International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

A Pilot Recommender System using K-Means Clustering to Find Desirable Paths in Aircraft Takeoffs

Nurmohamad

Abstract-This paper aims to study the geometric patterns produced by aircrafts on different takeoff paths and to establish a correlation between the fuel consumption and the geometry of the path. Based on the findings of the study, pilots could be advised to change their flying styles and strategy during take-off on a path that maximizes fuel conservation. This is validated by grouping the similar takeoff paths using the k-means clustering technique and by verifying linear relationship between the parameters of different paths in the clusters and their corresponding braking patterns. In addition, various runways are classified in order to study the variations in the takeoff paths.

Keywords:Takeoff paths, K-Means clustering, Turn angle, geographic distance and outliers.

I. INTRODUCTION

In order to improve the fuel efficiency and ease in the current air carrier operations, there is a growing emphasis on proactive pilot recommender systems that advise pilots how to fly the aircraft fuel-efficiently and along the most optimal paths [2]. These systems, by analyzing the historical flight takeoff data, helps in alerting the pilots when they reach a longer route with more bends to shift to a shorter and straighter path that may help to achieve the maximum possible fuel efficiency.

In this paper, we attempt to devise one such recommender system by analyzing the ground movement of the air craft just before its takeoff. For the same, we cluster similar aircraft takeoff paths and understand its plausible dependency on the flight and runway attributes in order to advise the pilots appropriately. *R* programming language is used as the platform to perform data analysis on the given authentic input dataset.

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.
* Correspondence Author
Nurmohamad, Joint director of income tax, Ministry of finance, Department of revenue, Government of India.

Section II discusses the introduction of new variables on which clustering is performed and gives a brief overview on data cleaning. In Section III, the ideas used to create the variables are presented. Section 1V deals with the clustering techniques employed to analyze the dataset and how the clusters can be used to find plausible patterns in the takeoff phase of the aircraft.

Finally, in SectionV, the study is concluded by explaining the patterns obtained by clustering, and by discussing ways to build models upon the different clusters.

II. NEW VARIABLES AND DATA CLEANING

A. New variables to be defined

The data set contains information regarding different takeoff instances of an aircraft from different airports. Each data instance contains variables that define the status of the aircraft at every second of its movement from the hangar to runway, viz., GPS position, date, time, speed and brake pressure.

A new data set is created by taking the required information from the original data set, where each takeoff instance of the aircraft is saved and transformed to new variables. In the remaining part of this Section, these new variables are introduced as depicted in Figure 1, and aspects of data cleaning are mentioned as they are important to avoid abnormal clustering results.

The *First variable*, ***Flight_Instance***, is the unique id corresponding to the flight instance.

The *second variable*, ***Turn_Angle,*** captures the total degree of turns made by the aircraft on its path during the particular flight instance.

The *third variable*, ***Total_Geo_Dist***, accounts for the total distance that the aircraft has travelled on the corresponding takeoff instance.

Finally, the *fourth* and the last variable in our transformed dataset, which is a categorical variable, is the ***Runway_ID*** that is calculated by classifying the position of the aircraft at different instances just before its takeoff. Basic assumptions are made for defining this variable which are discussed in the next Section.

FLIGHT_INSTANCE	TURN_ANGLE	TOT_GEO_DIST	RUNWAY_ID
1.CSV	14408.36	5540.608	1
2.CSV	9534.475	4221.388	2
4.CSV	15096.52	5946.23	1
5.CSV	4632.775	3677.281	3
6.CSV	15363.55	6246.571	4
7.CSV	13793.13	5702.402	5
8.CSV	21351.35	7331.488	1
9.CSV	8188.21	4896.738	6
10.CSV	32974.91	10204.22	7
11.CSV	4978.92	2243.771	8
12.CSV	5852.782	2625.55	1
13.CSV	4739.005	2539.737	5
14.CSV	4643.011	2500.358	9
15.CSV	2140.805	1833.785	10
16.CSV	17957.89	6630.666	9
17.CSV	10602.62	4439.816	11
18.CSV	6664.671	2831.119	1
19.CSV	5433.295	2493.89	8

Figure 1. An overview of the newly defined dataset

Before defining values for these variables, it is imperative to clean the original data-set since outliers on the original data-set can manifest themselves in the new data-set, and clustering techniques, especially K-Means, performs less efficiently in the presence of outliers.

