Journal of Aviation/Aerospace Education & Research

Volume 33 | Number 2 Article 10

2024

Integrated Organizational Machine Learning for Aviation Flight Data

Michael J. Pritchard Ph.D.
Kansas State University, mjp001@ksu.edu

Austin T. Walden Ph.D.
Kansas State University, austinwalden@ksu.edu

Paul J. Thomas Ph.D.
Purdue University, pjthomas@purdue.edu

Abstract

Increased availability of data and computing power has allowed organizations to apply machine learning techniques to various fleet monitoring activities. Additionally, our ability to acquire aircraft data has increased due to the miniaturization of small form factor computing machines. Aircraft data collection processes contain many data features in the form of multivariate time series (continuous, discrete, categorical, etc.), which can be used to train machine learning models. Yet, three major challenges still face many flight organizations: 1) integration and automation of data collection frameworks, 2) data cleanup and preparation, and 3) developing an embedded machine learning framework. Data cleanup and preparation have been a well-known challenge since database systems were first invented. While integration and automation of data collection efforts within many organizations is quite mature, there are special challenges for flight-based organizations (i.e., the automatic and efficient transmission of aircraft flight data to centralized analytical data processing systems). Furthermore, this creates additional constraints for the operationalization of embedded machine learning methods for classical tasks such as classification and prediction; and magnifying design challenges for the more novel *prescriptive based* architectures. Our research is focused on the application of a design pattern for a) the integration and automation of data collection and b) an organizationally embedded ensemble machine learning method.

Keywords: machine learning, unsupervised teaching, integrated systems, theory of polymorphic learning, flight data, Garmin G1000®

Introduction

Airline fleet monitoring processes are not new or novel. Since the advent of commercial air travel, organizations have implemented such procedures with the goal of comparing fleet performances in terms of aircraft deterioration and fuel consumption (Taylor, 1969). More recently, studies have been conducted utilizing data analytics frameworks and machine learning frameworks with the goal of detecting anomalies in the physical aircrafts themselves (Gorinevsky et al., 2012) to reduce delays that occur as a part of turnaround operations (Wu, 2008), improve operational efficiency (Sumathi et al., 2017), or to predict failures or life of aircraft (Zaccaria et al., 2018). However, there is dearth of literature pertaining to the standards for operationalizing a machine learning framework in the context of flight data. This manuscript details an approach for standardizing and cleaning flight data, and its subsequent incorporation in the proposed unsupervised ensemble machine learning framework, that leverages Principal Component Analysis (PCA) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Analysis of data collected from 65,525 flight logs across six Cessna 172 aircraft is presented. This manuscript is guided by the research question: how can an unsupervised ensemble machine learning architecture be used to standardize quantitative parameters for flight operations and flight outlier detection? Additionally, the manuscript details what can be predicted from this architectural design pattern.

Background

Airplane monitoring systems are not new or novel with variants of systems in use for several decades (Miligan, 1995; Taylor, 1969). Airplane monitoring systems serve the role of capturing data from sensors pertaining to its structure, engine, cabin environment, and inflight entertainment systems (Gao et al., 2018). Data is routed from sensors, usually through wires,

though wireless transmission has also been implemented in some aircraft. Monitoring systems also aid with improving efficiency of aircraft maintenance through predictive maintenance (Zelenika et al., 2020). Instruments have also been developed that can evaluate aircraft takeoff performance in real-time that aid pilots in making decisive choices as to whether to takeoff or not (Miligan et al., 1995).

Airplane fleet monitoring systems play an integral role in overall operating costs for commercial airlines by allowing more efficient and effective maintenance to be done on airplanes (Dupuy et al., 2011). Aircraft operators use either a preventative or condition-based approach towards maintenance. Prevalence of monitoring systems and the prompt analysis of data collected from fleets can allow for more timely and effective maintenance activities, which will reduce aircraft downtime while also reducing operational costs arising from maintenance (Dupuy et al., 2011). There has been a trend for applying statistical techniques to data collected from fleets of commercial aircraft to identify aircraft anomalies or abnormal trends (Gorinevsky et al., 2012; Sumathi et al., 2017). Application of technologically supported data analytics can have positive impacts in terms airplane maintenance, fleet management, and operations (Sumathi et al., 2017). Use of such technology can also allow for aircraft operators to take proactive measures to reduce delays and identify root causes of delays (Wu, 2008). While big data analytics and statistical analysis techniques have been utilized in the context of aircraft fleet data, there is a dearth of literature in the context of utilizing machine learning to analyze datasets obtained from flights.

Unsupervised Machine Learning

Machine learning refers to the domain of computing centered around algorithms and statistical models that allow systems to accomplish tasks without specifically being programmed

for that task (Mahesh, 2020). Machine learning has been widely applied across various industries to enable systems to learn from data and to extract useful patterns or information from large sets of data. Unsupervised machine learning (ML) facilitates "analysis of raw datasets, thereby helping in generating analytic insights from unlabeled data" (Usama et al., 2019, p. 65580). This in turn eliminates the need for manual feature engineering or labeling of data. Unsupervised machine learning techniques require no prior training. They identify features from a given dataset and respond to new data based on previously learned features (Mahesh, 2020). Examples of unsupervised learning algorithms include, but are not limited to, k-means clustering and principal component analysis.

