Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1315~1329
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i3.pp1315-1329

Optimum partition in flight route anomaly detection

Mohammad Yazdi Pusadan, Joko Lianto Buliali, R.V. Hari Ginardi
[1]Department of Informatics, Institute Technology of Sepuluh Nopember, Surabaya, Indonesia
[2]Department of Informatics, University of Tadulako, Sulawesi Tengah, Indonesia

Article Info

Article history:

Received Aug 06, 2018
Revised Dec 23, 2018
Accepted Jan 21, 2019

Keywords:

FIR
Partition
Segment

ABSTRACT

Anomaly detection of flight route can be analyzed with the availability of flight data set. Automatic Dependent Surveillance (ADS-B) is the data set used. The parameters used are timestamp, latitude, longitude, and speed. The purpose of the research is to determine the optimum area for anomaly detection through real time approach. The methods used are: a) clustering and cluster validity analysis; and b) False Identification Rate (FIR). The results archieved are four steps, i.e: a) Build segments based on waypoints; b) Partition area based on 3-Dimension features P1 and P2; c) grouping; and d) Measurement of cluster validity. The optimum partition is generated by calculating the minimum percentage of FIR. The results achieved are: i) there are five partitions, i.e: (n/2, n/3, n/4, n/5) and ii) optimal partition of each 3D, that is: for P1 was five partitions and the P2 feature was four partitions.

Corresponding Author:

Mohammad Yazdi Pusadan,
Department of Informatics,
Institute Technology of Sepuluh Nopember,
Jl. Teknik Kimia, Gedung Informatika, Kampus ITS Sukolilo Surabaya, Indonesia.
Email: yazdi.diyanara@gmail.com

1. INTRODUCTION

Research on flight anomaly detection has begun in the last decade, such as using a regression model with the data set used from FOQA (Flight Operations Quality Assurance) with 26 parameters [1], anomalous detection on FDR flight data was presented in [2] in which density-based clustering (DBSCAN) technique was used to cluster the entire FDR data on one flight thus could not detect flight anomaly early.

Our interest in this research topic stems from the phenomenon of aviation problems such as the case of QZ8501 Surabaya (SUB)-Singapore (SIN) Air Asia plane crash in December 2014. Based on the investigation results of the National Transportation Safety Commission in 2015, that there were anomalies on the flight route (there was significant deviation from the position of latitude and longitude coordinates that the flight should be [3]. Figure 1 shows anomalies on latitude and longitude of this flight. The dashed line is the trajectory that the plane should have passed through, while the red line is the actual trajectory that the plane passed through, resulting in a crash. The problem in this phenomenon is that the detection of anomalies were non-preventive.

Based on the above flight problems and previous research studies, we present real-time flight anomaly detection using distance-based clustering technique (K-Means). The data source used as training data is data from the Automatic Dependent Surveillance Broadcasting (ADS-B) obtained for thirty days. The obtained ADS-B data are grouped into two 3-dimensional features, is a feature P1 = (latitude, longitude, speed) and feature P2 = (latitude, longitude, travel time). The clustering process to obtain anomaly detection is based on the pattern of flight habits in a call sign. The process continues on determining cluster performance based on internal validity tests based-on silhouette index. The contribution of this research is the determination of optimum partition on 3-Dimensional ADS-B features. The optimum size is the minimum percentage of identification error (False Identification Rate) in each 3-dimensional feature.

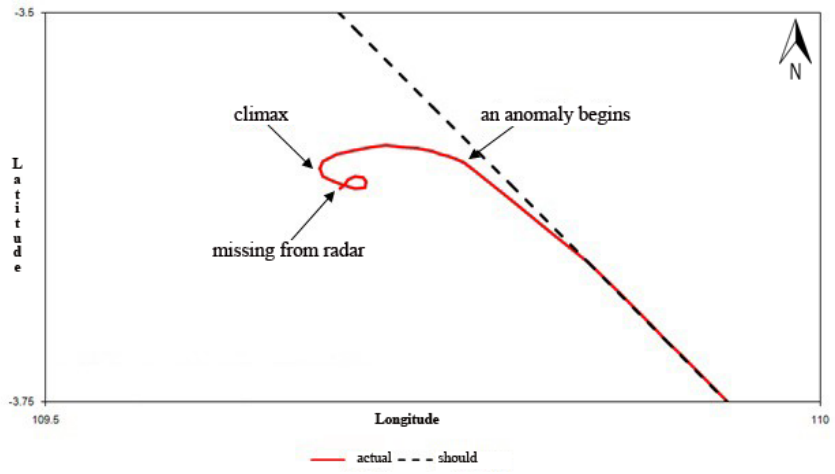

Figure 1. Flight route anomaly on AirAsia QZ8501 (SUB-SIN)

Research related to anomaly detection can be categorized into one dimensional analysis and multi-dimensional analysis. In one dimension analysis, statistical approach is used based on the deviation value, such as [4] in which three methods (Interquartile Range, Moving Average, and Median Absolut Deviation) are used in detecting anomaly in IP address that access the network.

Multi-dimensional analysis can be done by clustering method. In [5], automatic initial number of center k was proposed for K-Means to form clusters on big data by using map reduce paradigm. In [6], one of cluster quality measurement methods (i.e., index silhouette) was used to measure the quality of document group which have financial risks. Silhouette index is used because of its simplicity to measure how well a document is placed in a cluster. In [7], euclidean distance is used to measure the similarity of distance between training data and testing data in two classes. The result is that the percentage of accuracy testing data based on Euclidean values is found in a class.

Flight anomaly detection using clustering techniques has been the focus of several papers [2] used DBSCAN to group FOQA (Flight Operation Quality Assurance) flight data. The results obtained were the data that were detected as anomaly based on a certain density. There are several anomaly criteria in flight, including: high and low energy states, unusual pitch excursions, abnormal flap settings, high wind conditions. The flight phase specified is taking off and landing phase. The proposed method, are: 1) the conversion of time-series data set into vectors. 2) Reduction of data dimension using Principal Component Analysis method in aviation phase. The takeoff phase from 6188 to 77, while in the landing phase from 6279 to 95. 3) cluster-based density analysis (DBSCAN) is used to obtain the outlier area in each flight phase. The study was developed in 2015 to detect abnormal flight [8], so it could assist experts in handling anomalous conditions in operational aviation. The difference from previous research is that this study did not require of flight anomaly criteria. The proposed method is ClusterAD-Flight. It is based on the use of data mining techniques in observing and analyzing common patterns. It is possible that all flights have a common pattern. If there are deviant conditions, it is possible to experience anomaly and can be used as a reference in aviation safety management.

Based on the above research, it can be concluded that there are several important aspects that need to be developed. The development is time-based anomaly detection. Analysis is evaluative and events then the data obtained from the process is analyzed. Resulted in the early detection process of flight anomalies was not found. In addition, the cluster-based method of density did not become the main approach to clustering.

It is because there is still a minimum point determination (MinPts). Therefore, the distance-based cluster approach can be an alternative solution. An anomaly detection approach in real-time can be developed in this research, as a solution in the field of aviation to develop anomaly aviation of early warning detection system.

Related research on flight clustering techniques is fault / anomaly detection [9] by knowing the abnormal circumstances that occur [8]. The technique used was to perform parameter reduction before data was clustered [2]. Furthermore, the data was clustered according to density [8], [10], so that areas outside the density were potentially anomaly. Below in Figure 2 shows anomaly-based detection stages based on density.

