even if we select the best fit matrices. Using actual flight data, we compute the minimum error for every combination of matrices $A \in \mathcal{M}_{d,K}^{A}$ and $B \in \mathcal{M}_{d,K}^{B}$ at each time step. The minimum error gives the lower limit (the greatest lower bound) of the error by an arbitrary matrix selection method.
2. *Choice of appropriate parameter values*: There are two parameters when the matrices are extracted: one is the length d of the time interval and the other is the number K of clusters. Based on the error analysis, we try to find appropriate values of the two parameters d and K.
3. *Sector Demand Prediction*: We develop methods to predict the sector demand at time step $k+1$ from the past history of sector demands $x(0), \cdots, x(k)$ and the inputs $u(0) \cdots u(k)$ up to time step k.

3.2 Evaluation of Prediction Errors

The prediction error is defined as discrepancy between the predicted sector demand $\tilde{x}(k+1)$ and the actual sector demand $x(k)$. We here use the normalized L2-norm of the difference of two vectors:

$$\varepsilon(\tilde{x}, x) := \frac{\|\tilde{x} - x\|_2}{\|x\|_2} \tag{14}$$

Since the sector demand prediction is done all at once by a single prediction model, we adopt error evaluation as vectors. This measure gives discrepancy between a predicted state vector and the actual state vector. Since the total number of aircrafts in airspace varies significantly with time, the norm of the difference between two vectors is divided by the norm of the actual state vector reflecting the total number of aircrafts.

The following procedure was repeated for time intervals $d = 600, 900, 1800, 3600, 7200$ (seconds).

1. Using CARATS Open data from 2012 to 2016, the sets of center matrices $\mathcal{M}_{d,K}^{A}$ and $\mathcal{M}_{d,K}^{B}$ for $K = 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100$ were computed. The flight trajectory data used for model construction is those in high demand time (8:00–20:00). The number of available matrices for each d is

$$\lfloor (20-8) \times 3600/d \rfloor \times 7(\text{days}) \times 6(\text{weeks}) \times 5(\text{years}).$$

2. CARATS Open data in 2017 is used for the evaluation. At each time step k, $x(k)$, $x(k+1)$ and $u(k)$ were computed from the event logs. Then the predicted sector demand $\tilde{x}(k+1)$ at time step $k+1$ was computed by the state Eq. (13) for every combination of matrices $A \in \mathcal{M}_{d,K}^{A}$ and $B \in \mathcal{M}_{d,K}^{B}$. Then the minimum value of $\varepsilon(\tilde{x}(k+1), x(k+1))$ is treated as the prediction error.

The target center was set to Tokyo Control, the largest center in Fukuoka FIR. The results of the evaluation are summarized as follows:

- The minimum error is computed at every time step. Figure 3 shows its average value in the evaluation period 8:00–20:00. The prediction error decreases as K increases, but the decreasing rate is not constant. The error increases as d increases. The errors for $d = 3600$ and $d = 7200$ are almost the same. For small d, the number of aircrafts that traverse sectors is small. Since the error is defined as a relative value to the norm of the state vector, it becomes small.
- To estimate a sufficient number K of clusters, we introduce the following number. Let $K_1 = 5, K_2 = 10, K_3 = 15, \cdots, K_{13} = 100$ and let E_{K_i} denote the average prediction error when the number of clusters is K_i. Then for $i = 1, 2, \cdots, 12$, let

$$\delta_{K_{i+1}} := (E_{K_{i+1}} - E_{K_i})/(K_{i+1} - K_i) \tag{15}$$

$\delta_{K_{i+1}}$ denotes decrement of the prediction error per addition of one matrix. We can assume that $|\delta_{K_i}|$ is large when the number of matrices is insufficient. Figure 4 shows this value for every combination of parameter values. It is observed that δ_{K_i} becomes almost constant after $K_i = 30 \sim 40$. Therefore, we conclude that around 30 is the appropriate number K of clusters.
- Not all matrices are used in the best combination of matrices, i.e., the pair of matrices that gives the minimum error, at each time step. Let K^{95} be the number of matrices from $\mathcal{M}_{d,K}^{A}$ that covers 95% of the total time steps. Figure 5 shows K^{95}/K for every d and K. It is observed that the value K^{95}/K is almost constant and under 0.6. The result for $\mathcal{M}_{d,K}^{B}$ is similar.

3.3 Sector Demand Prediction Methods

We propose two method for the sector demand prediction. The first method is based on matrix selection, and the second method is based on aggregation of predictions weighted by posterior probabilities.

