Journal of Aerospace Information Systems, Vol. 15, Issue 5, 2018, pp. 297-306
DOI:10.2514/1.I010577

A New Application of Data Analysis using Aircraft Fault Record Data

Chang-Hun Lee*, Hyo-Sang Shin†, Antonios Tsourdos‡, and Zakwan Skaf§
Cranfield University, Cranfield MK43 0AL, UK

I. Introduction

In recent years, there have been great and continuous efforts in producing new values using data in various industries [1]. It is recognized that the aerospace industry is one of the areas having greatest potential in exploiting the values of data [2]. The operation and maintenance of aircraft generate various types of data in extensive quantities such as the digital flight data recorder (FDR) data [3], aircraft condition monitoring system (ACMS) data, weather data, maintenance data, fault record data, aircraft safety reports, and pilot reports (PIREP) [4]. The value of these data comes from the fact that they contain evidence of potential issues in the aircraft operation and maintenance. This implies that the results of data analysis can be used to improve aircraft stability and aircraft operation/maintenance process.

Thanks to its potential, airline companies have a number of dedicated departments for data analysis, endeavouring to fully exploit values of data. However, many of these companies used to manually handle data and find difficulties in analyzing vast amount of data. Automating data analysis in the aerospace industry has drawn a great attention as it could greatly assist engineers and operators. Accordingly, in recent years, research has been actively conducted to improve aircraft operation and maintenance process by automatically analyzing extensive amounts of data. In previous research [1, 5–8], the detection of abnormal operations or flights due to mechanical or human factor problems was performed using the FDR data. Data analysis development has been proposed to detect unexpected trends in maintenance demands using maintenance history log data [9]. In previous study [10], the authors have suggested the data analytics to detect failure modes using ACMS dataset as well as maintenance report dataset. Data analysis method has been proposed to estimate required man-hours for certain types of maintenance tasks [11]. In the air safety report [12] and maintenance text message [13], a text mining technique has been devised to extract new information. In previous works [14, 15], a text mining algorithm has been devised to classify maintenance reports. Using repair messages, the author of [16] proposed a data analysis approach to investigate the relationship between the nature of damage and the time span of repair services. The authors in [17] suggested a method of identifying recurring faults and classifying maintenance reports.

Accordingly, based on the above-mentioned aspect, this study aims to propose automatic data analysis framework

*Research Fellow, School of Aerospace, Transport and Manufacturing, College Road, Cranfield, Bedfordshire, MK43 0AL, UK
†Reader, School of Aerospace, Transport and Manufacturing, College Road, Cranfield, Bedfordshire, MK43 0AL, UK
‡Professor, School of Aerospace, Transport and Manufacturing, College Road, Cranfield, Bedfordshire, MK43 0AL, UK
§Lecturer, Integrated Vehicle Health Monitoring (IVHM) Centre, College Road, Cranfield, Bedfordshire, MK43 0AL, UK

for a new type of data (i.e., fault history record data of aircraft fleets called airman data provided by TAP Portugal), with the purpose of supporting aircraft operation and maintenance process. Generally, certain types of faults occur frequently on certain aircraft if there are potential problems with relevant aircraft or if there exist potential problems in the corresponding maintenance process. Because the fault history record data are most likely to contain the evidence for this information, the potential problems can be indirectly identified through analysis of the occurrence pattern of faults. Since the analysis results of this data can provide different information from the existing viewpoint, it can be used as a complement to conditional based maintenance (CBM) or prognostics health management (PHM) approaches to improve aircraft operation and maintenance process.

In this paper, we extend the concept of the research work [9] and develop new methods to extract useful information from the fault history record data. The key enablers for revealing the useful information will be selection of key features and data analysis approaches. Therefore, the focus of our analysis is design of key features and selection of appropriate data analytics, given the fault record data. The baseline approach adapted in this study is pattern analysis of the fault occurrences. The pattern in this study refers to the distribution of aircraft fault frequency across the aircraft fleet for each fault type. The analysis then identifies the candidates of fault types exhibiting abnormal patterns in the distribution. To this end, the pattern is first modeled as a probability mass function (PMF), and the similarity between PMFs is measured using the probability-based distance such as the Bhattacharyya distance [18]. Then, similar patterns and unusual patterns are analyzed using the DBSCAN (density-based spatial clustering of applications with noise) [19] clustering technique. The pattern analysis results can provide information on whether there is a potential problem in the maintenance phase of the aircraft or whether a specific fault type is a fault associated with a specific aircraft only.

This study also performs correlation analysis: the correlation of occurrence time patterns for each fault type is investigated. First, the occurrence time patterns for each fault type is converted to a binary vector. The similarity between binary vectors is then measured using the correlation-based distance such as the Jaccard coefficient [20]. If a certain type of faults is correlated to other types of faults in time, the proposed correlation analysis should be able to uncover their correlation and also the time interval in their occurrences. Therefore, the correlation analysis will enable detection of the occurrence possibility of the correlated faults in advance and consequently introduction of proactive measures to mitigate impact of those faults. Finally, in this paper, we provide illustrative examples of the proposed analysis methods using real fault record data in order to show the validity of proposed methods.