Principal Component Analysis

Principal Component Analysis (PCA) is an example of an unsupervised ML framework (Usama et al., 2019). PCA is defined as a "mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set" (Ringnér, 2008). It is a multivariate statistical technique used to analyze data sets containing multiple correlated dependent variables (Abdi & Williams, 2010). PCA involves the transformation of dataset attributes or features into a set of uncorrelated attributes referred to as Principal Components (Howley et al., 2005). The input to PCA is a dataset with *n* dimensions. PCA rotates the dataset to ensure maximum variability while reducing dimensionality from *n* to *k*, such that 99% of variance present in the dataset is retained in *k* principal components (Usama et al., 2019). PCA has been used with a great degree of success in a wide variety of domains (Abdi & Williams, 2010; Cao et al., 2018; Usama et al., 2019; Wang & Zhai, 2017) and, as such, would be a good choice to analyze flight data which has been described as highly dimensional and multiple correlated data features (Memarzadeh et al., 2020).

Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is one of the most widely used and frequently referenced clustering methods (Ester et al., 1996; Scitovski & Sabo, 2020). DBSCAN is a non-parametric density-based clustering technique, which is not predicated on assumptions about the probability distribution; data can be gathered from samples whose distributions are not predetermined (Potvin & Roff, 1993). Based on spectral clustering algorithms, DBSCAN clusters together with lots of close neighbors and identifies point outliers that are isolated in low-density areas, whose nearest neighbors are too far away. The usage of DBSCAN for flight data is not without precedence. It has been applied to flight anomaly detection (Alhussein & Ali, 2020; Sheridan et al., 2020), rotary-wing data analysis (Shin & Hwang, 2016), and aviation operations using energy metrics (Puranik & Mavris, 2018).

Gaps in Research

Research in using specific machine learning models, or methods, for the analysis of flight data is reasonably developed across the research spectrum using the keywords: *machine learning architecture*, *flight data,* and *outlier detection*. However, there is a dearth of literature pertaining to the operationalization of a design pattern for an *unsupervised machine learning architecture* for outlier detection. In our survey of literature, only one source was found when searching for literature using *machine learning architecture*, *flight data,* and *outlier detection* (Cook et al., 2019). This specific study was however primarily centered around anomaly detection in the context of internet of things (IoT) devices.

The research presented in this manuscript is focused on the broader applications of the interstitial design gaps of an integrated organization machine learning system. In addition, the

findings contribute towards a theory of polymorphic learning, a machine learning method that works through a continuum of micro-tactical use cases through to macro-strategic use cases.

Methods

This study leverages a design science research methodology. Design science research should strike a balance between utility-driven research and the demands of methodological research rigor (Peffers et al., 2007). The fundamental divide is between knowledge questions and practical difficulties. A knowledge question is a lack of information that they desire to reduce, but a practical problem is a difference between stakeholder goals and experiences that they wish to reduce. For instance, asking what the relationship is between team communication structure and systems integration build failures is a knowledge question, as opposed to asking how to reduce the frequency of build failures in integrated systems engineering projects.

All forms of scientific investigation are inextricably entwined with contemporary problems. For illustration, an empirical research question is a knowledge issue that necessitates further investigation for the researcher to respond. However, carrying out empirical research is a practical matter in and of itself. Examining the research problem, planning, and validating the research are important. To establish whether a solution design would solve the problem, a problem solver must predict what would occur if it were implemented in the problem domain. This is a question of knowledge. The reciprocal recurrence of practical difficulties and knowledge questions may cause confusion, making it simple for the researcher to overlook important problems to solve or questions to pose. Addressing practical problems, which are knowledge questions, the design science researchers should define the knowledge space using C-K Theory (Hatchuel & Weil, 2003; Ondrus & Pigneur, 2009).

To precisely define a design scenario is the central tenet of C-K theory. An architectural artifact is an incomplete description of something that is largely undeveloped. This defines the proposed architecture as a concept. This first step of C-K Theory establishes a formal boundary between a concept space known as *C* and knowledge space known as *K*. A design's creative component is the outcome of two different expansions: C-expansions, which are sometimes referred to as "new ideas," and K-expansions, which are essential to authenticate these ideas or to extend them into effective designs (Hatchuel & Weil, 2009).

Artifacts created or researched in software engineering research include algorithms, methodologies, approaches, tools, notations, and even conceptual frameworks used in the field. Practical problems with the creation, construction, or maintenance of integrated systems are a constant in the field of systems engineering. In this regard, and for purposes specific to this research, the *unsupervised ensemble machine learning architecture* is the design science *artifact*. When it comes to artifact design, empirical research can be applied in two different ways: to validate an artifact before it is used, or to evaluate how well a design is implemented after it has been used (Wieringa, 2010). This also holds true for the field of aerospace engineering (Vincenti, 1993). Consequently, research questions under the auspices of design science can take on the following forms for a given design artifact: How to operationalize said artifact? What is the design prediction? What design trade-offs can be ascertained? How is the design valued? What is the design's effectiveness towards a given domain? At its core, our design science research question is congruent with the first two questions in the list: How can our proposed unsupervised ensemble machine learning architecture be used to standardize quantitative parameters for flight operations and flight outlier detection? Additionally, what can be predicted from this operationalization of this architectural artifact?