3.3.1 Method 1

The first method is based on selecting matrices $A \in \mathcal{M}_{d,K}^{A}$

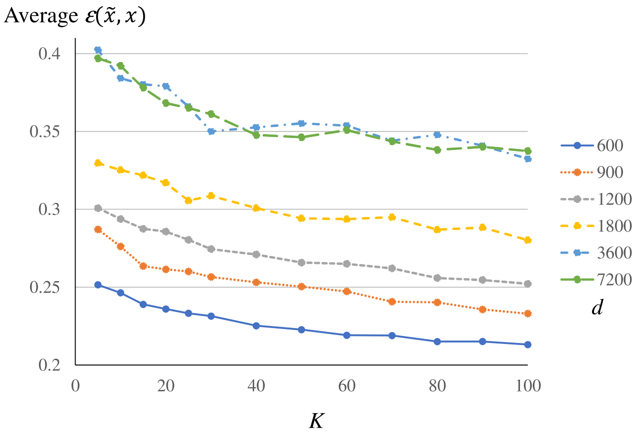

Fig. 3 Average prediction error.

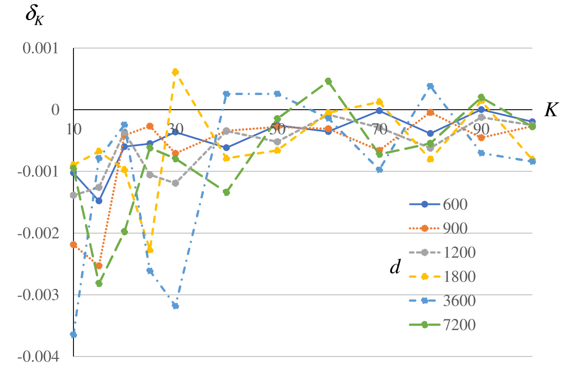

Fig. 4 Decrement of prediction error per addition of one matrix.

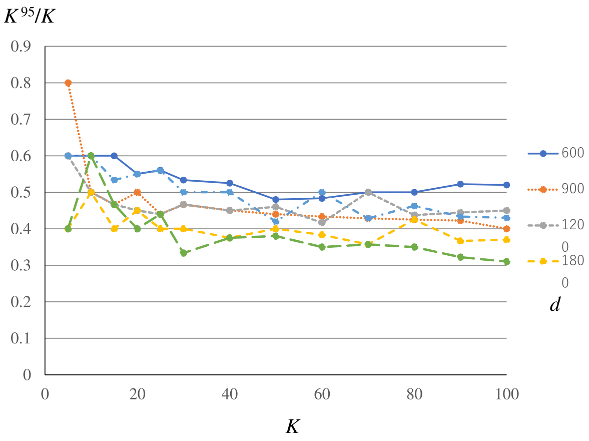

Fig. 5 Ratio of matrices that covers 95% of the total time steps.

and $B \in \mathcal{M}^B_{d,K}$. In [7], the hypothesis H_i that gives the highest probability $Pr(H_l \,|\, X_k)$ is selected. Then the matrices A and B are derived from the hypothesis. Without computing the posterior probabilities, we can directly compute the most fit matrices to the current situation. This is possible because the proposed approach is fully discretized. At each time step k, we compute the sector demand for every combination of matrices $A \in \mathcal{M}^A_{d,K}$ and $B \in \mathcal{M}^B_{d,K}$, and we obtain matrices A^* and B^* that minimizes the prediction error $\varepsilon(\tilde{x}(k), x(k))$. Then the predicted sector demand at $k+1$ is given by $\tilde{x}(k+1) = A^*x(k) + B^*u(k)$. If the change in the situation is small during two adjacent time steps, then we expect this method works well.

3.3.2 Method 2

The second method uses the n-gram model for past history of the best fit matrices A^* and B^*. Let $A^*(j)$ and $B^*(j)$ denote matrices that give the minimum prediction error at time step j. Then we have two sequence $A^*(0), \cdots, A^*(h_d)$ and $B^*(0), \cdots, B^*(h_d)$ of matrices, where h_d is the last time step in the target period, e.g., when the target period is 8:00–20:00, h_d is the time step just before 20:00. Note that h_d depends on parameter d and increases as d decreases. For a given positive integer n, we use the conditional probabilities

$$\begin{aligned} & p_A(A_0 \mid A_1, \cdots, A_n) := \\ & Pr(A^*(j+1) = A_0 \mid \\ & \qquad A^*(j) = A_1, \cdots, A^*(j-n+1) = A_n) \end{aligned}$$

and

$$\begin{aligned} & p_B(B_0 \mid B_1, \cdots, B_n) := \\ & Pr(B^*(j+1) = B_0 \mid \\ & \qquad B^*(j) = B_1, \cdots, B^*(j-n+1) = B_n) \end{aligned}$$

where $A_0, A_1, \cdots, A_n \in \mathcal{M}^A_{d,K}$ and $B_0, B_1, \cdots, B_n \in \mathcal{M}^B_{d,K}$. We here assume that the conditional probabilities are time-invariant, i.e., they do not depend on the time step j.