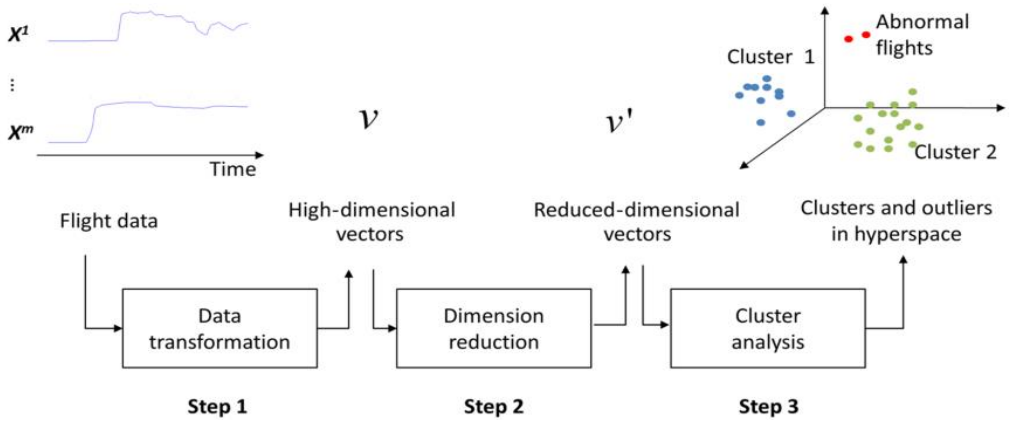

Figure 2. Anomaly Detection Stage with density-based clustering technique (DBSCAN)

The stages taking place are: first, the process of flight data transformation (FDR) patterned time series into multidimensional vector. Second, parameter reduction was done to obtain the main parameters contained in the multidimensional vector by Principle Component Analysis (PCA) technique. The last stage, it performed anomaly detection based on DBSCAN clustering technique. The result obtained was an area with a certain minimum density of potentially anomalous.

Problems that occurred in density-based clustering techniques were that there are minPT parameters (minimum points) in the data set. It needed further analysis of the minPT provisions. The following illustration of density-based cluster technique (DBSCAN) is shown in Figure 3.

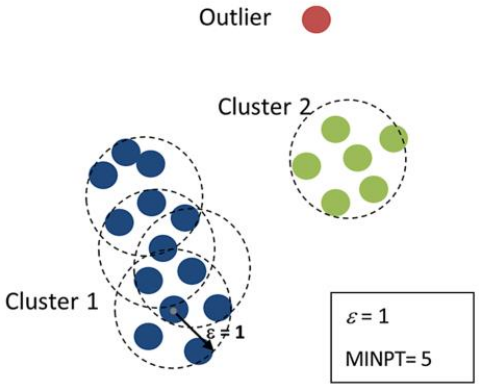

Figure 3. Illustration of DBSCAN cluster technique

It was shown that with two clusters with the number of members in each cluster were: cluster to-1 was 13 and cluster to-2 was 6. In addition, the minimum member of each cluster (minPT) was 5 and value ε = 1. Based on this cluster technique, it was obtained anomaly data with the amount of 1.

In addition, cluster analysis through the measurement of the validity / performance of the cluster was not done. So, it was not found the indicator to know the clusters well formed. In contrast to our research, the cluster technique used was a distance-based cluster technique K-Means [11]. After the cluster was formed, the measurement of cluster validity was based on internal validity. The internal method of validity with the maximum index value used is called the silhouette index [12].

The correlation with previous research that the flight data used had a time series pattern [2]. So, the flight data recording (FDR) parameters [10] received per record take place during the timestamp period. The FDR parameters include altitude, speed, latitude, longitude, and time. For our research, FDR is based on Automatic Dependent Surveillance Broadcasting (ADS-B) [13] [14]. One of the advantages of ADS-B is accessibility, that data access uses the http protocol. So it facilitates the ATC as a secondary FDR (in addition to radar data).

2. PROPOSED METHOD

There are several studies related to anomaly detection in flight such as anomaly detection by using clustering method [2], [8]. Furthermore, the thing to note in aviation anomaly detection is the determination of waypoint [15] which becomes the specific location of the aircraft accurately flying past several points before reaching the destination. Furthermore, between one Waypoint and the other waypoint formed an area called a segment [16]. The provisions of the segment are: waypoint1 ≤ segment1 ≤ waypoint2; waypoint2 ≤ segment2 ≤ waypoint3; until waypoint7 ≤ segment7 ≤ waypoint8. For initial segment determination starting from segment2, because the ADS-B data used is when the plane's position is in cruise [1], so segment1 is unavailable.

Furthermore, the data set in the segment is done K-Means clustering process [17]-[20]. To get a good cluster results, it performed cluster analysis to get the value of the index validity. There are two criteria in the process of measuring validity, are: compactness (cohesion) and separation [21]. In the cluster evaluation technique in order to obtain the optimum cluster, an internal validity evaluation [22] is used. There are two value categories of validity measurement index, the maximum index value [19], [23]-[25] and the minimum index value [21], [23], [26].

The next stage, anomaly-based clustering detection. After the cluster on the segment is generated, it measures the distance between centroid one and the other centroid to detect potentially anomalous areas. Measurement distance used Euclidean distance [20]. The result, in the form of the largest centroid distance from other centroids is as a potential cluster / anomaly area.

2.1. Research Framework

The framework in this study is shown in Figure 4 so the real-time determination process of the segment-based anomaly is happened. The focus is on determining the detection area by determining the optimum partition based on the distance-based cluster.

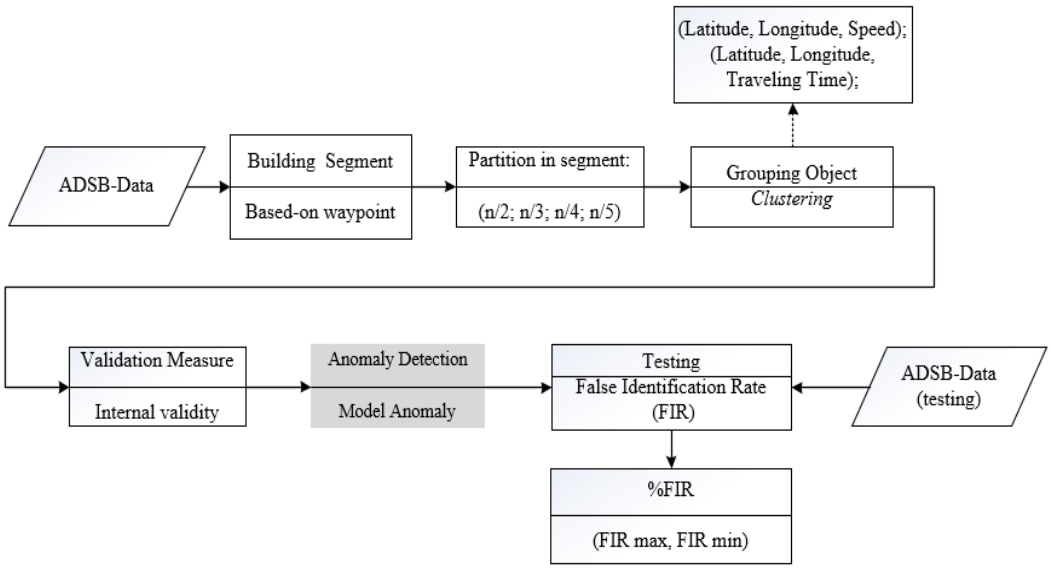

Figure 4. Framework research

The proposed method in this research is Building Segment, Partition, Grouping, and Validity Measurement (BsPGV). The target is anomaly detection model based on 3D data feature partition. First, the data source (data training) ADS-B is specified on several segments based on the waypoint point. Second, partition the data in a segment. Third, grouping with K-Means clustering method. Fourth, cluster validity measurements. In determining the optimum partition, it is done through testing data tested on BsPGV model. The result is an identification error (False Identification Rate/FIR).

