International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 156-164 March 2026-Sept 2026 www.csjournals.com

Flight Delay Prediction Based On Aviation Big Data and Machine Learning

Sahil[1], Dr. Vishal Verma[2], Sandeep Rathi[3]
[1]MCA Student, [2]Professor, [3]Asst Professor
Deptt. of Comp. Sc. and Appl, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Flight delay prediction is one of the crucial tasks in civil aviation. Flight delay leads to reduced efficiency in the operation, reduced satisfaction of the passengers and increased costs of the airlines . In this study, a machine learning framework is developed to predict the delays of Indian domestic flights with the help of aviation big data, which is attached. The aviation big data employed in the study combines the schedule information, operating indicators, airline performance data, airport characteristics and meteorological data.The data consists of 10,718 flights from February 2019 to December 2019, of which 10,633 were unique flights from 5 routes and 7 airlines upon removing the duplicate data. The study challenges the prediction of delays as a problem of classification with five classes: on-time, minor delay, moderate delay, significant delay and severe delay. The study is based on the class imbalance problem in the Random Forest training with SMOTE method.Random Forest with SMOTE is consistently shown as the most effective of the models tested. However, the slightly higher numerical accuracy reported in one results subsection for the two models, Decision Tree and XGBoost, is an inconsistency that should be recognized.Finally, the paper concludes that class balancing, ensemble learning and multi-source feature fusion enhance the prediction of flight delays in the Indian domestic aviation system.

Introduction

One of the most constant operational disruptions in commercial aviation is flight delays, which impacts missed connections, gate congestion, crew scheduling issues, etc., along with increased fuel burn and direct financial loss.In this study a delay is defined as a departure or arrival more than 15 minutes after schedule.The report provides context for this issue, noting the more than 140 million passengers flying domestically in India in 2019, and the heavy daily traffic at key metropolitan corridors. With the advent of aviation big data, the report says more proactive predictive capabilities are possible in aviation, as opposed to classic delay management, which is more reactive, with the ability to use historic flight data, telemetry and weather data.It points out four fundamental challenges: firstly, delay is caused by many interacting reasons; secondly, the data is highly imbalanced; thirdly, many previous studies reduce the task to a binary classification; and fourthly, the different types of data sources mandate extensive data preprocessing and feature engineering.To fill this, the study suggests a multi-class approach to Indian domestic aviation, which incorporates features at flight-level, airline-level, airport-level, and weather-level.

Literature Review

The attached report is part of a wider literature, in which classical machine learning remains quite competitive for flight delay prediction.The report concluded that the accuracy of Random Forest binary classification for ADS-B data from eastern China was 90.2%, and the model did not suffer from overfitting problems when applied to a small dataset, which is not the case for LSTM.The report indicated that the accuracy of Random Forest binary classification for ADS-B data from eastern China was 90.2%, and it did not have the problem of overfitting as LSTM did when dealing with limited datasets (Gui et al., 2020).Moreira et al. (2018) obtained 78.02% of accuracy for Brazilian aviation data and pointed that preprocessing and feature engineering have a strong impact on performance .With an artificial neural network, Pamplona et al. (2018) have obtained an accuracy of 87% and they have concluded that the time of the day and the hour in which the block is located, as well as the route have been decisive factors.

The report also covers deep learning and hybrid models, which can be very effective, if they have sufficient training data and optimization.Kim et al. (Kim et al., 2016) reported the accuracy of 85% and 87.42% using two-stage deep learning architecture. They state that deep models require massive amounts of learning data to be good generalizers.To achieve high accuracy, Dai (2024) proposed the COWRF model with 24,986 JFK flights, using clustering, feature selection, and optimized weighted Random Forest design, resulting in an average accuracy of 97.2%.Xu et al. (2022) adopted a five-layer fully connected neural network and obtained a success rate of 92.39% while Paramita et al. (2022) showed that cluster computing with Random Forest led to a 38% increase in the processing speed of the prediction for large-scale data.

