AIAA AVIATION Forum
June 15-19, 2020, VIRTUAL EVENT
AIAA AVIATION 2020 FORUM

10.2514/6.2020-2869

# Data-Driven Approach for Aircraft Arrival Flow Investigation at Terminal Maneuvering Area

Lui, Go Nam[*1], Thierry.Klein[†2,3], and Liem, Rhea Patricia[‡1]
[1]*The Hong Kong University of Science and Technology (HKUST), Hong Kong*
[2]*Institut de mathématique, UMR5219; Université de Toulouse; CNRS, UPS IMT, F-31062 Toulouse Cedex 9, France*
[3]*ENAC - Ecole Nationale de l'Aviation Civile, Université de Toulouse, France*

**Recent air traffic management aims to provide a safety-first operation to support the aircraft approaching and landing procedures. Due to the complexity of air traffic in the terminal control area (also known as the terminal maneuvering area or TMA), simultaneous consideration of aviation economics, environmental concerns, and safety operations in decision makings can be challenging. To improve air traffic controllers' work efficiency and reduce the adverse environmental impact, it is crucial to establish a robust arrival strategy that incorporates weather conditions and flight trajectory configuration. The current state-of-the-art solutions for arrival sequencing and scheduling problem focus more on the operation research aspect, which neglects the airway configuration. Also, no wind condition is assumed to simplify the weather condition. Furthermore, many research efforts have not properly considered practical phenomenon such as holding patterns in their arrival sequencing model, which affects the accuracy of fuel burnt consumption. In this work, we will construct a study on aircraft arrival flow based on historical data at Hong Kong International Airport (HKIA). By extracting features from the data, our results include the spatiotemporal pattern recognition for aircraft arrival transit time and congestion inside HKIA TMA. Besides delivering the statistical analysis on the HKIA aircraft arrival flow, an arrival transit time prediction based on random forest regression is also converted. Results show that our methodologies are not only advantageous in extracting crucial hidden information from historical data for air traffic controllers but also can increase the accuracy of arrival transit time prediction under most of the circumstances.**

## Nomenclature

| | | |
|---|---|---|
| TMA | = | terminal maneuvering area |
| HKIA | = | Hong Kong International Airport |
| ETA | = | estimated time of arrival |
| ATM | = | air traffic management |
| ATC | = | air traffic control |
| STAR | = | standard instrument arrivals |
| ILS | = | instrument landing system |
| FCFS | = | first-come first-serve |
| ASP | = | arrival sequencing problem |
| CPS | = | constraint position shifting |

## I. Introduction

In the past few decades, air transportation plays an increasingly important role in promoting global business interaction and economic development. The positive impacts of the world aviation industry on social and economic development are far more extensive and richer than the profits realized by the industry. In the first half of 2019, the total revenue of system-wide global commercial airlines is 865 billion US dollars [1]. The figure below (Fig. 1) illustrates the trend of

[*] PhD student, Department of Mechanical and Aerospace Engineering, gnlui@connect.ust.hk
[†] Professor, thierry.klein@math.univ-toulouse.fr or thierry01.klein@enac.fr
[‡] Assistant Professor, Department of Mechanical and Aerospace Engineering, rpliem@ust.hk, AIAA Member

the number of carried passengers in the global air transportation from 1980–2019, based on data obtained from the World Bank Open Data. Recently, the global aviation industry is facing a significant challenge due to the resulting travel restrictions for the COVID-19 pandemic. Starting in 2020, the local trough like the ones in 2001-2003 and 2007-2009 is expected to appear. Even though the pandemic will decrease the volume of air transportation temporally, the industry is expected to recover in a short period of time. Therefore, air transportation is still facing a series of challenges such as increasing demand, safety issues, and sustainable development from a long-term perspective. Data collected by the US Bureau of Transportation Statistics shows that 20.29% of arrival flights were delayed in 2019, up from 18.75% in 2018, 18.14% in 2017, and 17.16% in 2016 in U.S [2]. To provide an organized and safe air transportation environment, air traffic management (ATM) coordinates and manages different systems inside air transportation. The International Civil Aviation Organization (ICAO) first proposed the importance of new air navigation systems and procedures for the growth of the civil aviation industry in 1983 [3]. Nonetheless, the recent ATM systems are still under technological and operational limitations which contribute to undesirable delays and additional negative environmental impacts, such as emission and noise. FAA (Federal Aviation Administration), ICAO (International Civil Aviation Organization), and Eurocontrol established novel projects and plans successively for ATM upgrades (NextGen, ASBU, SESAR) in the late 2010s almost simultaneously.

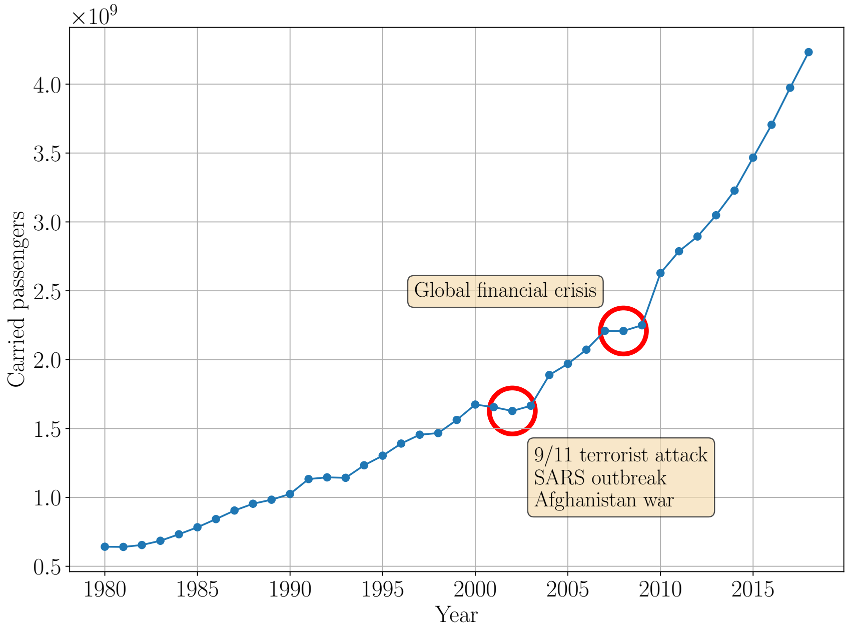


**Fig. 1 The trend of the number of carried passengers in air transportation from 1980 to 2018.**

To further improve the efficiency of the air traffic management system, novel technologies such as data-driven approaches have recently been explored. Generally, a data-driven approach can automatically analyze and obtain patterns from data and derives rules to predict unknown data. Because a large number of statistical theories are involved in learning algorithms, a data-driven approach is closely related to inferential statistics and large scale of data. In the context of ATM, the use of data-driven models has been explored in trajectory reconstruction, delay prediction, and fuel burnt computation [4–6].

