

## Journal Pre-proof

Exploratory data analysis for airline disruption management

Kolawole Ogunsina, Ilias Bilonis, Daniel DeLaurentis

PII: S2666-8270(21)00051-7  
DOI: <https://doi.org/10.1016/j.mlwa.2021.100102>  
Reference: MLWA 100102

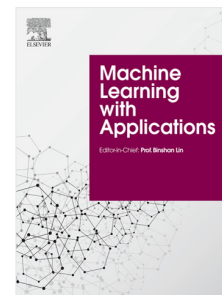
To appear in: *Machine Learning with Applications*

Received date: 6 February 2021  
Revised date: 6 June 2021  
Accepted date: 27 June 2021

Please cite this article as: K. Ogunsina, I. Bilonis and D. DeLaurentis, Exploratory data analysis for airline disruption management. *Machine Learning with Applications* (2021), doi: <https://doi.org/10.1016/j.mlwa.2021.100102>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# Exploratory Data Analysis for Airline Disruption Management \*

Kolawole Ogunsina<sup>1,1,\*</sup>, Ilias Bilonis<sup>1,2</sup>, Daniel DeLaurentis<sup>1,3</sup>

---

\*This article represents one full chapter from the corresponding author's completed doctoral dissertation.

\*Corresponding Author

Email addresses: kolawole08@gmail.com (Kolawole Ogunsina), ibilion@purdue.edu (Ilias Bilonis), ddelaure@purdue.edu (Daniel DeLaurentis)

<sup>1</sup>School of Aeronautics and Astronautics, Purdue University, United States.

<sup>2</sup>Department of Mechanical Engineering, Purdue University, United States.

<sup>3</sup>School of Aeronautics and Astronautics, Purdue University, United States.

---

**Abstract**

Reliable platforms for data collation during airline schedule operations have significantly increased the quality and quantity of available information for effectively managing airline schedule disruptions. To that effect, this paper applies macroscopic and microscopic techniques by way of basic statistics and machine learning, respectively, to analyze historical scheduling and operations data from a major airline in the United States. Macroscopic results reveal that majority of irregular operations in airline schedule that occurred over a one-year period stemmed from disruptions due to flight delays, while microscopic results validate different modeling assumptions about key drivers for airline disruption management like turnaround as a Gaussian process.

*Keywords:* airline disruption management, data analysis, machine learning

---

**1. Introduction**

Despite overcoming numerous financial and technical challenges over the last century through continued drive towards innovation and productivity, a complete solution to irregular operations in the airline industry has remained elusive. A major driver that has significantly stunted the progress in developing a full solution for airline disruption management is poor data integration and integrity.

Internal data containing real-time information about an airline's resources and their scheduled utilization over time, and external data for factors such as current and future weather forecasts, competitor activities, and air traffic control, are necessary for efficient operations (Nathans, 2015). These bits of information and data must be readily available and accessible to represent drivers and constraints for scenarios induced by irregular operations, so as to facilitate the development of effective self-governing platforms for airline disruption management. In addition, whenever a new airline system is replaced or upgraded, new data sources are typically integrated into the existing framework

(Gershkoff, 2016). The new data must be maintained for both existing and new applications, and thus present cost-intensive challenges for mitigating disruption because many facets of the airline infrastructure are impacted.

### 20 1.1. The Problem

While there have been consistent improvements to the existing decision support systems used by human controllers in the Airline Operations Control Center (AOCC), two factors have continued to limit the performance of the disruption resolutions that are applied. First, the decision support systems do not explicitly proffer solutions to specific schedule disruptions. As such, human controllers in the AOCC are required to be reactive in addressing disruptions by using their best judgement based upon their prior experience from resolving the same (or similar) disruption. Second, majority of the decision support (computer) systems used by multiple departments (including the AOCC) in an airline and other air transportation stakeholders (e.g. airports) are not designed or developed at the same time nor by the same vendor (Nathans, 2015). As such, information and data are required to be entered into multiple computer systems thereby exposing human controllers in the AOCC to data input problems and errors. Therefore, information entered into a decision support system for disruption management may be out of sync with other systems and yield incorrect decisions due to lack of data integrity.

Furthermore, extant literature on the adoption of machine learning for decision support systems have only focused on addressing irregular operations for system-wide disruption management in the national airspace system as a whole. In other words, existing literature address irregular operations for system-wide disruption management by exploiting airport data from multiple airlines in the air transportation system. For instance, Liu et al. (2019) used machine learning to analyze air traffic management actions for incidences of ground delay programs across major airports in the United States. The authors employed a two-stage framework, such that the first stage used support vector machines to correlate incidences of ground delay programs with regional convective (weather)

activity while the second stage trained logistic regression and random forest models by using a weather score obtained from the first stage. In addition, Ye et al. (2020) presented an approach for estimating aggregate flight departure delay at airports in China through supervised learning models. They applied four separate types of airport-related aggregate characteristics (including weather) to predict expected departure delays at a major airport in China, by employing linear regression, support vector machines, randomized trees, and LightGBM. While the models presented in both research studies provided acceptable predictive capacity for disruption management at the airport level in the respective air transportation networks, they (i.e. the models) do not describe the disruption management proclivities of a decision support system commissioned by a specific airline in the air transportation network.

In the bid to improve data integrity and fidelity for existing decision support systems, airlines have significantly invested in creating better localized data collection platforms within their respective organizations, which can amass information from different sources within and outside the organization that is easily accessible through a centralized data server (Amadeus IT Group, 2016). As such, there is a need to fully leverage the ubiquity and accessibility of information (data) collected by existing platforms in the AOCC to enhance agile decision-making capabilities of the AOCC during airline disruption management. To that effect, this paper provides a comprehensive discussion on exploratory analysis administered on historical scheduling and operations recovery data supplied by a major airline in the United States. This exploratory analysis serves as the basis for the development of credible predictive and prescriptive models for airline disruption management.

### *1.2. Contributions*

To the best of our knowledge, we provide the first literature that strictly employs the statistics of big data and machine learning to demonstrate the characterization and evaluation of a functional role (i.e. domain) in the AOCC of a major U.S. airline for disruption management. Thus, in contrast and com-

plement to existing literature, our work provides research on irregular operations and disruption management accustomed to a specific U.S. airline based upon data provided by the airline. To achieve the aforementioned objectives, this paper enhances prior research and literature on irregular operations and airline disruption management through the following contributions:

1. We introduce and explore several abstraction methods for applying, enhancing, and sequestering raw features and labels in a historical data set on airline scheduling and operations recovery from a major U.S. airline, to readily identify relevant cognitive patterns for key drivers during airline disruption management.
2. We investigate the application of appropriate machine learning techniques for revealing patterns, pertinence, and properties of abstracted data features, which provide necessary a priori information (i.e. beliefs) for Bayesian and pseudo-Bayesian methods. These methods are subsequently employed in future work for developing functional models in an intelligent multi-agent system for airline disruption management.

### 1.3. Paper Organization

The next section in this paper provides an overview of the elements of historical scheduling and operations recovery data retrieved from a major U.S. airline, followed by a section that expansively discusses several relevant and interrelated processes for exploratory data analysis. We conclude with a summary of pertinent findings and areas for further research in Section 4.

## 2. Data Overview

The raw data utilized for demonstrating the exploratory analysis discussed in this paper was provided by Southwest Airlines. Like many major U.S. airlines, Southwest Airlines employs an integrated AOCC organization wherein all functional roles share the same physical space (at the airline's headquarters in Dallas) and are hierarchically dependent on AOCC supervisors for multiple

105 problem dimensions in airline operations recovery (Hagel et al., 2017). As the  
 largest carrier in the United States in terms of originating domestic passen-  
 110 gers boarded with more than 4,100 flight schedule operations daily to over 100  
 destinations, the supervisors (and controllers) at Southwest Airlines Network  
 Operations Control (SWA-NOC) seek to use technology to see the impact of  
 their decisions to make better ones for improved disruption management. For  
 many years, the controllers at SWA-NOC relied on gut instincts to track and  
 understand how their disruption resolution actions cascaded throughout the air-  
 line's network, but could not inform their instincts with data. To address this  
 issue, upper management at SWA-NOC created the Baker workgroup; an in-  
 115 tegrated team of supervisors and software developers dedicated to improving  
 decision-making during disruption management by developing and enhancing a  
 suite of computerized decision support systems called the Baker tool. In order  
 to better support the Baker tool, the workgroup created an autonomous data  
 collection platform to record flight schedules that are subject and not subject  
 120 to different disruption incidents in the Southwest Airlines route network.