2.2. Data set

The parameters contained in the ADS-B data are: timestamp, UTC, call sign, position, altitude, speed, and direction. The following ADS-B data for call sign LNI860 / SUB-PLW are shown in Table 1.

Table 1. ADS-B Data Set for LNI860 Flight

Timestamp	Date	Time	Traveling Time (seconds)	Latitude	Longitude	Altitude	Velocity (v)
1480582105	12/01/2016	15:48:25	0	-6.601	115.329	37000	456
1480582099	12/01/2016	15:48:19	6	-6.609	115.318	37000	456
1480582063	12/01/2016	15:47:43	36	-6.643	115.249	37000	456
1480582002	12/01/2016	15:46:42	61	-6.695	115.131	37000	462
1480581939	12/01/2016	15:45:39	63	-6.749	115.005	37000	466

Based on the ADS-B data source in the Table 1, 3-Dimensions (3D) feature was proposed in this study. The first feature (P1) is parameters: latitude, longitude, and speed. While the second feature (P2) with parameters: latitude, longitude, and traveling time.

2.3. Waypoint

This research uses Surabaya (SUB) flight route to Palu (PLW) (call sign LNI860). The route passes through eight waypoints represented by latitude and longitude coordinates. The waypoints coordinates are indicated through the following latitude and longitude coordinates.

a) Waypoint1: Surabaya with coordinates latitude, longitude is (-7.373, 112.772);
b) Waypoint2: Fando with coordinates latitude, longitude is (-6.973, 113.985);
c) Waypoint3: Kasol with coordinates latitude, longitude is (-6.568, 115.173);
d) Waypoint4: Dasty with coordinates latitude, longitude is (-6.173, 116.330);
e) Waypoint5: Endog with coordinates latitude, longitude is (-5.877, 117.202);
f) Waypoint6: Gamal with coordinates latitude, longitude is (-2.863, 118.038);
g) Waypoint7: Rudal with coordinates latitude, longitude is (-2.662, 118.711);
h) Waypoint8: Palu with coordinates latitude, longitude is (-0.885, 119.962).

The following is shown in Figure 5, the distribution of waypoints in the latitude and longitude coordinate representations.

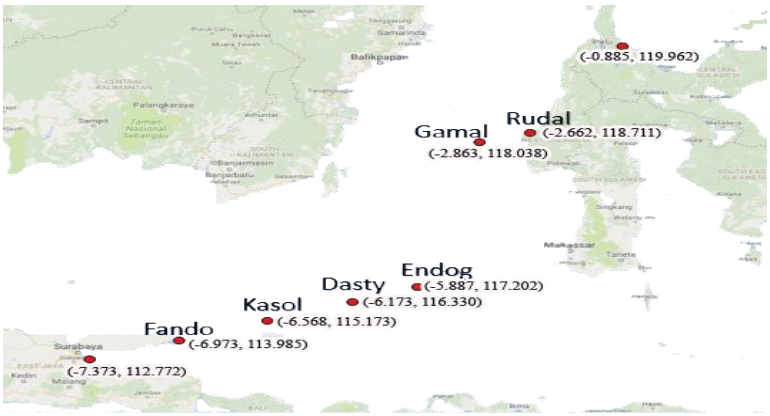

Figure 5. Distribution of waypoints for LNI860 flight

2.4. Segment

A segment is an area between a single waypoint and another waypoint. The formulation is as follows: $Wn \leq SEGMENTn \leq Wn + 1$ (n = 1, 2, k), so the range of area obtained is shown in Table 2.

Table 2. Segments in LNI860 Flight

Segment	Range Area Waypoint
$SEGMEN_1$	$Waypoint_1 \rightarrow Waypoint_2$
$SEGMEN_2$	$Waypoint_2 \rightarrow Waypoint_3$
$SEGMEN_3$	$Waypoint_3 \rightarrow Waypoint_4$
$SEGMEN_4$	$Waypoint_4 \rightarrow Waypoint_5$
$SEGMEN_5$	$Waypoint_5 \rightarrow Waypoint_6$
$SEGMEN_6$	$Waypoint_6 \rightarrow Waypoint_7$
$SEGMEN_7$	$Waypoint_7 \rightarrow Waypoint_8$

The following in Figure 6 shows the distribution of ADS-B training data for 30 days on each segment formed and waypoints.

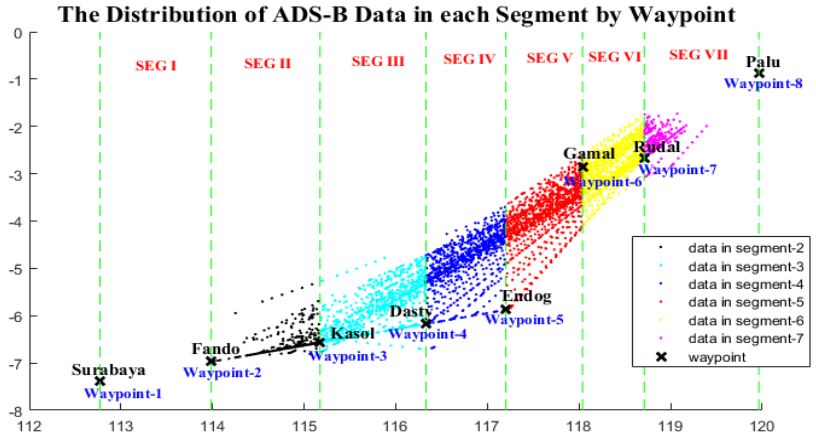

Figure 6. Distribution of ADS-B training data while 30 days for LNI860 flight

2.5. Partitions (sub segments)

Partitions (sub segment) is the process of dividing the area on each segment. To go to the clustering done partition for the nature of data on a cluster can be known specifically. Partitions are determined by the following 3:

$$n/2, n/3, n/4, n/5 \tag{3}$$

Partitioning process takes place in two features: 3D feature with position parameters, speed (latitude, longitude, speed) and 3D feature with position parameters, travel time (latitude, longitude, traveling time).

2.6. False Identification Rate (FIR)

False Identification Rate (FIR) is an indicator of error rate measurement in the identification process of data testing located in the nearest cluster. Figure 7 shows the FIR scenario that takes place in data testing with multiple clusters.