The composition of this paper is as follows. In Section II, the data preprocessing and the notations are explained. Section III provides the pattern analysis method. In Section IV, the correlation analysis method is described. In Section V, illustrative examples of the proposed methods are provided. Finally, in Section VI, we conclude our study.

RECORD ID	EVENT_DATE	TAIL NUMBER	⋯	FAULT TYPE
1	1/01/2016	AIRCRAFT #1	⋯	FAULT #2
2	2/01/2016	AIRCRAFT #3	⋯	FAULT #3
3	2/01/2016	AIRCRAFT #1	⋯	FAULT #1
4	3/01/2016	AIRCRAFT #3	⋯	FAULT #2
⋮	⋮	⋮	⋮	⋮
N	1/01/2017	AIRCRAFT #5	⋯	FAULT #10

Fig. 1 A sample of fault record dataset

II. Data Preparing and Notations

A. Data Preparing

The dataset to be analyzed contains history of faults of commercial aircraft fleet in the past several years. This study uses only fault records for a specific period of interest. Fig. 1 shows a sample of dataset: each row represents a fault record and each column an attribute of the data. These data attributes include when a fault occurred, the type of fault, and aircraft tail number related to the fault.

Two types of data errors are observed in the raw dataset: missing values and incorrect values associated with the time and date of the fault. It also contains a log record of checking part status that is not relevant to the goal of the data analysis. From the raw dataset, it is evident that there exist other data attributes that can additionally provide time and date information. Therefore, data errors related to time and date can be handled with the technique called imputation using companion data. Also, analysis on the raw dataset revealed that text messages for unnecessary start with a specific word. Thus, we use a string comparison technique to remove the unnecessary data.

The raw dataset has several hundreds of fault types. Since there are not enough records for some types of faults, analyzing these types of faults may lead to incorrect information. Therefore, it would be necessary to sample fault types that have a sufficient number of records to provide meaningful information. We first sort the number of records for each fault type in descending order and then select N_F fault types that capture p% of the number of records, i.e.

$$\frac{\sum_{i=1}^{N_F} \alpha_i}{\sum_{j=1}^{N_T} \alpha_j} > p \tag{1}$$

where N_T is the total number of fault types, α_i is the number of records for the i-th fault type. In this study, we choose the value of p as 0.9.

B. Notations of Dataset

This section defines several notations of the dataset using set theory to illustrate the proposed method. First, let us define the dataset D. Each row of D, called a record, is defined as $d \in \mathbb{D}$, where $\mathbb{D}$ represents the data space. The record d can be a column vector, and each item in this vector corresponds to each data attribute. The dataset can then be expressed as

$$D = \{ d | \, d \in \mathbb{D} \} \tag{2}$$

The size of D (i.e., $|D|$) is denoted by N_D, which is the number of records. The notation of d_i is the i-th record in the dataset. Similarly, a set of fault types F and a set of fleet aircraft A are defined as follows

$$F = \{ f | \, f \in \mathbb{F} \} \tag{3}$$

$$A = \{ a | \, a \in \mathbb{A} \} \tag{4}$$

The sizes of F and A are denoted as N_F and N_A, respectively. The parameters f_i and a_j represent the i-th fault type and the j-th aircraft in the fleet.

Let $g : D \rightarrow F$ be defined as a mapping function from a record $d \in D$ to a fault type $f \in F$, i.e.:

$$g(d) = f \tag{5}$$

This function indicates which fault type is involved in a given record. In the same way, let $h : D \rightarrow A$ be defined as a mapping function from a record $d \in D$ to an aircraft $a \in A$, that is:

$$h(d) = a \tag{6}$$

The role of this function is to indicate which aircraft is involved in a given record.

III. Pattern Analysis for Fault Record Data

A. The Concept of Pattern Analysis

The pattern considered in this study is the distribution of aircraft fault frequency across the aircraft fleet. The term of aircraft fault frequency is defined as the ratio of the total number of fault records for an individual aircraft to that for all aircraft. The pattern analysis is performed to identify the candidates of fault types that exhibit abnormal patterns

among all fault types. The fundamental rationales behind the pattern analysis are given as

1) The fault frequency for a certain aircraft is proportional to the usage of that aircraft.
2) Although the quality of maintenance for the same fault type slightly varies depending on the aircraft in practice, the probability of fault occurrence for the same fault type is similar across the aircraft fleet in the average sense in normal circumstance.
3) Although the cycles of fault occurrence for each fault type are different, the distributions of fault frequency across the aircraft fleet have similar pattern for all fault types in average sense as the fault records are accumulated over a specific time period, under the premises of (1) and (2)

Note that these assumptions can be accepted as common sense. Based on these assumptions, therefore, if a specific fault type f exhibits greater pattern difference, compared with other fault types, the type f could be considered as abnormal. This result means that the specific fault type f is mainly caused in the specific aircraft, which is not simply explained in common sense. It implies that there might be a systematic potential problem in the maintenance phase of the aircraft associated with a specific fault type. Therefore, the proposed method provides an early warning for this potential problem and the airline companies can use this information to improve the maintenance process.