The key point for the future is an integrated system that involves all flight conditions, decision phases, various locations, and weather conditions. However, due to political reasons and technical obstacles, constructing a whole-scale ATM data study is still challenging. The ATM research community defines several research areas for air transportation to reduce the level of complexity. Based on the flight phase, the research can be separated into airport stage, approach stage, en route stage, and regional stage [7]. On the other hand, ATM research could be classified by the air traffic flow management phase based on the period, including strategy phase, pre-tactical phase, and tactical phase [8]. The strategy phase starts from a year before the real operation day up to the week before, the pre-tactical phase represents the week before the real operation day, whereas the tactical phase stands for the real-time operation. For the weather impact on the air traffic flow management, seasonal patterns will impact the strategy phase of management. The precise continuous weather prediction data, on the other hand, can be used for the tactical phase air traffic management. Terminal control area, which includes airport stage and approach stage, is regarded as the bottleneck of the air traffic management system because of its natural high traffic density and complexity. From the fuel burn perspective, more efficient ATC (Air traffic control) could save 5-10% of aviation fuel by avoiding holding patterns and indirect airways[1]. The possible factors that contribute to aircraft delays inside TMA are local weather conditions, airspace capacity, and arrival sequencing. Various local weather conditions could conduct a complicated impact on aircraft approaching and landing. Rain, ice, snow or hail affect aircraft's occupied time on runways. Thunderstorm will reduce the airspace capacity and cause congestion, and clouds will affect the visibility for pilots' operation. For airspace capacity, limited manpower and lack of automation are causing holdings and congestion. Also, an optimum arrival sequencing strategy can reduce the undesirable extra airborne time due to aircraft wake vortex. This research is focused on the aircraft arrival flow inside the HKIA terminal control area, to investigate the correlation between spatio-temporal factors and aircraft arrival on-time performance. To further include the local weather forecast information, we will focus more at the tactical phase. In the next section, a brief literature review over this topic will be established. The proposed approach will be described in Section. III and the statistical analysis are presented in Section. IV. The arrival transit time prediction will be discussed in Section. V. Finally, the conclusion of the present work will be presented in Section. VI.

## II. Literature review

This section provides an overview of related research areas included air delay analysis, arrival sequencing, and data-driven approach in ATM research. Research on aircraft arrival sequencing has always closely tied with aircraft delay analysis studies. There have been a recent surge of data-driven investigations in this area. The abundance of aviation data has increasingly been used to study the spatio-temporal patterns of air traffic performance.

### A. Aircraft delay research

In 1976, the Federal Aviation Administration (FAA) published an Advisory Circular of Airport Capacity and Delay. This circular instructed the government to compute airport capacity and delay for airport planning and design. In chapter 3, the FAA defined many components of the performance of airport capacity and delay, and also introduced a method to calculate these factors. This Advisory Circular is considered as one of the earliest instructions for analyzing airport capacity and delay [9]. Erzberger, H. presented the design principle and algorithm for building a real-time scheduler in automated air traffic management concluded that a large amount of delay allocation can have adverse effects [10]. Due to the negative impacts of aircraft delay, researchers started to study the possible causalities for the aircraft delay and investigated ways to reduce delay. On the air traffic congestion aspect, research on delay propagation and air traffic network effect over multiple origin-destination (OD) pairs had shown some success [11–17]. Among these research, the authors focused mainly on applying different methods to describe the system-wise delay interaction and propagation, with much simplifications on the aircraft mission procedure and flight performance.

Besides the delay propagation research, there is a growing body of literature on the aircraft delay inside the terminal area. Some of the research focused on the causality analysis [18–20]. Several operational factors such as time of the day have been found to have a crucial impact on arrival delays, while some analyzed the negative economic impacts due to terminal delay [21, 22]. The economic loss due to terminal delay can be analyzed by evaluating the extra fuel cost and operational cost due to delay. However, evaluating the economic loss accurately is challenging if we only rely on low fidelity flight simulation and fuel burn quantification. Further improvement in this aspect can be incorporate with a high fidelity fuel consumption model, such as those presented in [6, 23, 24].

### B. Aircraft arrival sequencing research

When aircraft are approaching the near-terminal control area, air traffic controllers will decide on the landing runway and instruct the aircraft landing while making sure that the aircraft's horizontal separation limits are satisfied. The most commonly used procedures are the standard instrument arrivals (STAR), the instrument landing system (ILS), and the first-come first-serve (FCFS) strategy. The FCFS strategy is intuitively regarded as the best strategy for ATC which can reduce the operation workloads and the airspace complexity. However, the inefficiency and low runway usage rate of FCFS have recently been recognized as its limiting factors.

[1] The Economist. "Air-traffic control is a mess," Jun 15th, 2019 edition.

To maximize the runway throughput and minimize delay time, researchers are seeking an alternative solution for the arrival sequencing problem (ASP). Constraint Position Shifting (CPS), the methodology that aims to increase the runway throughput rate, was first developed by Dear in 1976 [25]. Compared to FCFS, CPS has a dynamic property that allows aircraft in sequence to shift their position under a maximum point shifting limitation. The performance of CPS under different circumstances have been investigated in [26–29], and new strategies such as Weighted-CPS [30] has been developed. Besides CPS, alternative solutions on aircraft dynamic scheduling problems have also been proposed during the last few decades [31, 32].

Other than regarding ASP as a dynamic scheduling problem, people also tried to solve it as a static case, which can incorporate multiple runways and more aircraft with less computational costs. Beasley *et al.* first presented their static solution on single and multiple runways in 2000 [33]. A similar methodology called time decomposition approach with Simulated Annealing (SA) was developed by Ma *et al.* to solve the aircraft landing sequencing problem afterward [34]. Three kinds of conflict including link conflict (wake separation regulation), node conflict, and runway conflict were defined. The proposed method reduced computational time compared to using FCFS. Based on their previous work, further expansion on the integrated airspace and airport operation system was established [35]. Three types of arrival strategies tested in this work are FCFS, optimized sequencing strategy (wake separation optimized), and optimized sequencing strategy with runway assignment. The objective was to minimize the summation of airway/taxiways overload and flight delay. However, in their research, they assumed that every flight could follow the STAR and ILS, which would not work with the adverse weather conditions and special airspace events.

From the industrial aspect, recent arrival sequencing tools will generate the estimated time of arrival for each flight and trajectory prediction for sequencing decision making based on runway occupation. The integration of environmental impact and seasonal variation into the short/medium term aircraft demand forecast is still lacking. Data-driven approaches might be useful in this scenario [7]. A brief introduction of the data-driven approach in ATM research is established in the next section. The limitation of the previous ASP research is lack of weather consideration and abnormal situation consideration. Furthermore, the additional cost calculation is simplified.