Thus, the raw data generously provided to us for the research discussed in  
 this paper contains approximately 1.1 million instances of direct flight sched-  
 ules from Southwest Airlines route network operations recorded from September  
 2016 to September 2017; of which there are 620,000 flight schedules that were not  
 125 subject to disruptions, over 430,000 flight schedules that were subject to flight  
 delays, and approximately 26,000 flight schedules that were either cancelled or  
 diverted. The instances of disrupted flight schedules (i.e. delayed, diverted  
 and cancelled flight schedules) are distributed across eleven separate functional  
 roles in SWA-NOC (i.e. the AOCC) that represent primary disruption reso-  
 130 lution domains for aircraft, crew, and passenger problem dimensions in airline  
 disruption management. Table 1 reveals a list of functional disruption resolu-  
 tion domains in the Southwest Airlines Network Operations Control, including  
 the corresponding problem dimensions they seek to address, and the class of  
 disruption and the number of instances of different effects of a disruption class  
 135 for a specific functional domain. The disruption class defines the origination of

Table 1: Disruption outlook for functional domains in Southwest Airlines network operations control from September 2016 to September 2017

Functional Domain	Affected Problem Dimension	Disruption Class	Delayed Flight Schedule Instances	Cancelled Flight Schedule Instances	Diverted Flight Schedule Instances
<i>Customer Hold</i>	Aircraft and Passenger	Controllable	46,870	0	0
<i>Dispatch CSC</i>	Aircraft and Crew	Controllable	17,468	0	0
<i>Flight Operations</i>	Crew	Controllable	36,370	1,099	909
<i>Fuel Management</i>	Aircraft	Controllable	4,841	0	0
<i>Ground Operations</i>	Aircraft and Passenger	Controllable	168,375	2518	460
<i>Inflight</i>	Crew	Controllable	79,444	984	67
<i>Maintenance</i>	Aircraft	Controllable	33,518	55	0
<i>NAS</i>	All	Uncontrollable	22,644	2,619	433
<i>Security</i>	Passenger	Controllable	2,955	6	11
<i>Technology</i>	All	Controllable	8,953	0	0
<i>Weather</i>	All	Uncontrollable	12,659	12,156	4905



a specific disruption, and as such, disruptions resolved by functional domains in SWA-NOC with a “controllable” disruption class indicate that all instances of disrupted flight schedules associated with those domains were caused (or could have been avoided) by the airline. Conversely, disruptions resolved by functional domains in SWA-NOC with an “uncontrollable” disruption class indicate that all instances of disrupted flight schedules affiliated with those domains were not caused (nor could have been avoided) by the airline. A brief description of the functional disruption resolution domains (or roles) in SWA-NOC highlighted in Table 1 is as follows:

1. **Customer Hold:** This functional domain addresses disruptions related to holding aircraft for passengers on inbound flight connections and holding aircraft to accommodate passengers off cancelled and delayed flights. As such, the customer hold functional domain resolves the aircraft and passenger problem dimensions in airline disruption management. Disruption instances for the customer hold domain accounted for about 11% of delayed flight schedules in the Southwest Airlines route network over the one-year period (i.e. September 2016 to September 2017).
2. **Dispatch CSC:** This functional domain manages disruptions related to flight dispatch activities by the airline that also includes holding flights to accommodate international flight schedule slot times. To that effect, the Dispatch CSC functional domain addresses the aircraft and crew problem dimensions during disruption management, and disruption instances related to Dispatch CSC represented 4% of delayed flight schedules in the airline operations between September 2016 and September 2017.
3. **Flight Operations:** This functional domain resolves disruptions defined by Pilot (cockpit crew) scheduling activities as they relate to Pilot tardiness and normal aircraft readiness, and addresses the crew problem dimension of airline disruption management. Between September 2016 and September 2017, disruption instances related to Flight Operations represented about 8.5% of delayed flight schedules, 5.7% of cancelled flight

schedules, and 13.5% of diverted flight schedules in Southwest Airlines operations.

4. **Fuel Management:** This functional role in SWA-NOC manages disruptions related to aircraft fueling and other energy administration activities, and addresses the aircraft problem dimension during disruption management. Disruption instances related to Fuel Management between September 2016 and September 2017 represented 1.1% of delayed flight schedules in Southwest Airlines operations.

5. **Ground Operations:** This functional domain in SWA-NOC manages disruptions defined by several activities ranging from passenger boarding and aircraft provisioning to ramp services and aircraft towing, and as such, resolves the aircraft and passenger problem dimensions in airline disruption management. Over the one year period of airline operations, disruptions related to Ground Operations accounted for the largest percentage of total flight schedule delays of 39%, and the third highest percentage of total flight schedule cancellations of 13%.

6. **Inflight:** Similar to Flight Operations, Inflight resolves disruptions defined by Flight Attendant (cabin crew) scheduling activities as they relate to Flight attendant tardiness and normal aircraft preparedness, and thus addresses the crew problem dimension of airline disruption management. Between September 2016 and September 2017, disruption instances related to Flight Operations represented about 18.5% of delayed flight schedules, 5.1% of cancelled flight schedules, and about 1% of diverted flight schedules in Southwest Airlines operations.

7. **Maintenance:** This functional domain resolves disruptions defined by aircraft maintenance and inspection activities, and as such, addresses the aircraft problem dimension of airline disruption management. Disruption instances related to Maintenance represented 7.8% of delayed flight schedules and about 0.3% of cancelled flight schedules during Southwest Airlines

195 operations from September 2016 to September 2017.

8. **NAS**: This adopted functional role in SWA-NOC manages disruptions defined by air traffic control activities related to gate hold for congestion at departure and arrival airport stations. As such, the NAS functional domain addresses uncontrollable disruptions representing all problem dimensions during airline disruption management. Disruption instances associated with NAS represented 5.3% of delayed flight schedules, 13.5% of cancelled flight schedules and 6.4% of diverted flight schedules during Southwest Airlines operations from September 2016 to September 2017.

9. **Security**: This functional domain addresses disruptions defined by security measures enforced to ensure the safety and convenience of passengers at airports prior to aircraft boarding. Its responsibilities includes managing disruptions due to baggage screening by TSA (Transportation Security Administration) at the skycap or ticket counter. As such, the Security functional domain resolves the passenger problem dimension during airline disruption management. Between September 2016 and September 2017, disruption instances related to Security represented the least percentages of total delayed, cancelled, and diverted flight schedules of 0.7%, 0.03%, and 0.16%, respectively, in Southwest Airlines operations.

10. **Technology**: This functional role manages all disruption activities defined by system-wide technology outages, and thus aims to resolve all problem dimensions during airline disruption management. Disruption instances related to Technology accounted for 2.1% of all delayed flight schedules in Southwest Airlines operations between September 2016 and September 2017.

11. **Weather**: Similar to NAS, this adopted functional domain in SWA-NOC manages all kinds of uncontrollable disruption defined by inclement weather activities. To that effect, the Weather functional role aims to resolve the aircraft, crew and passenger problem dimensions during dis-

ruption management. Disruption instances associated with the Weather functional domain accounted for the highest percentage of cancelled and diverted flight schedules (62.6% and 72.8% respectively) among all functional domains in SWA-NOC between September 2016 and September 2017. In addition, delayed flight instances related to Weather represented 2.9% of the total delayed flight instances addressed by all functional domains in SWA-NOC over the one year data collation period.

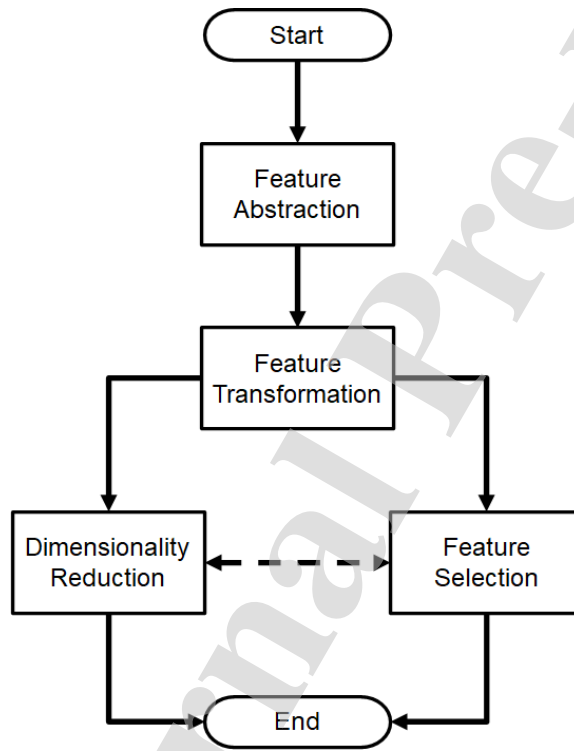


Figure 1: Procedure for exploratory data analysis in a data-driven paradigm for airline disruption management

### 3. Data Analysis

The previous section provided a macroscopic overview of disruption activities for different functional roles in SWA-NOC and revealed that about 42% of all

flight schedules for Southwest Airlines route network operations from September  
 235 2016 to September 2017 were disrupted. As a result, the functional disruption  
 resolution domains in SWA-NOC were most likely to address irregular oper-  
 ations due to delayed flight schedules, which represent approximately 94% of  
 all disrupted flight schedules recorded from September 2016 to September 2017.  
 Furthermore, there are two separate chunks of data which are defined by the oc-  
 240 currence of disruption during flight schedule execution. The first chunk, which  
 is known as the *non-disrupted* data set, represents the larger chunk that con-  
 tains instances of flight schedules that executed without any disruption. The  
 smaller chunk, also known as the *disrupted* data set, contains instances of flight  
 schedules that executed with disruption, and thus represent instances of flight  
 245 schedule execution due to irregular operations. As such, the major difference  
 between the *non-disrupted* data set and *disrupted* data set is the existence of  
 additional data features (i.e. disruption features) in the *disrupted* data set that  
 indicate different types of disruption. However, there is a small subset (6%) of  
 the *disrupted* data set that represents instances of canceled and diverted flight  
 250 schedules, which have less data fields with sparse data entries that present signif-  
 icant challenges for machine learning applications. To that end, we restrict the  
 scope of the research presented in this paper to irregular operations based upon  
 delayed flight schedules and ignore flight cancellations and diversions which are  
 primarily limited to the Weather functional domain. Henceforth, irregular oper-  
 ations only represent controllable and uncontrollable disruptions due to delayed  
 255 flight schedules.