B. Process and Technique of Pattern Analysis

1. Feature Generation

The pattern analysis uses the fault frequency as features. To this end, we first define a set of records for the i-th fault type and the j-th aircraft, denoted by $D_i^j \subset D$, as follows:

$$D_i^j = \left\{ d \middle| \, d \in \mathbb{D} \text{ and } g(d) = f_i \text{ and } h(d) = a_j \right\} \tag{7}$$

The size of D_i^j is defined as N_i^j. In a similar way, a set of records for the i-th fault type, denoted by $D_i \subset D$, is defined as follows:

$$D_i = \{ d \middle| \, d \in \mathbb{D} \text{ and } g(d) = f_i \} \tag{8}$$

We then define the number of records in D_i as N_i

$$N_i = \sum_{j=1}^{N_A} N_i^j \tag{9}$$

By definition, the total number of records N_D should be

$$N_D = \sum_{i=1}^{N_F} N_i \tag{10}$$

Then, we define the fault frequency of the j-th aircraft with respect to the i-th fault type as follows.

$$w_i^j = \frac{N_i^j}{N_i} \tag{11}$$

For each fault type and each aircraft, this feature can be obtained. Here, the feature w_i^j can be considered as the probability that a record in the set D_i belongs to the set D^j as follows:

$$\Pr\left[d \in D^j \middle| d \in D_i \right] = w_i^j \tag{12}$$

The physical meaning of this feature is the probability that the i-th fault type occurs in the j-th aircraft. By definition, this parameter should satisfy the following condition.

$$\sum_{j=1}^{N_A} w_i^j = 1 \tag{13}$$

2. Pattern modeling

This section discusses pattern modeling using the probability mass function (PMF) and the features obtained. First, suppose that the state space is discrete and the discrete states are the index of the aircraft in the fleet as follows:

$$\underline{x} = [1, 2, \ldots, j, \ldots, N_A] \tag{14}$$

Next, we regard the obtained features as weights of discrete states. Then, the pattern of fault records across the aircraft fleet can be modeled as PMF as shown in Fig. 2, which is assumed to be a discrete probability distribution of the form:

$$p_i\left(\underline{x}\right) = \sum_{j=1}^{N_A} w_i^j \delta\left(\underline{x} - j\right) \tag{15}$$

where $\delta(x)$ denotes the Dirac peak. Here, the pattern for the i-th fault type is given by Eq. (15). We obtain N_F different PMFs for each fault type.

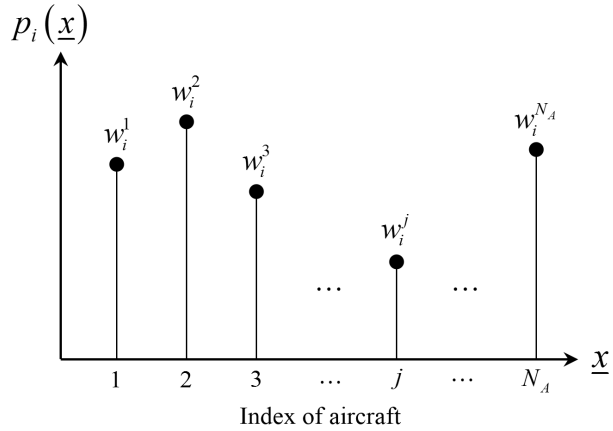

Fig. 2 The pattern modeling using probability mass function

3. Pattern Analysis

The pattern analysis is performed by comparing similarities of each PMF. In this study, the Bhattacharyya distance [18] is used to measure the similarity of two PMFs. For given two PMFs $p_i\left(\underline{x}\right)$ and $p_j\left(\underline{x}\right)$ over the same domain $\underline{x}$, the Bhattacharyya distance is defined as follows:

$$D_B\left(p_i, p_j\right) = -\ln\left[BC\left(p_i, p_j\right)\right] \tag{16}$$

where $BC\left(p_i, p_j\right)$ represents the Bhattacharyya coefficient for two PMFs, which is defined as

$$BC\left(p_i, p_j\right) = \sum_{x \in X} \sqrt{p_i\left(\underline{x}\right) p_j\left(\underline{x}\right)} \tag{17}$$

where X is the discrete state space. The physical meaning of this coefficient is a measure of the amount of overlap between two PMFs. If this coefficient is unity, then the Bhattacharyya distance is 0. On the other hand, when this coefficient is 0, then the Bhattacharyya distance becomes infinity. From the definition, the Bhattacharyya coefficient can be expressed by using the weights of discrete states of the proposed pattern model, i.e.

$$BC\left(p_i, p_j\right) = \sum_{k=1}^{N_A} \sqrt{w_i^k w_j^k} \tag{18}$$

We then determine the distance $\delta_{i,j}$ between the pattern of the i-th fault type and the j-th fault type as follows

$$D_B\left(p_i, p_j\right) = \delta_{i,j} = -\ln\left[\sum_{k=1}^{N_A} \sqrt{w_i^k w_j^k}\right] \tag{19}$$

This distance is obtained for each fault type.