### C. Data-driven approach in ATM Research

With the abundance of data generated in the air transportation industry, there have been research efforts to utilize the probabilistic and statistical methodology to describe, study, and improve the ATM system and regulations. Data features that are relevant to ATM research include flight track (e.g., position, vertical and horizontal speed), flight information (e.g., departure/arrival time, runway), weather condition (e.g., wind speed, visibility), and aircraft characteristic (e.g., aircraft type, engine type).

Flight trajectory data typically contain spatio-temporal information of flights. Researchers first investigated and discussed the emerging of spatio-temporal data mining in 1999 [36]. Various clustering algorithms have been applied and tested to merge flights into air flow to extract spatio-temporal information and features. K-means, hierarchical clustering and density-based spatial clustering of application with noise (DBSCAN) have been applied, either individually or combined, in airspace abnormal detection, TMA characteristic comparison and en-route traffic optimization [37–42]. With the help from clustering algorithms, the deep features of flight tracks can be used as inputs to learning models for air flow performance prediction purposes. In this study, the clustering algorithm is also applied in the arrival transit time prediction section.

## III. Proposed approach

Previous research has focused more on the standard route procedure which sometimes ignores holding patterns, abnormal air paths and weather conditions. In our research, based on the flight trajectory data, we will perform a data-analytic study on the historical data. In particular, we will derive a novel data-enhanced methodology to support aircraft arrival sequencing strategy inside the terminal control area that can take into account weather information and abnormal situation. The proposed approach procedure is as follows:

1) Analyze historical flight trajectory data to study the spatio-temporal patterns. In this process, we incorporate trajectory simplification and an algorithm for holding pattern detection. Statistical analysis will be constructed based on the historical data and the extracted features.
2) Following the data analytic results, random forest regression [43] is employed in our prediction study, the accuracy of the model for different size of data and different preprocessing procedures are evaluated. The input variables include the time of day, entry altitude, entry groundspeed, entry vertical rate and entry heading angle.

### A. Data description

Flight information data are collected from the HKIA online API. We can obtain the detailed information of flights that lands and departs at HKIA for the past three months. A new flight information schedule for the next two weeks is also available, which can be used for model validation purposes.

Flight trajectory data in this research are collected by the Automatic dependent surveillance-broadcast (ADS-B) technology. ADS-B is a technology that tracks the aircraft's position and periodically broadcasts it. There are several online ADS-B resources such as Flighradar24 [44], FlightAware [45], and OpenSky Network [46]. We incorporate the OpenSky Network data for its easy accessibility. We apply the Python tool "Traffic" for data extraction and rearrangement from the OpenSky Network [47]. 30 days (May 1st to May 30th) of arrival and departure flight trajectories are used in this research. The information included in the ADS-B data is shown below (Table 1).

**Table 1 Example features from ADS-B data.**

| Features | Examples |
|---|---|
| Callsign | CAL921 |
| Icao24 | 8991b3 |
| Latitude, longitude | 22.36°, 117.49° |
| Altitude | 32000 feet |
| Vertical rate | 64 feet/min |
| Heading angle | 249.44° |
| Groudspeed | 452.64 knots |
| Timestamp | 2019-05-25 14:56:37+00:00 |

The data include real-time geographical information and operating information for one flight, also the identity code recorded by ICAO. After filtering the data, 8760 of flights will be used in the arrival transit time prediction study.

### B. Trajectory simplification

The large amount of points to represent one flight trajectory can lead to high computational cost and complexity. Trajectory simplification algorithm aims to overcome this issue by reducing points that are redundant in representing a flight trajectory. In this study, we use Ramer-Douglas-Peucker (RDP) algorithm [48] for points reduction.The RDP algorithm was first introduced as a recursive simplification method in cartography, which has also been employed on vision system [49], computational geometry [50], and road modeling [51], etc. This method is demonstrated on a flight trajectory example (CAL 921, May 26th, 2019), which is illustrated in Fig. 2. Blue dots represent the original trajectory data, whereas red dots represent the remaining trajectory points obtained based on the RDP algorithm. In our example below, the trajectory simplification algorithm reduces the number of points from 2471 to 34, i.e., only 1.4% of data retained. From this simple illustration, we can observe a significant reduction in the number of points used to represent the flight trajectory, and yet the trajectory profile maintaining accuracy. This observation demonstrates the efficiency of RDP algorithm in reducing the computational cost in the context of flight trajectory modeling.

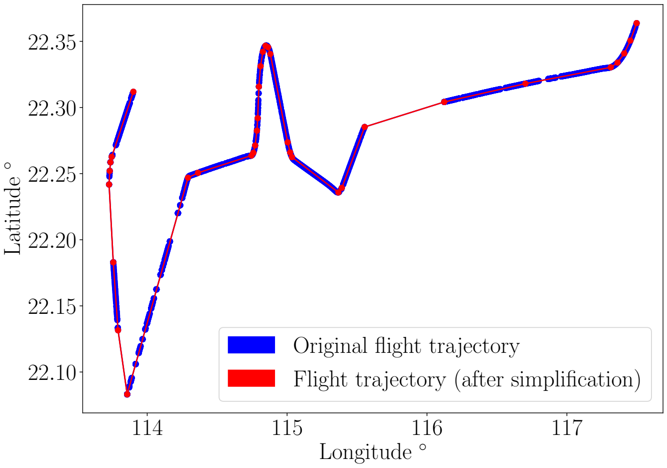
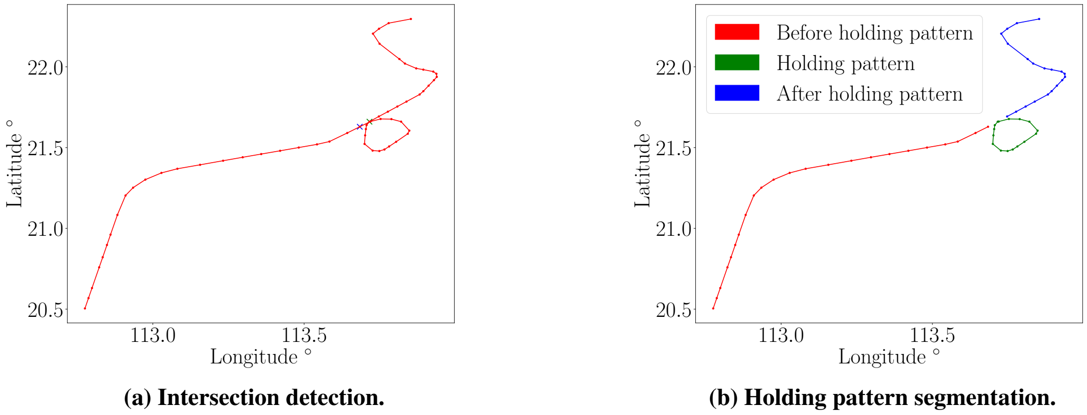