By the maximum likelihood estimation, the conditional probabilities are estimated from the past history of the best fit matrices. Suppose that a set of sequences $A^*(0), \cdots, A^*(h_d)$ is given. For a sequence of matrices $A_{i_1}, \cdots, A_{i_r}$ ($r \le h_d + 1$), let $\#(A_{i_1}, \cdots, A_{i_r})$ denote the number of its occurrences as a subsequence in the set. Then

$$p_A(A_0 \mid A_1, \cdots, A_n) = \frac{\#(A_n, \cdots, A_1, A_0)}{\#(A_n, \cdots, A_1)}$$

The probability p_B is similarly given. Using these conditional probabilities, the predicted sector demand $\tilde{x}(k+1)$ is computed as the sum of the predicted demand for each selection of the matrix weighted by the conditional probability for the matrix.

$$\begin{aligned} & \tilde{x}(k+1) := \\ & \sum_{A_i \in \mathcal{M}^A_{d,K}} p_A(A_i \mid A^*(k), \cdots, A^*(k-n+1)) \cdot A_i x(k) \\ & + \\ & \sum_{B_j \in \mathcal{M}^B_{d,K}} p_B(B_j \mid B^*(k), \cdots, B^*(k-n+1)) \cdot B_j u(k) \end{aligned} \tag{16}$$

Method 2 is based on a similar idea to that of [7], but has the following difference. The method in [7] uses conditional probability $Pr(x_k|X_{k-1}, H_l)$ on the state space. In Method 2, each state is replaced with the most fit matrix and the conditional probability is defined on the finite set of matrices. This makes the procedure simpler because we do not have to care about the conditional probability on the state space. Considering each matrix as a symbol, a sequence of matrices can be treated as a sequence of symbols. We use n-gram model, which is often used in natural language processing, to derive the conditional probability for predicting the next symbol.

3.4 Evaluation of the Methods

The two methods were implemented and the prediction errors by the methods were computed for CARATS Open data in 2017. The target center was Tokyo Control consisting of 39 sectors/subsectors, i.e., $N = 39$. Note that each sector may consists of several subsectors. In the construction of LDSM model, each subsector is treated as a sector. Figure 6 shows the average prediction errors for $K = 5$ and $K = 30$. The lower limits introduced in the previous subsection are also indicated. The indicated lower limits are computed using the same dataset as Method 1 and Method 2. Therefore, these lower limits show the error bounds that can be achieved by any method using the same LDSM model. We have the following observations.

- Method 1 shows stable performance for every combination of parameter values.
- The difference between the lower limit and the errors by the two method increases as d increases.
- Method 1 and Method 2 give similar prediction errors for small d, but Method 2 slightly outperforms Method 1, e.g., 0.274 (Method 1), 0.273 (Method 2 ($n = 1$)), 0.269 (Method 2 ($n = 2$)) and 0.277 (Method 2 ($n = 3$)) when $d = 600$ and $K = 30$.
- For large d, Method 1 outperforms Method 2. This means that the matrices do not change drastically when d is large.
- Figure 7 depicts the best method for every combination of parameter values. When $d = 600, 900, 1200$, Method 2 ($n = 3$) is the best for small K, and Method 2 ($n = 2$) becomes the best as K increases. Method 2 ($n = 1$) or Method 1 becomes the best for large K. When $d \geq 1800$, Method 1 is the best in most cases. The reason why Method 2 ($n = 3$) is not good for large d and large K is that the number of sequences used for composing the n-gram model is insufficient. In fact, most 4-grams appear only once in the data set when K becomes large. This is a weakness of using n-grams because the number of n-grams increases exponentially in n, but the amount of data is limited.