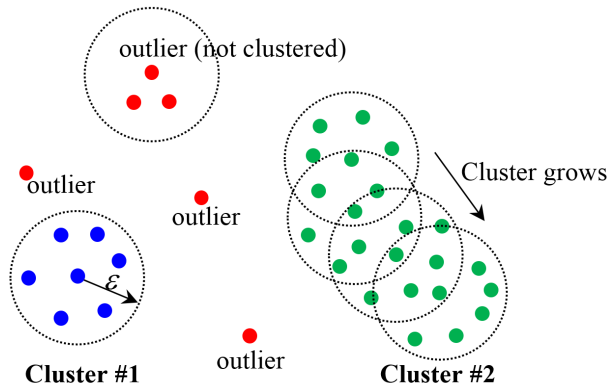

Fig. 3 Illustration of the DBSCAN clustering process

Next, a clustering algorithm called the DBSCAN (density-based spatial clustering of applications with noise) [19] is applied in order to identify fault types with significant deviations. The DBSCAN is a density-based clustering algorithm that can simultaneously provide clustering of similar patterns and detection of abnormal patterns. Fig. 3 shows the illustration of the DBSCAN clustering process. The design parameters are two: the minimum number of points N_{min} and the radius ϵ, which decide the density criteria. Then, clusters are identified by a selected density criteria. If at least N_{min} points are placed within ϵ radius of a circle, a cluster is formed as can be seen in Fig. 3. The cluster then grows by adding adjacent points that meet the density criteria (see cluster 2 in Fig. 3). If there are no other points to be added, then the cluster stops growing. A group of unclustered points are classified as outliers, which is considered as a point that exhibits different pattern compared with other points.

Remark 1. The premises of this analysis can be generally acceptable for fault record data in aircraft as well as in other systems. Thus, the proposed method can be applicable to any type of fault record data.

Remark 2. When new fault records are added over time, the absolute patterns for each fault type will change. However, since the proposed method is based on measuring the relative difference of the patterns between each fault type, the patterns that change with time do not affect the effectiveness of proposed method.

Remark 3. In the pattern modeling, the patterns obtained change according to the index of PMF. However, Bhattacharyya coefficient which is used for the similarity measure for given two patterns is the same even though the index is differently labeled. Accordingly, changing of index label does not affect the effectiveness of proposed method.

IV. Correlation Analysis of Fault Record Data

A. The Concept of Correlation Analysis

The fundamental rationale behind the correlation analysis is that if the occurrence of fault type f_A accompanies to the occurrence of the fault type f_B at the same time or in the near future, there is a correlation in the two fault types. Note

that such correlation information can be used for a high-level fault prognosis. If the fault type f_A occurs, we can predict whether or not the fault type f_B is likely to occur at the same time or in the near future from the correlation information. Based on this rationale, this paper analyzes the correlation of occurrences of individual fault types. Since the future occurrence event time of f_B with respect to the occurrence of f_A is unknown, a process of matching different occurrence event times of faults is additionally considered in this study compared to the conventional correlation analysis [21].

B. Process and Technique of Correlation Analysis

1. Feature Generation

The feature chosen is a binary vector $\underline{y}_i^j$ indicating occurrences of the fault as follows:

$$\underline{y}_i^j = \left[y_1^{(i,j)}, y_2^{(i,j)}, \ldots, y_k^{(i,j)}, \ldots, y_{N_k}^{(i,j)} \right] \tag{20}$$

where i and j represent the i-th fault type and the j-th aircraft and each entry of this vector has a binary value, that is $y_k^{(i,j)} \in \{0, 1\}$, according to the occurrences of faults. N_K is the length of vector. For the j-th aircraft, N_F binary vectors for each fault type can be obtained.

2. Time Shifting of Binary Vector

For convenience, when analyzing correlation of fault type f_B with respect to fault type f_A, we define f_A as the reference fault type and f_B as the dependent fault type. Hereafter, we define the notation i_r as the index of the reference fault type and the notation i_d as the index of the dependent fault type.

Since the difference of occurrence event times between f_A and f_B is unknown, a time shift is first applied to the dependent vector. The window size for the time shift is chosen as $W \geq 0$ to account for future occurrence of dependent fault type. At the j-th aircraft, a time shifted vector with W, $\underline{z}_{i_d}^j$, is obtained for each dependent vector.

$$\underline{z}_{i_d}^j = \left[z_1^{(i_d,j)}, z_2^{(i_d,j)}, \ldots, z_k^{(i_d,j)}, \ldots, z_{N_K-W}^{(i_d,j)} \right], \quad \text{for} \quad i_d = 1, \ldots, N_F \tag{21}$$

where

$$z_k^{(i_d,j)} = y_{k+W}^{(i_d,j)}, \quad \text{for} \quad k = 1, \ldots, N_K - W \tag{22}$$

Note that determining the window size is not straightforward because it is unknown at which time the dependent fault type will occur. In this study, for given a reference fault type and a dependent fault type, the window size W is iteratively determined, which will be discussed in the following section.

3. Correlation Analysis

Finally, the correlation analysis is performed to determine the Jaccard coefficient [20] between the reference vectors (i.e. $\underline{y}^j_{i_r}$, for $i_r = 1, \ldots, N_F$) and the time shifted dependent vectors (i.e., $\underline{z}^j_{i_d}$, for $i_d = 1, \ldots, N_F$) with every W.

$$J^j_{i_d|i_r} = \frac{N_{11}^{(i_d|i_r,j)}}{N_{11}^{(i_d|i_r,j)} + N_{01}^{(i_d|i_r,j)} + N_{10}^{(i_d|i_r,j)}} \tag{23}$$

where $N_{11}^{(i_d|i_r,j)}$, $N_{10}^{(i_d|i_r,j)}$, and $N_{01}^{(i_d|i_r,j)}$ are defined as the number of pairs being represented by $(1, 1)$, $(1, 0)$, and $(0, 1)$, respectively. The Jaccard coefficient physically represents a measure of overlap between the two binary vectors. If the two binary vectors are fully correlated, then $J^j_{i_d|i_r} = 1$. If there is no correlation, then $J^j_{i_d|i_r} = 0$. For given a reference fault type and dependent fault type, the process described is repeatedly performed by incrementally changing W. Then, the Jaccard coefficient of the maximum value is selected the correlation coefficient.