**Fig. 2 Trajectory simplification example.**

### C. Automatic holding pattern detection algorithm

Holding pattern is an alternative flight trajectory representation. While pilots told by the ATC to fly into a holding pattern, they will control the aircraft to loiter inside a race track loop. Many research has neglected the holding patterns while investigating the behaviors and characteristics flight trajectory data. Inspired by our previous work [52], a more robust holding pattern detection algorithm is developed. To incorporate the automatic holding pattern detection, the algorithm will go through each line segment representing a flight trajectory, where each line segment is defined by two adjacent points obtained through the RDP algorithm. Each line segment is evaluated whether it intersects any other line segments of the same flight trajectory. The algorithm will label all the intersection points, by then we can recognized the start point and endpoint for the holding pattern. This procedure is illustrated in Fig. 3. Fig. 3a indicates where the self-intersection is identified, and Fig. 3b shows the segmented sub-trajectories, where the portions of the flight trajectory before, during, and after the racetrack holding patterns are identified with different colors.



**(a) Intersection detection.** **(b) Holding pattern segmentation.**

**Fig. 3 Automatic holding pattern detection.**

# IV. Statistical analysis

## A. Temporal pattern recognition

By employing the automatic holding pattern detection algorithm, we first observe the statistics of 30 days data (May 2019). Fig. 4 illustrates the arrival transit time distributions of all available arrival flights we grabbed with and without holding patterns. Arrival transit time is the representation of the time aircraft spends inside the TMA before landing on the runways. This value is calculated from the flight trajectory data, by taking the difference between the time of entering HKIA TMA and the time of the aircraft first reaches its lowest altitude. As the figure illustrated, even within the same TMA, the arrival transit time of flights could vary between 500 seconds to 4000 seconds. In other words, some flights spend almost eight times longer airborne time inside the terminal control area than others. The figure indicates that those with holding patterns are likely to spend more time in TMA, although they occur less often than those without racetrack holding patterns. The overlap between the flights with and without holding patterns suggests that some flights will need to follow turning maneuvers or vectoring to wait for their turn to land instead of holding. Furthermore, holding planes in the air, either by following the race track patterns or by vectoring them and making them fly farther wastes fuel [53]. The hourly feature of arrival flight numbers is crucial since that hourly capacity reveals the airport capacity. According to ICAO, the capacity of an Air Traffic Services (ATS) unit depends on the level and type of ATS provided, the complexity of the sector/area/aerodrome and the associated route structure, and ATC workload (including control and coordination tasks to be performed) [7]. The time-of-the-day feature is considered as the most important factor for delay classification [5]. The box-plot shown in Fig. 5 infers the variation of the number of flights within each hour in May 2019. There seems to be a clear separation between busy and non-busy hour periods since the airport capacity for these two periods is notably different. Also, more variation is observed during the busy period, which supports that more operation needs to be done during a busy period. Before we really incorporate weather data into our study, we pick two days with different weather condition to construct a comparative study. Fig. 6 shows the hourly variations in the proportion of flights with and without holding patterns, the arrival transit time distribution and the corresponding rainfall contours. The accumulated rainfall contours demonstrates the specific higher rainfall on HKIA's arrival airways (Fig. 6a, Fig. 6c) on May 26$^{th}$. The data we used here are obtained from the public available JAXA Global Rainfall Watch database [54]. Data from May 26th (Fig. 6b), with heavy rainfall, and May 30th (Fig. 6d), with light rainfall are selected. Two figures are plotted with the same ranges of value along the y-axis. With heavier rainfall, we observe a significantly higher proportion of flights with holding patterns, especially during peak hours. In particular, these patterns are observed between noon to 15:00, and at 19:00. The increased proportion of flights with racetrack holding patterns correlate well with the higher arrival transit time, as shown in the scatter distribution plots.