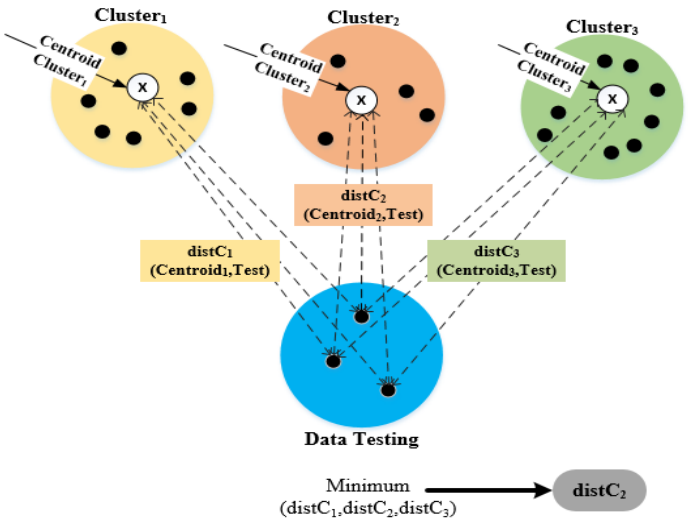

Figure 7. Scenario 1 FIR (distance measurement in data testing)

The first scenario, data testing measured the distance with the centroid of each cluster. Performed measurements between centroid clusters with some data testing. Data testing used comes from ten days of training data (in-set testing). Minimum size of distance generated, then allows the data testing is in the cluster distance acquisition minimum. In Figure 8 shows the process of checking the testing data on the nearest cluster, along with the determination of FIR value.

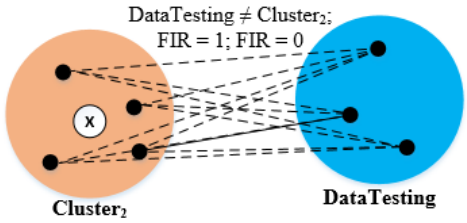

Figure 8. Scenario2 FIR (checking data testing on a cluster)

The second scenario, checking the number of possible data testing is located on the cluster that the minimum centroid distance. FIR = 0, if the data is in the cluster and FIR = 1, if the data is not in the cluster. For the application of the computational FIR is represented in the following pseudocode algorithm.

```
Algorithm-Compute FIR
Input: dataTesting
1:    //distance measurement
2:    distC1(test,CentC1);
3:    distC2(test,CentC2);
4:    distC3(test,CentC3);
5:    minDist←distC1;
6:    dist←[distC1,distC2,distC3];
7:    for i←1 to length(dist) do
8:          if dist(i) ≤ minDist then
9:                minDist←dist(i);
10:               cluster←i;
11:         endif
12:   endfor
13:   //compute FIR
14:   for j←1 to length(test) do
15:         if test(j) == cluster(j) then
16:               FIR(j)←0;
17:         else
18:               FIR(j)←1;
19:         endif
20:   endfor
```

The False Identification Rate (FIR) algorithm is grouped into two processes. First, the calculation of the distance between the centroid and the data testing. Measurement distance with Euclidean distance on a function distC1, distC2, and distC3 based on number of cluster centroid (k = 3). Minimum distance is obtained, indicating the data testing is likely to be on the cluster. There are two variables, minDist variable to store the minimum distance, and cluster variable that indicates a certain cluster. Second, the FIR calculation by checking the test data is in the selected cluster. FIR = 0 (variable FIR(j) ← 0) if the test data is not in the selected cluster. While FIR =1 (variable FIR(j) ←1) if the test data is in the selected cluster.

3. RESEARCH METHOD

3.1. Clustering

The clustering method used is a distance-based cluster, called K-Means. This method is chosen, because the data grouped on one cluster is determined by distance (Euclidean distance) and produces centroid for each cluster. For former initialization of the centroid was done by a random value. In the implementation of clustering, a cluster per segment was performed. Furthermore, the cluster process took place on the 3 Dimensional feature. The 3 Dimensions feature is the ADS-B parameters involving three parameters. The parameters are based on features P1 and P2. P1 = (latitude, longitude, speed) and P2 = (latitude, longitude, traveling time).The cluster result is formed in Figure 9 for cluster P1, and Figure 10 for cluster P2

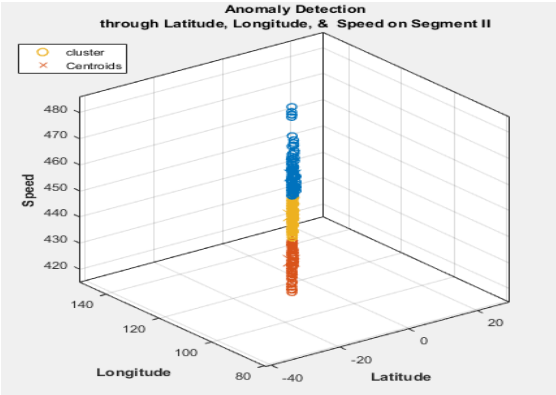

Figure 9. Clustering on P1

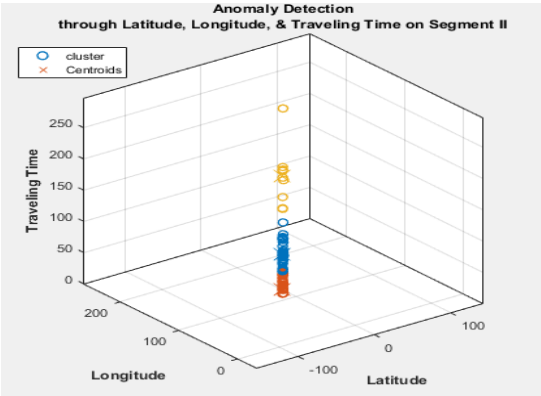

Figure 10. Clustering on P2

For P1 centroid cluster is Centroid1 = (-5.779, 115.796, 422.570); Centroid2 = (-5.911, 115.748, 459.472); Centroid3 = (-5.930, 115.767, 444.552). There is a large amount of data per cluster (nCluster), which are: nCluster1 = 151; nCluster2 = 354; nCluster3 = 433. For P2, centroid cluster is Centroid1 = (-5.871, 115.757, 73.57); Centroid2 = (-6.088, 115.498, 332); Centroid3 = (-5.927, 115.788, 11.26). There is a large amount of data per cluster (nCluster), which are: nCluster1 = 531; nCluster2 = 19; nCluster3 = 388.

The clusters formed are represented on the x, y, z Cartesian axis determined by 3-dimensional features. These features are P1 (latitude, longitude, speed) and P2 (latitude, longitude, traveling time). Cluster results on feature P1 shown in Figure 5 is: data in cluster to-1 marked by blue dots are 151. Data in cluster to-2 marked by orange dots are 354. For data in clusters to -3 marked by red dots are 433. Next to the centroid point at P1 in each cluster is the centroid point of the cluster 1: (-5.779, 115.796, 422.570). Centroid point in the second cluster is: (-5.911, 115.748, 459.472). Then the centroid point in the 3rd cluster is: (-5.930, 115.767, 444.552).

For cluster results in feature P2 shown in Figure 6 are: data in cluster to-1 marked by blue dots are 531. Data in the 2nd cluster marked by orange dots are 19. For data in the cluster to-3 marked by red dots are 388. Next to the point of centroid on P2 in each cluster is the centroid point clustered to-1 is: (-5.871, 115.757, 73.57). Centroid point in the 2nd cluster is: (-6.088, 115.498, 332). Then the centroid point in the 3rd cluster is: (-5.927, 115.788, 11.26).