B. Data Cleaning

The following steps for data cleaning have been taken up:

- If missing values are encountered, the corresponding row is removed completely. It is observed that "na" values are not frequent in the given data set, and hence removing such rows does not seem to affect the efficiency of clustering in a considerable way [1].
- Corrupted files and empty files are removed.
- The ***outliers*** in the dataset are detected using box-plot technique and are averaged out using their neighboring values.
- All instances with zero flight speed are removed. It was observed that the files did not have any takeoff patterns, instead were making irrelevant movements (irrelevant for our scope: which is takeoff patterns) through the runways.

In the next Section, the techniques used for assigning values to the newly defined four variables are explained.

III. DEFINING NEW VARIABLES

The approach employed to define the three non-trivial variables in the transformed dataset are explained below.

The most important non-trivial variable in the transformed dataset is ***TURN_ANGLE*** which as mentioned in the last Section gives a quantitative measurement of the type of path that the aircraft took from the hangar to runway just before its takeoff. It is trivial for the human eye to differentiate between two takeoff paths, but in order to define it in a generic way one needs to extract non-trivial and significant features hidden in these paths.

Using simple plotting techniques, we get a takeoff path (which includes the two-dimensional GPS-coordinates of all the points forming the path) formed by the aircraft in motion, similar to the one given in Figure 2.

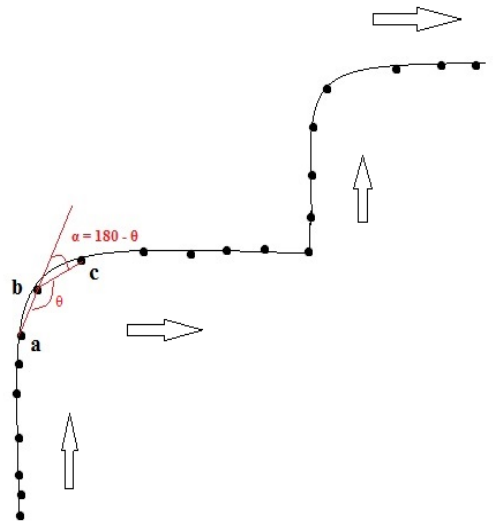

Figure 2. A typical flight path

In Figure 2, points '*a*', '*b*' and '*c*' are three adjacent points in a particular takeoff path of the aircraft. Line-segments *ab* and *bc* are formed by joining the adjacent points to each other. Here, the aircraft travelling through the path *ab* made a turn of 180-θ degrees in order to shift its direction to the path *bc*, where θ is the angle between *ab* and *bc*. This 180-θ degrees is defined as the turn-angle between *ab* and *bc*. In order to calculate the total turn-angle made by the aircraft, one needs to sum up all such turns made by the aircraft during its takeoff. This sum total of the turn angle value is represented by the variable ***TURN-ANGLE***.

Note that:$\theta = \cos^{-1}\left[\frac{\alpha^2+\beta^2-\gamma^2}{2\alpha\beta}\right]$ [3], where, α, β and γ are the geographic distance of the paths corresponding to '*ab*', '*bc*' and '*ca*', respectively.

The next main variable in the transformed dataset is ***Total_Geo_Dist***, which accounts for the total distance travelled by the aircraft on the corresponding travel instance. Given the GPS locations of two points *a* and *b*, the geographic distance between them can be found by using the Haversine formula [4] which calculates the great-circle distance between two points, ie., the shortest distance between the two points over the earth's surface. The Haversine formula is used as it is numerically better-conditioned for calculating smaller distances similar to the ones that are attempted to measure here [4].

Haversine formula:

$a = \sin^2(\frac{\Delta\phi}{2}) + cos\phi_1 . cos\phi_2 . \sin^2(\Delta\lambda/2)$; where, $\Delta\phi=(\phi_2 - \phi_1)$ and $\Delta\lambda = (\lambda_2 - \lambda_1)$

$c = 2.atan2(\sqrt{a}, \sqrt{(1-a)})$

$d = R.c$

Where ϕ_2 and ϕ_1, and, λ_2 and λ_1 are the latitude and longitude values of the two points respectively,R is earth's radius and d is the distance between the two points that needs to be calculated. Note that the angles need to be in radians to pass the trigonometric functions.

The last non-trivial variable, ***Runway***, is acategorical variable. As discussed in the last Section, it is defined by comparing the last latitude and longitude values of each of the travel instances given in the original data set. Few basic assumptions were made in defining this variable, for example, no two takeoff points on the same runway are at a distance greater than 1 mile and no two takeoff points on two different runways are at a distance less than or equal to 1 mile.