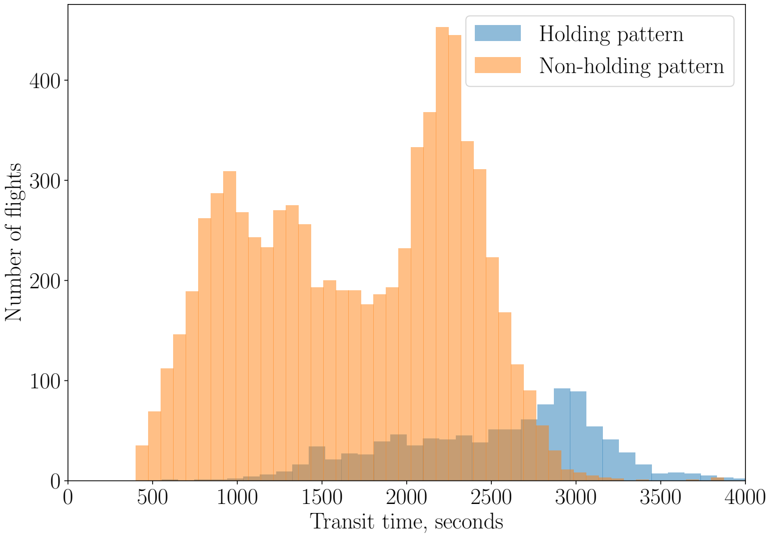


**Fig. 4 Flight time distributions of flights with and without holding patterns arriving at HKIA in May 2019.**

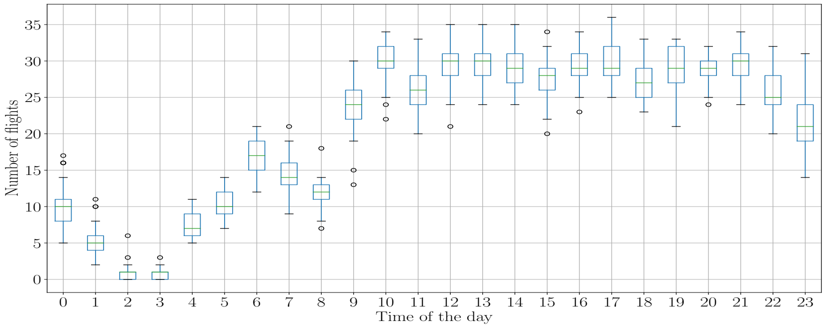
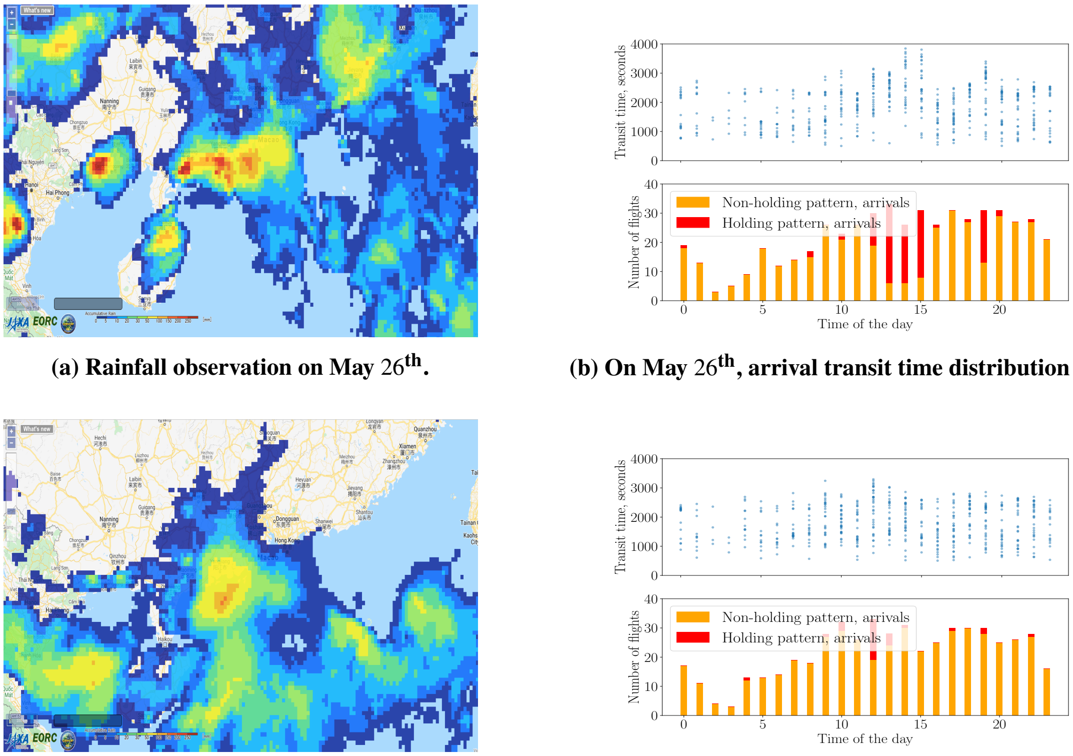


**Fig. 5 Hourly distribution of actual arrival flight number at HKIA (May 2019).**



**(a) Rainfall observation on May 26th.**

**(b) On May 26th, arrival transit time distribution hourly.**

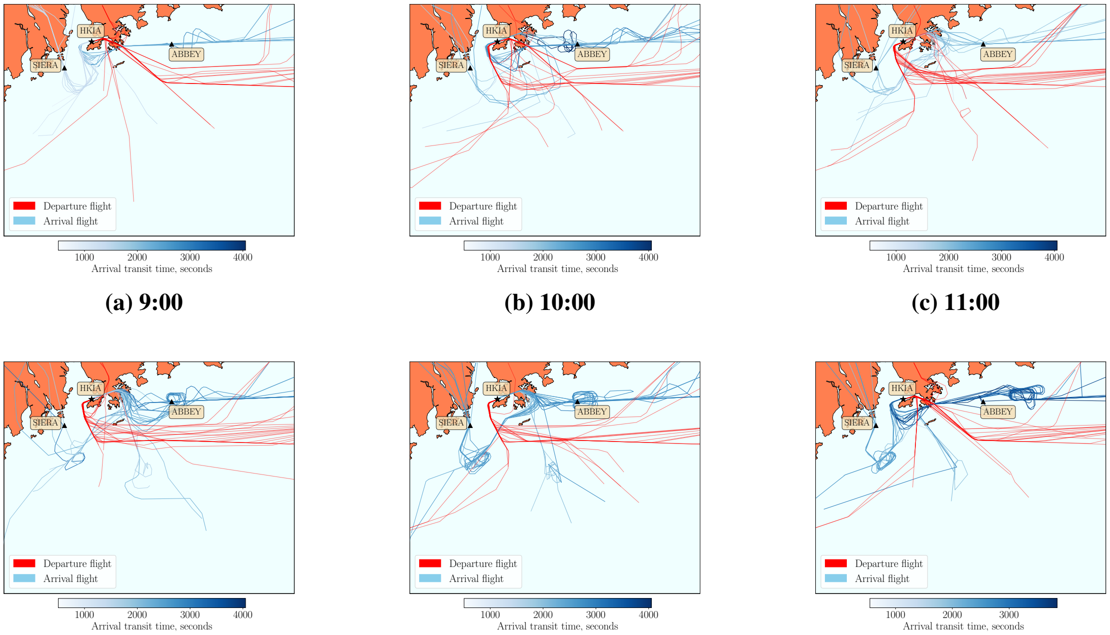
**(c) Rainfall observation on May 30th.**

**(d) On May 30th, arrival transit time distribution hourly.**

**Fig. 6 Hourly variations in the proportion of flights with and the proportion of flights without holding patterns, and the corresponding transit time distribution along with rainfall observation.**

### B. Abnormal congestion observation

Based on our previous observation, we continue our research by looking into the spatial arrival flow performance for each hour, to discover the possible reasons for large amount of holding patterns and arrival airborne delays. The data we used in this section are from 9:00 to 14:00 on May 26th. Due to the extreme rainfall and thunderstorm, there were numerous delays and holdings on May 26th. By projecting the arrival transit time on the arrival flight trajectory, we can observe the possible causes for longer arrival airborne time. Fig. 7 shows the variety of aircraft flight trajectories pattern from 9:00 to 14:00 on May 26th. Before 10:00 (Fig. 7a), the air traffic were operating in the normal mode. For HKIA, the normal operation mode is segregation, which means arrival and departure are using segregated runway. Starting from 10 am (Fig. 7b), the operation mode became to mixed-mode, which is not typical at HKIA. Meanwhile, some flights are instructed to fly in holding patterns, which increase the landing time. From 11:00 to 13:00 (Fig. 7c, Fig. 7d) the operation turned back to segregated mode but with reverse runway usage. The possible reason for runway operation transformation is the extreme rainfall and thunderstorm at certain area. Noticeable, severe delays and lot of holding patterns appeared from 14:00 to 15:00 (Fig. 7f), where ATC were changing the reverse segregated mode back to normal segregated mode. This information proves the propagation of delay and holding pattern are existed inside terminal control area, and not just happened in air traffic network in a wider sense.



**(a) 9:00**

**(b) 10:00**

**(c) 11:00**

**(d) 12:00**

**(e) 13:00**

**(f) 14:00**

**Fig. 7 Flight trajectory representation in HKIA TMA within a hour**

## V. Arrival transit time prediction

After performing statistical analysis on the historical data, we employ a machine learning technique for the arrival transit time prediction. Machine learning methodology is an application of artificial intelligence which is capable to learn from data automatically without human assistance. In this study, we incorporate the random forest regression [43] for our arrival transit time prediction. Random forest is one of the ensemble learning method which was first introduced in 1995 [55]. Random forest applies the typical technique of bootstrap aggregating, or bagging, to train the dataset. Random forest can be described as a combination of bagging and multiple decision-tree models. For our study, we apply random forest regression as our learning model. The output feature for our model is arrival transit time, and the input variables for the model are:

1) Altitude at entry point,
2) Groundspeed at entry point,

3) Vertical rate at entry point,
4) Heading angle at entry point,
5) Time of the day (HK hour) when enters HKIA TMA.

The five input variables are regarded as the simplest case in our case. In addition to these five basic variables, we further refine our models by including two identifiers: one to indicate whether a flight undergoes the racetrack holding pattern or no, and another one to indicate which cluster a flight belongs to. In particular, we performed three refinements to the model: by including the holding pattern identifier, the cluster identifier, and both. Clustering is also a machine learning algorithm which can automatically label clusters by the given input variables (i.e. geometrical information). To cluster the flight trajectories, we use Hausdorff distance as our distance metric:

$$d_{\mathrm{H}}(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\} \tag{1}$$

where $sup$ represents the supremum and $inf$ the infimum. Generally, it is the greatest of all the distances from a point in one set to the closest point in the other set. We choose to apply K-medoids to cluster the flight trajectories spatially [56]. At this preliminary stage of the work, we only consider two clusters for illustration purposes, as shown in Fig. 8. As shown here, we separate incoming flights from the East from the rest of the flight trajectories. Further studies with more refined clustering will be performed at a later stage to provide more computationally rigorous and conclusive results.

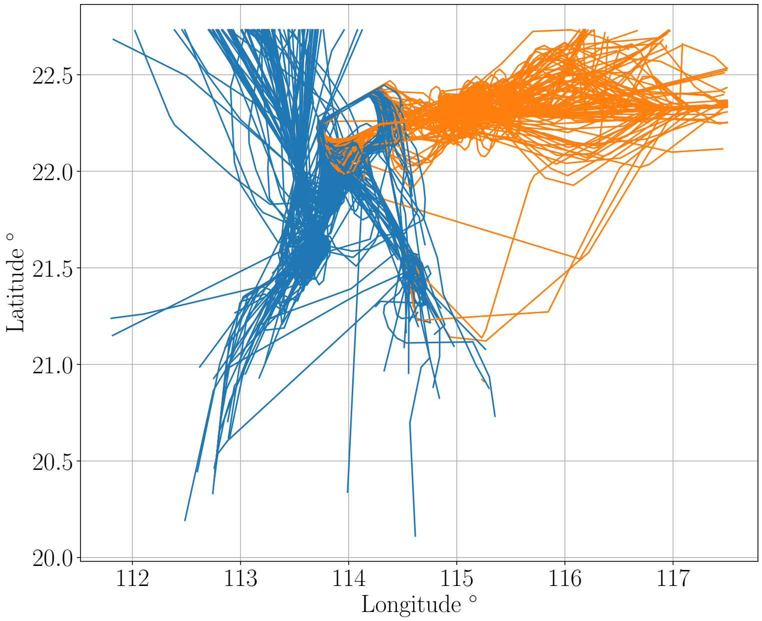


**Fig. 8 Spatial flight trajectory clustering with K-medoids (May 26th, 2019).**

The k-medoids, or partitioning around medoids (PAM) algorithm, is a clustering algorithm reminiscent of the k-means algorithm. This method attempt to minimize the Hausdorff distance between trajectory labeled to be in a cluster and a trajectory designated as the center of that cluster. Our experiment will be constructed on different size of dataset, which are one-day data, one-week data, and one-month data. The hyperparameter tuning procedure will be discussed next.

## A. Hyperparameter tuning

For a specific learning algorithm, hyperparameter tuning is crucial for selecting a set of optimal hyperparameters for the model. For this work, we choose the random search algorithm [57] for our tuning process. The advantage of random search algorithm is its gradient-free property. For the random forest model, there are 6 hyperparameters for tuning:

1) Number of estimators (trees) in random forest,
2) Minimum number of samples required to split a node,
3) Minimum number of samples required at each leaf node,
4) Method of selecting samples for training each tree (bootstrap or not),
5) Number of features to consider at every split,
6) Maximum depth (maximum number of levels in tree).

Using 3 fold cross validation, search across 100 different combinations to define the optimal hyperparameter set. Using one-day data as example, the mean absolute percentage error can be reduced from 17.06% to 14.94% with the tuned hyperparameter set.

### B. Results and discussion

The proportion of training set and testing set is 8:2. Three different size of data are examined, which are one day data (May 26th, 523 flights), one week data (May 24th to May 30th, 2917 flights), and one month data (May 1st to May 30th, 8760 flights). Because the computational time is acceptable for all cases, we did not perform evaluations from that aspect in this work. Root mean square error (RMSE) and mean absolute percentage error (MAPE) is used as the evaluation metrics in this experiment. For one day data, the best performance for random forest regression on this problem is 13.77% MAPE with both holding pattern and clustering features added. Using the one-month dataset, the accuracy can reach 0.07% while only add the holding pattern feature on the simplest case. For our results, we are convinced that holding pattern detection and clustering algorithm are beneficial for increasing the prediction accuracy in most cases.

**Table 2 Results for arrival transit time prediction experiment.**

| | One day data | | One week data | | One month data | |
|---|---|---|---|---|---|---|
| | RMSE (sec) | MAPE % | RMSE (sec) | MAPE % | RMSE (sec) | MAPE % |
| 5 variables | 449.95 | 19.19 | 57.32 | 0.83 | 13.59 | 0.14 |
| 5 variables + 1 (HP) | 343.50 | 14.61 | 40.06 | 0.8 | 8.51 | 0.07 |
| 5 variables + 1 (Clustering feature) | 357.98 | 15.17 | 31.63 | 0.81 | 18.57 | 0.14 |
| 5 variables + 2 (HP & Clustering) | 316.99 | 13.77 | 44.19 | 0.77 | 9.91 | 0.10 |

## VI. Conclusion

This paper presents a series of statistical analysis and an arrival transit time prediction for the aircraft arrival flow investigation inside HKIA TMA. Based on the one month ADS-B data, to better extract relevant features from the raw flight trajectory data, we developed an automatic holding pattern detection algorithm and prove its usage in both statistical analysis and arrival transit time prediction. Detailed statistical analysis reveals the apparent temporal patterns for arrival flights. Furthermore, the arrival transit time mapping on spatial trajectory visualization also shows the propagation characteristic inside the terminal control area. When unexpected airspace restriction or severe local weather condition happens, the highest arrival delays might appear few hours later. Not only can this crucial information be useful for air traffic controllers, it can also provide inspiration and direction for the future TMA air traffic modeling research.