V. Illustrative Examples

This section provides illustrative examples of the proposed data analysis using the real fault record data for Airbus A330 fleet. The fault data records from the past 2012 to 2015 are used for the analysis.

A. Example of Pattern Analysis

In the pattern analysis, the design parameters of DBSCAN are first set based on the sensitivity analysis in a similar way to the one in [8]. Fig. 4 depicts the detection rate of outliers with respect to changes in the design parameters of DBSCAN. In this figure, the x-axis represents neighborhood radius and the y-axis is the detection rate of outliers (fault types having unusual patterns). The sensitivity analysis indicates that the detection rate of outliers identified is insensitive to N_{min}. On the other hand, the selection of ϵ significantly affects the detection rate of outliers. In the analysis, we set $N_{min} = 5$ and $\epsilon = 0.3$ to find top 5% outliers based on the sensitivity analysis results shown in Fig. 4.

Then, we perform DBSCAN with those parameters to identify similar patterns and abnormal patterns, respectively. In the proposed pattern modelling, the similarity between each PMF can be viewed through the weights of each PMF. Then, the weights of each PMF should be presented in a single graph to compare each PMF graphically. However, since it is difficult to visualize the weights of each PMF (which are given by high-dimensional vectors) using an ordinary graph, a stacked bar graph is adopted in this study to visualize the weights of each PMF in a single graph. In this graph, the x-axis represents the index of each fault type and the y-axis represents the cumulative sum of weights of each PMF. Namely, the distribution of each PMF is represented by color composition in the stacked bar graph. PMFs with a similar distribution have a similar color composition.

Clustered patterns for each fault type are illustrated using the stacked bar graph shown in Fig. 5. Through pattern analysis, two clusters denoted by Cluster #1 and Cluster #2 are identified as shown in Fig. 5. These results imply that

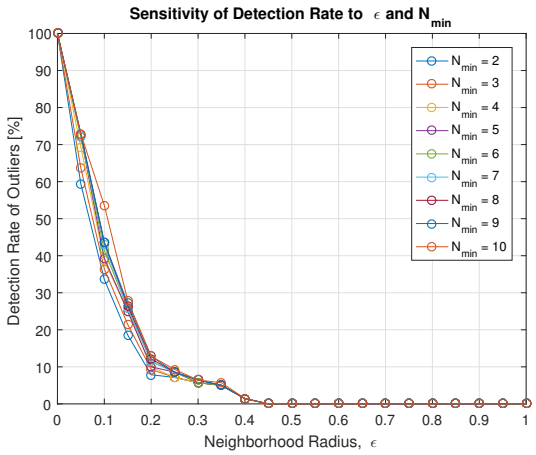

Fig. 4 Sensitivity analysis to radius and minimum points

those fault types have more similar characteristics compared to other fault types. Additionally, 8 outliers are identified through this analysis as shown in Fig. 6. These are Fault Type #31, #42, #64, #35, #54, #36, #60, and #65, respectively. From the results, we can readily observe that these fault types have frequently occurred at the specific aircraft mainly, which is not common. Therefore, these results imply that there may be potential problems associated with these fault types in relevant aircraft or there may be potential problems in the maintenance process associated with such fault types.

B. Example of Correlation Analysis

Now, let us examine correlations between each fault type using the time series of fault occurrence for the aircraft fleet. Fig. 7 shows an example of the time series of fault occurrence for all fault types on the Aircraft #1. In this figure, the x-axis represents the time unit and the y-axis represents the index of each fault type. A circle marker indicates that a specific fault type occurred at a certain time. Through computing the Jaccard coefficient given in Eq. (23), we can quantify correlations between each fault type. Here, by imposing a threshold on the Jaccard coefficient, correlated fault types can be identified. In this study, when the time series of fault occurrence of two fault types has a similarity of 60% or more, the two fault types are considered to be correlated with each other. As a result, a total of 9 pairs of correlated fault types are identified. As a selective example, Fig. 8 provides the time series of correlated faults (Fault Type #56 and #58, Fault Type #99 and #102, Fault Type #21 and #75, Fault Type #77 and #81). As shown in Fig. 8, accordingly, we can readily observe that these occurrence patterns are very similar each other, which confirm the performance of the proposed correlation analysis technique. Although we describe in this paper only the Aircraft #1 results, the correlation analysis can be applied in the same way for other aircraft in the fleet. As a result of applying to all aircraft, fault types with a total of 31 pairs of correlations are identified. As discussed, these results can be utilized for a high-level prognosis of fault occurrence.