In Figure 1, the first entry can be interpreted as follows: the travel instance defined in file 1.csv of the original dataset corresponds to the aircraft moving through the runway classified as runway number *1* and the total geographic distance covered is 5540.608 meters with a total turn angle of 14408.36 degrees in its motion from the hangar to the runway.

In the next Section, we describe the clustering mechanism applied on the variables ***Turn_Angle*** and ***Total_Geo_Dist***.

IV. CLUSTERING ANALYSIS

Here, the ***K-means*** clustering technique has been employed to classify similarly distributed variables. K-means clustering [3] requires that the number of clusters to be extracted needs to be specified in advance. Different techniques can be adopted to find the optimal number of clusters. For the same, the total within-groups sums of squares is plotted against the number of clusters in a K-means solution [3]. A bend in the graph suggests the appropriate number of clusters.

On applying ***K-means*** on the variables ***Turn_Angle*** and ***Total_Geo_Dist***, takeoff paths that amount to similar turn-angle and geographic distance get clustered together. After performing ***K-means clustering***, the optimal number of clusters were found to be 5. Note that both the ***Turn_Angle*** *and* ***Total_Geo_Distance*** variables were normalized before applying ***K-means*** algorithm on them. It is important to normalize these variables since Euclidean distance was used as the distance measure in clustering, and normalization is required here to prevent the covariate with the highest variance from driving the clustering algorithm. After performing ***K-means clustering***, the optimal number of clusters were found to be 5, each cluster containing flight takeoff-patterns with similar turn angles and geographic distances.

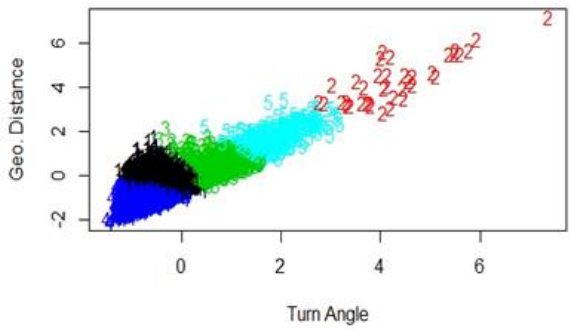

Figure 3. Clustering patterns

Each color in Figure 3 represents patterns of different clusters. It can be observed that travel instances present in cluster number *4* contains the least attribute values (***Turn_Angle*** *and* ***Total_Geo_Distance*** values). Notice that cluster number *2* contains paths with extreme attribute values.

One can perform regression analysis on these clusters by transforming each of the clusters into a two-dimensional vector (or a point in a 2-D plane) containing the mean ***Turn_Angle***and the mean of the *total brake pressure* of all the instances (or all the takeoff paths) in each cluster, where *total brake pressure* of an instance is the sum total of all the different brake pressure applied in the given instance (or takeoff path). Performing a simple linear regression on the above defined 2-D data points may give us a trend on the total amount of brake pressure applied, for example, if more ***Turn_Angle*** accounts to an increase in the amount of brake pressure applied on the aircraft, it is the same as saying that more angle of turns causes more brake pressure. Note that more brake pressure accounts for more aircraft fuel wastage.

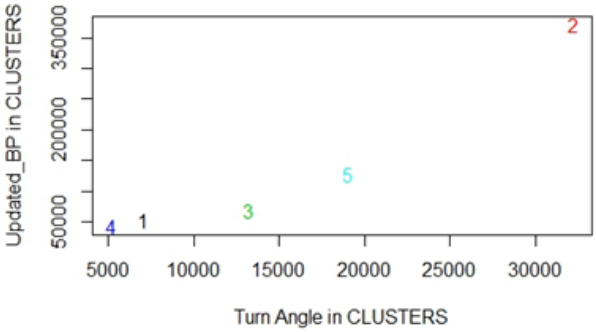

Figure 4. Average Turn angle vs. Average *Total brake pressure* (Cluster-wise)

Figure 4 shows a two- dimensional plot on the average***Turn_Angle***against the average brake pressure applied in each of the *5* clusters and a desired linear correlation is observed between the two variables similar to the one given in Figure 3 that depicts all the five different clusters. In fact, this indicates that the takeoff instances in each cluster share similar brake pressure thereby validating the claim that brake pressure applied by the aircraft in its takeoff path is correlated with

Turn_Angle* and *Total_Geo_Dist.