The architecture for this research starts with the aircraft themselves. We randomly selected six Cessna 172 aircraft from our fleet of university aircraft (K-State, 2022). Each aircraft selected contained a Garmin G1000 electronic flight instrument systems (EFIS) device. Additionally, all aircraft are part of the university's professional pilot training program (i.e., these are aircraft strictly used for student pilot training). Each aircraft maintains an active flight log. These flight logs are stored on Secure Digital (SD) memory cards. The G1000 stores flight logs in a series of flat files via a structured comma-separated values format (CSV). The data on each card is manually extracted, loaded, and transferred (ETL) into a Microsoft SQL Server database engine (see Figure 1).

Figure 1

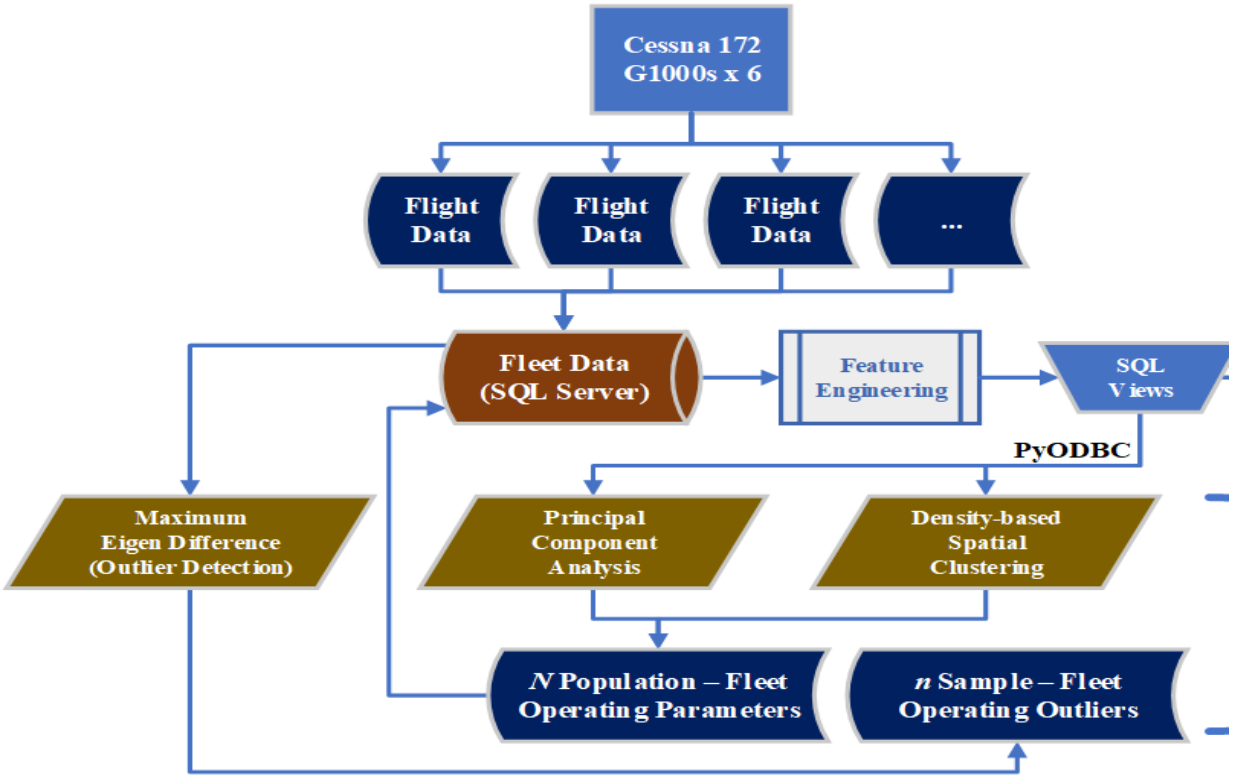
Conceptual Space, The Concept of Operation

Next, using C-K Theory, we move from the conceptual *C*-Space to the physical *K*-Space (Knowledge Space). The *K*-Space considers the actual implementation of the proposed architecture. In the *K*-Space architecture, we have more couple more steps that we need to describe. First, we bring the flight data into the SQL Server database as raw data. Second, we

perform feature engineering on the data to make it palatable for the machine learning framework. This primarily consisted of a) converting the raw textual datatypes into numerical decimal floating-point values and b) replacing null data values with zeroes using an ISNULL() SQL function (e.g., Structured Query Language function). The researchers found 3.71% of missing data across all data features in this study. Data analysis where less than 5% of the data is missing is statistically inconsequential (Shafer, 1999). A multitude of imputation methods for missing data exist as our data was not greater than the 5% threshold (Bennett, 2001; Dong & Peng, 2013). Once feature engineering is completed, the flight logs are unified using *SQL Views*. These are *virtual tables* that allow us to rapidly iterate on the data output without disturbing the base physical data within the database. This also allows us to easily split the data up into Location Data and Operations Data (see Figure 2).

