Weather Variability Control in Three Colombian Airports

Karen Vargas-Daza, Giovanny Misat-Gómez, and Dionisio Neira-Rodado

Universidad de La Costa, Barranquilla, Colombia
{kvargas17, gmisat1, dneira1}@cuc.edu.co

Abstract. The aeronautic sector has been economically affected by the closure of its operations with the appearance of the Covid-19. For reducing the impact of weather variables at airport operations, we present a predictive model for better planning. Better planning reduces operative costs and increase the level of client satisfaction. This paper uses hourly observation from 2011 to 2018 at three Colombian airports: The Dorado airport in Bogota, the Olaya Herrera airport in Medellin, and the Matecana airport in Pereira. We build prediction models with deep learning and machine learning methods. These models aim to forecast horizontal and vertical visibility variables with minimum errors. The Random Forest decision tree model performs better predicting these variables in one, six, and twenty-four hours. This model has better results with the horizontal variable visibility forecasting for the three airports giving errors among 4% and 8%. This algorithm gave a flexible solution, and any airport can implement it.

Keywords: Random forest · Horizontal visibility · Vertical visibility

1 Introduction

One of the most economically affected sectors with the appearance of the Covid-19 is the aeronautic sector. As a result, many countries decided to stop aerial activity to prevent the virus propagation. To reduce economic effects in a stage of new normality, we propose a predictive model to control the shift of main weather variables in three Colombian airports. This model will be useful for airline planning in the short and long term. Consequently, better planning will reduce operative costs and will increases client satisfaction levels. Weather variability can impact airport operations. For instance: flight cancellations or changes on the route due to unsuitable conditions.

Closing airports means that the aeronautic sector would lose around 250,000 million dollars worldwide. As a result, most airlines decided to enter in financing law [1]. Even though, last year the Colombian aeronautical industry performed well, around 41,2 million passengers used aerial transport. It means an increase of 9,1% compared to the 37,6 million passengers who traveled in 2018. As a result of 40 new routes and 4 new airlines [2].

On the other hand, the Colombian government is planning in 2020 to reopen national and international flights by September. Then, airlines will meet challenges,
such as recovering their finances and make investments in adapting their system to the biosecurity protocol.

Hence, to improve operational costs, airlines must consider two essential meteorological variables in their planning: horizontal and vertical visibility. These variables will affect flight security and operation. Then, if these variables are above or under their limits, it will affect their landing or take-off schedule. For this reason, airlines need a good predictive model to improve short- and long-term planning. This research aims to present an algorithm that helps to forecast these variables. We test this model in three Colombian airports.

2 Related Work

Impunctuality and high operational costs affect airlines due to weather restrictions. Then, airlines will need a good predictor of weather variables to reduce this negative impact. Techniques like machine learning and classification models including Random Forest and Neural Networks, provided good results in the forecasting of weather variables [3–5].

[3] uses tree decision techniques to predict a categorical variable that indicates security procedures at airports taking as reference their capacity due to low visibility. This variable merge vertical and horizontal variable; This categorical variable indicates if it is possible for airplanes landing or taking off. These authors test four models at the Vienna International Airport between September 2007–2012, for 1-h predictions. They use four models: Boosting, Bagging, Random Forest, and an ordered logistic regression. The Boosting model produces better results.

As well as, [4] uses statistical models, including a machine-learning algorithm to predict vertical and horizontal visibility variables. These authors employ Three Classifiers and Random Forest. The Three Classifier model results with more weight within the final forecast model, obtaining values between 0.3 and 0.4.

Similarly, [5] applies deep learning techniques to predict the visibility variable using hourly information for ten years at the Urumqi International Airport. The research proves that when the visibility is more than 1000 m, the absolute error is about 706 m per hour; If the visibility is less than 1000 m, the absolute error is approximately 325 m. This model has better performance when this variable is less than 1000 m.

Also, [6] presents exploratory research using machine learning algorithms to classify and predict visibility variables. They collect the data from a meteorological station near the Orlando International Airport in Florida. These authors employ four classifiers, and the neural network algorithm gets 89.71% of accuracy.