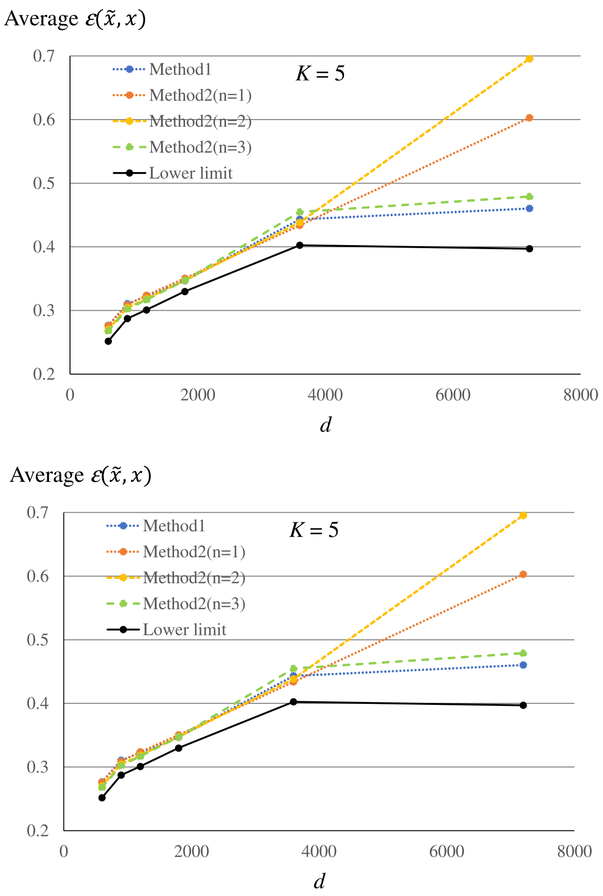

Fig. 6 Average prediction error by the two method.

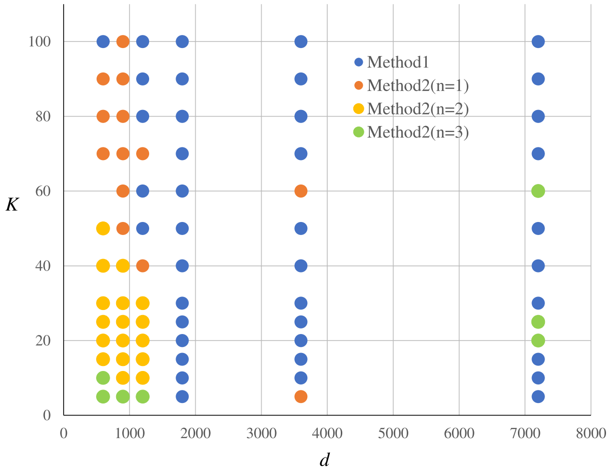

Fig. 7 Best method for each parameter setting.

4. Concluding Remarks

We have shown a method to construct LDSM from CARATS Open data, and have evaluated error bounds of the model by comparison with the actual data. Next we have proposed two methods for sector demand prediction based on the model. The proposed two prediction methods have been shown to achieve accurate sector demand prediction to some extent. It is difficult to compare the accuracy of the proposed methods with that in [7], because the situation in the airspace is different (one is in Japan and the other is in United States) and the detail of the methods and the traffic data is not described. The contribution of this paper is summarized as follows: (i) We have presented a complete workflow to construct LDSM from CARATS Open data. (ii) The lower limit of prediction errors has been obtained. Showing the lower limit is useful for the objective evaluation of arbitrary prediction methods which will be developed in the future. (iii) Model selection methods based on exhaustive search and the past history of the best fit models have been proposed. This can be seen as another realization of the idea in [7].

The followings are ideas toward improvement of the proposed methods.

- The clustering algorithm is applied to the set of all matrices derived from the data set. Before the clustering, we can divide the set of matrices into several subsets according to the situation such as weather, time zone (morning, noon, afternoon, evening), and the seasons. Then the set of center matrices is computed for each of the subsets. Considering the current situation, appropriate set of representative matrices can be selected. By using this idea, we expect that smaller K can give comparable performance.
- The definition of states in the state equation is the vector of sector demand at each time step. In addition to the current sector demand, we know the past history of passed sectors for each aircraft. Incorporating such information in the model, we can probabilistically estimate the next sector with the time for aircrafts having the same history. This may improve the accuracy of the prediction.

Of course, using flight plan of individual aircraft may increase the accuracy, but the model becomes very close to the multi-agent simulation. This is not what we intend to do.

Acknowledgments

This research is partly supported by Grant-in-Aid for Scientific Research of MEXT under Grant No. 20K04544. The authors also thank to Civil Aviation Bureau, MLIT for providing CARATS Open data.