3.2. Cluster Validation with Silhouette index

To get a good cluster results, it is necessary to measure the cluster validity. This stage is called cluster analysis by conducting the cluster truth test process formed. Silhouette index is a measure of how similar an object to its cluster (cohesion) is compared to other clusters (separation). Range of silhouette values ranges is from -1 to +1. If it shows the maximum value, then the object is on its cluster and not in other clusters. If many objects have a high value, then the clustering configuration is complete. If multiple points have low or negative values, then the clustering configuration may be too many or too few clusters.

Silhouette can be calculated by using distance matrix, such as Euclidean Distance, as well as formulated in mathematics as follows. 1 and 2 shows the Silhouette-index calculation formula.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, if\ a(i) < b(i) \\ 0, \qquad if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, if\ a(i) > b(i) \end{cases} \tag{2}$$

Where,

a) a(i): the average distance from i to all points on one cluster;
b) c(i): the average distance from i to all points on the other cluster (cluster neighbors);
c) b(i): minimum c(i).

4. RESULTS AND DISCUSSION

The results achieved in this study is formed cluster on each partition that was determined based on 3D features. Furthermore, computational analysis on each cluster is formed based on the number of iterations, replication, total distance (total sum of distance), and travel time. Cluster analysis process was done as optimal size of cluster. This is done through the internal process of cluster validity by the method of calculating the value of the silhouette index. The specified range ranges from -1 to +1 or -1 ≤sh≤ + 1. The last stage is the testing process. The data testing used comes from training data (in-set training data) with 10 days sample. Furthermore, with the error identification method or False Identification Rate (FIR), then the minimum% error obtained on each partition of each 3D feature as a determinant of an established partition is optimal. Here are the results of experiments and analysis conducted, then obtained anomaly detection model through 3D features P1 and P2.

4.1. Features 3-Dimension

In the 3-Dimensional feature that proceeds, two models are generated through the features P1 = (latitude, longitude, speed) and P2 = (latitude, speed, traveling time). Overall the process in 3D, features P1 and P2 is a cluster-based anomaly detection model. In the model that formed, there are three things that determine, namely: First, there is a potential anomaly area based on the furthest distance and the smallest amount of data from each cluster generated. Second, computing aspect calculations based on parameters: replication, iteration, total distance, and travel time per partition.

4.2. Features 3-Dimension P1

The result of grouping with K = 3 as anomaly detection model on P1 is shown in Figure 11.

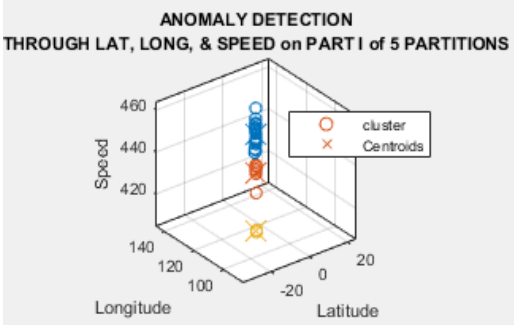

Figure 11(a). Clustering in partition I of five partitions

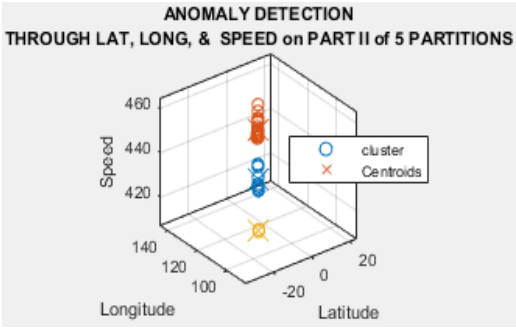

Figure 11(b). Clustering in partition II of five partitions

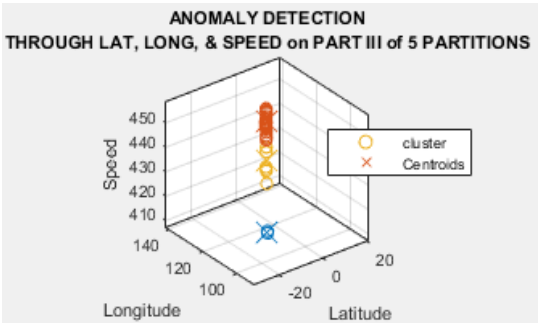

Figure 11(c). Clustering in partition III of five partitions

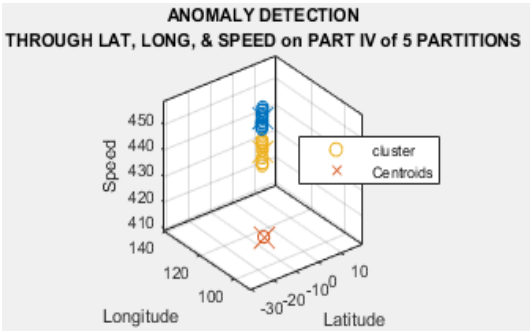

Figure 11(d). Clustering in partition IV of five partitions

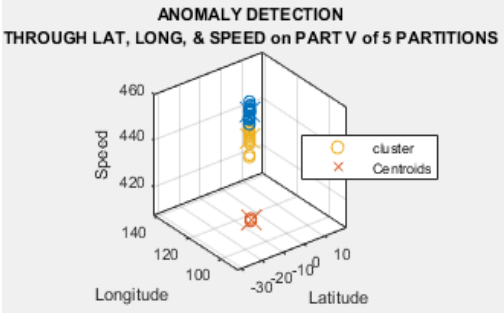

Figure 11(e). Clustering in partition V of five partitions

The computational analysis in each partition on feature P1 are shown in the following Table 3a and Table 3b. For internal measurement of cluster validity based on the silhouette index value on P1, it is shown in the silhouette chart in Figure 12.

Table 3a. Computational Aspect in Each Partition on P1 (Two Partitions and Three Partitions)

Computational Aspect	Two Partitions		Three Partitions		
Partition	I	II	I	II	III
Replication	1	1	1	1	1
Iteration	4	2	1	4	2
Total sum of Distance	2231	1026	1607	979	777

Table 3b. Computational Aspect in Each Partition on P1 (Four Partitions and Five Partitions)

Computational Aspect	Four Partitions				Five Partitions				
Partition	I	II	III	IV	I	II	III	IV	V
Replication	1	1	1	1	1	1	1	1	1
Iteration	2	10	3	4	5	1	5	1	3
Total sum of Distance	1381	817	453	566	1091	608	562	304	566