Figure 2

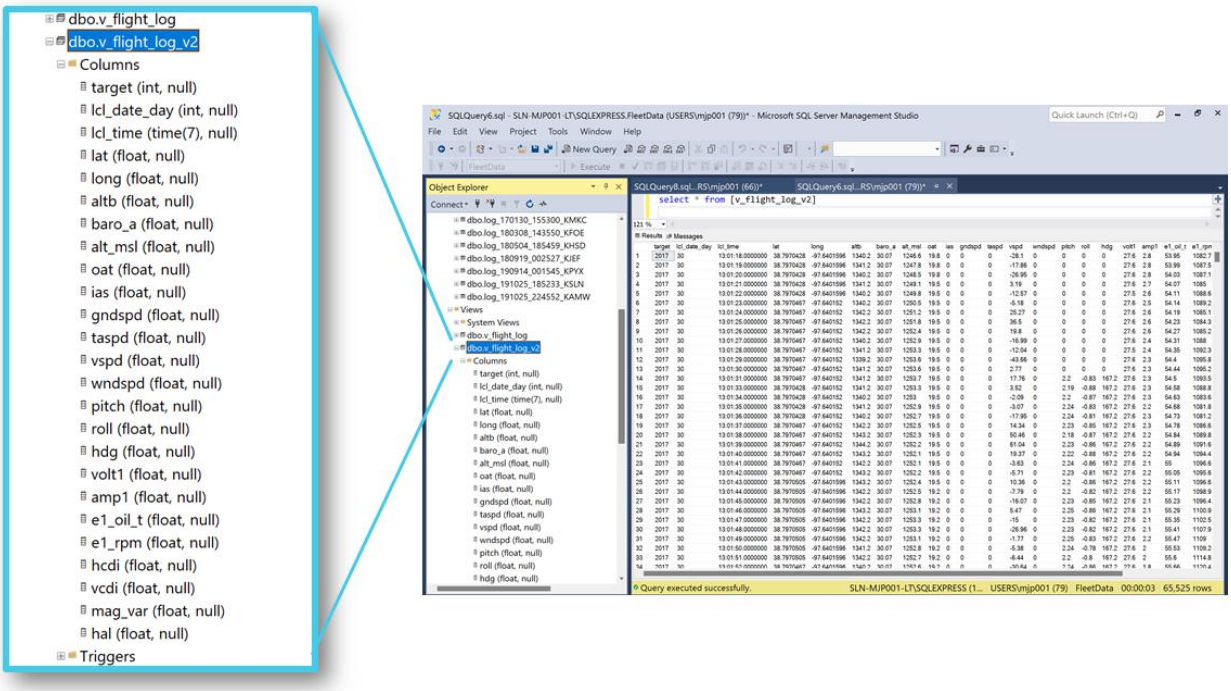
Knowledge Space, The Physical Architecture



After the data has been standardized (and centralized) within the SQL Server engine, we can easily access the data using Python's database connectivity interface library (i.e., PyODBC database connection library). The centralized data view 25 data features (see Figure 3). With the aim of forming groups, cluster analysis is a set of useful exploratory techniques that can be used whenever we want to confirm the existence of similar behavior between observations (aircraft operating parameters) or the detection of outliers that deviate from the existence of similarly clustered dimensionally reduced data features (Fávero & Belfiore, 2019).

Figure 3

Definition of Centralized Data View



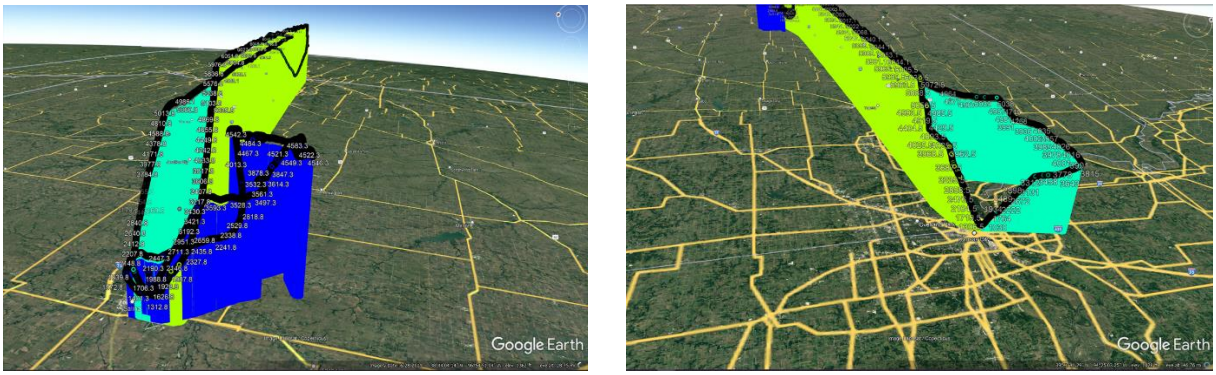
At this stage of the architecture, we utilized machine learning capabilities such as Principal Component Analysis (PCA) and Density-Based Spatial Clustering of Application with Noise (DBSCAN). The rationale behind this was that SQL (Structured Query Language), while great for data engineering, does not have the same maturity as Python in terms of machine learning capabilities (Blacher et al., 2022). At this point, PCA and DBSCAN can be leveraged from within Python to determine standard operating characteristics of the fleet sample as well as to identify operating outliers based on the dimensionally reduced eigenvalues.

The last step of the process was focused on data verification. We first verified the location data by geo-mapping the flight paths to determine that the data we imported was indeed correct. After reviewing the location data output on Google Earth, we were able to determine that the data translations from the prior feature engineering processes did not impact the data output

required to make this an operational machine learning architecture (see Figure 4). Consequently, this test allowed us to move forward with unsupervised machine learning portion of our architecture.

Figure 4

Location Data, Salina (SLN) to Kansas City (MCI)



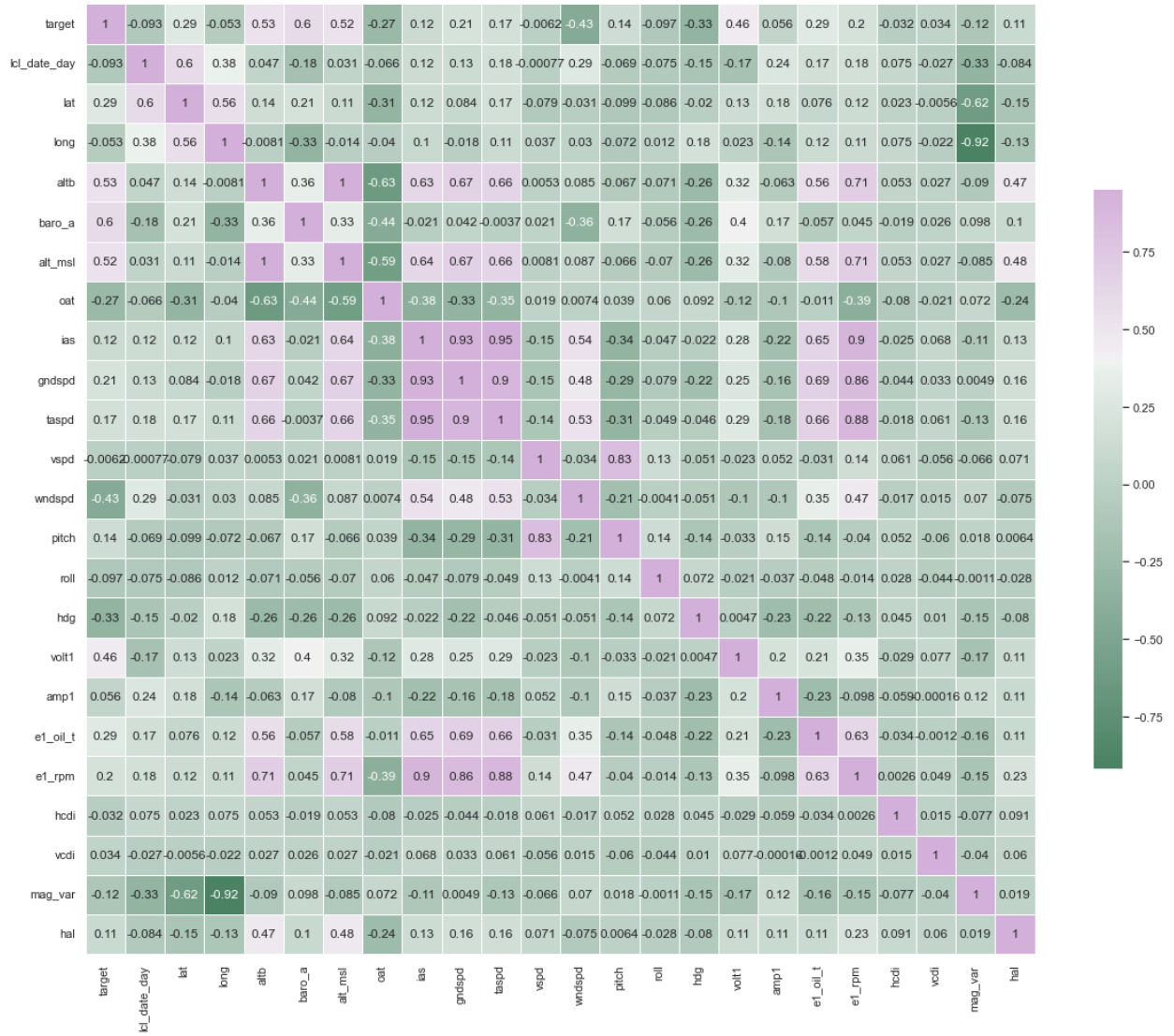
Results

After processing and unifying 65,525 flight log entries, we were able to determine accurate operating parameters using a standard correlation plot (see Figure 5). The rank order column within the correlation plot is greyed due to it not being a derived data feature and not part of the operational flight data features. Correlations between altitude (AltB and Alt_MSL), oil temperature (E1_Oil_T), and engine revolutions per minute (E1_RPM) were positive. Indicated air speed (IAS) is positively correlated with oil temperature (E1_Oil_T) and engine revolutions per minute (E1_RPM). Additionally, outside air temperature (OAT) is negatively correlated with indicated air speed (IAS), oil temperature (E1_Oil_T) and engine revolutions per minute (E1_RPM). These correlations were measured using Pearson Correlation Coefficients (see

Figure 5). The correlation coefficient is a standardized coefficient. If the absolute value is exactly 1, then all data points fall on a straight line and a linear equation accurately captures the relationship between X and Y. The regression slope determines the direction of the correlation: a value of +1 indicates that all data points lie on a line where Y increases as X increases, and a value of -1 indicates the opposite. A value of 0 suggests that the variables are not linearly dependent on one another.

Figure 5

Feature Correlations



These coefficients are exploratory and provide us with a way towards a) selecting important data features and b) implementing PCA and DBSCAN. In looking at specific relationships within the PCA model, we were able to get excellent variance values in our model metrics across many data features (see Table 1). According to the UCLA: Statistical Consulting Group (2021), variance values above 0.70 are considered fair, values above 0.80 are considered good, and values above 0.90 are considered excellent.

Table 1

PCA Explainable Variance

Data Features	**PCA Variance Explained**	**DBSCAN Silhouette Score**
['ias', 'e1_oil_t', 'e1_rpm']	0.967	0.761
['vspd', 'pitch', 'roll']	0.944	0.843
['oat', 'ias', 'e1_oil_t', 'e1_rpm']	0.897	0.760
['amp1', 'volt1', 'vspd']	0.739	0.843
['mag_var', 'hdg', 'vspd']	0.728	0.799
Scale Range	*0 to 1*	*-1 to 1*

In evaluating DBSCAN, we can use what is called the Silhouette Coefficient (also known as the Silhouette Score). Since we cannot use any type of visualization to validate clustering when dimensions are greater than 3, the silhouette score is quite helpful when dealing with higher dimensions. The silhouette value gauges an object's cohesion with its own cluster in comparison to other clusters (separation). A high number on the silhouette implies that the object

is well matched to its own cluster and poorly matched to nearby clusters. The silhouette has a range of -1 to +1 (Shahapure & Nicholas, 2020). While largely dependent on use case and the configuration of DBSCAN, silhouette values above 0.50 are considered fair, values above 0.60 are considered good, and values above 0.70 are considered excellent (Ogbuabor & Ugwoke, 2018; Trivedi & Patel, 2020). In looking over the results, the Silhouette Scores illustrate the viability of DBSCAN for analyzing the operational characteristics of our flight data.

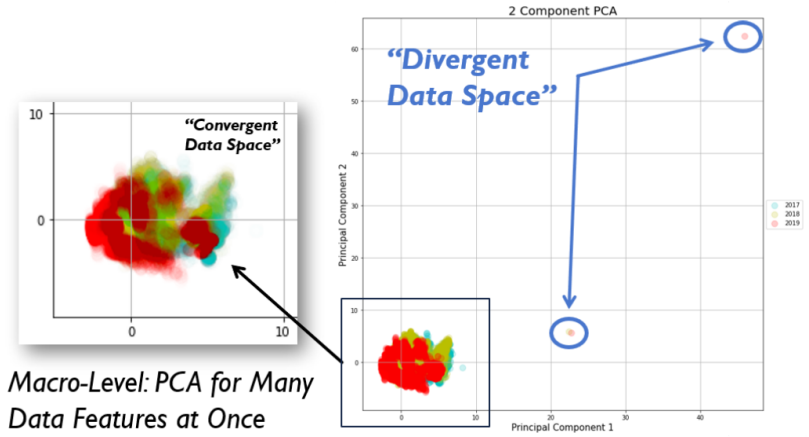
Discussion

The original premise of this research is focused on the operationalization of an integrated organizational machine learning architecture for aviation flight data to be used towards a) standardizing operational flight parameters using unsupervised learning methods, and b) identification of operational outliers using unsupervised learning methods. Thirdly, the scope of this research also includes a design pattern prediction for the proposed architectural artifact. The proposed machine learning architecture was not just proposed but tested using real flight data, from a real training fleet. Early in our data capture (at roughly 5,000 flight log samples) it was becoming clear that more data was needed to standardize the flight operational parameters for principal component analysis (PCA). An interesting observation we found in our data analysis is we approximated a single anomaly detection hit rate for every 20,000 flight log samples.

Given our total sample size of 65,525 flight log entries under analysis, using all data features, PCA was able to detect 3 anomalous outliers (see Figure 6). Overall, the architecture's use of an unsupervised machine learning method was able to standardize operational flight parameters.

Figure 6

PCA, Macro-Strategic Analysis of Features



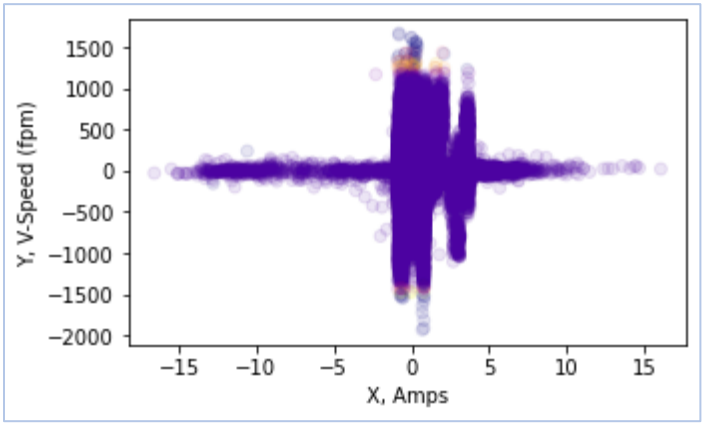
Regarding identifying operational outliers using unsupervised learning methods, we developed and implemented the architecture. New terminology was also developed to distinguish between convergent data behavior and divergent data behavior. Convergent data behavior is non-anomalous behavior (e.g., operational parameters are within the normal *converged data space*). Divergent data behavior is a candidate for anomalous behavior (e.g., operational parameters are within the abnormal *divergent data space*). Additionally, this allowed us to leverage PCA as a Macro-Level detection framework. This architecture distinguished itself in its ability to a) standardize operating parameters into dimensionally reduced eigenvalues that are easily captured, and b) identify anomalous behavior within the divergent data space.

Finally, the proposed organizational machine learning architecture illustrated a useful design pattern with a predicted system capability to standardize operational fleet parameters and detect operational outliers. An emergent system property of the architecture (Weinstock, 2010) was the ability of the architecture to perform standardization and outlier detection at both the

macro-strategic and micro-tactical levels (see Figure 7). This demonstrates that unsupervised machine learning methods, which do not require training data, have utility for analyzing flight data.

Figure 7

Micro-Tactical Analysis of Features



Limitations, Conclusion, and Future Work

The research has some limitations. First, the machine learning architecture was built on data from a fleet of Cessna 172s from within a training environment. If one were to standardize operational fleet parameters based on flight log data, the data model would need to be extended to take this into account. Eigenvalue standardization is specific to a given aircraft and flight environment. An additional area for improvement is the wireless transmission of data from the fleet to the to the centralized database. Current technologies are very costly and are usually cloud-based, further increasing their subscription costs for data processing. Automated data transfers can be done much cheaper if the fleet operator has the willingness to engage in

infrastructure development as well as the research efforts in unsupervised machine learning methods.

In implementing a design science research methodology, we focused on validating the architectural artifacts and evaluating it following implementation. It can be inferred from the results that the proposed architecture provides value at multiple levels. The implementation of this proposed architecture isn't very difficult provided sufficient fluency in data engineering, python programming, structured query language (SQL), and unsupervised machine learning. One way to measure the effectiveness of a given architecture is to evaluate its scalability. The proposed architecture was implemented on a single laptop (Intel i7 processor, 32 gigabytes of random-access memory, and 1 terabyte of hard drive space) which is a testament to architecture's scalability and portability. The addition of more data to the database did not result in any discernable performance problems. This architecture is highly modular and can easily be implemented in a cloud environment should an enterprise build-out be needed.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Alhussein, I., & Ali, A. H. (2020, September). Application of DBSCAN to anomaly detection in airport terminals. *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*,pp. 112–116. IEEE. https://doi.org/10.1109/IICETA50496.2020.9318876

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Blacher, M., Giesen, J., Laue, S., Klaus, J., & Leis, V. (2022, January 9-12). *Machine learning, linear algebra, and more: Is SQL all you need?* [Paper]. 12th Annual Conference on Innovative Data Systems Research (CIDR '22), Santa Cruz, CA, United States https://www.cidrdb.org/cidr2022/papers/p17-blacher.pdf

Cao, W., Czarnek, N., Shan, J., & Li, L. (2018). Microaneurysm detection using principal component analysis and machine learning methods. *IEEE Transactions on NanoBioscience*, *17*(3), 191–198. https://doi.org/10.1109/TNB.2018.2840084

Cook, A. A., Mısırlı, G., & Fan, Z. (2019). Anomaly detection for IoT time-series data: A survey. *IEEE Internet of Things Journal*, *7*(7), 6481–6494. https://doi.org/10.1109/JIOT.2019.2958185

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data for researchers. *Springerplus*, *2*, 1–17. https://doi.org/10.1186%2F2193-1801-2-222

Dupuy, M. J., Wesely, D. E., & Jenkins, C. S. (2011). Airline fleet maintenance: Trade-off analysis of alternate aircraft maintenance approaches. In K.A. Neeley (Ed.), *2011 IEEE Systems and Information Engineering Design Symposium* (pp. 29–34). IEEE. https://doi.org/10.1109/SIEDS.2011.5876850

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (pp. 226–231). Association for the Advancement of Artificial Intelligence.

Fávero, L. P., & Belfiore, P. (2019). Chapter 11 - Cluster Analysis. *Data Science for Business and Decision Making* (pp. 311–382). Academic Press. https://doi.org/10.1016/B978-0-12-811216-8.00011-2

Gao, S., Dai, X., Hang, Y., Guo, Y., & Ji, Q. (2018). Airborne wireless sensor networks for airplane monitoring system. *Wireless Communications and Mobile Computing*, *2018*. https://doi.org/10.1155/2018/6025825

Gorinevsky, D., Matthews, B., & Martin, R. (2012, October). Aircraft anomaly detection using performance models trained on fleet data. In K. Das, N. V. Chawla, & A. N. Srivastava (Eds.), *2012 Conference on Intelligent Data Understanding (CIDU 2012)* (pp. 17–23). IEEE. https://doi.org/10.1109/CIDU.2012.6382196

Hatchuel, A., & Weil, B. (2003). A new approach of innovative design: An introduction to C-K theory. In A. Folkeson, K. Gralen, M. Norell, & U. Sellgren (Eds.), *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design*. The Design Society. https://www.designsociety.org/publication/24204/

Hatchuel, A., & Weil, B. (2009). C-K design theory: An advanced formulation. *Research in Engineering Design, 19,* 181–192. https://doi.org/10.1007/s00163-008-0043-4

Howley, T., Madden, M. G., O'Connell, M. L., & Ryder, A. G. (2005). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In A. Macintosh, R. Ellis, & T. Allen (Eds.), *Applications and Innovations in Intelligent Systems XIII* (pp. 209–222). Springer. https://doi.org/10.1007/1-84628-224-1_16

K-State Aerospace and Technology Campus Fleet. (2022). Kansas State University. https://www.salina.k-state.edu/academics/degree-options/professional-pilot/fleet.html

Mahesh, B. (2020). Machine learning algorithms - A review. *International Journal of Science and Research, 9*(1), 381–386.

Memarzadeh, M., Matthews, B., & Avrekh, I. (2020). Unsupervised anomaly detection in flight data using convolutional variational auto-encoder. *Aerospace, 7*(8), 1–19. https://doi.org/10.3390/aerospace7080115

Miligan, M.W., Zhou, M. M., & Wilkerson, H. J. (1995). Monitoring airplane takeoff performance prototype instrument with learning capability. *Journal of Aircraft, 32*(4), 768–772. https://doi.org/10.2514/3.46789

Ogbuabor, G., & Ugwoke, F. N. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology, 10*(2), 27–37. https://doi.org/10.5121/ijcsit.2018.10203

Ondrus, J., & Pigneur, Y. (2009). C-K design theory for information systems research. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (pp.1–2). Association for Computing Machinery. https://doi.org/10.1145/1555619.1555656

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Potvin, C., & Roff, D. A. (1993). Distribution-free and robust statistical methods: Viable alternatives to parametric statistics. *Ecology, 74*(6), 1617–1628. https://doi.org/10.2307/1939920

Puranik, T. G., & Mavris, D. N. (2018). Anomaly detection in general-aviation operations using energy metrics and flight-data records. *Journal of Aerospace Information Systems, 15*(1), 22–36. https://doi.org/10.2514/1.I010582

Ringnér, M. (2008). What is principal component analysis?. *Nature Biotechnology, 26*, 303–304. https://doi.org/10.1038/nbt0308-303

Scitovski, R., & Sabo, K. (2020). DBSCAN-like clustering method for various data densities. *Pattern Analysis and Applications, 23*, 541–554. https://doi.org/10.1007/s10044-019-00809-z

Shafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medicine*. *8*(1), 3–15. https://doi.org/10.1177/096228029900800102

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. In G. Webb, Z. Zhang, V. S. Tseng, G. Williams, M. Vlachos, & L. Cao (Eds.), *2020 IEEE 7th International Conference on Data Science and Advanced Analytics* (pp. 747–748). IEEE. https://doi.org/10.1109/DSAA49011.2020.00096

Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., & Mavris, D. N. (2020, January 6-10). *An application of DBSCAN clustering for flight anomaly detection*

during the approach phase [Paper]. AIAA Scitech 2020 Forum, Orlando, FL, United States. https://doi.org/10.2514/6.2020-1851

Shin, S., & Hwang, I. (2016, January 4-8). *Helicopter cockpit video data analysis for attitude estimation using DBSCAN clustering* [Paper]. AIAA Infotech@ Aerospace, San Diego, CA, United States. https://doi.org/10.2514/6.2016-0920

Sumathi, N., Gokulakrishnan, S., Kaushik Ramana, S., Muralitharan, R., & Kamal, C. V. (2017). Application of big data systems to airline management. *International Journal of Latest Technology in Engineering, Management & Applied Science*, *6*(12), 129–132. https://www.ijltemas.in/DigitalLibrary/Vol.6Issue12/129-132.pdf

Taylor, W. (1969, July 14-16). *Airline fleet performance survey* [Paper]. AIAA Aircraft Design and Operations Meeting, Los Angeles, CA, United States. https://doi.org/10.2514/6.1969-770

Trivedi, S., & Patel, N. (2020). Clustering students based on virtual learning engagement, digital skills, and e-learning infrastructure: Applications of K-means, DBSCAN, hierarchical, and affinity propagation lustering. *Sage Science Review of Educational Technology*, *3*(1), 1–13.

UCLA: Statistical Consulting Group (2021). *Principal Components (PCA) and Exploratory Factor Analysis (EFA) with SPSS*. https://stats.oarc.ucla.edu/spss/seminars/efa-spss/

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE Access*, *7*, 65579–65615. https://doi.org/10.1109/ACCESS.2019.2916648

Vincenti, W. G. (1993). *What engineers know and how they know it: Analytical Studies from Aeronautical History*. Johns Hopkins University Press. https://doi.org/10.56021/9780801839740

Wang, C., & Zhai, H. (2017). Machine learning of frustrated classical spin models. I. Principal component analysis. *Physical Review B, 96*(14). https://doi.org/10.1103/PhysRevB.96.144432

Weinstock, M. (2010). *The architecture of emergence: The evolution of form in nature and civilization.* John Wiley and Sons.

Wieringa, R. (2010). Design science methodology: Principles and practice. In J. Kramer, J. Bishop, P. Devanbu, & S. Uchitel (Eds.), *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering* (Vol. 2, pp. 493–494). Association for Computing Machinery. https://doi.org/10.1145/1810295.1810446

Wu, C.-L. (2008). Monitoring aircraft turnaround operations – Framework development, application and implications for airline operations. *Transportation Planning and Technology, 31*(2), 215–228. https://doi.org/10.1080/03081060801948233

Zaccaria, V., Stenfelt, M., Aslanidou, I., & Kyprianidis, K. G. (2018). Fleet monitoring and diagnostics framework based on digital twin of aero-engines. *Proceedings of the ASME Turbo Expo*. American Society of Mechanical Engineers (ASME). https://doi.org/10.1115/GT2018-76414

Zelenika, S., Hadas, Z., Bader, S., Becker, T., Gljušćić, P., Hlinka, J., Janak, L., Kamenar, E., Ksica, F., Kyratsi, T. and Louca, L., Mrlik, M., Osmanovic, A., Pakrashi, V., Rubes, O., Sevecek, O., Silva, J., Tofel, P., Trkulja, B., …Vrcan, Ž. (2020). Energy harvesting technologies for structural health monitoring of airplane components—A review. *Sensors, 20*(22), 1–57. https://doi.org/10.3390/s20226685