Recent studies also employ model of spatial delay structure and temporal delay structure rather than flights as isolated events.To achieve the highest accuracy, Li and Jing (2022) used a hybrid model that fused the spatial information from airport networks and temporal information from LSTM, achieving an accuracy of 92.39%.Based on the above frameworks, we extended existing methods to include propagation learning and applied them to a global dataset covering 136 countries and 370 airports.Inspired by these, we proposed network-wide spatiotemporal propagation learning and implemented it in FlightNet-ST using LSTM, GCN and 3D-CNN models with an MAE of 5.2 minutes and 94.6% binary accuracy on the US domestic dataset .The report summarizes four main findings in this literature: ensemble methods are always successful, multi-source feature integration is important, class imbalance is a major challenge, and deep learning requires more data and regularization.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 156-164 March 2026-Sept 2026 www.csjournals.com

Methodology

In this study, real-time data on the number of domestic flights operating in India during February to December 2019 is used.Airlines flight data was obtained from public airline performance records and combined with World Weather Online API hourly weather data.It has five routes covering Delhi- Hyderabad, Delhi- Mumbai, Delhi-Bengaluru, Delhi-Kolkata and Mumbai-Bengaluru, while there are seven airlines Indigo, Air India, Vistara, SpiceJet, Air Asia, and the new airline Go Ai(2025).

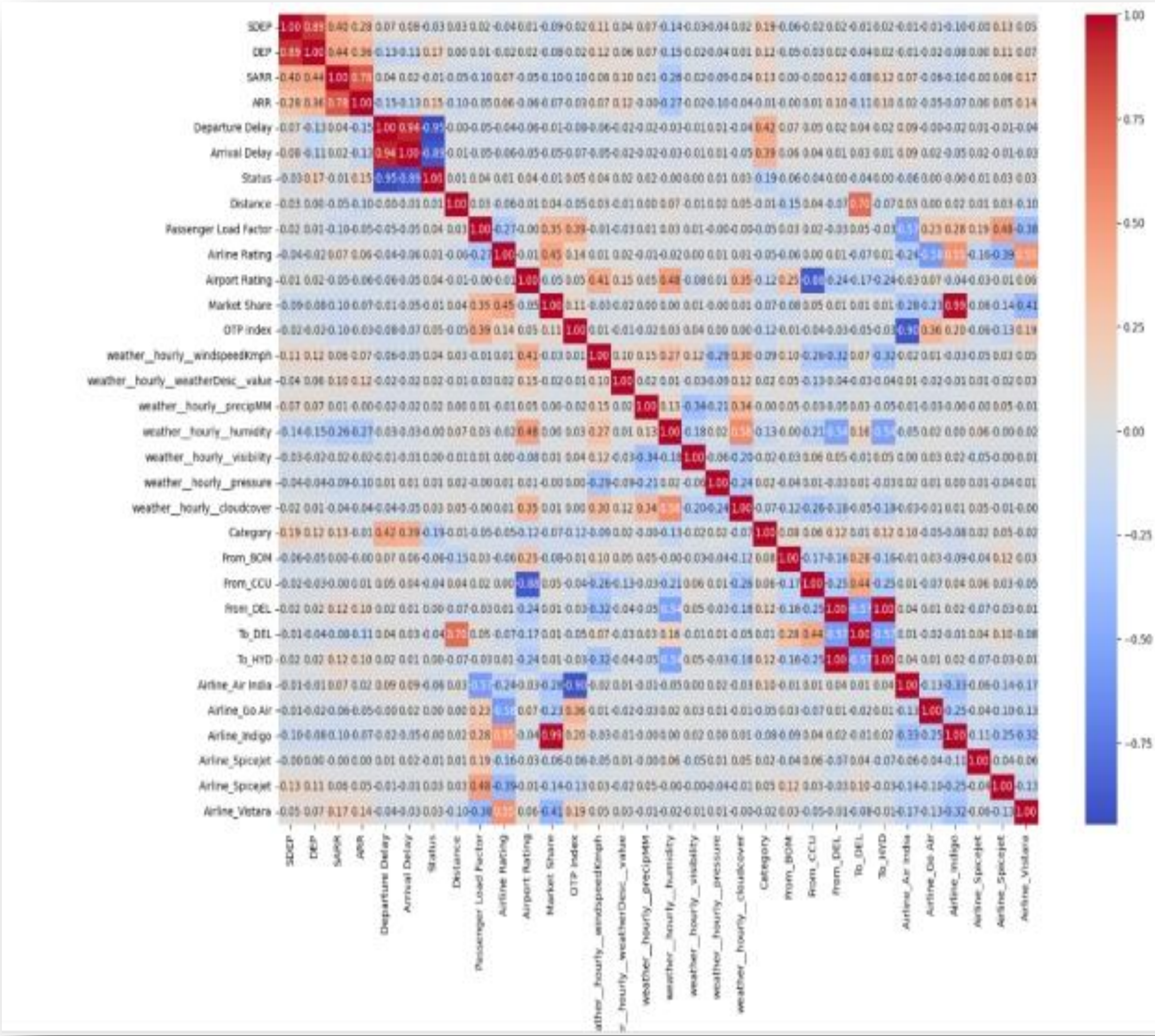
There are 10,633 distinct instances in this cleaned dataset and 29 features categorized into scheduling, airline performance, airport and meteorological.The target variable is a five level delay category according to the magnitude of the arrival delay, on-time, minor, moderate, significant, and severe.There is a significant class imbalance: 51.0% of flights are on time, and severe delays only 8.5% .The average delay in arriving at the hospital is 27.62 minutes, which led to the need to use SMOTE to create synthetic minority class instances to train the model.