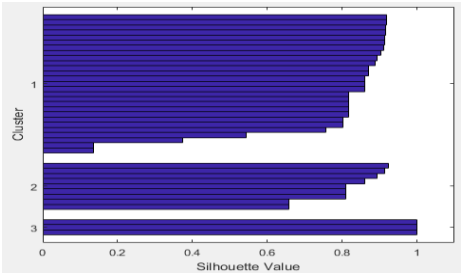

Figure 12(a). Silhouette index from cluster in partition I of five partitions

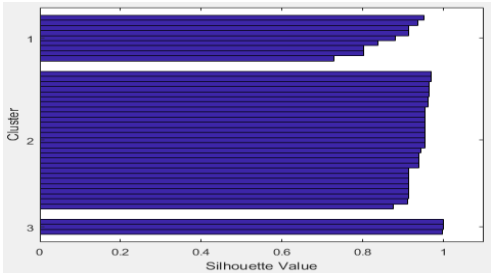

Figure 12(b). Silhouette index from cluster in partition II of five partitions

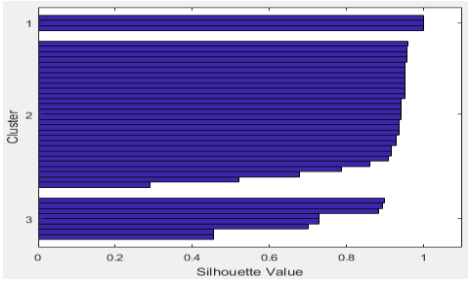

Figure 12(c). Silhouette index from cluster in partition III of five partitions

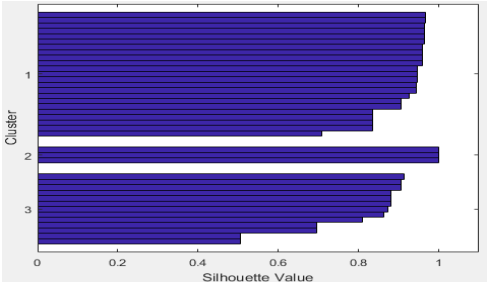

Figure 12(d). Silhouette index from cluster in partition IV of five partitions

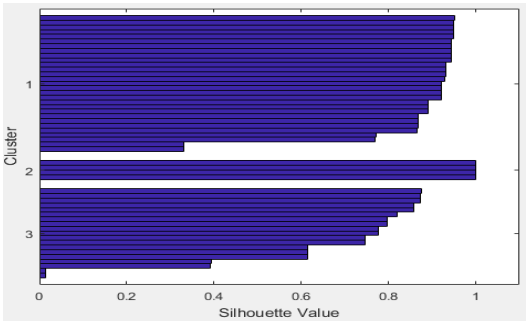

Figure 12(e). Silhouette index from cluster in partition V of five partitions

The silhouette index in each partition can be shown in the following Table 4a and Table 4b.

Table 4(a). Silhouette Index in Each Partitions on P1 (Two Partitions and Three Partitions)

	Two Partitions		Three Partitions		
Partition	I	II	I	II	III
Silhouette index	-0.001 ≤sh≤ 0.998	0.178 ≤sh≤ 1	0.123 ≤sh≤ 0.997	0.127 ≤sh≤ 0.99	-0.003 ≤sh≤ 1
Information	minus(-)	-	-	-	minus(-)

Table 4(b). Silhouette Index in Each Partitions on P1 (Four Partitions and Five Partitions)

	Four Partitions				Five Partitions				
Partition	I	II	III	IV	I	II	III	IV	V
Silhouette index	0.122 ≤sh≤ 0.997	0.079 ≤sh≤ 0.99	0.122 ≤sh≤ 1	0.013 ≤sh≤ 0.99	0.135 ≤sh≤ 0.99	0.728 ≤sh≤ 0.99	0.291 ≤sh≤ 0.99	0.506 ≤sh≤ 1	0.013 ≤sh≤ 0.99
Information	-	-	-	-	-	-	-	-	-

Features 3-Dimension P_2

The clustering results with K = 3 as anomaly detection model on P2 is shown in Figure 13. Computational anaysis on each partition on feature P2 are shown in the following Table 5a and Table 5b. Silhouette index in each partition can be shown in Table 6a and Table 6b.

Table 5(a). Computational Aspect in Each Partition on P2 (Two Partitions and Three Partitions)

Computational Aspect	Two Partitions		Three Partitions		
Partition	I	II	I	II	III
Replication	1	1	1	1	1
Iteration	5	1	3	2	1
Total sum of Distance	82379	2262	7210	47547	1969

Table 5(b). Computational Aspect in Each Partition on P2 (Four Partitions and Five Partitions)

Computational Aspect	Four Partitions				Five Partitions				
Partition	I	II	III	IV	I	II	III	IV	V
Replication	1	1	1	1	1	1	1	1	1
Iteration	2	2	1	1	1	1	1	1	1
Total sum of Distance	5086	39187	460	1756	1091	1250	4353	432	1756

Table 6(a).Silhouette Index in Each Partitions on P2 (Two Partitions and Three Partitons)

	Two Partitions		Three Partitions		
Partition	I	II	I	II	III
Silhouette index	-0.058≤sh≤0.901	0.452≤sh ≤1	0.293≤sh≤1	0.143≤sh≤1	0.413≤sh≤1
Information	minus (-)	-	-	-	-

Table 6(b). Silhouette Index in Each Partitions on P2 (Four Partitions and Five Partitions)

	Four Partitions				Five Partitions				
Partition	I	II	III	IV	I	II	III	IV	V
Silhouette index	0.197 ≤sh≤ 1	0.260 ≤sh≤ 1	0.594 ≤sh≤ 0.997	0.339 ≤sh≤ 1	0.235 ≤sh≤ 1	0.526 ≤sh≤ 0.986	0.215 ≤sh≤ 0.983	0.836 ≤sh≤ 1	0.339 ≤sh≤ 1
Information	-	-	-	-	-	-	-	-	-

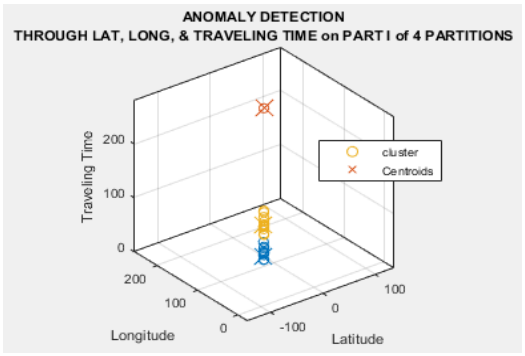

Figure 13(a). Clustering in partition I of four partitions

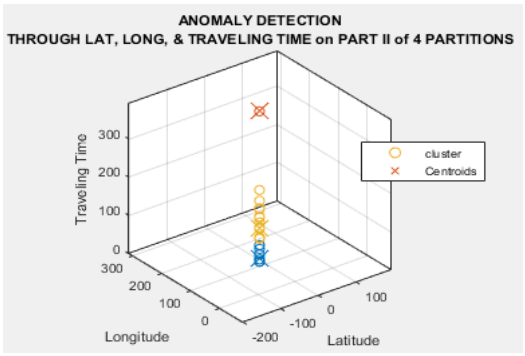

Figure 13(b). Clustering in partition II of four partitions

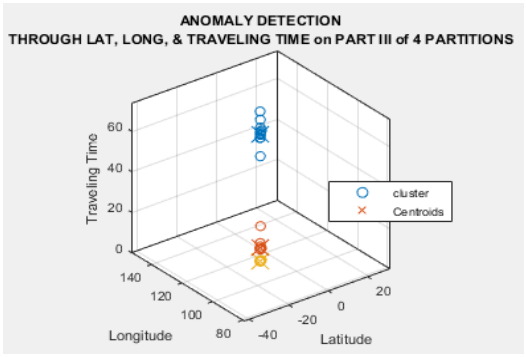

Figure 13(c). Clustering in partition III of four partitions

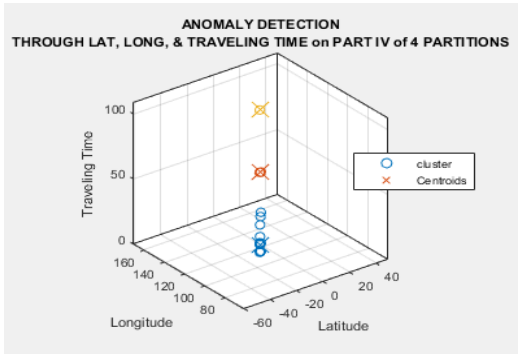

Figure 13(d). Clustering in partition IV of four partitions

For internal measurement of cluster validity based on the silhouette index value on P2, it is shown in the silhouette chart in Figure 14.