Fine details that highlight pertinent high-level patterns for elements (or fea-  
 tures) that define each flight schedule in the raw data set introduced in Section 2  
 can not be readily observed nor acknowledged through macroscopic inspection.  
 260 Hence, in order to mine relevant microscopic information from raw data, this  
 section elucidates the exploratory data analysis used to effectively generalize  
 pattern-finding schemes for consistent flight schedule features that are appli-  
 cable in all functional roles in SWA-NOC. Fig 1 shows the general procedure  
 adopted for performing exploratory data analysis on the raw data set. The pro-

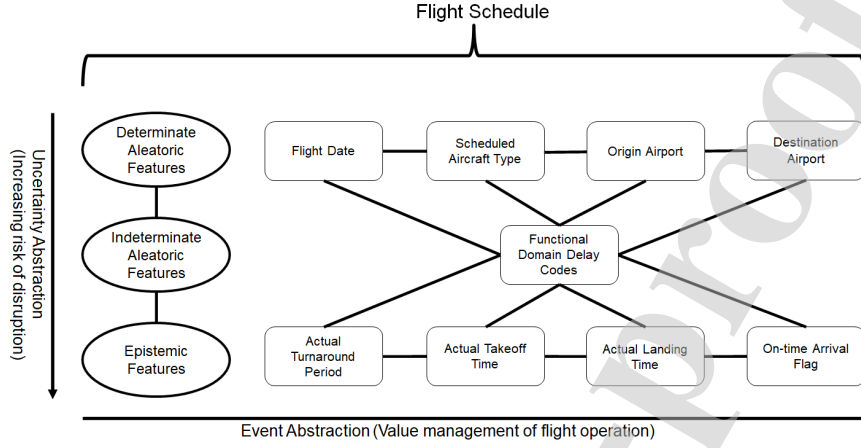


Figure 2: Sample knowledge abstraction of a basic flight schedule for airline disruption management

cess commences by abstracting separate data features that represent distinct properties of flight scheduling and operations, next raw data features are transformed into data forms that are readily decipherable by appropriate machine learning algorithms. The data transformation is necessary for applying separate methods for identifying critical data features and reducing the dimension space of the data set achieved by the feature selection and dimensionality reduction processes shown in Fig 1. The subsequent parts of this section provide more insight into the aforementioned processes as they relate to the historical scheduling and operations data adapted for the microscopic analysis techniques discussed in this paper.

### 3.1. Feature Abstraction

Feature abstraction (often referred to as data abstraction) is an effective technique for accommodating semantic relationships between features in a database (Ramakrishnan & Ullman, 1995). Feature abstraction ensures that features that define specific properties of fundamental tenets of flight scheduling and operations (embedded in each flight schedule from the raw data set) is generalized into abstract values (Huh et al., 2000). As such, a specific property can be

viewed as a specialized quality of an abstract value. The properties describing the tenets of flight scheduling and operations are described by two separate principles of abstraction that represent a knowledge abstraction (Abbott, 1987; Huh et al., 2000; Liskov, 1988). These related feature abstraction principles exist for creating semantic associations among data features for airline scheduling and disruption management and are namely: *event abstraction* and *uncertainty abstraction*.

- *Event abstraction*: Event abstraction serves two primary purposes. First, it characterizes the importance of planned airline activities and resources associated with a specific flight schedule, and how their importance can be subject to change prior to (or on) day of operation due to the risk of disruption from passenger-boarding at a departure station to aircraft gate-parking at an arrival station. Second, event abstraction defines the manner in which planned airline activities for a specific flight schedule varies during schedule execution based upon the impact of irregular operations. As such, the event abstraction principle is synonymous to *flight operation value abstraction*, wherein the specific value (or profit) that a particular flight schedule in the airline route network provides is appraised by how effectively the flight schedule features estimated prior to schedule execution align with flight schedule features that are realized during and after schedule execution. To that effect, flight schedule features that are determined prior to schedule execution represent low-value features for event abstraction, and flight schedule features estimated during and after schedule execution represent high-value features for event abstraction.
- *Uncertainty abstraction*: From an airline planning perspective, the uncertainty abstraction principle, also analogous to a *functional domain planning abstraction*, defines the manner in which the uncertainty for the risk of irregular operations during schedule planning and execution is quantified and propagated via various features in the data set. Most airlines, including Southwest Airlines, adopt a perspective on the scheduling process that

accentuates the internal planning approach of different planning departments as an iterative cycle of flight schedule development and assessment over a timeline horizon, such that the flight schedule is continually adjusted and optimized until a suitable schedule is obtained or the planning period is over (Grosche, 2009a). Thus, the primary purpose of *uncertainty abstraction* is to express the relationships among flight schedule features that are representative of the transitions through three iterative and interconnected airline planning phases namely: strategic planning, tactical planning and operational planning respectively (Grosche, 2009b). Strategic planning, otherwise known as “future scheduling”, focuses on long-term decision-making for the subsequent tactical and operational planning phases, such that a generic service plan consisting of essential and viable sets of serviceable routes without specific aircraft and crew assignments and tentative departure and arrival times are determined. Tactical planning or “current scheduling” focuses on creating a refined schedule of operations for the service plan based upon the resources that are actually expected to be available to the airline over a definitive time period. Lastly, the operational planning phase focuses on adjusting the schedule generated in the tactical planning phase with respect to changes in demand for air travel prior to executing the flight schedule, and also on executing the schedule with minimal penalty (cost) in the event of unexpected disruptions on the day of operation (i.e. rescheduling for disruption management) (Mathaisel, 1997).

By applying the event and uncertainty abstraction principles, we identify three separate classes of features in the raw data that can be used to enable robust airline disruption management and are described as follows:

1. *Determinate aleatoric features*: These represent flight schedule features that are determined during the strategic planning phase of airline planning and are required to remain unchanged during schedule execution on the day of operation. Examples of determinate aleatoric features include



flight date, origin (departure) station, destination (arrival) station, route originator indicator, route distance, etc. With respect to disruption management, determinate aleatoric features represent flight schedule features whose alternatives do not differ considerably each time the AOCC invokes a disruption management initiative. Thus, from a statistical perspective, determinate aleatoric features are flight schedule features that are subject to the least possible uncertainty for the risk of reassessment (or alteration) during irregular operations for disruption management, based upon inherent randomness of disruption events. For instance, airport identifiers and exact longitude and latitude coordinates that provide specific information for origin and destination stations are always assumed to remain unchanged, by the AOCC, during the recovery of a delayed flight schedule. However, if a human specialist in the AOCC chooses to divert the same delayed flight to another airport during schedule execution, then the airport information for the destination station changes to that of the airport where the flight is to be diverted. It is important to note that this scenario is unlikely, based upon our research scope, because we consider irregular operations for delayed flight schedules only.

2. *Indeterminate aleatoric features*: These are separate data features from flight schedule features that represent disruption types for different functional domains, which occur randomly during schedule execution on day of operation. Examples of indeterminate aleatoric features include delay codes for uncontrollable inclement weather and controllable maintenance inspections. From a disruption management perspective, indeterminate aleatoric features represent triggers for the need of the AOCC to address a specific disruption. As such, indeterminate aleatoric features are data features that can create the most uncertain responses in disruption management initiatives employed by the AOCC during schedule execution. From a statistical perspective, indeterminate aleatoric features are data features that are subject to the most possible uncertainty for the risk of

occurrence (or instantiation) of irregular operations during schedule execution, due to inherent randomness of disruption events. For example, inclement weather at a particular airport may require a human specialist in the AOCC to delay the departure of a specific flight at the (origin) airport and reassign some or all of its passengers to another flight with a later departure, while also reallocating the arrival of the original delayed flight to a different gate at the destination airport.

3. *Epistemic features*: These represent flight schedule features that are determined during the tactical and operational phases of airline planning and can be subject to change during schedule execution on day of operation. Examples of epistemic features include specific departure and arrival times during the day, aircraft type, delay periods, actual turnaround and block time periods. With regards to disruption management, epistemic features represent flight schedule features with considerable amount of alternatives for every time the AOCC initiates a disruption management plan. As such, from a statistical standpoint, epistemic features are flight schedule features that are subject to the most possible uncertainty for the risk of alteration during irregular operations for disruption management, due to lack of knowledge of the exact impact of their alteration. For instance, following a specific disruption like late arrival of flight crew for a scheduled flight, a human specialist in the AOCC may choose to delay the departure of the flight by a specific period of time after the original departure time. However, most times, the human specialist can not guarantee that the decision on applying a particular delay duration after scheduled departure will produce a specific recovery plan, due to the cascading effect of disruptions in large airline networks.

Fig. 2 shows a generic knowledge abstraction for airline disruption management based upon some specific flight schedule features. The horizontal axis in Fig. 2 represents event abstraction for defining the value of flight operations management based upon the perishable nature of a flight service during schedule

execution, while the vertical axis represents uncertainty abstraction for defining the risk of disruption instances and schedule alteration during flight schedule planning.

### 405 3.2. Feature Transformation

While the abstraction of raw flight schedule data features provides an excellent avenue for effectively representing latent planning capabilities in airline operations control, the quality of the knowledge extracted from the raw data can be enhanced through transformation to enable discernible representation and interpretation for machine learning algorithms (Kusiak, 2001; Liu & Motoda, 1998). These algorithms provide efficient means for easily recognizing useful patterns and relationships amongst flight schedule features in a data set. To this end, feature transformation is the process of converting flight schedule and disruption features in raw historical airline scheduling and operations data into relevant mathematical properties (or functions) that can be readily understood by machine learning algorithms. Every direct flight schedule in the raw data set is defined by forty separate data features (or attributes) that describe different resources, behaviors, and performance indicators that are observable during airline scheduling and disruption management. As such, raw flight schedule features can be separated into four distinct categories namely: geographical features, temporal features, categorical features, and continuous features.