References

[1] http://www.mlit.go.jp/report/press/kouku13_hh_000087.html
[2] C.R. Kaplan, E.S. Oran, N. Alexandrov, and J.P. Boris, "The monotonic lagrangian grid particle grid: A fast tracking methodology for air-traffic modeling," AIAA Paper 2009-1635, American Institute of Aeronautics and Astronautics, 2009.
[3] A. Bayen, P. Grieder, G. Meyer, and C.J. Tomlin, "Lagrangian delay predictive model for sector-based air traffic flow," Journal of Guidance, Control, and Dynamics, vol.28, no.5, pp.1015–1026, 2005.
[4] Á. Cámera, D. Castro, E. Oliveria, and P.H. Abreu, "Comparing a centralized and decentralized multi-agent approaches to air traffic control," Proc. 28th European Simulation and Modelling Conference (ESM'2014), pp.189–193, 2014.
[5] R. Breil, D. Delahaye, L. Lapasset, and É. Féron, "Multi-agent systems for air traffic conflicts resolution by local speed regulation and departure delay," Proc. IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016.
[6] https://www.fly.faa.gov/Products/Information/ETMS/etms.html
[7] B. Srighar, N.Y. Chen, and H.K. Ng, "An aggregate sector flow model for air traffic demand forecasting," Proc. 9th AIAA Aviation Technology, Integration, and Operations Conference, AIAA 2009-7129, 2009.
[8] C.F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," Transportation Research Part B: Methodological, vol.28, no.4, pp.269–287, 1994.
[9] C.F. Daganzo, "The cell transmission model, part II: Network traffic, transportation research part B: methodological," vol.29, no.2, pp.79–93, 1995.
[10] P.K. Menon, G.D. Sweriduk, and K.D. Bilimoria, "New approach for modeling, analysis, and control of air traffic flow," Journal of Guidance, Control, and Dynamics, vol.27, no.5, pp.737–744, 2004.
[11] P.K. Menon, G.D. Sweriduk, T. Lam, G. Diaz, and K.D. Bilimoria, "Computer-aided Eulerian air traffic flow modeling and predictive control," Journal of Guidance, Control, and Dynamics, vol.29, no.1, pp.12–19, 2006.
[12] D. Sun and A.M. Bayen, "Multicommodity Eulerian-lagrangian large-capacity cell transmission model for En route traffic," Journal of Guidance, Control, and Dynamics, vol.31, no.3, pp.616–628, 2008.
[13] H.M. Arneson and C. Langbort, "Distributed control design for a class of compartmental systems and application to Eulerian models of air traffic flows," 46th IEEE Conference on Decision and Control, pp.2876–2881, 2007.
[14] Q.K. Tran and K. Hiraishi, "An improved version of cell transmission model for air traffic flow," Proc. 3rd Int. Conf. Transportation Infrastructure and Sustainable Development (TISDIC2019), pp.335–343, 2019.
[15] S. Roy, B. Sridhar, and G.C. Verghese, "An aggregate dynamic stochastic model for an air traffic system," Proc. 5th Eurocontrol/Federal Aviation Agency Air Traffic Management Research and Development Seminar, Budapest, Hungary, 2003.
[16] B. Sridhar, T. Soni, K. Sheth, and G. Chatterji, "Aggregate flow model for air-traffic management," Journal of Guidance, Control, and Dynamics, vol.29, no.4, pp.992–997, 2006.
[17] C. Gwiggnwe and S. Nagaoka, "Recent models in the analysis of air traffic flow," Proc. 46th Aircraft Symposium, The Japan Society of Aeronautical and Space Sciences, 2B10, 2008.
[18] https://aisjapan.mlit.go.jp/
[19] S. Tokumaru and K. Hiraishi, "Sector identification for a large amount of airspace traffic data," IEICE Trans. Fundamentals, vol.E102-A no.5, pp.755–756, May 2019.
[20] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Ryosuke Mishima received B.S. in Engineering from National Institute of Technology, Matsue College in 2019, and M.S. in Information Science from Japan Advanced Institute of Science and Technology in 2021. From 2021, he is working at a Prop Tech company, GA technologies. His research interests are big data and predictive modeling.

Kunihiko Hiraishi received from the Tokyo Institute of Technology the B.E. degree in 1983, the M.E. degree in 1985, and D.E. degree in 1990. He is currently a professor at School of Information Science, Japan Advanced Institute of Science and Technology. His research interests include discrete event systems and formal verification. He is a member of the IEEE, IPSJ, and SICE.