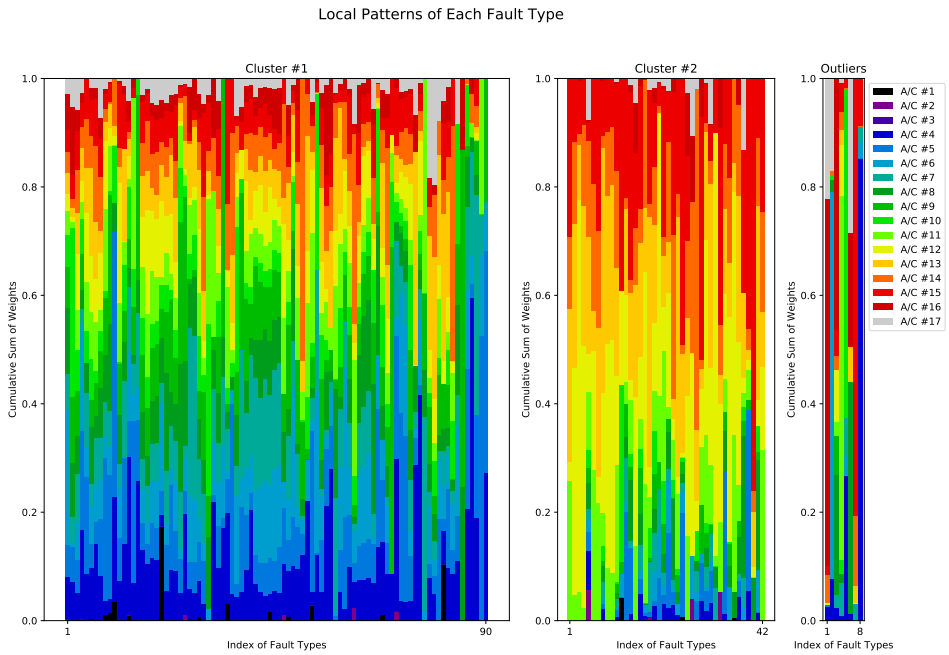

Fig. 5 Patterns for each fault type

VI. conclusions

In this paper, we proposed a new application of data analysis that detects potential problems in aircraft maintenance operations or in aircraft itself using aircraft fault record data. To do this, we proposed pattern analysis and correlation analysis as a data analysis technique for extracting important information contained in aircraft fault record data.

In the pattern analysis, abnormal patterns are identified by comparing the patterns of the fault frequency of aircraft across the aircraft fleet for each fault type. At this time, the pattern of fault frequency of aircraft was modeled by probability mass function, similarity between probability distribution was measured by the Bhattacharyya distance, and abnormal pattern was finally identified by using clustering method. This information implies that there is a potential problem in the aircraft maintenance operation and in the aircraft itself, and this information can be passed on to operators and used to improve the aircraft maintenance process. In the correlation analysis, time series data of the occurrence pattern was extracted for each fault type, and then this information was converted into a binary vector. The correlation of the two fault types was analyzed by comparing the similarity of the binary vector. In that case, the Jaccard coefficient was used as a method to measure the correlation between two fault types. Such correlation information of two fault types can be used as a high-level fault prognosis that predicts the possibility of another fault occurrence when one fault occurs.

In this paper, we investigated the effectiveness of the proposed application and the proposed method using the real fault record data. Through the proposed pattern analysis technique, it is confirmed that fault types with abnormal patterns are well identified. In addition, we confirmed that the proposed correlation analysis technique effectively