- *Geographical features*: These are flight schedule features which represent resources, behaviors, or performance indicators that require or enable the perception and property of geographic location (or position) during airline disruption management. Examples of geographical features in the raw data set are International Air Transport Association (IATA) codes for departure and arrival airport stations, and the identifier for the origin of the first departure flight of the day.
- *Temporal features*: These are flight schedule features that describe and enable the perception and property of time during airline scheduling and

disruption management. Temporal features are conceptualized by four different types of time (Shurkhovetskyy et al., 2018) namely:

1. *ordinal time*: This represents time points that occur one after another on day of operation. Examples of flight schedule features defined by ordinal time are time-of-day events such as aircraft pushback time, takeoff time, landing time, and aircraft gate-parking time.
  2. *interval time*: This represents time events that are measured on an interval scale with a specific duration (or length). Examples of flight schedule features characterized by interval time include the duration between aircraft pushback and aircraft gate-parking otherwise known as blocktime, the duration for boarding passengers and loading cargo onto an aircraft also known as turnaround, and the duration of any form of delay in airline operations during schedule execution.
  3. *cyclic time*: This describes cyclic or repeatable processes wherein the application of an ordered relation is inane. Flight date is an example of a flight schedule feature characterized by cyclic time.
  4. *branching time*: This represents time points that can occur in different branches or alternatives to describe several scenarios or processes. Thus, all temporal flight schedule features in the raw data set are defined by branching time.
- *Categorical features*: These are flight schedule features that represent fields in the raw data defined by discrete values which belong to a finite set of categories or classes. Categorical features can be text or numeric, and are separated into two classes namely nominal and ordinal, based upon the perception of ordering.
    1. *nominal*: Nominal categorical features represent flight schedule features for which there is no concept of ordering among different values of each feature. An example of a nominal categorical feature is A0,

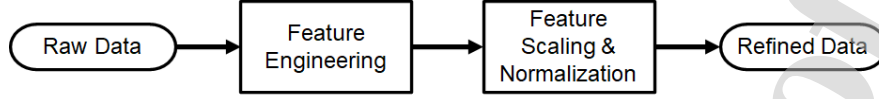


Figure 3: Feature transformation process of raw data for airline disruption management

which is a binary number (i.e. 0 or 1) indicating whether or not a flight schedule arrives exactly on time.

2. *ordinal*: Ordinal categorical features represent flight schedule features for which there is a strict adherence to the concept of ordering among different values of each feature. An example of an ordinal categorical feature in flight schedule data is aircraft type, which effectively characterizes the relevance of size and seat capacity for aircraft performance.

- *Continuous features*: These are flight schedule features that represent fields in the raw data, which have infinitely many alternatives between any two values. Examples of continuous features in raw flight schedule data include digital timestamps for different time-of-day events (such as takeoff time) during schedule execution.

Fig 3 reveals a two-layer data transformation process of raw flight schedule data features for airline disruption management. The first layer, known as feature engineering, enables the creation of additional data features from mathematical functions that characterize rudimentary properties of the airline operations control center. Next, these data features are combined with extant low-level data features in the raw data set and normalized by using fundamental statistical parameters in the second layer through a process called feature scaling.

- (i) **Feature engineering** represents a *first-degree* transformation of raw flight schedule data that defines the augmentation of properties associated with daily routines of functional roles in the AOCC by using mathematical

principles. Thus, flight schedule features representing geographic locations (i.e. geographical features) such as departure and arrival stations are first transformed into spherical directional vectors based upon the longitude and latitude coordinates of their corresponding airport stations, and subsequently transformed into the distance between the departure and arrival airports on an oblate spheroid Earth via the Vincenty geodesic equation (T. Vincenty, 1975). Ordinal temporal flight schedule features (such as departure and arrival times) are transformed into two separate periodic (i.e. sine and/or cosine) vectors of different amplitudes, based upon a 24-hr clock period and the percentage of 8-hr work shift completed (at the time of departure or arrival) by human specialists in the AOCC, respectively. The work shift characterization of time-of-day events, via a periodic vector, is intended to capture and represent daily disruption resolution proclivities of human specialists, which can be induced by how much time the specialists have to address a disruption before their work shift is complete. Cyclic temporal flight schedule features defined by Gregorian dates are transformed into four separate periodic vectors, whose periods are based upon the season of the year, month of the year, day of the week, and day of the year respectively.

Lastly, most categorical flight schedule features defined by texts are transformed into binary numbers by one-hot encoding (Seeger, 2018), wherein all  $n$  feature values are represented as a  $n$ -dimensional sparse vector with zero entries except for one of the dimensions for which the entry is one. However, categorical flight schedule feature values for aircraft type, which are defined by aircraft model codes, are transformed into discrete numbers based upon the total number of available seats in the aircraft. A secondary objective of feature engineering is to provide a precursor for continuous feature representation by ensuring that all data features are in numeric form. As such, feature engineering may not be applicable to flight schedule features that are already in continuous form in a raw data

set.

- (ii) **Feature Scaling** represents a *second-degree* transformation of raw numeric data features that include additional features created from feature engineering, such that a uniform statistical grounding basis is used to transform values for all data fields (i.e. flight schedule features) in the raw data set into bounded continuous values that describe a differentiable function. We explore three different approaches for enabling feature normalization (Pedregosa et al., 2011) namely: *Standard scaling*, *Range scaling*, and *Power scaling*.

- (a) *Standard scaling*: Standard scaling or standardization normalizes values for each flight schedule feature in the data set by removing the mean of the values and scaling to a unit variance, thus resulting in a standard score for each value. The standard score,  $z_i$ , of an arbitrary sample (i.e data feature value),  $x_i$ , in the data set is calculated as follows:

$$z_i = \frac{(x_i - u)}{s} \quad (1)$$

where  $u$  and  $s$  represent the mean and standard deviation, respectively, of all values for each flight schedule feature in the data set. As such, standardization provides a platform to ensure that each flight schedule feature in the data set follows a Gaussian distribution with zero mean and a variance of one.

- (b) *Range scaling*: Range scaling, otherwise known as min-max normalization, transforms each flight schedule feature in the data set by scaling the values of each feature by the difference between the maximum value and the minimum value (i.e. range). This results in adjusted values of a range (or distance) between zero and one for each flight schedule feature. The adjusted (range-scaled) value,  $y_i$ , of a characteristic flight schedule feature in the data set is calculated

as follows:

$$y_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (2)$$

where  $x_i$  and  $X$  represent an original value and set of all original values, respectively, for a flight schedule feature.

- (c) *Power scaling*: Power scaling involves adapting a family of parametric and monotonic transformations to convert flight schedule data values from any distribution to the closest possible representation of Gaussian distribution, so as to reduce variance and skewness in data. An appropriate power transformation of flight schedule and disruption features is the Yeo-Johnson transform (Weisberg, 2001), because it can be applied to all forms of numeric data just like standard and range scaling transforms. The Yeo-Johnson transform is given by:

$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i + 1) & \text{if } \lambda = 0, x_i \geq 0, \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases} \quad (3)$$

where  $x_i$  and  $\lambda$  represent an original data value and an arbitrary parameter that is determined through maximum likelihood estimation (Glas, 2017), respectively.

Completion of the feature transformation process shown in Fig. 3 results in a refined, continuous data set that can be readily comprehensible by suitable machine learning estimators.

### 3.3. Dimensionality Reduction

The efficacy of constructing and applying relevant machine learning algorithms for identifying and acknowledging high-level properties from data features (such as flight schedule and disruption data features) is dependent on the



form in which the data values are presented. To this end, feature transformation has a strong propensity to increase the number of elements in the flight schedule and disruption feature space that constitutes the problem dimensions for airline disruption management. As such, the intrinsic dimensionality of the refined data appropriated for airline disruption management is defined by the least number of flight schedule and disruption features required to delineate observed behavioral properties from AOCC routines. Hence, dimensionality reduction is the process of mitigating the curse of dimensionality (Verleysen & François, 2005) and other unwanted properties of high-dimensional feature space through classification, visualization, and compression of high dimensional data obtained as a result of feature transformation (Van Der Maaten & Hinton, 2008). In essence, dimensionality reduction aims to provide a rudimentary means to attain and observe the latent feature space of a refined data set for airline disruption management.

From a mathematical perspective, we assume that the refined flight schedule and disruption data set is represented in a  $n \times m$  matrix  $\mathbf{X}$ , which consists of  $n$  feature vectors  $\mathbf{x}_i (i \in \{1, 2, \dots, n\})$  with dimensionality  $m$ . Furthermore, we assume that the refined data set has an intrinsic dimensionality  $d$ , such that  $d < m$  and often  $d \parallel m$ . The intrinsic dimensionality property refers to points in the refined data set  $\mathbf{X}$ , which lie near a manifold with dimensionality  $d$  that is embedded in the  $m$ -dimensional feature space. To that effect, dimensionality reduction techniques transmute the refined flight schedule and disruption data set  $\mathbf{X}$  with dimensionality  $m$  into a new data set  $\mathbf{Y}$  with dimensionality  $d$ , while maintaining the geometry of the refined data set  $\mathbf{X}$  as much as possible. Typically, the intrinsic dimensionality  $d$  and the geometry of the manifold of the new data set  $\mathbf{Y}$  are unknown, and as such, most dimensionality reduction techniques require that certain assumptions about the properties (like intrinsic dimensionality) of the refined data set be made a priori. For the remainder of this section, we denote a high dimensional data instance for flight schedule and disruption (i.e. datapoint) by  $\mathbf{x}_i$ , such that  $\mathbf{x}_i$  is the  $i^{th}$  row of the refined  $m$ -dimensional data set  $\mathbf{X}$ . In complement, the low-dimensional equivalent of

570  $\mathbf{x}_i$  is expressed by  $\mathbf{y}_i$ , where  $\mathbf{y}_i$  is the  $i^{th}$  row of the new  $d$ -dimensional matrix  $\mathbf{Y}$ .

To demonstrate the usefulness of dimensionality reduction on refined flight schedule and operations data, we investigate two separate techniques that employ linear and nonlinear principles nicknamed *PCA* and *t-SNE*, respectively, by  
 575 utilizing delayed flight schedule and disruption instances for the Weather functional domain in SWA-NOC between September 2016 and September 2017. It is important to note that the flight schedule data for the Weather functional domain constitutes a subset (with 12,659 delayed flight schedule instances) of the full refined data set. For validation, we adopt min-max normalization for scaling  
 580 all feature values in the refined data set because both dimensionality reduction techniques strongly depend on Euclidean distances between refined high-dimensional datapoints  $\mathbf{x}_i$  and  $\mathbf{x}_j$  to obtain and simplify the gradient of their respective cost functions (Van Der Maaten & Hinton, 2008; Van Der Maaten et al., 2009).