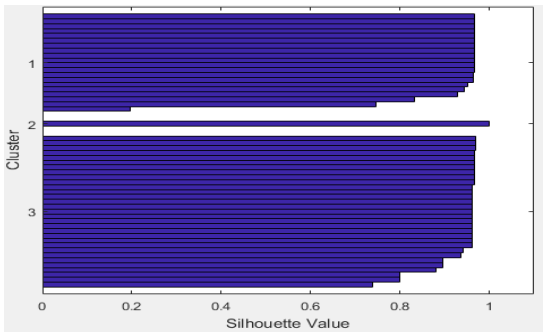

Figure 14(a). Silhouette index from cluster in partition I of four partitions

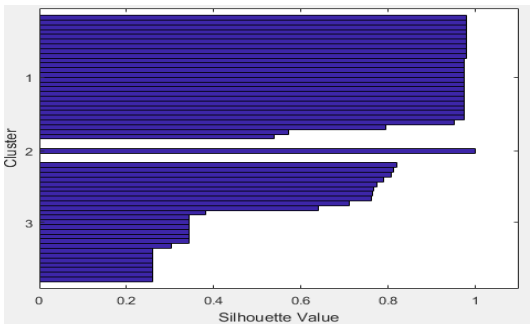

Figure 14(b). Silhouette index from cluster in partition II of four partitions

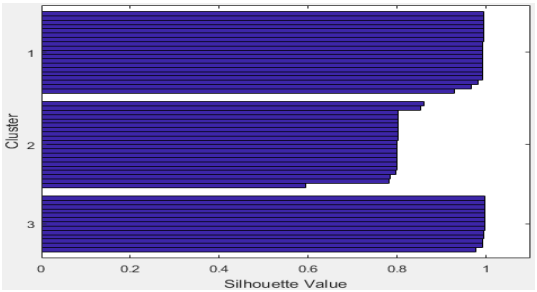

Figure 14(c). Silhouette index from cluster in partition III of four partitions

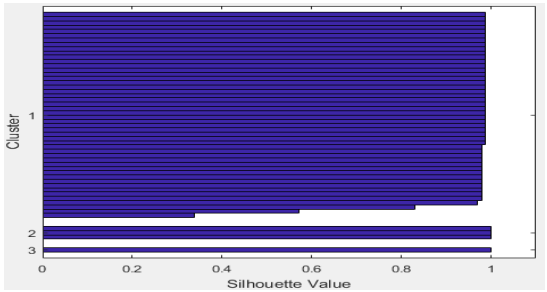

Figure 14(d). Silhouette index from cluster in partition IV of four partitions

For the determination of traveling time in each feature P1 and P2 are grouped by partition. In the P1 and P2 features obtained the same traveling time. The following in Table 7a and Table 7b shows the travel time per partition in each feature.

Table 7(a). Traveling Time Of P1 and P2 Per Partitions (Two Partitions and Three Partitions)

	Two Partitions		Three Partitions		
Partitions	I	II	I	II	III
Traveling Time per Partitions	7.5 (min)	2.8 (min)	3.9 (min)	3.9 (min)	1.7 (min)

Table 7(b). Traveling Time of P1 and P2 Per Partitions (Four Partitions and Five Partitions)

	Four Partitions				Five Partitions				
Partitions	I	II	III	IV	I	II	III	IV	V
Traveling Time per Partitions	3.1 (min)	3.6 (min)	1.7 (min)	0.7 (min)	2.2 (min)	1.7 (min)	1.9 (min)	1.1 (min)	0.7 (min)

4.3. False Identification Rate (FIR)

The False Identification Rate (FIR) measurement is applied to the 3-Dimension P1 and P2 features. It is for each feature there are two processes. First, calculate the minimum distance between the test data with the centroid in each cluster. Second, checking the test data in each record for ten days, whether it is appropriate to be in the selected cluster. If not true then FIR= 1 and correct then FIR = 0. The FIR graphic representation as a representation of the size of the identification error in the trial process. On the graph increases if the error identification is great, otherwise if the graph decreases indicates that the error identification is getting smaller. The following shows the FIR graph as the result of the error identification of the experiments that occurred on the 3D features.

For testing based on FIR percentage per partition for 3D features P1 is shown in the graph representation in Figure 15. For testing based on the percentage of FIR per partition for 3D P2 is shown in the graphical representation in Figure 16. Furthermore, the average percentage of FIR features 3D P1 and P2 per partition is shown in Figure 17.

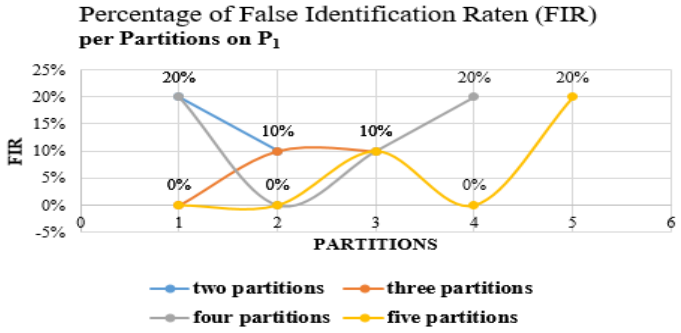

Figure 7. FIR graphics on each partition on 3D features P1

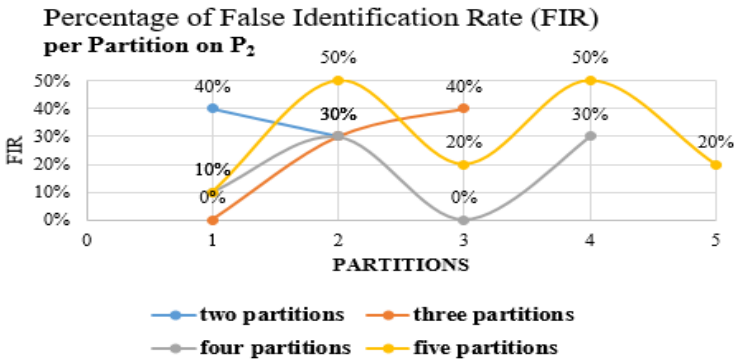

Figure 8. FIR graphics on each partition on 3D features P2

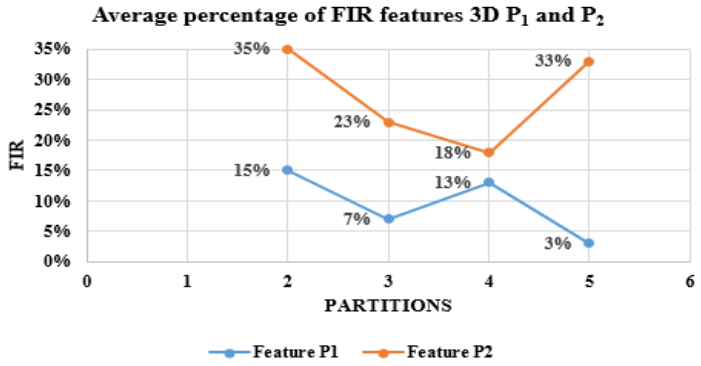

Figure 9. Graphics average Percentage of FIR P1 and P2

Based on the percentage of FIR per partition on the P1 feature, the optimal partition is obtained in five partitions. The optimal partition generated based on the minimum FIR percentage value is 3%. While the optimum partition in feature P2 occurs in four partitions with FIR percentage is 18%.