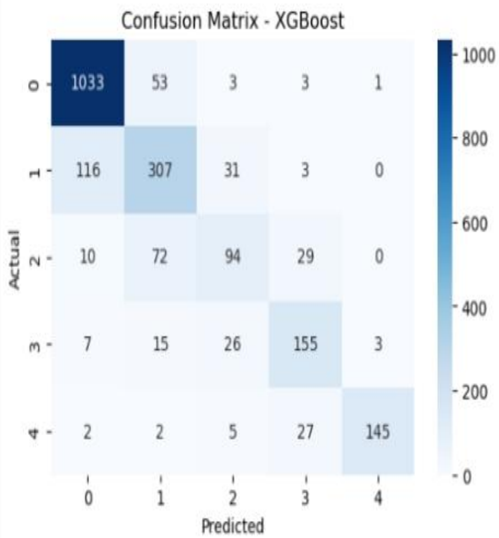
The preprocessing steps involved duplicate removal, the imputing of missing values with mean for continuous features and mode for categorical features, one-hot encoding for origin, destination and airline, and label encoding for textual weather descriptions.This resulted in the removal of raw temporal columns, and the encoding of timing variables was kept to maintain operational patterns.The data were divided into 80/20 stratified sampling with 8,506 training data and 2,127 test data.

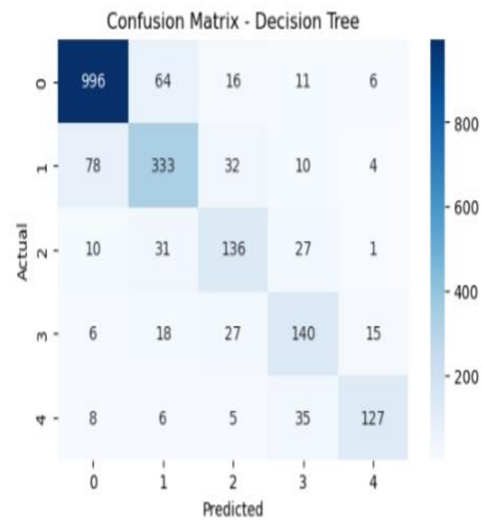
Three models were tested. Random Forest with 100 trees and SMOTE on train split .Decision Tree was pruned to be a baseline and allowed to grow without pruning.With default learning and regularization parameters, and no SMOTE applied, XGBoost employed 300 trees to test its native ability to handle imbalance.Evaluation was predominantly based on classification accuracy, and precision, recall, F1-score and confusion matrix analysis were also provided.