### 585 3.3.1. Principal Component Analysis (PCA)

Principal component analysis or *PCA* is a standard non-parametric tool in modern data analysis used for extracting relevant information from large and confusing data sets (Shlens, 2014). *PCA* is also a full spectral linear technique for dimensionality reduction that embeds data into a linear subspace of lower  
 590 dimensionality. In the lower dimension, the refined variables (or data features) in the data set are transformed into linear combinations of the data features, which are called principal components. With minimal effort, *PCA* provides a schema for reducing a fairly complex data set to a lower dimension in order to show simplified structures that often define it, by revealing as much of the variance  
 595 in the data as possible. As such, the first and second principal components are the orthogonal linear combinations of the refined data features that have the largest and second-largest possible variance (or inertia), respectively, in the refined data set. In mathematical terms, *PCA* aims to find a linear mapping  $\mathbf{X}$  that maximizes the variance (or minimizes the reconstruction error) defined

by  $\text{trace}(\mathbf{S}^T \text{cov}(\mathbf{X}) \mathbf{S})$ , wherein  $\text{cov}(X)$  is the sample covariance matrix of the refined data (Wold et al., 1987). Thus, the linear mapping created by the  $d$  principal components (or principal eigenvectors) are solutions to the eigenproblem defined as follows:

$$\text{cov}(\mathbf{X}) = \lambda(\mathbf{S}) \quad (4)$$

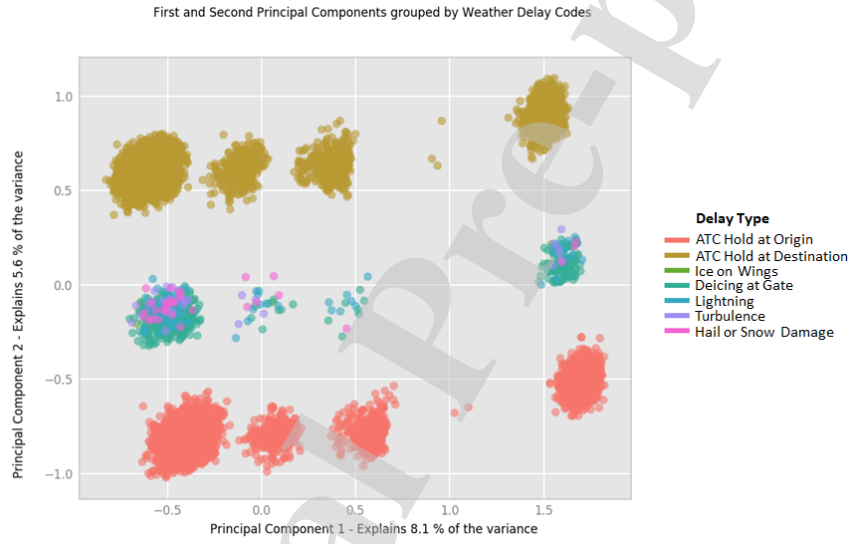


Figure 4: Principal component analysis of indeterminate aleatoric features for Weather domain

The lower dimensional representation of the refined flight schedule and disruption feature instances, defined by  $\mathbf{y}_i$  of  $\mathbf{x}_i$  datapoints, are computed by mapping them onto a linear basis  $\mathbf{Y} = \mathbf{X}\mathbf{S}$  that solves the eigenproblem for the  $d$  principal eigenvalues defined by  $\lambda$ , via the scikit-learn software (Pedregosa et al., 2011).

Fig. 4 presents a visualization of the analysis for the first two principal components describing indeterminate aleatoric features that represent delays for the Weather functional domain in SWA-NOC. The first and second principal components represent orthogonal linear combinations of the refined flight schedule

features that account for 8.1% and 5.6%, respectively, of the variance of indeterminate aleatoric features related to weather delays in the data set. Fig. 4 reveals that there are four major types of weather-related delays (*ATC Hold at Origin*, *ATC Hold at Destination*, *Deicing at Gate*, and *Hail or Snow Damage*) in the data set. *ATC Hold at Origin* and *ATC Hold at Destination* represent weather delays due to gate hold from air traffic control (ATC) at departure and arrival stations respectively. *Deicing at Gate* and *Hail or Snow Damage* represent weather delays due to deicing at the gate, and aircraft swap due to hail or snow damage respectively. Fig. 4 shows that the data set is divided into four separate clusters of the same delay type along the axis of the first principal component. This reveals that the axis of the first principal component (horizontal) represents linear combinations of flight schedule features that capture the seasonal behavior of weather-related delays, as each data cluster describes each weather season over the one-year data-collation period. Furthermore, the data set is divided into two polarizing (*ATC Hold at Origin* and *ATC Hold at Destination*) and two overlapping (*Deicing at Gate* and *Hail and Snow Damage*) clusters along the second principal component (vertical) axis. This shows that the axis of the second principal component represents linear combinations of flight schedule features that capture the difference in the types of indeterminate aleatoric features for weather-related delays in the refined data set.

The patterns and information gleaned from the results and observations from the *PCA* method can be appropriated for informing model development for airline disruption management. For instance, the seasonal relationship among the four predominant types of weather-related delays observed in Fig. 4 can be quantified via a linear combination of flight schedule features, which suggests that decision-making by human specialists in the AOCC is sensitive to weather seasons. In addition, the overlapping effect observed between *Deicing at Gate* and *Hail and Snow Damage* in Fig. 4 is representative of similarities in the type of disruption resolutions used for weather-related delays during the winter season, thus bolstering the significance of the effect of seasonal properties on decision-making by human specialists in the AOCC.

### 3.3.2. *t*-distributed Stochastic Neighborhood Embedding (*t*-SNE)

t-distributed Stochastic neighborhood embedding or *t*-SNE represents a recent advancement in clustering and visualization for dimensionality reduction that provides a nonlinear platform for transforming the Euclidean distances between refined values (i.e. datapoints) of flight schedule and disruption features into conditional probabilities that define similarities. As such, the similarity of a datapoint  $x_j$  to another datapoint  $x_i$  is the conditional probability ( $p_{j|i}$ ) that  $x_i$  will select  $x_j$  as its neighbor if neighbors are selected in proportion to their probability density under a Student-t distribution with one degree of freedom (i.e. Cauchy distribution) centered about  $x_i$  (Maaten, 2014; Van Der Maaten & Hinton, 2008). Thus,  $p_{j|i}$  remains comparatively high for datapoints in close proximity and insignificant for datapoints that are substantially separated. Mathematically, the objective of *t*-SNE is to minimize the Kullback-Leibler divergence (Joyce, 2011; Pérez-Cruz, 2008) between a joint probability distribution defined by  $P$  in the high-dimensional feature space and a joint probability distribution defined by  $Q$  in the low-dimensional feature space. Hence, the Kullback-Leibler divergence represents the cost function of the following optimization problem, which is solved via the scikit-learn software (Pedregosa et al., 2011):

$$\begin{aligned} \min \quad & \text{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \\ \text{s.t.} \quad & p_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} \exp(-||x_i - x_k||^2/2\sigma^2)} \\ & q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_i - y_l||^2)^{-1}} \end{aligned} \quad (5)$$

where  $p_{ii}$  and  $q_{ii}$  are set to zero, and  $p_{ij} = p_{ji}$  and  $q_{ij} = q_{ji}$  for all  $i, j$ .

Fig. 5 presents a strictly visual perception of the *t*-SNE nonlinear projection of the two-dimensional space for flight schedule features, which describes indeterminate aleatoric features that represent delays for the Weather functional domain in SWA-NOC. Similar to observations from *PCA*, the red and gold clusters in Fig. 5 reveal that weather-related delays due to *ATC Hold at Origin* and *ATC Hold at Destination* are the most prominent and oppositely related, based

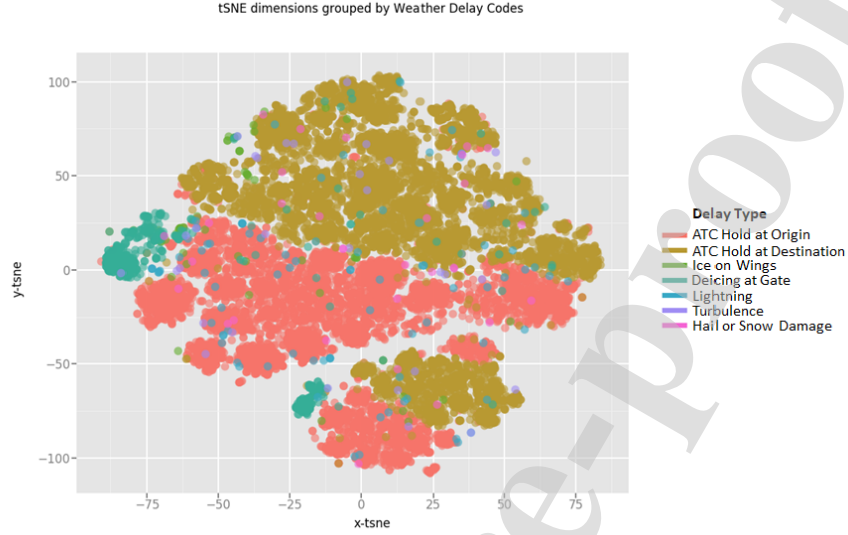


Figure 5: t-Distributed Stochastic Neighborhood Embedding analysis of indeterminate aleatoric features for Weather domain

upon the symmetry observed from the small and large blobs of red and gold clusters. As such, the polarizing effect observed between *ATC Hold at Origin* and *ATC Hold at Destination* in Figs. 4 and 5 can be attributed to the importance of the geographical location (i.e departure or arrival stations) on how weather-related disruption resolutions are applied in the AOCC. As such, flight schedule features associated with geographical location are relevant for creating robust models for airline disruption management.

### 3.4. Feature Selection

Although dimensionality reduction techniques provide an effective means to readily (i.e. visually) discern high-level patterns and properties associated with a data set, they are ineffectual in revealing detailed information on the specific importance of flight schedule and disruption features in a data set and their corresponding relationships (Koller & Sahami, 1996). To this end, feature selection presents simple fundamental methods for efficiently selecting and

investigating pertinent associations among data features in a refined data set, which can provide insightful knowledge (or a priori information) for developing useful data-driven models for robust airline disruption management. In essence, feature selection methods aim to proactively enhance model prediction performances by increasing generalization (i.e. minimize data overfitting) and decreasing model runtimes (Moran & Gordon, 2019). There are three major categories of feature selection methods namely: wrapper, filter, and embedded methods.