Arrival transit time prediction using random forest regression was also performed. The accuracy of the prediction model increases with the size of the data. The relatively large error retaining to one-day data can be attributed to the limited data, as compared to one-week data and one-month data. Any incidental variations in air traffic operators can be better absorbed in a longer time window. Note, however, that data included in this analysis are still limited. Further investigations with more data need to be performed for a more solid and robust result. The future plan from this preliminary work is a two-way investigation. From the statistical analysis aspect, a larger dataset is expected to construct a seasonal pattern recognition study. Also, discovering effective ways to convert significant information for air traffic controllers. On the other hand, how to increase the prediction's accuracy of arrival transit time with a smaller dataset is also worth focusing on. More algorithms should be tested, and detailed weather data will be involved for a more computationally rigorous study of how weather conditions affect the air traffic at the terminal control area.

## VII. Acknowledgement

Authors would like to acknowledge The PROCORE-France/Hong Kong Joint Research Scheme 2018/19, Project NO. F-HKUST 602/18. The third author thanks the support from the Hong Kong University of Science and Technology Start-Up Grant, Project NO. R9354.

## References

[1] IATA, "Industrial Statistics Fact Sheet," `https://www.iata.org/pressroom/facts_figures/fact_sheets/Documents/fact-sheet-industry-facts.pdf`, 2019. Accessed July 3rd, 2019.

[2] Bureau of Transportation Statistics, "On-Time Performance – Flight Delays at a Glance," `https://www.transtats.bts.gov/HomeDrillChart_Month.asp`, 2019. Accessed July 3rd, 2019.

[3] ICAO, "Global Air Navigation Plan for CNS/ATM Systems," `https://www.icao.int/publications/Documents/9750_2ed_en.pdf/`, 2002. Accessed July 3, 2019.

[4] Fernández, E. C., Cordero, J. M., Vouros, G., Pelekis, N., Kravaris, T., Georgiou, H., Fuchs, G., Andrienko, N., Andrienko, G., Casado, E., et al., "DART: A Machine-Learning Approach to Trajectory Prediction and Demand-Capacity Balancing," *Seventh SESAR Innovation Days*, 2017.

[5] Gopalakrishnan, K., and Balakrishnan, H., "A comparative analysis of models for predicting delays in air traffic networks," ATM Seminar, 2017.

[6] Yanto, J., and Liem, R. P., "Aircraft fuel burn performance study: A data-enhanced modeling approach," *Transportation Research Part D: Transport and Environment*, Vol. 65, 2018, pp. 574–595.

[7] Kistan, T., Gardi, A., Sabatini, R., Ramasamy, S., and Batuwangala, E., "An evolutionary outlook of air traffic flow management techniques," *Progress in Aerospace Sciences*, Vol. 88, 2017, pp. 15–42.

[8] Liang, M., "Aircraft Route Network Optimization in Terminal Maneuvering Area," Ph.D. thesis, Université Paul Sabatier (Toulouse 3), 2018.

[9] Young, S. B., and Wells, A. T., *Airport planning and management*, McGraw-Hill Education, 2019.

[10] Erzberger, H., "Design principles and algorithms for automated air traffic management," *Knowledge-Based Functions in Aerospace Systems*, Vol. 7, No. 2, 1995.

[11] Wang, P. T., Schaefer, L. A., and Wojcik, L. A., "Flight connections and their impacts on delay propagation," *Digital Avionics Systems Conference, 2003. DASC'03. The 22nd*, Vol. 1, IEEE, 2003, pp. 5–B.

[12] Rebollo, J. J., and Balakrishnan, H., "Characterization and prediction of air traffic delays," *Transportation research part C: Emerging technologies*, Vol. 44, 2014, pp. 231–241.

[13] Pyrgiotis, N., Malone, K. M., and Odoni, A., "Modelling delay propagation within an airport network," *Transportation Research Part C: Emerging Technologies*, Vol. 27, 2013, pp. 60–75.

[14] AhmadBeygi, S., Cohn, A., Guan, Y., and Belobaba, P., "Analysis of the potential for delay propagation in passenger airline networks," *Journal of air transport management*, Vol. 14, No. 5, 2008, pp. 221–236.

[15] Wong, J.-T., and Tsai, S.-C., "A survival model for flight delay propagation," *Journal of Air Transport Management*, Vol. 23, 2012, pp. 5–11.

[16] Xu, N., Donohue, G., Laskey, K. B., and Chen, C.-H., "Estimation of delay propagation in the national aviation system using Bayesian networks," *6th USA/Europe Air Traffic Management Research and Development Seminar*, FAA and Eurocontrol Baltimore, MD, 2005.

[17] Laskey, K. B., Xu, N., and Chen, C.-H., "Propagation of delays in the national airspace system," *arXiv preprint arXiv:1206.6859*, 2012.

[18] Allan, S., Gaddy, S., and Evans, J., "Delay causality and reduction at the New York City airports using terminal weather information systems," Tech. rep., Citeseer, 2001.

[19] Mueller, E., and Chatterji, G., "Analysis of aircraft arrival and departure delay characteristics," *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, 2002, p. 5866.

[20] Abdel-Aty, M., Lee, C., Bai, Y., Li, X., and Michalak, M., “Detecting periodic patterns of arrival delay,” *Journal of Air Transport Management*, Vol. 13, No. 6, 2007, pp. 355–361.

[21] Ayra, E. S., Insua, D. R., and Cano, J., “To fuel or not to fuel? Is that the question?” *Journal of the American Statistical Association*, Vol. 109, No. 506, 2014, pp. 465–476.

[22] Ryerson, M. S., Hansen, M., and Bonn, J., “Time to burn: Flight delay, terminal efficiency, and fuel consumption in the National Airspace System,” *Transportation Research Part A: Policy and Practice*, Vol. 69, 2014, pp. 286–298.

[23] Liem, R. P., Mader, C. A., and Martins, J. R. R. A., “Surrogate Models and Mixtures of Experts in Aerodynamic Performance Prediction for Mission Analysis,” *Aerospace Science and Technology*, Vol. 43, 2015, pp. 126–151. https://doi.org/10.1016/j.ast.2015.02.019.

[24] Lyu, Y., Yanto, J., and Liem, R. P., “Aircraft Reserve Fuel Study with High-fidelity Fuel Approximation Model,” *AIAA Aviation 2019 Forum*, 2019, p. 3509.

[25] Dear, R. G., “The dynamic scheduling of aircraft in the near terminal area,” Tech. rep., Cambridge, Mass.: Flight Transportation Laboratory, Massachusetts Institute of Technology, 1976.