Based on FIR analysis for parameters P1 and P2, the determination of the optimum partition for detection of anomaly flight routes tends to be in parameter P1 (latitude, longitude, speed) compared to parameter P2 (latitude, longitude, traveling time). Therefore we need further analysis on P2 parameters, one of them is through optimum clustering by optimizing centroid initialization so that K-means clustering runs optimally.

5. CONCLUSION

In this paper, new methods are produced to determine the real-time aviation of anomaly detection based on ADS-B data with three-dimensional features. There are several steps: First, the determination of segment on the flight route is based on waypoint. Second, doing the partitioning process in each segment based on the partition, ie: n/2, n/3, n/4, and n/5. Third, the distance-based clustering with k = 3, so it obtains the point of centroid. Fourth, the distance measurement (Euclidean distance) is between the centroid one and the other centroid. The distant centroid indicates that the cluster is a potentially anomalous area. Fifth, testing an anomaly detection model. The indicator used in this research is to identify the False Identification Rate (FIR). The result for the 3D feature (P1) of the optimal partition is in five partitions with the average percentage of FIR = 3%. While for 3D feature (P2) it is obtained optimal partition in four partitions with average percentage of FIR = 18%.

The next work, it is possible to decrease the percentage of FIR P2 whose value is more than 10%. The reduction of FIR percentage is possible to conduct by using optimization at centroid initialization, so that the optimal cluster and FIR reach the minimum. In addition, the research development is done by real time anomaly detection approach. The time domain aspect of determining the detection area becomes the deciding factor. The use of clustering methods is to determine the optimum cluster results. Finally, for the anomaly detection process that enables large computing, a parallel computing process is required.

ACKNOWLEDGEMENTS

This research is supported by Indonesia Air Navigation Service (AirNav) Discrete Djuanda Surabaya. Specific in the Navigation Surveillance (CNS) / Air Traffic Management (ATM) section.

REFERENCES

[1] E. Chu, D. Gorinevsky, and S. Boyd, "Detecting aircraft performance anomalies from cruise flight data," *AIAA Infotech Aerosp. Conf.*, no. April, p. 3307, 2010.
[2] L. Li, M. Gariel, R. J. Hansman, and R. Palacios, "Anomaly detection in onboard-recorded flight data using cluster analysis," *2011 IEEE/AIAA 30th Digit. Avion. Syst. Conf.*, p. 4A4-1-4A4-11, 2011.
[3] M. Sudibyo and Kementerian Perhubungan, "Apakah AirAsia Penerbangan QZ8501 Jatuh Oleh Awan Cumulonimbus," 2015. [Online]. Available:
[4] https://ekliptika.wordpress.com/2015/01/26/apakah-airasia-penerbangan-qz8501-jatuh-oleh-awan-cumulonimbus/.
[5] R. Rastogi, S. Nahata, P. Ghuli, D. Pratiba, and G. Shobha, "Anomaly detection in log records," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 10, no. 1, pp. 343–347, 2018.
[6] K. R. Nirmal and K. V. V. Satyanarayana, "Issues of K means clustering while migrating to map reduce paradigm with big data: A survey," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 6, pp. 3047–3051, 2016.
[7] I. Wahyudin, T. Djatna, and W. A. Kusuma, "Cluster analysis for SME risk analysis documents based on Pillar K-Means.," *Telkomnika*, vol. 14, no. 2, pp. 674–683, 2016.
[8] H. Sabrol and S. Kumar, "Recognition of tomato late blight by using DWT and component analysis," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 1, pp. 194–199, 2017.
[9] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, "Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations," *J. Aerosp. Inf. Syst.*, vol. 12, no. 9, pp. 1–12, 2015.
[10] P. A. Denis Kolev, "ARFA: Automated Real-time Flight Data Analysis using Evolving Clustering, Classifier and Recursive Density Estimation," pp. 91–97, 2013.
[11] Z. S. Chang-Hun Lee, Hyo-Sang Shin, Antonio Tsourdos, "Anomaly Detection of Aircraft Engine in FDR (Flight Data Recorder) Data," *IET 3rd Int. Conf. Intelligent Signal Process. (ISP 2017)*, vol. 2, pp. 2–7, 2017.
[12] D. P. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," *Int. J. Adv. Intell. Informatics*, vol. 2, no. 1, pp. 38–45, 2016.
[13] S. G. Rao and A. Govardhan, "Performance Validation of the Modified K-Means Clustering Algorithm Clusters Data," vol. 6, no. 10, pp. 726–730, 2015.
[14] Z. Jun, L. I. U. Wei, and Z. H. U. Yanbo, "Study of ADS-B Data Evaluation," *Chinese J. Aeronaut.*, vol. 24, no. 4, pp. 461–466, 2011.
[15] K.-Y. Baek and H.-C. Bang, "ADS-B based Trajectory Prediction and Conflict Detection for Air Traffic Management," *Int. J. Aeronaut. Sp. Sci.*, vol. 13, no. 3, pp. 377–385, 2012.
[16] Z. Liang, Q. Li, and Z. Ren, "Waypoint constrained guidance for entry vehicles," *Aerosp. Sci. Technol.*, vol. 52, pp. 52–61, 2016.
[17] M. Y. Pusadan, J. L. Buliali, and R. V. H. Ginardi, "Anomaly Detection of Flight Routes through Optimal Waypoint," *IOP Conf. Ser. J. Phys. Conf. Ser. 801*, 2017.
[18] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodoligies," *Artif. Intell. Rev.*, vol. 22, no. 1969, pp. 85–126, 2004.
[19] M. Halkidi, "On Clustering Validation Techniques," pp. 107–145, 2001.
[20] R. Cordeiro, D. Amorim, and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Inf. Sci. (Ny).*, vol. 324, pp. 126–145, 2015.
[21] Y. S. Thakare, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics," vol. 110, no. 11, p. 8887, 2015.
[22] F. Kovács, C. Legány, and A. Babos, "Cluster Validity Measurement Techniques."
[23] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of Internal Clustering Validation Measures," 2010.
[24] L. Vendramin and E. R. Hruschka, "On the Comparison of Relative Clustering Validity Criteria ∗," pp. 733–744.
[25] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," vol. 20. pp. 53–65, 1987.
[26] D. I. Komputer, FMIPA, and Institut Pertanian Bogor, "Perbandingan Metode Cluster Validity pada Jenis Data Numerik dan Kategorik," 2013.
[27] D. L. D. & D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979.