Wrapper methods involve algorithms that search the feature space for plausible subsets of features by assessing each subset after running a specific model. Typically, the model is validated on a test data set to estimate the model's error rate, before a score is registered for each feature subset and the feature subset with the best score is ultimately selected. Unlike computationally intensive wrapper methods, filter methods do not consider a model when searching the feature space for relevant subsets of the feature space, and rely on general statistical measures such as Pearson correlation coefficient (Benesty et al., 2009) and mutual information (Jiang et al., 2010). In this manner, filter methods are somewhat analogous to dimensionality reduction techniques, such that they are not customized to a particular type of predictive model and consume significantly less computational resources than wrapper methods. Embedded methods involve feature selection methods that are entrenched in a specific learning algorithm that performs classification (or regression) and feature selection concurrently. As such, embedded methods deliver the advantages of both wrapper and filter methods with medium computational expense.

To demonstrate the relevance of feature selection on refined flight scheduling and operations data, we apply two specific types of feature selection that belong to the filter and embedded categories, respectively, to identify flight schedule features that are pertinent for disruption management during turnaround. Turnaround is an airline process (or time period) primarily representative of loading, unloading and occasional servicing of aircraft, and is crucial for minimizing overall flight schedule delays. In addition to reducing overall flight delay,

most airlines typically aim to expedite the turnaround process as much as possible in order to avoid causing discomfort to passengers, stemming from long waits in the aircraft on the ground, thus invariably minimizing loss of passenger goodwill.

In that regard, the filter method that we apply is defined by mutual information and the embedded method is defined by a Gaussian process, such that actual turnaround duration is set as the target flight schedule feature. We do not consider wrapper methods in our discussion because of the significant computational expense required as compared to filter and embedded methods. Similar to the dimensionality reduction analysis discussed in Section 3.3, we utilize delayed flight schedule and disruption instances for the Weather functional domain in SWA-NOC between September 2016 and September 2017 for our analysis. For validation, we adopt standardization for the *second-degree* transformation (i.e. scaling) of all feature (i.e label and target) values in the refined data set, because the algorithms for both filter and embedded methods perform best with a zero-mean Gaussian distribution as prior instantiation for each feature space in the refined data set (C. E. Rasmussen & Williams, 2006; Jiang et al., 2010; Ross, 2014).

#### 3.4.1. Mutual Information Regression (MIR)

Mutual information is a non-negative measure from information theory that provides an excellent statistic for quantifying the degree of relatedness among flight schedule and disruption features in a refined data set. In that regard, mutual information is closely related to the entropy of a flight schedule feature based upon observing another flight schedule feature in a refined data set (Kraskov et al., 2004). In addition to the ability to readily identify relationships amongst data features, mutual information provides a fundamental metric for straightforward interpretation of the relationships among data features as shared information (i.e. shannons or bits) between data features. To that effect, mutual information is insensitive to the number of instances in a data set (Ross, 2014).



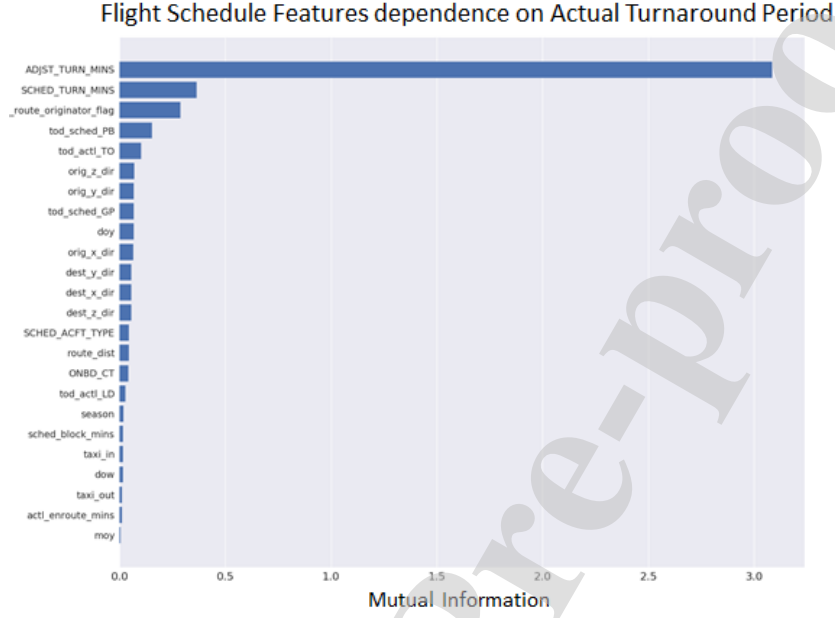


Figure 6: Mutual information of flight schedule data features for Weather Domain with respect to actual turnaround duration

Mathematically, mutual information,  $I$ , is expressed as:

$$I(X; Y) = \int_X \int_Y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (6)$$

where  $p(x, y)$  is the joint probability density function of  $X$  and  $Y$ , and  $p(x)$

and  $p(y)$  are the marginal probability density functions of  $X$  and  $Y$  respectively.

Thus, for feature selection, the objective of *MIR* is to maximize the mutual information between a subset of flight schedule features defined by  $\mathbf{X}_s$  and a target flight schedule feature defined by  $y$  as represented by the following optimization problem:

$$s^* = \arg \max_s I(\mathbf{X}_s; y) \quad \text{s.t.} \quad |s| = k \quad (7)$$

where  $k$  is the number of features that are to be selected. A non-parametric regression algorithm based upon entropy estimation from k-nearest neighbor

distances is used to solve the NP-hard optimization problem via the scikit-learn software (Pedregosa et al., 2011), for a set of possible combinations of data features that increases exponentially (Gao et al., 2015; Kraskov et al., 2004).

Fig. 6 shows the mutual dependence, in decreasing order, of flight schedule features (i.e. determinate aleatoric and epistemic features) on actual turnaround period for instances of flight delay from the Weather functional domain. Fig. 6 reveals that turnaround duration adjusted during schedule execution has the highest mutual information of over 3, thus implying that it has the strongest mutual dependency on the decision-making for estimating actual turnaround duration during disruption management. In addition, turnaround period estimated prior to schedule execution and a flights capacity to be a route originator (i.e. first departure flight of the day) have the second and third highest mutual information of 0.4 and 0.3 respectively, thereby revealing a weak mutual dependency on the estimation of actual turnaround schedule for managing weather-related delay of flight schedule. Month of the year (*moy*) can not be selected as a significant predictor for estimating actual turnaround duration because of zero mutual dependency as shown in Fig. 6.

#### 3.4.2. Gaussian Process Regression (GPR)

A Gaussian process is a stochastic process (i.e. random variables indexed by time or space) where a finite collection of random variables have a multivariate normal distribution (C. E. Rasmussen & Williams, 2006). As such, Gaussian Process Regression or *GPR* is the inference of continuous feature values with a Gaussian process (or distribution) prior, such that the marginal likelihood of the data is maximized (C. E. Rasmussen & Williams, 2006). Converse to *MIR*, which is a model-free method for dimensionality reduction and feature selection, *GPR* is an embedded method that offers nonlinear and non-parametric regression properties that enable the natural decomposition of flight schedule features in an airline data set for simultaneously attaining high fidelity dimensionality reduction and feature selection. *GPR* provides an appropriate medium to obtain the sensitivity and importance of flight schedule and disruption features neces-

sary for informing the development of appropriate models for airline disruption management.

Following the nomenclature in C. E. Rasmussen & Williams (2006), consider  
 785 a training data set defined by  $\mathcal{D}$  with  $n$  observations, such that  $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\}$ , where  $\mathbf{x}$  represents an input vector of flight schedule features of dimension  $D$  and  $y$  is a scalar target or output feature. As such, the column input vector of flight schedule features for all  $n$  instances are collected in a matrix  $\mathbf{X}$  of size  $D \times n$ , and scalar outcomes of the target feature are collected in the vector  
 790  $\mathbf{y}$  to yield  $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ . GPR assumes that outcomes of a target feature defined by  $\mathbf{y}$  are noisy observations of an unknown function of the input flight schedule features  $\mathbf{x}$  such that  $y = f(\mathbf{x}) + \epsilon$ , where  $\epsilon$  represents independent and identically distributed zero mean Gaussian random variables with unknown variance  $\sigma_n^2$ . Thus, a Gaussian process prior is placed over all values of  $f(\mathbf{x})$  before  $y$  is  
 795 observed, such that  $f(\mathbf{x}) | \theta \sim \mathcal{GP}(m(\mathbf{x}; \theta), K(\mathbf{x}, \mathbf{x}; \theta))$  and  $m(\mathbf{x})$ ,  $K(\mathbf{x}, \mathbf{x})$  and  $\theta$  represent a mean function, a covariance function, and all hyperparameters that influence the mean and covariance functions, respectively. A combination of Bayes' rule and Gaussian identities (C. E. Rasmussen & Williams, 2006) provides a method to retrieve the posterior distribution of the values of  $f(\mathbf{x})$  after  
 800 observing  $y$ , such that  $\sigma_n^2 \in \theta$  albeit the mean and covariance functions are independent of  $\theta$  (Lee et al., 2020). As such, the posterior distribution is defined by the analytical solution for  $\mathbf{f}$ :

$$\mathbf{f} | \mathbf{y}, \mathbf{X} \sim \mathcal{N}(\mu + K K_n^{-1}(\mathbf{y} - \mu), K - K K_n^{-1} K), \quad (8)$$

where  $\mu$  is the prior mean of  $\mathbf{f}$ ,  $K$  represents the prior covariance function evaluated at  $\mathbf{X}$ , and  $K_n^{-1} = (K + \sigma_n^2 I)^{-1}$ . In order to obtain predictions of target outcomes for new test cases represented by a collection of input flight schedule features  $\mathbf{X}^*$ ,  $\mathbf{f}^*$  is defined as:

$$\begin{aligned} \mathbf{f}^* | \mathbf{y}, \mathbf{X}, \mathbf{X}^* &\sim \mathcal{N}(\mu^* + K(\mathbf{X}^*, \mathbf{X}) K_n^{-1}(\mathbf{y} - \mu), \\ &K(\mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X}) K_n^{-1} K(\mathbf{X}, \mathbf{X}^*)), \end{aligned} \quad (9)$$

where  $\mu^*$  is the prior mean function evaluated at  $\mathbf{X}^*$ .