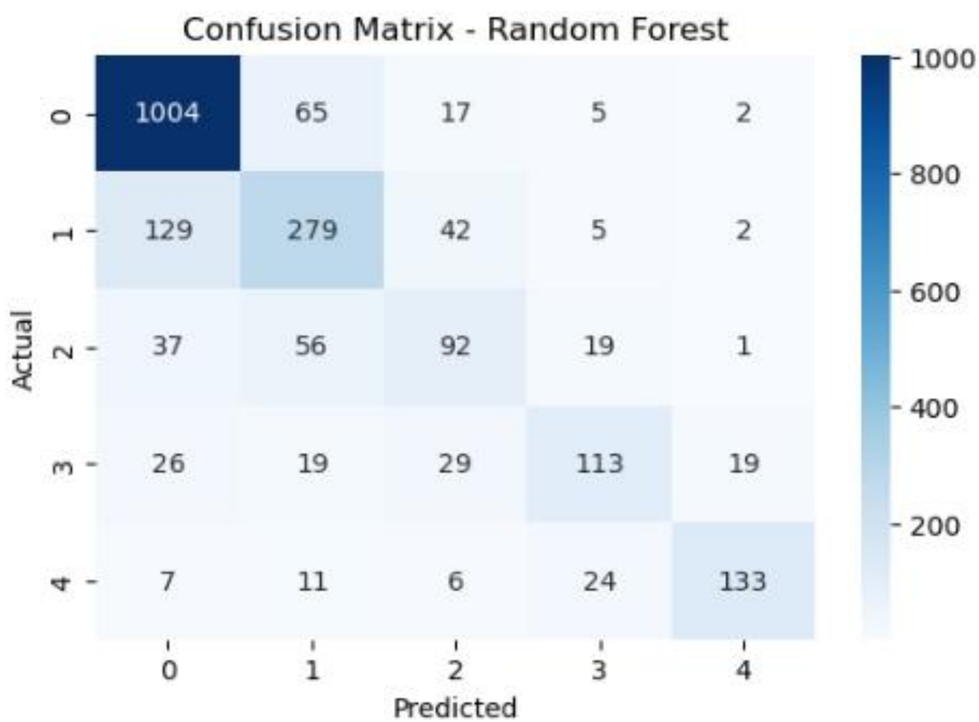
Results and Discussion

This is an important inconsistency that should be explicitly acknowledged in an academic draft in the results section. One part says the accuracy of Decision Tree was 80.86% and XGBoost 80.95% with Random Forest being at 75.68% meaning that the numerical best would be XGBoost.All abstract, summary table, discussion and conclusion, however, state that the Random Forest with SMOTE had the best or highest generalization performance.There are internal contradictions that indicate that the actual reporting was incorrect or that there was a difference in what is being reported and what is being interpreted.


SDEP
DEP
SARR
ARR
Departure Delay
Arrival Delay
Status
Distance
Passenger Load Factor
Airline Rating
Airport Rating
Market Share
OTP Index
weather_hourly_windspeedKmph
weather_hourly_weatherDesc_value
weather_hourly_precipMM
weather_hourly_humidity
weather_hourly_visibility
weather_hourly_pressure
weather_hourly_cloudcover
Category
From_BOM
From_CCU
From_DEL
To_DEL
To_HYD
Airline_Air India
Airline_Go Air
Airline_Indigo
Airline_Spicejet
Airline_Spicejet
Airline_Vistara
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75

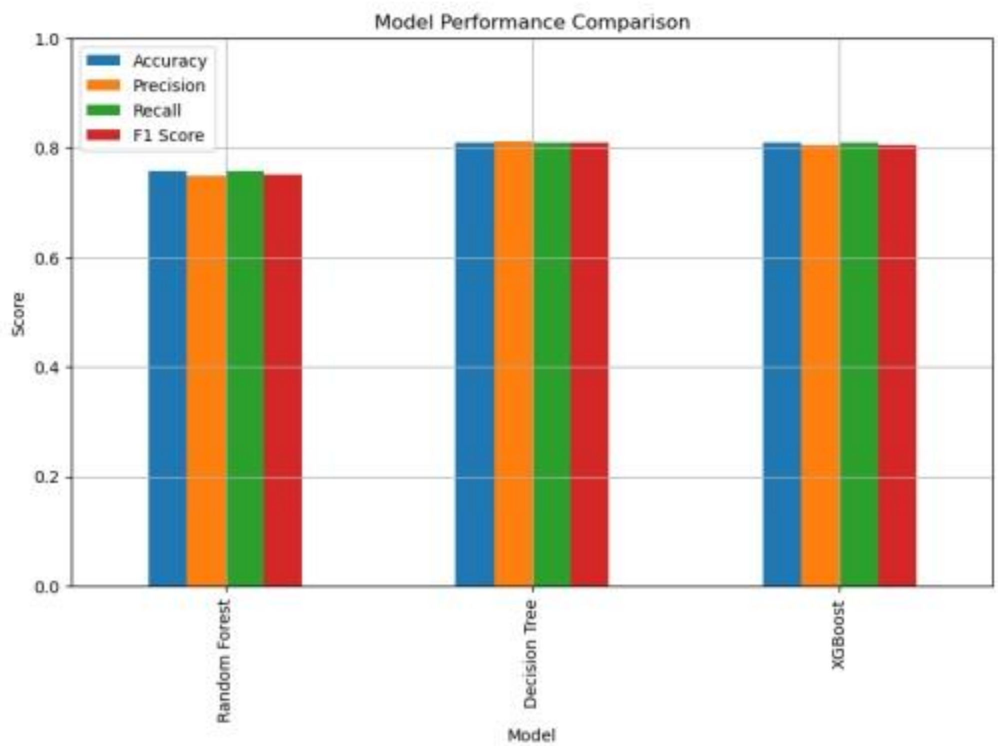

Confusion Matrix - XGBoost
Actual
Predicted
0 1 2 3 4
1033 53 3 3 1
116 307 31 3 0
10 72 94 29 0
7 15 26 155 3
2 2 5 27 145
1000 800 600 400 200 0


Confusion Matrix - Decision Tree
Actual
Predicted
0 1 2 3 4
996 64 16 11 6
78 333 32 10 4
10 31 136 27 1
6 18 27 140 15
8 6 5 35 127
800 600 400 200


Confusion Matrix - Random Forest
Actual
Predicted
0 1 2 3 4
1004 65 17 5 2
129 279 42 5 2
37 56 92 19 1
26 19 29 113 19
7 11 6 24 133
1000 800 600 400 200

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 156-164 March 2026-Sept 2026 www.csjournals.com

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	0.756769	0.748680	0.756769	0.750447
1	Decision Tree	0.808590	0.810827	0.808590	0.809106
2	XGBoost	0.809524	0.804677	0.809524	0.805021



The discussion, however, provides some reasonable interpretations of model behaviour. The report suggests that Random Forest has the advantage of having diversity in the ensemble and SMOTE has the advantage of balancing training data to enhance the recognition of minority classes.It further states that XGBoost can beat the competitors due to the sequential boosting and regularization, which aids it to control complex and imbalanced patterns.The Decision Tree

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 156-164 March 2026-Sept 2026 www.csjournals.com

baseline is said to be more susceptible to overfitting when trained to complete maturity on high dimensional, imbalanced multi-class data ..

The design of the feature is one of the best aspects of the study. Airlines' performance indicators like OTP Index, Airline Rating, and Market Share are considered significant predictive factors in addition to the weather factors like wind speed, visibility, humidity, precipitation, and cloud cover.There are also structural differences between flight corridors and airport congestion patterns that are captured by the route-level variables.The report also highlights the difficulty in distinguishing between classes of moderate and significant delays, as well as between nearby classes, a typical problem in multi-class delay prediction.

Model or Study	**Setting**	**Reported Outcome**	**Main Takeaway**
Gui et al. (2020)	China, ADS-B, binary classification	90.2% accuracy	Random Forest strong on limited data .
Dai (2024)	JFK, 24,986 flights, hybrid model	97.2% accuracy	Optimization and clustering improved performance .
Zhong et al. .	US domestic, spatiotemporal binary task	94.6% accuracy, MAE 5.2 min	Spatial-temporal modeling improves prediction .
Attached study	India, 10,633 flights, five classes	Best model reported inconsistently	Multi-class Indian setting is the key contribution .

Limitations and Conclusion

The study attached has a number of obvious shortcomings. It is limited to five routes, and approximately 11 months of observations, so it is not generalizable to the Indian aviation network or years .The report itself points out that seasonality, like monsoon disruptions, winter fog, etc., could lead to temporal nonstationarity that the model would not fully account for.Some of its drawbacks are that it only uses accuracy as the metric for a class-imbalanced multi-class problem, and class-wise recall and macro-averaged metrics would give a fuller picture.The most significant methodological issue is the lack of consistency across reported best model performances that should be addressed prior to submission. Despite these limitations, the attached report is useful as it addresses the problem of prediction of flight delays in domestic flights as a multi-class machine learning problem with integration of operational and weather data.The core of its findings are that ensemble methods, imbalance handling and multi-source features play a pivotal role in practical delay prediction.The report suggests sequence modeling using LSTM, network level modeling using GCN, real time ADS-B and weather streams, feature selection, clustering and multi year data expansion for future work.

References

[1] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. IEEE Transactions on Vehicular Technology, 69(1), 140-150.

[2] Dai, M. (2024). A hybrid machine learning-based model for predicting flight delay through aviation big data. Scientific Reports, 14, 4603. https://doi.org/10.1038/s41598-024-55217-z

[3] Zhong, Q., Yu, Y., Huang, Y., & Zhang, T. .Prediction and optimization of civil aviation flight delays based on machine learning algorithms. International Journal of Computational Intelligence Systems, 18, 189. https://doi.org/10.1007/s44196-025-00932-2

[4] Prudhvi Raj, S., Abhishek, P., Pavan Sai, P., Akshitha, P., & Nandini, P. (2023). Flight delay prediction based on aviation big data and machine learning. International Journal of Scientific Research in Engineering and Management (IJSREM), 7(12), 1-16.

[5] Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. IEEE Digital Avionics Systems Conference, 1-6.

[6] Moreira, L., Dantas, C., Oliveira, L., Soares, J., & Ogasawara, E. (2018). On evaluating data preprocessing methods for machine learning models for flight delays. IEEE International Joint Conference on Neural Networks, 1-8.

[7] Li, Q., & Jing, R. (2022). Flight delay prediction from spatial and temporal perspective. Expert Systems with Applications, 205, 117662.

[8] Li, Q., Guan, X., & Liu, J. (2023). A CNN-LSTM framework for flight delay prediction. Expert Systems with Applications, 227, 120287.

[9] Khan, R., Akbar, S., & Zahed, T. A. (2022). Flight delay prediction based on gradient boosting ensemble techniques. IEEE International Conference on Open Source Systems and Technologies (ICOSST), 1-5.

[10] Hatipoglu, I., Tosun, O., & Tosun, N. (2022). Flight delay prediction based with machine learning. LogForum, 18(1).

[11] Xu, H., Shi, J., & Wang, T. (2022). Departure flight delay prediction model based on deep fully connected neural network. Journal of Computer Applications, 42(10), 3283.

[12] Paramita, C., Supriyanto, C., Syarifuddin, L. A., & Rafrastara, F. A. (2022). The use of cluster computing and random forest algorithm for flight delay prediction. International Journal of Computer Science and Information Security, 20(2).

[13] Wu, Y., Yang, H., Lin, Y., & Liu, H. (2023). Spatiotemporal propagation learning for network-wide flight delay prediction. IEEE Transactions on Knowledge and Data Engineering, 1-1.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[15] Hao, L., Hansen, M., Zhang, Y., & Post, J. (2014). New York, New York: Two ways of estimating the delay impact of New York airports. Transportation Research Part E, 70, 245-260.