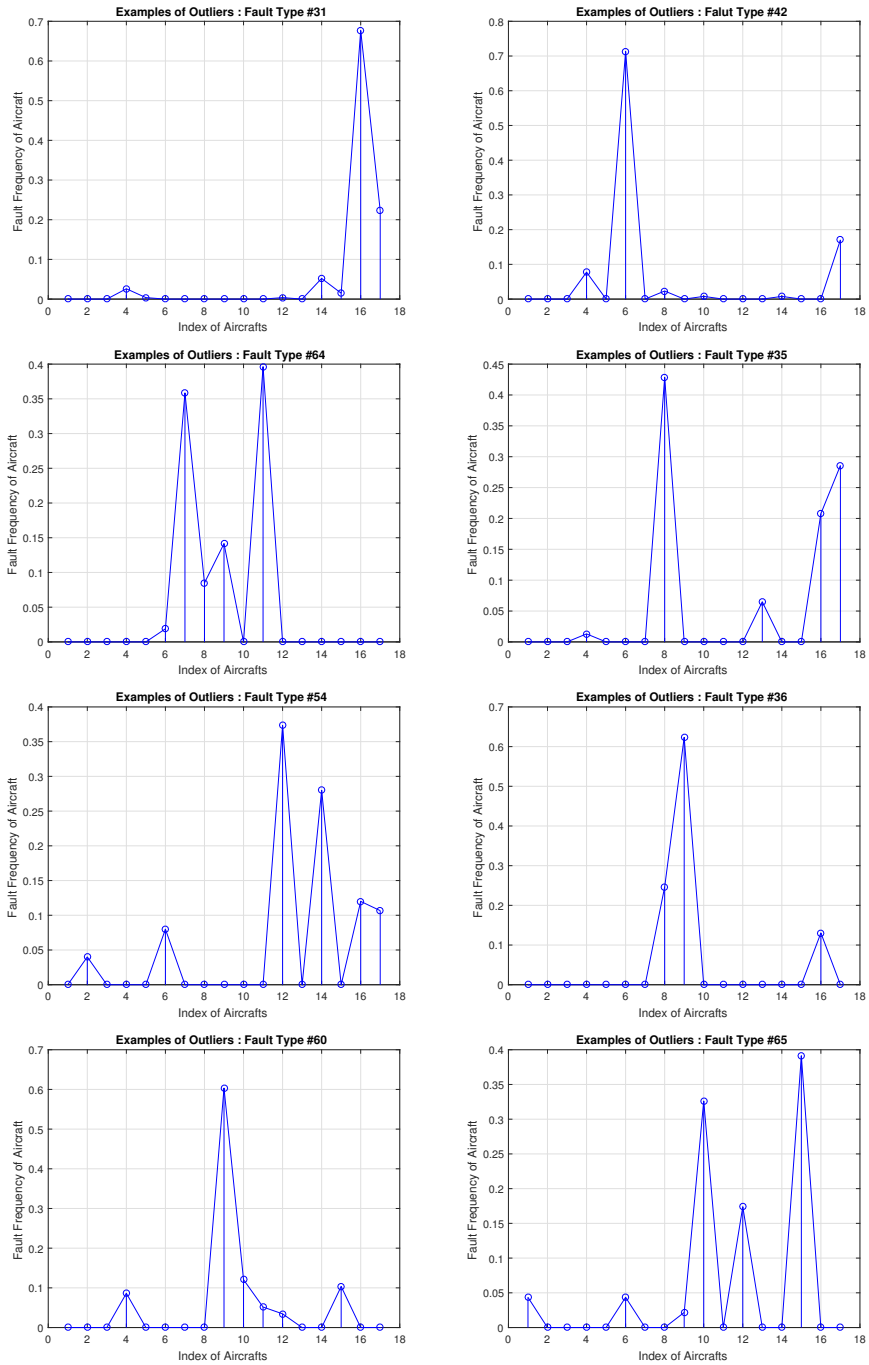

Fig. 6 Examples of outliers obtained from clustering analysis

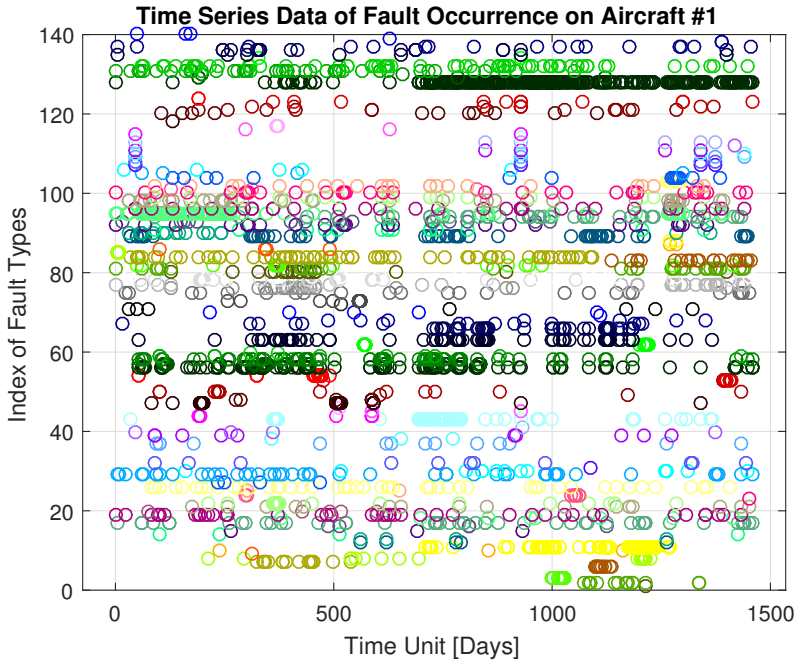

Fig. 7 Time series data of fault occurrence for all fault types on the Aircraft #1

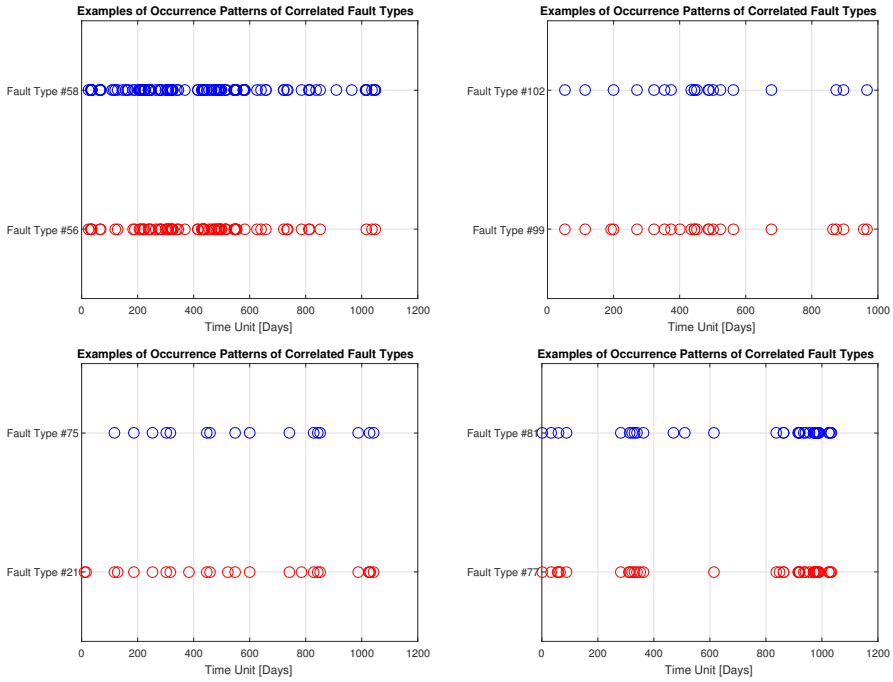

Fig. 8 Examples of occurrence patterns of correlated fault types on the Aircraft #1

identifies similar fault types with time series data of occurrence patterns.

Acknowledgment

This work has been performed under the AIRMES (Airline Maintenance Operations implementation of an E2E Maintenance Service Architecture and its enablers) project. This project has received funding from the Clean Sky 2 Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 681858.

References

[1] Chen, H., Chiang, R. H., and Storey, V. C., "Business intelligence and analytics: From big data to big impact," *MIS quarterly*, Vol. 36, No. 4, 2012, pp. 1165–1188.

[2] Larsen, T., "Cross-platform aviation analytics using big-data methods," *Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013*, IEEE, 2013, pp. 1–9. doi:10.1109/ICNSurv.2013.6548579.

[3] Stephenson, E. A., "Aircraft flight data recorder data acquisition system," , Apr. 7 1987. US Patent 4,656,585.

[4] Schwartz, B., "The quantitative use of PIREPs in developing aviation weather guidance products," *Weather and Forecasting*, Vol. 11, No. 3, 1996, pp. 372–384. doi:10.1175/1520-0434(1996)011<0372:TQUOPI>2.0.CO;2.

[5] Budalakoti, S., Srivastava, A. N., and Akella, R., "Discovering atypical flights in sequences of discrete flight parameters," *Aerospace Conference, 2006 IEEE*, IEEE, 2006, pp. 1–8. doi:10.1109/AERO.2006.1656109.

[6] Das, S., Matthews, B. L., Srivastava, A. N., and Oza, N. C., "Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study," *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2010, pp. 47–56. doi:10.1145/1835804.1835813.

[7] Sriastava, A., et al., "Discovering system health anomalies using data mining techniques," .

[8] Li, L., Das, S., John Hansman, R., Palacios, R., and Srivastava, A. N., "Analysis of flight data using clustering techniques for detecting abnormal operations," *Journal of Aerospace Information Systems*, Vol. 12, No. 9, 2015, pp. 587–598. doi: 10.2514/1.I010329.

[9] Dubrawski, A., and Sondheimer, N., "Techniques for early warning of systematic failures of aerospace components," *Aerospace Conference, 2011 IEEE*, IEEE, 2011, pp. 1–9. doi:10.1109/AERO.2011.5747589.

[10] Mack, D. L., Biswas, G., Koutsoukos, X. D., Mylaraswamy, D., and Hadden, G., "Deriving Bayesian Classifiers from Flight Data to Enhance Aircraft Diagnosis Models," *Annual Conference of the Prognostics and Health Management Society*, 2011.

[11] Mathur, A., "Data mining of aviation data for advancing health management," *AeroSense 2002*, International Society for Optics and Photonics, 2002, pp. 61–71. doi:10.1117/12.475495.

[12] Pena, J. M., Famili, F., and Létourneau, S., "Data mining to detect abnormal behavior in aerospace data," *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2000, pp. 390–397. doi:10.1145/347090.347173.

[13] McKenzie, A., Matthews, M., Goodman, N., and Bayoumi, A., "Information extraction from helicopter maintenance records as a springboard for the future of maintenance text analysis," *Trends in Applied Intelligent Systems*, 2010, pp. 590–600. doi:10.1007/978-3-642-13022-9_59.

[14] Clements, R., Morse, J., Darr, D., and Laskowski, B., "Meta-data mining for optimized aircraft repair and overhaul," *International symposium on NDT in aerospace 2010*, NDT in Aerospace, 2010.

[15] Yu, J., and Gulliver, S., "Improving aircraft maintenance, repair, and overhaul: A novel text mining approach," *International conference on intelligent computing and intelligent systems*, 2011.

[16] Fu, M., Lu, R. F., Storch, R. L., and Kirkham, C. J., "Extracting and analyzing information from a large volume of aircraft repair messages," *POMS 21st annual conference*, 2010.

[17] Srivastava, A. N., and Zane-Ulman, B., "Discovering recurring anomalies in text reports regarding complex space systems," *Aerospace conference, 2005 IEEE*, IEEE, 2005, pp. 3853–3862.

[18] Bhattachayya, A., "On a measure of divergence between two statistical population defined by their population distributions," *Bulletin Calcutta Mathematical Society*, Vol. 35, No. 99-109, 1943, p. 28.

[19] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*, Vol. 96, 1996, pp. 226–231.

[20] Jaccard, P., "The distribution of the flora in the alpine zone," *New phytologist*, Vol. 11, No. 2, 1912, pp. 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x.

[21] Runkler, T. A., *Data Analytics*, Springer, 2012.