Computationally, *GPR* estimates the maximum marginal log likelihood of the distribution for the target flight schedule feature in a training data set, such that hyperparameters (or lengthscales) associated with all input flight schedule features that define different drivers of disruption management by the AOCC are optimized with respect to a certain kernel (covariance) function.

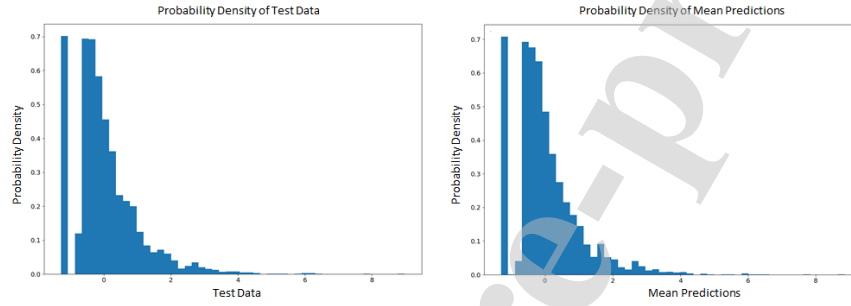


Figure 7: Probability densities of test turnaround duration and mean predictions of turnaround duration for delayed flight schedule instances in Weather domain

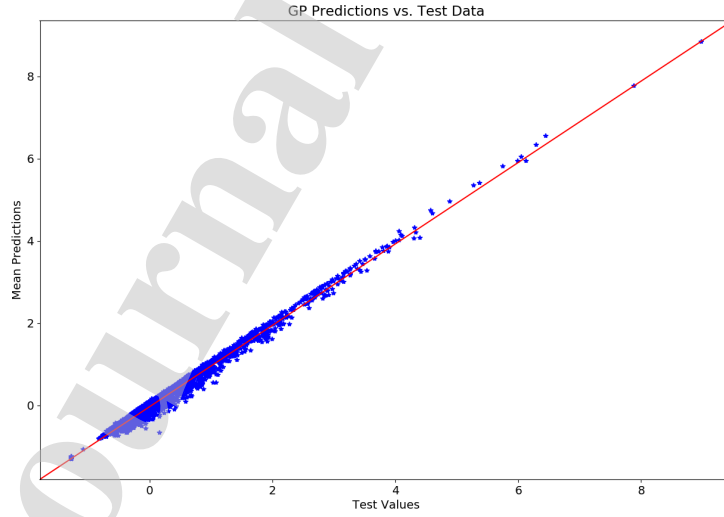


Figure 8: Mean predicted turnaround duration data vs. test turnaround duration data

As previously mentioned, a subset of the data set defined only by flight  
 810 schedules delayed by weather incidents is used for the *GPR* demonstration.  
 This subset of data is split into two separate sets of training and test (new)  
 data respectively, such that the training data (70% of the data subset randomly  
 selected) is used to fit the *GPR* model for actual turnaround duration and  
 the test data (i.e. remaining 30% of the data subset that is unseen) is used  
 815 to validate the model by verifying that the test data is consistent with mean  
 predictions from the model. Plotting the probability density function of the test  
 data, revealed a lognormal distribution of the actual turnaround duration in the  
 data set, as evidenced by Fig. 7. Hence, the Matern32 kernel function, which is  
 a combination of Gamma and Bessel functions correlated by an hyperparameter  
 820 of  $3/2$ , is selected to fit the *GPR* model by means of a Gaussian process software  
 named GPpy (SheffieldML, 2014).

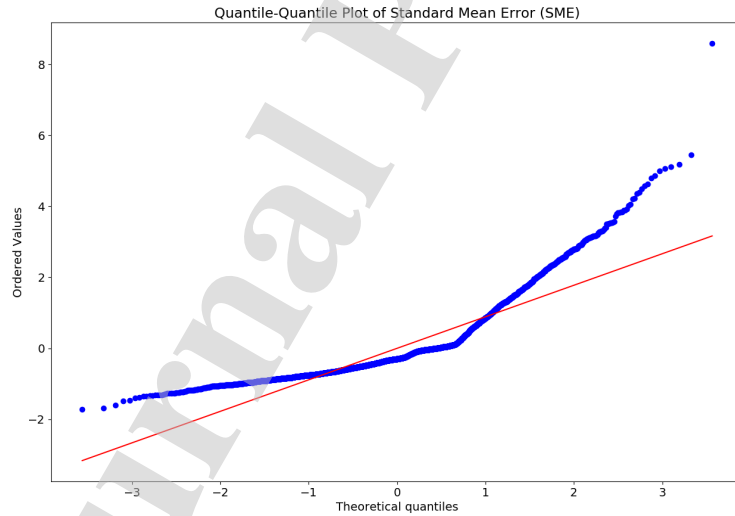


Figure 9: Quantile-Quantile plot of standard mean error between predicted and test data for actual turnaround duration

Fig. 8 shows the plot of the mean predictions of the actual turnaround du-

Table 2: Lengthscales of refined data features for predicting actual turnaround duration through Gaussian process regression

GPR.Mat32.lengthscale	Feature Class	Feature Name
1738.68	Determinate Aleatoric	<i>sin_date</i>
1650.37	Determinate Aleatoric	<i>cos_date</i>
1857.16	Determinate Aleatoric	<i>orig_x_dir</i>
1808.49	Determinate Aleatoric	<i>orig_y_dir</i>
1652.50	Determinate Aleatoric	<i>orig_z_dir</i>
1750.13	Determinate Aleatoric	<i>ONBD_CT</i>
1193.69	Determinate Aleatoric	<i>SCHED_TURN_MINS</i>
44.77	Epistemic	<i>ADJST_TURN_MINS</i>
1367.11	Determinate Aleatoric	<i>schd_acft_type</i>
1367.11	Epistemic	<i>actl_acft_type</i>
1406.15	Epistemic	<i>SWAP_FLT_FLAG</i>
975.98	Indeterminate Aleatoric	<i>ATC Hold at Origin</i>
972.46	Indeterminate Aleatoric	<i>ATC Hold at Destination</i>
1.00	Indeterminate Aleatoric	<i>Deicing at Gate</i>
156.75	Indeterminate Aleatoric	<i>Ice on Wings</i>
1.00	Indeterminate Aleatoric	<i>Lightning Strike</i>
1.00	Indeterminate Aleatoric	<i>Turbulence</i>
1.00	Indeterminate Aleatoric	<i>Hail or Snow Damage</i>

ration from the *GPR* model versus the actual turnaround duration from the test data, for which the turnaround duration values in both axes are scaled to a unit variance from the mean of the data values. The red diagonal line in Fig. 8 represents the 45-degree line, while each blue star represents a coordinate of the mean *GPR* prediction and test data describing actual turnaround duration for each datapoint (i.e. instance of weather-delayed flight schedule). Fig. 8 shows that the coordinates for the datapoints follow the trend of red diagonal line almost perfectly (root mean square error of 9%), which indicates that the *GPR* model is able to effectively predict “unknown” actual turnaround duration. Each coordinate that falls on the red line implies an exact prediction of the test data by the *GPR* model, and as such, Fig. 8 shows that the *GPR* model perfectly predicts actual turnaround periods that lie over six standard deviations away from the mean.

Table 2 shows the values of the optimized hyperparameters (i.e. length-scales) of refined flight schedule and disruption features for estimating actual turnaround time. Lower values in Table 2 indicate higher importance of features for predicting actual turnaround period. Similar to the result from *MIR*, turnaround duration adjusted during schedule execution (i.e. *ADJUST\_TURN\_MINS*) is the most significant epistemic flight schedule feature for predicting actual turnaround duration, as indicated by its low lengthscale value of approximately 45. Of all the aleatoric features (determinate and indeterminate), disruption features representing deicing at the gate, lightning strike, turbulence, and hail and snow damage (all with lengthscale values of 1) are the most significant for accurately estimating actual turnaround period during schedule execution.

Fig. 9 shows the quantile-quantile (QQ) plot of the standard mean error (SME) between the mean predictions from the *GPR* model and the test data for actual turnaround duration. The straight red line in Fig. 9 represents the trend line for a standard normal distribution, and as such, the bilinear trend for standard mean error (portrayed by the blue dots) in Fig. 9 confirms that the distribution of actual turnaround period for weather-delayed flight schedules is lognormal. Categorically, the overlapping trend between the 45-degree red line

and the spread of the coordinates in Fig. 8, coupled with the lognormal distribution trend noted from the QQ plot in Fig. 9, validates that the turnaround process for weather-delayed flight schedules is indeed a Gaussian process. To that effect, the relationship between the data features (shown in Table 2) and the actual turnaround duration during schedule execution (i.e. target feature) can be described by a Matern32 covariance function.

#### 4. Conclusion

This paper provided macroscopic and microscopic analysis of the historical airline scheduling and operations data necessary for addressing disruption management at a major airline in the United States. Through macroscopic analysis, we identified that over 94% of irregular operations over a one-year period occurred due to different forms of flight schedule delays. To that end, we investigated crucial drivers and properties for effectively managing flight schedule delays through microscopic analysis of weather-delayed flight schedule data, which also demonstrated a toolbox for applying appropriate machine learning techniques to enable data-driven schedule recovery during airline disruption management.