[26] Balakrishnan, H., and Chandran, B., “Scheduling aircraft landings under constrained position shifting,” *AIAA guidance, navigation, and control conference and exhibit*, 2006, p. 6320.

[27] Chandran, B., and Balakrishnan, H., “A dynamic programming algorithm for robust runway scheduling,” *2007 American Control Conference*, IEEE, 2007, pp. 1161–1166.

[28] Psaraftis, H. N., “A dynamic programming approach to the aircraft sequencing problem,” Tech. rep., Cambridge, Mass.: Massachusetts Institute of Technology, 1978.

[29] Lee, H., and Balakrishnan, H., “A study of tradeoffs in scheduling terminal-area operations,” *Proceedings of the IEEE*, Vol. 96, No. 12, 2008, pp. 2081–2095.

[30] Xu, B., Ma, W., Huang, H., and Yue, L., “Weighted Constrained Position Shift Model for Aircraft Arrival Sequencing and Scheduling Problem,” *Asia-Pacific Journal of Operational Research*, Vol. 33, No. 04, 2016, p. 1650028.

[31] Beasley, J. E., Krishnamoorthy, M., Sharaiha, Y. M., and Abramson, D., “Displacement problem and dynamically scheduling aircraft landings,” *Journal of the operational research society*, Vol. 55, No. 1, 2004, pp. 54–64.

[32] Zhan, Z.-H., Zhang, J., Li, Y., Liu, O., Kwok, S., Ip, W., and Kaynak, O., “An efficient ant colony system based on receding horizon control for the aircraft arrival sequencing and scheduling problem,” *IEEE Transactions on Intelligent Transportation Systems*, Vol. 11, No. 2, 2010, pp. 399–412.

[33] Beasley, J. E., Krishnamoorthy, M., Sharaiha, Y. M., and Abramson, D., “Scheduling aircraft landings—the static case,” *Transportation science*, Vol. 34, No. 2, 2000, pp. 180–197.

[34] Ma, J., Delahaye, D., Sbihi, M., and Mongeau, M., “Merging Flows in Terminal Moneuvering Area using Time Decomposition Approach,” *7th International Conference on Research in Air Transportation (ICRAT 2016)*, 2016.

[35] Ma, J., Delahaye, D., Sbihi, M., Scala, P., and Mota, M. A. M., “Integrated optimization of terminal maneuvering area and airport at the macroscopic level,” *Transportation Research Part C: Emerging Technologies*, Vol. 98, 2019, pp. 338–357.

[36] Roddick, J. F., and Spiliopoulou, M., “A bibliography of temporal, spatial and spatio-temporal data mining research,” *ACM SIGKDD Explorations Newsletter*, Vol. 1, No. 1, 1999, pp. 34–38.

[37] Gariel, M., Srivastava, A. N., and Feron, E., “Trajectory clustering and an application to airspace monitoring,” *IEEE Transactions on Intelligent Transportation Systems*, Vol. 12, No. 4, 2011, pp. 1511–1524.

[38] Li, L., Gariel, M., Hansman, R. J., and Palacios, R., “Anomaly detection in onboard-recorded flight data using cluster analysis,” *2011 IEEE/AIAA 30th Digital Avionics Systems Conference*, IEEE, 2011, pp. 4A4–1.

[39] Murça, M. C. R., Hansman, R. J., Li, L., and Ren, P., “Flight trajectory data analytics for characterization of air traffic flows: A comparative analysis of terminal area operations between New York, Hong Kong and Sao Paulo,” *Transportation Research Part C: Emerging Technologies*, Vol. 97, 2018, pp. 324–347.

[40] Marzuoli, A., Gariel, M., Vela, A., and Feron, E., “Data-based modeling and optimization of en route traffic,” *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 6, 2014, pp. 1930–1945.

[41] Chao, W., Xiaohao, X., and Fei, W., "ATC Serviceability Analysis of Terminal Arrival Procedures Using Trajectory Clustering [J]," *Journal of Nanjing University of Aeronautics & Astronautics*, Vol. 1, 2013.

[42] Olive, X., and Morio, J., "Trajectory clustering of air traffic flows around airports," *Aerospace Science and Technology*, Vol. 84, 2019, pp. 776–781.

[43] Liaw, A., Wiener, M., et al., "Classification and regression by randomForest," *R news*, Vol. 2, No. 3, 2002, pp. 18–22.

[44] Flightradar24, "Data," https://www.flightradar24.com/data.com, 2019. Accessed July 3, 2019.

[45] Flightaware, "Tracklog," https://flightaware.com, 2019. Accessed July 3, 2019.

[46] Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., and Wilhelm, M., "Bringing up OpenSky: A large-scale ADS-B sensor network for research," *Proceedings of the 13th international symposium on Information processing in sensor networks*, IEEE Press, 2014, pp. 83–94.

[47] Olive, X., "traffic, a toolbox for processing and analysing air traffic data," *Journal of Open Source Software*, Vol. 4, 2019, p. 1518. https://doi.org/10.21105/joss.01518.

[48] Douglas, D. H., and Peucker, T. K., "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: the international journal for geographic information and geovisualization*, Vol. 10, No. 2, 1973, pp. 112–122.

[49] Ramer, U., "An iterative procedure for the polygonal approximation of plane curves," *Computer graphics and image processing*, Vol. 1, No. 3, 1972, pp. 244–256.

[50] Rote, G., "The convergence rate of the sandwich algorithm for approximating convex functions," *Computing*, Vol. 48, No. 3-4, 1992, pp. 337–361.

[51] Visvalingam, M., and Williamson, P. J., "Simplification and generalization of large scale data for roads: a comparison of two filtering algorithms," *Cartography and Geographic Information Systems*, Vol. 22, No. 4, 1995, pp. 264–275.

[52] Maulydiana, S. F., Guajardo, J. C., and Liem, R. P., "Probabilistic approach in flight trajectory modeling for fast and efficient noise contour generation," 2018.

[53] Spinardi, G., "Up in the air: barriers to greener air traffic control and infrastructure lock-in in a complex socio-technical system," *Energy research & social science*, Vol. 6, 2015, pp. 41–49.

[54] Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., Kawamoto, N., et al., "Global Satellite Mapping of Precipitation (GSMaP) products in the GPM era," *Satellite precipitation measurement*, Springer, 2020, pp. 355–373.

[55] Ho, T. K., "Random decision forests," *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1, IEEE, 1995, pp. 278–282.

[56] Park, H.-S., and Jun, C.-H., "A simple and fast algorithm for K-medoids clustering," *Expert systems with applications*, Vol. 36, No. 2, 2009, pp. 3336–3341.

[57] Bergstra, J., and Bengio, Y., "Random search for hyper-parameter optimization," *Journal of machine learning research*, Vol. 13, No. Feb, 2012, pp. 281–305.