Furthermore, [7] develop a model to predict climatic variables using data from the meteorological tower at the University of Chennai in India to mitigate the impact of natural disasters. This model gathers important variables as cloud cover, which refers to horizontal and vertical visibility. This study uses two classifiers, and the Random Forest model gets more accuracy, obtaining an 87.1% of success.
3 Analysis and Data Collection

Vertical and horizontal visibilities are the climatic variables with more impact on airline operations. If there are weather variability, airports will have restrictions in their takeoff and landing. Therefore, these are the dependent variables in our forecast model. This research presents a Random Forest model for the forecasting in short: one to six hours, and medium-term: twenty-four hours. We test this model in three Colombian airports: The Dorado Airport, Bogota (SKBO); Jose María Cordova Airport, Medellin (SKRG); and the Matecana Airport, Pereira (SKPE).

We follow the next data mining steps to process the data [6]:

**Data Collection.** We use the database of the Iowa State University (ISU) presented in the Iowa Environmental Mesonet. This data is on an hourly basis between January of 2011 until December of 2018. The parameters included were: temperature in Farenheit (tmpf), humidity (relh), visibility (vsby), etc.

**Data Transformation.** We call the data using a CSV file format to feed the model.

**Mining Tool.** We identify missing data in the 29 variables available in the data base from 2011–2018, developing a R code.

**Data Preprocessing.** The data cleansing process is completed. For example, If the skyc2 variable is null, we assume that the sky is clear. For dwpf and tmpf null variables, we take the hour before. With the drct and sknt variables, we replace the null value with cero. In addition, we set a time series with the date.

**Feature Extraction.** We consider 22 of the 29 variables. We consider these parameters considering uncorrelation between them.

**Data Mining.** We analyze the given datasets with the Random Forest and with the Neural Network algorithms. Then, we choose the one who gave better results: The Random Forest Algorithm. We consider 2 to 10 h before the actual time to make predictions with that one (Table 1).

| Table 1. | Variable description |
|----------|----------------------|
| Variable | Description          |
| Station  | Three- or four-character site identifier |
| Valid    | timestamp of the observation |
| Tmpf     | Air Temperature in Fahrenheit |
| Dwpf     | Dew Point Temperature (Fahrenheit) |
| Relh     | Relative Humidity in % |
| Drct     | Wind Direction in degrees from north |
| Sknt     | Wind Speed in knots |
| Vsby     | Visibility in miles |
| Mslp     | Sea Level Pressure in millibar |
| Skyl1    | Sky Level 1 Altitude in feet |

(continued)
4 Methodology

Previous researchers [3, 5] use deep learning and machine learning to forecast meteorological variables; they use Neural Networks models and Decision Trees due to their accuracy regard to traditional methods. We use three forecast techniques: Multiple Regression, Neural Network, and Random Forest.

First, our data set is taken from the database of the Iowa State University (ISU) database from the Colombia Asos Network. Then, we proceed to select the cities where the study takes place: Bogota, Medellin, and Pereira. Also, we include all available weather variable data. Finally, we select our specific date range between January 2011 and December 2018 (Fig. 1).

Second, we select the variables presented in the three airports. Third, we construct three forecast models one for each airport, and one for each type of model: two for the short term: 1 h and 6 h, and the third one for the medium term: 24 h; We aim to forecast the horizontal and vertical visibility variables. We work on an hourly basis to make the forecast. We make a considerable quantity of simulations for each method. We select the Random Forest to predict our predictor variables. This model gave better results with our data set than with the neural networks and the multiple regression model. The Random forest model is a classifier that groups several classification trees.

| Variable | Description               |
|----------|---------------------------|
| Skyl2    | Sky Level 2 Altitude in feet |
| SkyC2    | Sky Level 2 Coverage       |
| Metar    | Unprocessed reported observation |
| Gust     | Wind Gust in knots         |
| Alti     | Pressure altimeter in inches |

Fig. 1. Iowa State University (ISU) Data Base [6]
This model builds each tree individually, and then, variables are selected randomly. This model weight selected variables and then calculates the trees most voted class generated. Finally, the algorithm makes the prediction [7]. First, a sample enters on a tree, and then a binary test series is made at each node. After that, at the training stage, the algorithm optimizes parameters obtained in a split function of training samples [8]. Random Forest fits perfect with a large amount of data, for instance our set is a large amount of hourly data between 2011–2018 [9].

We make 10, 20, 50, and 100 repetitions. The Random Forest algorithm fits better with 100 repetitions. Then, we calculate the error metrics with three different measures: Mean Absolute Percentage Error (MAPE), the Theil’s U statistic, and the Root Mean Square Error (RMSE). The Root Mean Square Error (RMSE) follows the next equation:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}$$  

(1)

Where $n$ are the quantity of error samples and $e_i$ ($e_i = 1, 2, ..., n$) the error samples [11]. $e_i$ value is equivalent to $P_i - O_i$. $P$ are the predicted values and $O$ are the observed values.

We use the next equation to obtain the Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{\sum |\frac{A-F}{A}| * 100}{N}$$  

(2)

Where $A =$ current value, $F =$ forecast, $N =$ amount of data. MAPE measures the error size in percentage terms.

Finally, we use the following formula to calculate the Theil’s U statistic,

$$U = \sqrt{\frac{\sum_{t=1}^{n-1} \left( \frac{\hat{Y}_{t+1} - Y_{t+1}}{\hat{Y}_t} \right)^2}{\sum_{t=1}^{n-1} \left( \frac{Y_{t+1} - Y_t}{\hat{Y}_t} \right)^2}}$$  

(3)

This measure compares forecast results with historical data. $Y_t$ is the current variable value on a specific time with the notation $t$, and $n$ is the total number of observations at that time, $\hat{Y}$ is the forecast value.

With these three-error metrics, we compare the results of the models within the different sets of time: 1, 6, and 24 h.

We use as error estimators the RMSE, Theil’s U statistic and MAPE to evaluate the model performance. We calculate the RMSE to measure the accuracy of forecast or classification. Accuracy refers to the degree of correspondence mean between individual data of predicted values and observed values [10]. Observed values are those obtained at the meteorological measurement station. Finally, we explain if the data fits well with the forecast model. We select these three-error metrics because of its accuracy in their results.
5 Results

Airlines need a tool that allows them to reduce error predicting weather forecast, therefore they will improve their planning in the short and medium-term. As a result, airlines will reduce their operative costs and increase client satisfaction levels. This paper presents an efficient technique to predict the most critical weather variables: horizontal and vertical variability. We test this model in three Colombian airports. This algorithm brings an efficient solution in terms of error and visualization using Decision Trees.

We use deep and machine learning techniques, such as Neural Networks and Random Forests. The model with better results is the Random Forest that uses decision trees to make predictions. We develop an algorithm using R for three airports: El Dorado, Olaya Herrera and Matecana, more over any airport can implement it. This empiric study finds relevant hour one, two, and three before the prediction of one hour, six hours, and twenty-four hours. In Table 2 and Table 3, we present the Random Forest error metrics results of the vertical and horizontal visibility:

### Table 2. Error metrics vertical visibility (skyl1)

| Airports                  | 1 h forecast | 6 h forecast | 24 h forecast |
|---------------------------|--------------|--------------|---------------|
|                           | MAPE | RMSE | U | MAPE | RMSE | U | MAPE | RMSE | U |
| Olaya Herrera, Medellín   | 16.9 | 252.967 | 0.34 | 18.3 | 277.478 | 0.37 | 16.03 | 229.358 | 0.31 |
| El Dorado, Bogotá         | 17.3 | 154.851 | 0.24 | 17.91 | 178.325 | 0.27 | 16.8 | 144.729 | 0.22 |
| Matecana, Pereira         | 21.0 | 127.42 | 0.08 | 16.97 | 929.151 | 0.78 | 12.19 | 927.591 | 0.78 |

### Table 3. Error metrics horizontal visibility (vsby)

| Airports                  | 1 h forecast | 6 h forecast | 24 h forecast |
|---------------------------|--------------|--------------|---------------|
|                           | MAPE | RMSE | U | MAPE | RMSE | U | MAPE | RMSE | U |
| Olaya Herrera, Medellín   | 6.37 | 58    | 0.04 | 5.5  | 53   | 0.04 | 4.8  | 47.93 | 0.038 |
| El Dorado, Bogotá         | 23.8 | 130   | 0.07 | 18.4 | 113  | 0.06 | 16.9 | 107   | 0.05 |
| Matecana, Pereira         | 19.03 | 127   | 0.08 | 12.97 | 100  | 0.06 | 5.6  | 60    | 0.037 |

This model presented better results for the 1 and 6 h of horizontal visibility prediction of the Olaya Herrera Airport, with an 4% error. Also, for vertical visibility, the 1-h forecast fits better with an 8% error at the Matecana Airport. In addition, for the 24-h prediction horizontal variable, the Olaya Herrera Airport got a 4% error and the Matecana Airport a 5% error. To conclude, the random forest model fits better with the horizontal visibility variable in the 3-time horizons at all the airport’s samples.
6 Conclusions

Due to the high impact of the vertical and horizontal weather variables in airline planning operation, they need to use an accurate forecast model. The aeronautic sector needs to use a forecast model with high level of precision to reduce their costs and increase the level of client’s satisfaction. This paper presents the application of a forecast model using random forest decision trees to predict these variables in three Colombian Airports using 7 years data for the model’s prediction of these two variables. This model can be used in any airport, however it will be necessary to understand particular features of each data and compare with other models such as neural network or multiple regression to analyze their error results.

References

1. Semana: ¿En qué consiste la ley de quiebras a la que se sometió Avianca en EE. UU.? Rev. Sem. (2020)
2. Aerocivil: En 9,1 por ciento aumentó el tráfico de pasajeros movilizados vía aérea en 2019, Grup. Comunicación y Prensa - Unidad Adm. Espec. Aeronáutica Civ. vol. 2019, pp. 2019–2021 (2020)
3. Dietz, S.J., Knieringer, P., Mayr, G.J., Zeileis, A.: Correction to: forecasting low-visibility procedure states with tree-based statistical methods (Pure Appl. Geophys. 176(6), 2631–2644 (2019)). https://doi.org/10.1007/s00024-018-1914-x), Pure Appl. Geophys. 176(6), 2645–2658 (2019). https://doi.org/10.1007/s00024-018-1993-8
4. Herman, G.R., Schumacher, R.S.: Using reforecasts to improve forecasting of fog and visibility for aviation. Weather Forecast. 31(2), 467–482 (2016). https://doi.org/10.1175/WAF-D-15-0108.1
5. Zhu, L., Zhu, G., Han, L., Wang, N.: The application of deep learning in airport visibility forecast. Atmos. Clim. Sci. 07(03), 314–322 (2017). https://doi.org/10.4236/acs.2017.73023
6. ISU Department of Agronomy: Iowa Environmental Mesonet (2020). https://mesonet.agron.iastate.edu/request/download.phtml?network=CO__ASOS.
7. Medina-Merino, R.F., Nique-Chacón, C.I.: Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases (010), 165 (2017). https://doi.org/10.26439/interfases2017.n10.1775
8. Neira-Rodado, D., Nugent, C., Cleland, I., Velasquez, J., Viloria, A.: Evaluating the impact of a two-stage multivariate data cleansing approach to improve the performance of machine learning classifiers: a case study in human activity recognition. Sensors (Switzerland) 20(7) (2020). https://doi.org/10.3390/s20071858
9. Ali, J., Khan, R., Ahmad, N., Maqsood, I.: Random forests and decision trees. Int. J. Comput. Sci. Issues 9(5), 272–278 (2012)
10. Vargas, K., Gonzalez, A., Silva, J.: The Effect of Global Political Risk on Stock Returns: A Cross-Sectional and a Time-Series Analysis BT - Intelligent Computing, Information and Control Systems, pp. 540–548 (2020)
11. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7(3), 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014