FLIGHT DATA OF AIRPLANE FOR WIND FORECASTING

Astha Sharma
Dept. of Computer Science
University of New Orleans
New Orleans, USA
asharma6@uno.edu

Md Tamjidul Hoque*
Dept. of Computer Science &
Canizaro/Livingston Gulf States
Center for Environmental Informatics,
University of New Orleans
New Orleans, USA
thoque@uno.edu
*Corresponding author

Elias Ioup
Center for Geospatial Sciences
Naval Research Laboratory
Stennis Space Center, MS, USA
elias.ioup@nrlssc.navy.mil

Mahdi Abdelguerfi
Dept. of Computer Science &
Canizaro/Livingston Gulf States
Center for Environmental Informatics,
University of New Orleans
New Orleans, USA
mabdelgu@uno.edu

IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9323718

ABSTRACT

Understanding and predicting weather behavior is vital for informing pilots about changing flight conditions. This paper presents a new approach towards forecasting one component of weather information, wind speed, from data captured by airplanes in flight. We compare two datasets for prediction suitability, and a collinearity analysis between these datasets reveals a better model performance with smaller test error with one of them. We then apply machine learning and a genetic algorithm to process this data further and arrive at a competitive error rate. Finally, we create an offline software for wind prediction using the best performing classifier.

Index Terms— Machine Learning, Weather Forecasting, Genetic Algorithm, kNN imputation, Linear Regression, Extreme Gradient Boosting, Sliding Window

1. INTRODUCTION

Air transport is the most popular form of transportation for long-distance travel in the US [1] and is a good resource for data. The data collected from an airplane during flight includes information about the aircraft's position as well as meteorological and environmental measurements. These data are of great value for analyzing and predicting various natural conditions, like turbulence, which can assist the pilot and flight crew in making decisions about the future and avoid any possible mishap. In addition, these data can be used to monitor the flight progress and provide improved arrival and departure estimates to passengers.

However, there are still limitations with the wind forecasts used in flight planning. Most of the available wind forecasts for US flights are based on the Wind Aloft Program from the US National Oceanic and Atmospheric Administration (NOAA) [2]. This program collects data via the recurring release of weather balloons and radar. The forecasts models then use linear interpolation to combine information from the available measurements [3]. However, there is evidence that this NOAA data may not be sufficient for making accurate predictions [4].

The NOAA data come from weather models that are fed with measurements from ground stations along with data from weather balloons, satellites, and other instruments. The wind data are available at different altitudes ranging from 6,000 to 53,000 feet. They include information from 9 different regions of America: Northeast, Southeast, Northcentral, South Central, Rocky Mountain, Pacific Coast, Alaska, Hawaii, and West Pacific [5].

Alternatively, the Atmospheric Carbon and Transport-America (ACT-America) campaign from NASA covers 4 seasons and 3 regions of the central and eastern United States and is based heavily on direct in-flight measurements. Using a variety of instruments, airplanes record their positional data as well as meteorological and environmental readings across a variety of surface and atmospheric conditions. The dataset includes 118 days of data with a temporal resolution of 1 second. There is a total of 34 different features, including latitude, longitude, altitude, ground speed, air temperature, and wind speed and direction [6].

Our objective is to choose the quality dataset from the set of above explained two groups, as a step forward in the direction of accurately forecasting the wind speed.

In this paper, we describe the results of model performances based on linear regression. Then we further cleanse the dataset and apply machine learning algorithms to derive useful information about the wind speed. We then calculate the wind velocity 10 minutes ahead and finally evaluate our performance and compare the results with some related works.

2. BACKGROUND AND RELATED WORK

A significant effort was put into understanding the relevant atmospheric phenomena and the methods employed by NOAA and NASA to collect and publish data for their respective programs. Historically, audio reports from pilots have been used in weather models for over half a century. However, the past 20 years have seen the trend of employing data from commercial aircraft to produce more accurate predictive models [7]. Efforts have also been made to develop machine learning models for the predictive analysis of airplane data to improve upon the existing NOAA forecasts. One of the major inspirations for this research is a similar project from Microsoft [3], in which they conduct a comparative analysis of possible approaches for wind prediction. We tried our best to find and use the identical dataset for a fair comparison. However, it was not publicly available. Therefore, we used NASA's dataset, which was closely comparable and also open source. The results obtained from different approaches mentioned in the Microsoft Project are described in Table I.

TABLE I. RESULTS FROM THE MICROSOFT PAPER

Approach	RMS Error
NOAA data	51.53
Gaussian Process Estimate	50.93
Gaussian Process + Airplane Data	43.66

3. METHODOLOGY

3.1. Dataset – NOAA data vs. Airplane data

Two different sets of data were considered in our research: NOAA's Wind Aloft Program and NASA's ACT-America project. In order to determine the superiority of one over the other, a comparative analysis was performed using the correlation coefficients of both dataset's features with respect to wind speed. We applied Pearson's Correlation Coefficient as a matrix to find the Root Mean Square Error (RMSE) for training and testing cases. The dataset with the minimum test error was selected for further analysis.

In the case of the NASA dataset, the original set of 34 features was reduced down to 6 by selecting only the features most strongly correlated to wind speed, wiz., Mach Number, Ground Speed, Track Angle, Drift Angle, Static Air Temperature and Wind Direction. For the NOAA data, since there are only a few features provided (direction, temperature, latitude, longitude, and altitude), all were included. The NOAA data was trained using 10 fold cross-validation at all 170 different site locations at 30,000 feet height. A sample data of the first 3 days from the airplane data was also trained using 10 FCV. The RMSE, using Linear Regression, for both sources are given in Table II. Since the airplane data look much promising and have more features and observations than the NOAA data, our research thereafter focused only on airplane data.

TABLE II. RMSE FOR TRAINING AND TEST DATASETS

Source	#Observation	RMSE
Wind Aloft	170	20.0503
ACT-America	45126	31.9042

3.2. Data Analysis and Cleansing

We began preprocessing our airplane dataset by sampling data from 5 days (selected randomly) having a reasonable number of input rows. The random sampling approach was employed because it gives an equal probability of selection for each element in the full dataset, thereby reducing the probability of biased results.

A deeper analysis of the available dataset revealed that more than 72% of the total rows had one or more missing fields. Almost 23–41% of the columns had missing values. This suggested that the available dataset is noisy. Simply dropping the rows with missing values would have been undesirable since it would mean losing a sizeable fraction of the data and potentially decreasing overall accuracy. We thus required a technique that could address gaps in data without losing samples.

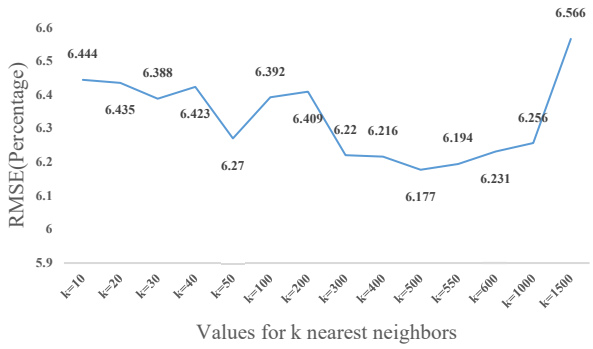

Fig. 1. Corresponding RMSE for different k values.

We adopted the very popular technique of replacing missing values called kNN imputation. Based on the kNN algorithm, kNN imputation is widely known because of its

great performance in machine learning applications. Here, the average of the *k* nearest neighbors at a fixed distance is used as the imputation estimate. We used Euclidean distance as the fixed distance parameter. The value for *k* was decided after computing the root mean squared error (RMSE) for a range of different values, from *k* = 10 to 1,500, as shown in figure 1. Our result showed the minimum RMSE of 6.177 at *k* = 500. Therefore, *k* = 500 was used for imputing the missing values in the dataset.

3.3. Feature Selection and Filtering

Careful feature selection and filtering was a critical step of this research as we wanted to retain only the useful variables that are most related to the wind speed feature. For the feature selection process, we used the powerful genetic algorithm (GA) approach. Although we had calculated the Pearson's Correlation Coefficient in the previous step, we still use GA as it gives a clear idea of feature selection without requiring expertise about the project's domain and inclination. For instance, we can determine whether the Mach number of an airplane is highly correlated to wind speed without necessarily understanding the principles behind that variable.

Two algorithms — Extreme Gradient Boosting (XGBoost) and Linear Regression — were used to analyze the fitness function, and the better algorithm was selected based on the output MSE. Our GA ran for 300 generations for both fitness function algorithms. The following standard parameters were set for our GA: *Population Size* of 20, *Crossover Rate* of 80%, *Mutation Rate* of 5%, and *Elite Rate* of 10%.

At the end of 300 generations, XGBoost gave a total of 6 fittest chromosomes: indicated airspeed, Mach number, track angle, roll angle, potential temperature, and wind direction. Linear regression gave a total of 10: latitude, GPS altitude, ground speed, vertical speed, true heading, pitch angle, static pressure, sun azimuth, partial pressure water vapor, and saturated vapor pressure H2O.

Since XGBoost reduced the number of chromosomes to 6 and obtained a fitness score (28.91) far better than linear regression (42.24), it was the better performer. Therefore, for prediction, we examined only these 6 features plus the wind speed.

3.4. Sliding Window

After filtering and noise reduction, we approached the time series forecasting with the sliding window technique. This approach takes a set of observations sequential in time and creates a model to fit in historical data. The model then predicts future outputs based on factual evidence.

The first step consisted of selecting the sliding window size. We considered a set of window sizes, ranging from 2 to 15. Again, RMSE was the deciding factor. After calculating the RMSE using two regression algorithms, XGBoost, and linear regression, we again decided on using XGBoost, in which the least RMSE was obtained at window size 9, figure 2. The least RMSE using linear regression was obtained at window size 10. For consistency, however, and considering the 1-second resolution of the data, we settled on a window size of 9 for both algorithms.

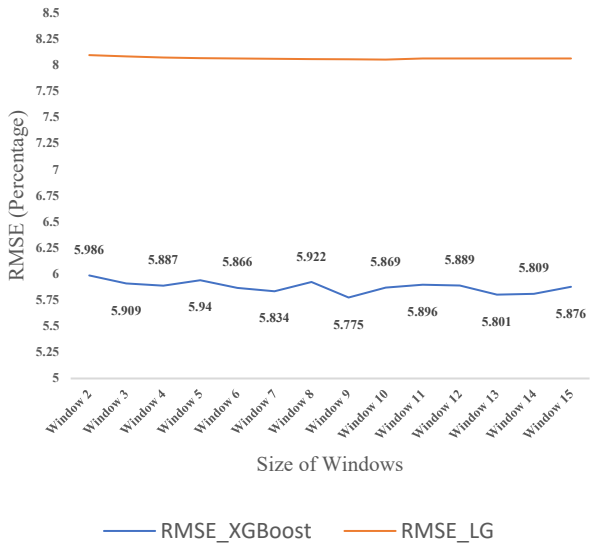

Fig. 2. RMSE values obtained from using XGBoost and Linear Regression.

3.4. Prediction

Through this research, we wanted to predict the wind speed in the near future that can be used by pilots or flight schedulers to make decisions about flights ahead of time. Since our dataset had a resolution of one second, rescaling was needed. So we took an average of every six hundred lines (10 × 60 seconds) in each file from the best performing window size 9 and generated a new dataset, which then had a resolution of ten minutes. The decision to choose a standard period of ten minutes was made as a result of retaining a decent number of points even after scaling. The total data points at the end of rescaling were 2127.

We then used the newly created dataset to train our prediction model with different classifiers. The models were fed with features from one datapoint and mapped to the windspeed of the following datapoint. The classifiers were trained with the wind speed of one step ahead in time. The RMSE in each case is as illustrated in Table III. The best

performing classifier (i.e., Random Forest) was used to create an offline tool[1] to predict wind speed.

TABLE III. MODEL PERFORMANCE ON TRAINING DATA

Classifier	RMSE
Linear Regression	6.538
KNN	10.066
Random Forest	5.540
Bagging	6.236

4. RESULT AND COMPARISON

In this paper, we presented a method of improving dataset quality towards predicting the speed of one of the important atmospheric phenomena, wind. Our approach is valuable and general enough for use in similar cases and datasets. We were able to create a basic forecast model that can predict the wind speed ten minutes ahead of time. We had intended to further the project by creating a basic predictive model for at least 30 minutes ahead. Still, there was a limitation in a number of data points as a result of averaging with better accuracy. Nevertheless, we continue to get competitive results at each step, as demonstrated in Table IV.

TABLE IV. RMSE AT DIFFERENT STAGES OF THE PROJECT

State	Datapoints	RMS Error
Initial State	45126	31.904
After kNN Imputation	78023	6.177
After Sliding Window	1595422	5.775
After Training Model	2127	5.540

Our RMSE obtained at different stages of the project show a significant improvement. Recall that the best RMSE obtained by Microsoft's project discussed earlier was 43.66. This is a good indication that our project is headed in the right direction.

5. CONCLUSIONS AND FUTURE WORK

The number of commercial and military aircraft flying each day is massive and only expected to increase. Applying the in-flight data these aircraft collect to wind speed prediction can be efficient and cost-effective.

This wind model is also closely related to turbulence experienced in flight, which depends on the wind speed at a particular position and altitude. We can, therefore, extend this project to create a predictive model that can be used to optimize flight time based on wind speed. Improving wind speed models also has applications in the creation of more fuel-efficient aircraft designs. Nevertheless, the result of this project is impeccable when compared to that of the NOAA and Microsoft models (Table V).

TABLE V. OUTPUT COMPARISON

Projects	RMS Error
Wind Aloft from NOAA	51.53
Microsoft Research Project	43.66
Our Project	5.54

Current airplane flight planner applications are using weather information from the NOAA-based Wind Aloft program, which is quite noisy and less accurate. With a better system in place, keeping track of flights can help manage arrivals and departures more efficiently and assist in making decisions about flight schedules.

6. REFERENCES

[1] "Air traffic by the number." Federal Aviation Administration, Jun-2019.

[2] Federal Aviation Administration (FAA)/Aviation Supplies & Academics (ASA), *Aviation Weather Services: FAA Advisory Circular 00-45G, Change 1*. 2010.

[3] A. Kapoor, Z. Horvitz, S. Laube, and E. Horvitz, "Airplanes aloft as a sensor network for wind forecasting," presented at the Proceedings of the 13th international symposium on Information processing in sensor networks, 2014, pp. 25–34.

[4] C. Roberts, "America has gotten bad at predicting weather – but there's a plan to fix it," *Observer Media Group Inc.*, 26-Aug-2019.

[5] National Oceanic and Atmospheric Administration, "Wind/Temp Forecast," *Aviation Weather Center*, 2019. [Online]. Available: https://www.aviationweather.gov/windtemp/help.

[6] M. YANG, J. BARRICK, C. SWEENEY, J. DIGANGI, and J. BENNETT, "ACT-America: L1 Meteorological and Aircraft Navigational Data," *ORNL DAAC*, 2018.

[7] National Oceanic and Atmospheric Administration, "Aircraft Data Web." [Online]. Available: https://amdar.noaa.gov/.

[1] https://github.com/astha1015/Wind_Predict