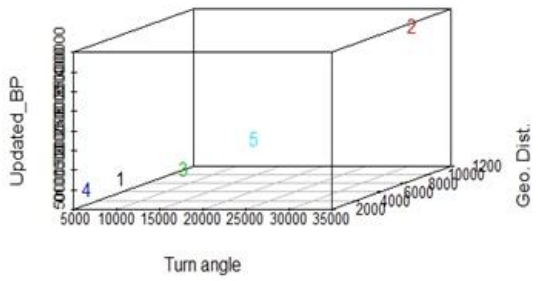

Figure 5. Average Turn Angle vs. Average Total Geographic Distance vs. Average *Total Brake pressure* (Cluster-wise)

Figure *5* illustrates a three-dimensional plot on the variables average***Turn_Angle***and average ***Total_Geo_Dist*** and average *total brake pressure***,** and in each of the 5 clusters we observe a linear correlation to the distribution of the clusters depicted in Figure 3 just as in the previous case.

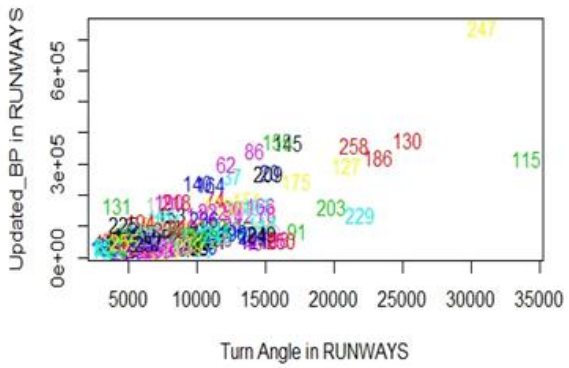

Figure 6. Average Turn Angle vs. Average Total brake pressure(Runway-wise)

Figure 6 shows a two- dimensional plot on the average***Turn_Angle***against the average *total brake pressure* in each of the runways and it shows a nearly linear correlation to the distribution of variables given in Figure 3 which depicts all the five different clusters.

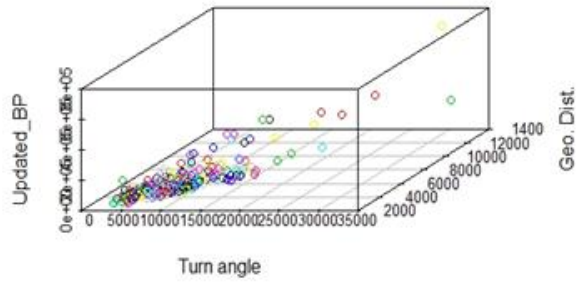

Figure 7. Average Turn Angle vs. Average Total Geographic Distance vs. Average Total brake pressure(Runway-wise)

Finally, Figure 7 shows a three-dimensional plot of the variablesaverage***Turn_Angle***and average ***Total_Geo_Dist*** and average *total brake pressure*in each of the runways, and this also indicates a nearly linear correlation.

V. CONCLUSION AND FUTURE WORK

The paper presents an overview of the ways to classify features from the given authentic dataset in order to identify any hidden patterns in them, if they exist. It is observed that there is a linear correlation between the turn angle and brake applied which reveals that braking patterns, (hence the fuel consumption), are dependent on the distance travelled by the aircraft and the total degree of turns made by the aircraft in its takeoff path.

The study can be extended by performing various data modelling techniques on the different clusters formed. One could also try to find the effect caused by updating the value of the total brakes applied after removing the amount of brakes applied during idle time (no change in the GPS location indicates idle time).

Addition of new parameters in the input dataset could help provide further interesting analysis. For example, with the information of the unique ID of the pilot responsible for each travel instances, the running styles corresponding to different pilots can be clustered and this information can be used to guide them to improve their piloting skills towards minimizing the fuel consumption. Needless to say, this data could also help to provide maximum comfort to passengers and other occupants by taking paths with shorter distance and shorter total turn angle.

REFERENCE

1. A first course in statistical programming with R (Braun & Murdoch).
2. Li, L. (2013). Anomaly detection in airline routine operations using flight data recorder data (Doctoral dissertation, Massachusetts Institute of Technology).
3. Gariel, M., Srivastava, A. N., & Feron, E. (2010). Trajectory Clustering and an Application to Airspace Monitoring. Arxiv preprint arXiv:1001.5007, 1–15.
4. Sinnott, R.W, 1984. Virtues of the Haversine. Sky and Telescope 68(2): 159.

AUTHORS PROFILE

Nurmohamad, Working at Income tax department as Joint director of income tax(systems),Ministry of finance, Department of revenue, government of India.
PHD(Part time) joined in Vignan University,Vadlamudi
Expectation: i) To obtain advice and guidance to enable them to assume a high degree of independence and control over their work. ii) Assess understanding of their skills and discuss their aspirations with supervisor officers in revenue generations.