Furthermore, machine learning techniques such as feature transformation and dimensionality reduction provided excellent platforms for verifying empirical processes and validating domain knowledge for airline disruption management. The patterns and information gleaned from the results and observations from visual techniques (such as *PCA* and *t-SNE*) identified the viability of seasonal (temporal) and geographical features in the data set as suitable predictors for artificial intelligence (AI) models that describe disruption management tendencies of the airline. The results obtained from qualitative methods (such as *MIR* and *GPR*) established the importance of the predictors and the underlying processes that define their combination for high fidelity AI models. As such, the findings from our research reveal that the statistics of machine learning can be used to characterize and assess the functional parts of the AOCC for intelligent



decision support systems.

While this paper provides high fidelity exploratory data analysis for airline  
 885 disruption management, there exist a few areas for further research. First,  
 the data utilized for the analysis was provided by a single U.S. airline that  
 operates a point-to-point route network and may not completely represent the  
 disruption management propensities of other airlines in the United States. As  
 such, there is a need for exploratory analysis on a data set from an airline  
 890 that operates a hub and spoke route network structure. Second, the machine  
 learning techniques demonstrated in this paper are only analyzed with a data  
 set of weather-delayed flight schedules that represent the Weather functional  
 domain. Hence, in a sequel to this paper, we extensively discuss the processes  
 and routines used to obtain high fidelity AI models through analysis of the  
 905 performance of machine learning techniques on multiple data sets that represent  
 different functional domains in the AOCC.

### Acknowledgement

The authors would like to thank Blair Reeves, Chien Yu Chen, Kevin Wiecek,  
 Jeff Agold, Dave Harrington, Rick Dalton, and Phil Beck, at Southwest Airlines  
 900 Network Operations Control (SWA-NOC), for their expert inputs in abstracting  
 the data used for this work.

### Conflict of Interest

All authors have no conflict of interest to report.

### References

905 Abbott, R. J. (1987). Knowledge abstraction. *Communications of the ACM*,  
 30, 664–671. doi:10.1145/27651.27652.

Amadeus IT Group (2016). Airline Disruption  
 Management. URL: <http://www.amadeus.com/>

- documents/airline/airline-disruption-management/  
 910 airline-disruption-management-whitepaper-2016.pdf.
- Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. doi:10.1007/978-3-642-00296-0{\\_}5.
- C. E. Rasmussen, & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*.
- 915 Gao, S., Ver Steeg, G., & Galstyan, A. (2015). Estimating mutual information by local Gaussian approximation. *Uncertainty in Artificial Intelligence - Proceedings of the 31st Conference, UAI 2015*, (pp. 278–285).
- Gershkoff, I. (2016). Shaping the future of Airline Disruption Management (IROPS), . (pp. 1–32). URL:  
 920 <https://amadeus.com/documents/en/airlines/white-paper/shaping-the-future-of-airline-disruption-management.pdf>.
- Glas, C. A. (2017). Maximum-likelihood estimation. In *Handbook of Item Response Theory: Volume Two: Statistical Tools*. doi:10.1201/9781315373645.
- Grosche, T. (2009a). *Computational Intelligence in Integrated Airline Scheduling*.  
 925
- Grosche, T. (2009b). Integrated airline scheduling. *Studies in Computational Intelligence*, 173, 59–171. doi:10.1007/978-3-540-89887-0{\\_}4.
- Hagel, J., Brown, J. S., De Maar, A., & Wooll, M. (2017). Southwest Airlines: Baker workgroup - Reducing disruption and delay to accelerate performance.
- 930 Huh, S. Y., Moon, K. H., & Lee, H. (2000). Data abstraction approach for query relaxation. *Information and Software Technology*, 42, 407–418. doi:10.1016/S0950-5849(99)00100-7.
- Jiang, A. H., Huang, X. C., Zhang, Z. H., Li, J., Zhang, Z. Y., & Hua, H. X. (2010). Mutual information algorithms. *Mechanical Systems and Signal Pro-*

- 935 *cessing*, 24, 2947–2960. URL: <http://dx.doi.org/10.1016/j.ymssp.2010.05.015>. doi:10.1016/j.ymssp.2010.05.015.
- Joyce, J. M. (2011). Kullback-Leibler Divergence. In *International Encyclopedia of Statistical Science*. doi:10.1007/978-3-642-04898-2{\\_}327.
- Koller, D., & Sahami, M. (1996). Toward Optimal Feature Selection. In *Inter-*  
 940 *national Conference on Machine Learning*.
- Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, . doi:10.1103/PhysRevE.69.066138.
- Kusiak, A. (2001). Feature transformation methods in data mining. *IEEE*  
 945 *Transactions on Electronics Packaging Manufacturing*, 24, 214–221. doi:10.1109/6104.956807.
- Lee, T., Bilonis, I., & Tepole, A. B. (2020). Propagation of uncertainty in the mechanical and biological response of growing tissues using multi-fidelity Gaussian process regression. *Computer Methods in Applied Mechanics and*  
 950 *Engineering*, . doi:10.1016/j.cma.2019.112724.
- Liskov, B. (1988). Data Abstraction and Hierarchy. *ACM SIGPLAN Notices*, 23, 17–34. doi:10.1145/62139.62141.
- Liu, H., & Motoda, H. (1998). Feature transformation and subset selection. *IEEE expert*, 13, 26–28. doi:10.1109/mis.1998.671088.
- 955 Liu, Y., Liu, Y., Hansen, M., Pozdnukhov, A., & Zhang, D. (2019). Using machine learning to analyze air traffic management actions: Ground delay program case study. *Transportation Research Part E: Logistics and Transportation Review*, . doi:10.1016/j.tre.2019.09.012.
- Maaten, L. V. D. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15, 3221–3245. URL: [http://jmlr.org/papers/v15/vandermaaten14a.html%5Cnfiles/1017/JMLR-van\\_der\\_](http://jmlr.org/papers/v15/vandermaaten14a.html%5Cnfiles/1017/JMLR-van_der_)  
 960

- Maaten-2014-Accelerating\_t-SNE\_using\_Tree-Based\_Algorithms.  
pdf%5Cnfiles/1019/JMLR-van\_der\_Maaten-2014-Accelerating\_t-SNE\_  
using\_Tree-Based\_Algorithms-Supplement.pdf%5Cnhttp://lvdmaa.  
doi:10.1007/s10479-011-0841-3.
- Mathaisel, D. F. (1997). Decision support for airline schedule planning. *Journal of Combinatorial Optimization*, 1, 251–275. doi:10.1023/A:1009776309890.
- Moran, M., & Gordon, G. (2019). Curious Feature Selection. *Information Sciences*, . doi:10.1016/j.ins.2019.02.009.
- Nathans, D. (2015). Efficient operations: Building an operations center from the ground up. In *Designing and Building Security Operations Center* chapter 1. (pp. 1–24). Elsevier. URL: [http://www.sabreairlinesolutions.com/images/uploads/Efficient\\_Operations\\_Brochure.pdf](http://www.sabreairlinesolutions.com/images/uploads/Efficient_Operations_Brochure.pdf)<https://linkinghub.elsevier.com/retrieve/pii/B978012800899700001X>.  
doi:10.1016/B978-0-12-800899-7.00001-X.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, . (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, .
- Pérez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. In *IEEE International Symposium on Information Theory - Proceedings*. doi:10.1109/ISIT.2008.4595271.
- Ramakrishnan, R., & Ullman, J. D. (1995). A survey of deductive database systems. *The Journal of Logic Programming*, 23, 125–149. doi:10.1016/0743-1066(94)00039-9.
- Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE*, . doi:10.1371/journal.pone.0087357.

- Seger, C. (2018). An investigation of categorical variable encoding techniques  
 990 in machine learning: binary versus one-hot and feature hashing. *Degree  
 Project Technology*, (p. 41). URL: [http://www.diva-portal.org/smash/  
 get/diva2:1259073/FULLTEXT01.pdf](http://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf).
- SheffieldML (2014). GPY Documentation, . URL: [http://gpy.readthedocs.  
 io/en/deploy/](http://gpy.readthedocs.io/en/deploy/).
- 995 Shlens, J. (2014). A Tutorial on Principal Component Analysis, . URL: [http:  
 //arxiv.org/abs/1404.1100](http://arxiv.org/abs/1404.1100). doi:10.1.1.115.3503.
- Shurkhovetsky, G., Andrienko, N., Andrienko, G., & Fuchs, G. (2018). Data  
 abstraction for visualizing large time series. *Computer Graphics Forum*, 37,  
 125–144. doi:10.1111/cgf.13237.
- 1000 T. Vincenty (1975). Direct and Inverse solutions of geodesics on the ellipsoid  
 with application of nested equations. *Survey Review*, .
- Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal  
 of Machine Learning Research*, .
- Van Der Maaten, L. J. P., Postma, E. O., & Van Den Herik, H. J. (2009). Dimen-  
 1005 sionality Reduction: A Comparative Review. *Journal of Machine Learning  
 Research*, 10, 1–41. doi:10.1080/13506280444000102.
- Verleysen, M., & François, D. (2005). The Curse of Dimensionality in Data  
 Mining. *Analysis*, . doi:10.1007/11494669{\\_}93.
- Weisberg, S. (2001). Yeo-Johnson Power Transformations. *Department of Ap-  
 1010 plied Statistics, University of Minnesota*, .
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component anal-  
 ysis. *Chemometrics and Intelligent Laboratory Systems*, . doi:10.1016/  
 0169-7439(87)80084-9.

Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting  
1015 aggregate flight departure delays in airports based on supervised learning,  
*Sustainability (Switzerland)*, . doi:10.3390/su12072749.

### **CRedit Author Statement**

**Kolawole Ogunsina:** Conceptualization, Methodology, Software, Validation, Visualization, Investigation, Formal Analysis, Writing-Original draft preparation.

**Ilias Bilonis:** Writing-Reviewing and Editing.

**Daniel DeLaurentis:** Writing-Reviewing and Editing.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: