JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A STUDY ON AIR TRAFFIC DATA FOR AIR TRAFFIC FLOW ANALYSIS UTILIZING MACHINE LEARNING AND REGRESSION APPROACHES

Keerthana .C shekhar
Department Information of Science Engineering
R. V. College of Engineering,
Bangalore, India
keerthanacs.is16@rvce.edu.in

Priya D.
Department. Information of Science Engineering
R. V. College of Engineering,
Bangalore, India
priyad@rvce.edu.in

ABSTRACT

The science of managing air traffic flow necessitates extensive study into the topic of traffic flow prediction. There is currently a lack of attention paid to the impact of topological aviation linkages on traffic flow at a single airport in the existing literature. The Air Force's escalating toil Transportation networks are becoming more complicated due to the increasing use of unmanned aerial vehicles and general aviation aircraft (GAAs). It is now possible to follow and monitor aerial vehicles in real time and with high accuracy thanks to ADS-B technology, which has been developed to the highest level. Machine learning techniques can be used to count and estimate how much air traffic will be moving between cities using a dataset that has already been analyzed and information that has been extracted and mapped to the routes.

Keywords: Air traffic flow management, Machine Learning, Regression, surveillance-broadcast (ADS-B), Air Traffic analysis etc.

INTRODUCTION

Increasing numbers of people are turning to air travel as their primary form of transportation because of advancements in civil aviation. Traveling by plane is becoming more and more difficult as more people take to the skies. Flight surveillance systems will face new challenges as a result of this increased demand on air traffic control (ATC). In order to assist the flight control department in timely and suitable analyses of whether aircraft traffic is approaching the upper limit, ATFM's primary purpose is to aid in the utilization of available airspace capacity.. The primary goal of ATFM Having precise and trustworthy data on air route flow is critical for the ATFM system's future data processing, data analysis, and visualization needs. These out-of-date radar identification devices have been utilized by air traffic controllers and allied organizations for decades since they're cheap to operate but fail miserably at doing their job. An ADS-B message stream is examined and used to derive air traffic flow data in this article. Long-term support vector regression (LSVR), short-term and long-term storage of information (LSTM), and long-term memory (LTM) are all examples of machine learning approaches that can be used to predict flow. When evaluated on a large dataset, the LSTM-based model beat the SVR-based model. Both models are capable of accurately forecasting aviation traffic.[1]

- **Air Traffic Flow Management:**

Aviation recommends deploying ATFM in airspaces when the declared capacity of the air traffic control services is or is expected to be exceeded. The International Civil Aviation Organization (ICAO) has recognized the need of ensuring that all air navigation service providers (ANSPs) have a common understanding of what ATFM is all about. ICAO's definition of "ATFM," which follows, specifies that "ATC" must be used to the fullest degree possible and that "traffic volume" must be in accordance with capacities indicated by the appropriate air traffic services (ATS). On a diverse array of time scales, it is feasible to strike a balance between the amount of traffic and the available capacity (ATFM phases). In the event that the amount of traffic demand constantly exceeds the capacity of the ATC, the ANSP ought to make preparations to increase capacity.

Because there are fewer and fewer instances of good operating conditions, pre-tactical and tactical ATFM processes are going to have to be implemented. When capacity is reduced, airports and airspaces that are already operating at or near their maximum capacity will have a demand-capacity imbalance. As a result of increasingly accurate demand forecasts, the application of ATFM has grown widespread in many regions of the world.[2]

PROBLEM STATEMENT:

Understanding that air traffic flow management (ATFM) will be vital in the future crowded air traffic is essential. The ATFM has had to deal with an upsurge in demand for unmanned aerial vehicles and light aircraft. Overcrowding in traffic is a big challenge in this project (funding problem is also included but at a secondary level). From air travel to freight carriers, there can be a congestion. As a result, additional issues have arisen as a result of the continued development of roadways, runways, and so on

REVIEW OF LITERATURE:

(Gui, n.d.) [3] Air traffic flow management, often known as ATFM, is a subject that becomes of the utmost significance whenever there is a significant amount of congestion in the air. As a result of an increase in the number of unmanned aerial vehicles and aircraft used for general aviation, the resources available to the ATFM are being pushed to their absolute limit. It is now possible to track and keep an eye on aerial aircraft in real time while also maintaining a high level of precision thanks to the ADS-B (automatically dependent surveillance-broadcast) method. This makes it possible for an advanced ATFM design, which in turn makes it possible for a more advanced ATFM design. A Big Data platform for aviation is developed in this piece of writing with the support of ADS-B ground stations and ADS-B communications. By utilizing two distinct machine learning methods, it is possible to count as well as make predictions regarding the flow of air traffic that occurs between cities. After conducting an analysis of the newly formed dataset and mapping the information that was retrieved to the routes, these approaches are put into action. The performance of traffic flow prediction models that are built on long short-term memory (LSTM) can be improved when unexpected characteristics of traffic control are taken into consideration in the experimental results. This becomes abundantly clear if an analysis of the data is performed.

(Sun, n.d.) [4] The Air Traffic and Flight Management Authority (ATFM) is now under a significant amount of pressure, and this load is only likely to increase as the market for unmanned aerial vehicles and general aviation aircraft continues to expand. ADS-B is an upgraded type of automatic dependent surveillance broadcast that allows for the detection and tracking of aerial vehicles in real time and with pinpoint accuracy. This is made possible by the technology. The technology known as ADS-B makes this a real possibility. Utilizing the ADS-B ground stations and the communications capabilities of the ADS-B system will be the foundation upon which we will construct an aviation Big Data platform as the first stage.

(Dhariwal, 2020) [5] Human air traffic controllers are still used in today's aviation practice to keep track of and direct the many planes passing through their designated airspace zone.

(Jurado, 2022) [6] It is becoming increasingly difficult for the ATM system to keep up with the increased demand for aircraft operations. As a result, a number of initiatives have been launched to create new technology to enhance the

ATM system's capability. Three-dimensional traffic density estimation and analysis are now possible in at least one area of air traffic control.

(Aghdam, 2021) [7] As a result of the complication and sensitivity of the situation, this idea is now being examined by a number of stakeholders, including passengers, airlines, regulatory bodies, and others. The challenges posed by air traffic management (ATM) have been surmounted with the assistance of neural networks and other data mining technologies, which has delayed the elimination of the majority of the issues that have been encountered up until this point. In addition to this, statistical and fuzzy techniques have been utilized.

(Zhu, 2022) [8] Improved algorithms based on data exercise and a mathematical model of intelligent motion coordinates were developed in this work to adapt to modern aviation's intelligent traffic flow management. These researchers investigate how the equation of motion for six-degrees-of-freedom motion in the coordinate system used in civil aviation can be used to describe a rigid body that has a certain mass and distribution of that mass.

(Yang, 2022) [9] In the realm of air traffic control, predicting how busy an airport will be is a critical part of the investigation. There is a lack of study in this area that takes into account the topological airport network's impact on overall traffic flow predictions.

(Shuaida Xiang, 2020) [10] By using a non-iterative prediction model of air traffic flow time series based on the ELM algorithm, a new understanding of air traffic flow has been gained. An advantage of using SVR over iterative methods is that it eliminates the necessity for data iteration.

(Banavar Sridhar, 2020) [11] Air Traffic Management (ATM) is growing more and more dependent on machine learning techniques (MLT) (ATM). Cloud computing, open source software, and MLT's breakthroughs in automation, consumer behavior, and finance, all of which rely on massive databases, have piqued the curiosity of many individuals recently. Many benefits and drawbacks are identified in this paper's examination of the current level of machine learning application in aircraft operations. Models that provide operational information have been used in the study of aviation operations for many years now.

(Opitz, n.d.) [12] Modern surveillance networks can provide trajectories over a wide area or around the world for a wide variety of ships and planes. AIS and ADS-B have long been the most widely used air and maritime surveillance systems in the United States, respectively. The two systems complement one other effectively. Emerging technologies like the Internet of Things (IoT), digitalization, automobiles, smart cities, and decentralization will enable ground monitoring in the near future (block chain).

(Chen, 2022) [13] We formed elliptical dots made of cobalt and perm alloy using a process called self-aligned shadow deposition. The dots had a thickness of 25 nm, a length of 1 mm, and a width of 220 nm. Brillouin prisms have been used to investigate the link between temperature-induced magnetic Eigen mode frequency and external magnetic field intensity.

PROPOSED SYSTEM:

- **Dataset:** There is a lot of data here about air traffic, so we can use it as we see fit. This strategy makes use of data from aviation traffic. Data is available on: [14]

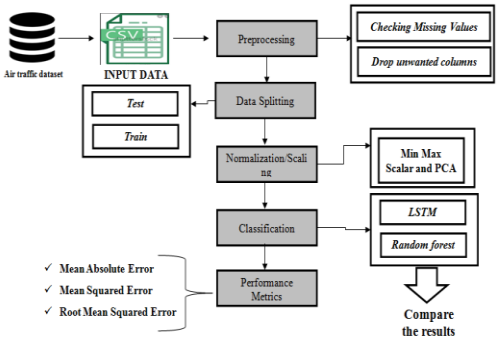


Figure 1: Process Diagram

Here are the top 5 rows of a dataset:

Figure 2: Rows of 5 dataset

Descriptive statistics of dataset:

Statistics that summarize a dataset's central tendency, dispersion and form (excluding Nan values) are known as descriptive statistics.

	Activity Period	Passenger Count	Adjusted Passenger Count	Year
count	15007.000000	15007.000000	15007.000000	15007.000000
mean	201045.073366	29240.521090	29331.917105	2010.385220
std	313.336196	58319.509284	58284.182219	3.137589
min	200507.000000	1.000000	1.000000	2005.000000
25%	200803.000000	5373.500000	5495.500000	2008.000000
50%	201011.000000	9210.000000	9354.000000	2010.000000
75%	201308.000000	21158.500000	21182.000000	2013.000000
max	201603.000000	659837.000000	659837.000000	2016.000000

Figure 3: Statistics Description

- **Python modules:**

Import the necessary library. We typically utilize Numpy, pandas, matplotlib, sea born, and sklearn for easy data processing.

"NumPy":

NumPy, a Python package for array manipulation, can be used. Using this software's linear algebra and matrices capabilities is a cinch. NumPy was created in 2005 by Travis Oliphant. There are no restrictions on using it because it is an open-source project.

Pandas:

It's easy and intuitive to work with relational or labelled data with Pandas, an open-source library. In order to manipulate numerical data or time series, several data structures and operations are provided. The NumPy library serves as a foundation for this library. Pandas is lightning-fast, and its customers may expect top-notch output.

Matplotlib:

For 2D representations of arrays, Python's Matplotlib is a wonderful visualisation package to use. Using NumPy arrays, Matplotlib is a cross-platform data visualization library for the SciPy stack. In 2002, John Hunter became the first person to use it.

Via visualizations, we may access vast volumes of data in an easily digestible form. This is one of the main advantages of visualizations. There are a variety of plots to choose from in Matplotlib.

Sklearn:

Python's machine learning capabilities are greatly enhanced with the Scikit-learn library (Sklearn). Approaches such as classification, regression, clustering, and dimensionality reduction are examples of some of the machine learning and statistical modeling methods that can be implemented. All of these can be accessed with the use of a Python consistency interface. The majority of this toolkit's code is written in Python, and its core components include NumPy, SciPy, and Matplotlib.

Pre-processing: Module for pre-processing the dataset in order to remove unnecessary data and extract information. Cleaning up the dataset by removing any and all null values.

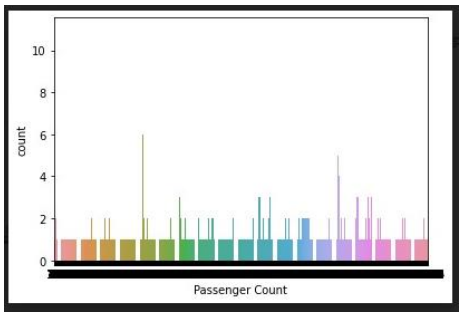
Label encoding:

- When working with datasets in machine learning, it is common to have many labels in a single column, or even multiple columns. These things can be labeled with either words or numbers. To make it easier to interpret, a lot of the training data is labeled with words.

It is called label encoding to encode the labels into a numeric form that can be read by a machine. These labels can then be used by machine learning algorithms to make better judgements. Pre-processing the structured dataset is an important step in supervised learning.

Another prominent method for dealing with categorical data is One-Hot Encoding. Simply by counting the number of unique values in a categorical feature, it generates new features. A feature will be introduced for each and every one of the category's specialties. Create dummy variables with One-Hot Encoding.

Encoding all features using label encoder.



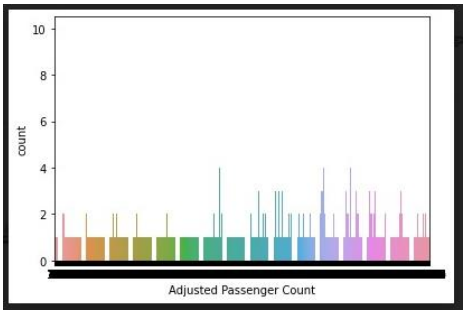


Figure 4: Count-Plot of total passenger

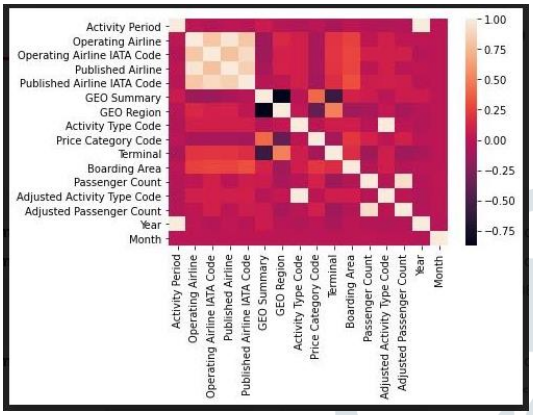


Figure 5: Correlation matrix of all features

Data splitting:

In order for machine learning to take place, data is needed. The algorithm's performance must be evaluated using both training and test data. 70% of the input dataset was defined as training data, while 30% was designated as testing data in our approach. In cross-validation, data splitting is the process of partitioning a dataset into two separate subsets. A subset of the data is used to build predictive models, and a different subset is used to assess the models' performance. Separating the data used to train and test data mining models is a necessary step in the evaluation process. When a data set is separated into two groups, one for training and the other for testing, the vast majority of the data is used for training and the minuscule amount for testing.

ALGORITHMS:

- **RNN**

Adding loops to feed forward NN creates RNN a more advanced version of RNN. It is possible for an RNN to learn from data by analyzing a set of time-correlated samples. Hidden nodes are used to add parameters to the network and subsequently release the state depending on the values of the input. For example, RNN activates the states based on network events, resulting in better performance. The bias and weight of a typical RNN node are both a single value. The gated recurrent unit and the long-term memory are used to evaluate the RNN. With this network design, each input's time step is used to produce an output with the same time step. In contrast to the single bias and weight of a standard RNN node, the LSTM node has four biases or weights:

- **Forget gate layer**
- **Input gate layer**
- **Output gate layer**
- **State gate layer**

1. **Random Forest Regressor:**

We will be utilizing the Random Forest Regressor function, which is a component of the sklearn package, during the training phase for our random forest regression model. This function can be found in the sklearn library. The documentation for the Random Forest Regressor presents a wide choice of parameters for our model that can be selected. These parameters are available for selection. The table that follows provides an overview of some of the more significant parameters:

- **n_estimators** — the number of decision trees that will be used in the model that you will be running.

- **criterion** — With the help of this variable, you will be able to choose the criterion or loss

- function that will be used to decide the model outputs. Loss can be used as a selection criteria.

- **Functions:** Including the mean squared error, often known as MSE, and the mean absolute error (MAE). The default value is MSE.

- **max_depth** — this sets the maximum possible depth of each tree

- **max_features** — the maximum number of features the model will consider when determining a split
- **bootstrap** — the default value for this is True, meaning the model follows bootstrapping principles (defined earlier)
- **max_samples** — This parameter assumes bootstrapping is set to True, if not, this parameter doesn't apply. In the case of True, this value set the largest size of each sample for each tree.Other important parameters are min samples **split, min leaf, n_jobs**, and others that can be read in the sklearn's Random Forest Regressor documentation.

2. LSTM:

For analyzing and anticipating events in time series with long inter- and delay intervals, the LSTM is a particularly effective recurrent neural network (RNN). This neural network is immune to the gradient vanishing problem, unlike recurrent neural networks, which are. When it comes to interpreting genuine language, picking out specific targets, and identifying sounds, the LSTM really shines.

LSTM cells have a gated neuron feature that is distinct. For temporal sequence prediction problems, the LSTM's structure allows it to store large short-term memories, making it perfect.

1. To create the LSTM network, set the imp units, lstm units, op units, and optimizer (L) to 3 to normalize the dataset (Di).
2. Decide on the training window's dimensions (tw) and then set up Di to reflect that.
3. follow these directions, taking into consideration the number of epochs and batch size
4. Educate the People Around You (L)
5. Use L for Predictions
6. Compute a loss function by following the loss function.

RESULTS:

In this results, here we have discussed the various parameters and their accuracy which we have find out during the code . we have achieved good results with LSTM & Random forest Regression Using best parameters which described below the table.

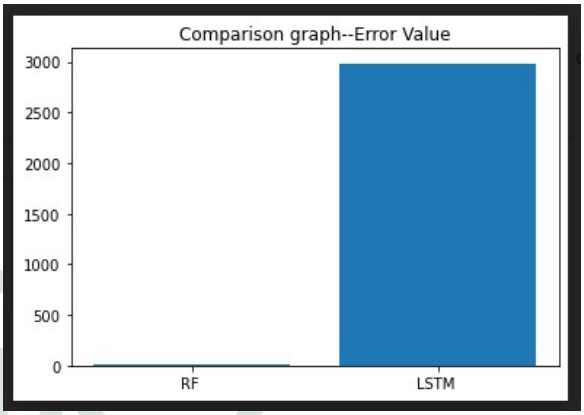


Figure 6: Error value Comparison

TABLE-1 Accuracy Table

	MODEL	MAE
0	**Random Forest Regression**	**4.7832**
1	LSTM	3000

CONCLUSION:

It is one of the functions of flow management systems to control aircraft locations and flight plans at critical moments. An additional duty is to monitor the flow in key areas. The introduction and processing of dynamic data include such things as air traffic control radar, navigation telegrams, and flight data. Since the flow management application system relies on the transmission of information, it is critical that the problem of information transmission be examined as soon as possible. Using a geographic data model, an intelligent management model is created for civil aviation traffic flow. As a result, future civil aviation traffic flow management models can be guided by the model. Due to its success in simulation, civil aviation can now exploit the air traffic management model's impacts on the ground. Because this model was designed for civil aviation, it is

dependent on geographic data. This new technique will increase the controller's coordination time because it needs releasing the point sequence time for each aircraft. In the future, air traffic management will be used as a tactic by the controller. The actual performance of the control can be affected by a wide range of variables. Weather, route, and aircraft characteristics are just a few examples of factors that can affect a flight controller's ability to deploy. There is a need for more research into the modeling of such complicated components.

REFERENCES

[1] D. Munot, "Air Traffic Flow Analysis Based on Aviation Big Data using Machine Learning," *IJERT*, vol. 08, no. 11, 2021, [Online]. Available:

[2] L. S.p.a., "Implementing Air Traffic Flow Management and Collaborative Decision Making", [Online]. Available:

[3] G. Gui, "Machine Learning Aided Air Traffic Flow Analysis Based on Aviation Big Data," *scinapse*, [Online]. Available:

[4] J. S. Sun, "Air Traffic Flow Analysis Based on Aviation Big Data", [Online]. Available:

[5] P. Dhariwal, "Air Traffic Control using Big Data Analysis and Machine Learning," *SSRN*, 2020, [Online]. Available:

[6] R. D.-A. Jurado, "3-D Prediction of Air Traffic Density in Atc Sectors Based on Machine Learning Models," *SSRN*, 2022, [Online]. Available:

[7] M. Y. Aghdam, "Optimization of air traffic management efficiency based on deep learning enriched by the long short-term memory (LSTM) and extreme learning machine (ELM)," *springer link*, 2021, [Online]. Available:

[8] X. Zhu, "Computational Intelligence Systems for Vehicular Ad Hoc Networks," *HINDAWI*, 2022, [Online]. Available:

[9] H. Yang, "A Deep Learning Approach for Short-Term Airport Traffic Flow Prediction," 2022, [Online]. Available: [10] J. G. Shuaida Xiang, "Research on Air Traffic Flow Forecast Based on ELM Non-Iterative Algorithm," *springer link*, 2020, [Online]. Available: https://link.springer.com/article/10.1007/s11036-020-01679-0

[11] Banavar Sridhar, "Lessons Learned in the Application of Machine Learning Techniques to Air Traffic Management," *ARC*, 2020, [Online]. Available:

[12] F. Opitz, "Data Analytics and Machine Learning based on Trajectories", [Online]. Available:

[13] D. Chen, "Novel Machine Learning for Big Data Analytics in Intelligent Support Information Management Systems," *ACM Trans. Manag. Inf. Syst.*, vol. 13, no. 01, 2022, [Online]. Available:

[14] "link of dataset", [Online]. Available: