A Data Quality Scorecard for Assessing the Suitability of Asset Condition Data for Prognostics Modeling

Sarah Lukens[1], Damon Rousis[2], Dominic Thomas[3], Travis Baer[4], Michael Lujan[5], and Marshall Smith[6]

[1,2,3,4,5,6] *LMI, Tysons, VA, 22102, USA*
sarah.lukens@lmi.org
damon.rousis@lmi.org
dothomas@lmi.org
tbaer@lmi.org
mlujan@lmi.org
msmith@lmi.org

Abstract

High efficacy algorithm development for prognostics requires quality data from sensors and other contextual sources, such as maintenance, usage and inspection data. Data quality challenges, such as lack of sensor-based history (depth) across the entire fleet of components (breadth), can prohibit the ability to develop algorithms which are both cost-effective and useful. Therefore, the first step in prognostics modeling is determining the sufficiency of the data required to support the development of predictive algorithms. We present an assessment process for determining data suitability in the development of prognostic models based on available data that determines which modeling approaches are feasible, allowing for a first determination of decision-making for data adequacy. The assessment process follows a full data quality framework which also identifies where data eligibility and quality may be further enhanced using advanced technologies for data quality improvement approaches such as imputation, increasing the probability of obtaining the required data needed for the successful development of predictive algorithms. Use of this framework maximizes the quantity of quality data harvested from industrial data sources, increasing the probability of obtaining the required data needed for the successful development of predictive algorithms. Additionally, repeating this assessment as further data becomes available enables further expansion of the set of usable prognostic models as data availability grows.

1. Introduction

Time series data, collected from sensors on industrial assets, provides a means for measuring and monitoring asset health. The amount of sensor data is increasing with the reduction of costs around the acquisition and installation of sensor technology, resulting in data stores of immense volume. The combination of increased availability of sensor data, development of data science tools and emerging technologies have led to opportunities for developing quantitative approaches to assist decision making. While there have been significant advances in prognostics modeling approaches over the past two decades, much of the published algorithm development has been focused on benchmark data sets which lack many characteristics of raw sensor data collected from the field. Challenges and considerations around handling field data include considerations of the size (high volume), structure, complexity and quality of the data.

One component of the solution is the development of a data pre-processing pipeline which supports best practices for consuming asset condition data from different sources in preparation for prognostics model building. Understanding the suitability of the data for model development is another important component. The condition of the raw data must first be evaluated to identify if the data can support training prognostics models. Any quality issues in the raw data must be identified which can lead to reduced accuracy of model predictions. In the context of model development for prognostics and health management (PHM), reduced accuracy can mean notification of an event which will *not* occur in the near future (false positive), failure to notify an event before it occurs (false negative) or notification of an event but without enough advanced notice to take appropriate action.

In practice, data quality issues often are present in the data and can directly impact model performance. Data quality issues can arise due to a number of reasons, including:

- how data is captured, transmitted, and stored;
- missing or incomplete data;

- inconsistencies between data sources;
- limitations in the data model; or
- systemic features of the problem domain.

Analysis of sensor data alone is not enough to accomplish this task. Appropriate context must be given to the sensor readings data from two additional categories: maintenance and operating time data. Maintenance data is typically transactional records representing remove, replace, and repair events that provide additional information about the health of sensor measurements. Furthermore, when transaction maintenance records are recorded with units of calendar time, operating time, translates these units into an aging unit more appropriate for prognostics modeling. These three data categories, often from disparate data systems, must be aligned using join keys (such as asset IDs or serial numbers) to build the data set for training prognostics models.

In this paper, we present a rubric for calculating the quality of sensor data. A qualitative assessment of the implications on prognostic model performance is provided through a simplified case study, which is also used to demonstrate how the rubric may be applied to a real problem. The case study is based on open source simulated C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) data for turbofan engine degradation (Saxena, Goebel, Simon, & Eklund, 2008).

Our research goals. The purpose of data quality scorecard presented in this paper is to formalize the process of preparing data for building prognostics models. By systematizing data analysis, we hope to address the sometime ad-hoc approach to Exploratory Data Analysis (EDA) specifically in the context of PHM that often comprises the early stages of projects. The discipline of building machine learning models has widely accepted frameworks, metrics (e.g. accuracy, precision, recall, F-score), and approaches for improving those metrics. EDA, on the other hand, remains a somewhat artful endeavor. The goal is to bring a scientific rigor to EDA and offer metrics that quantify suitability of entering into the model building phase of PHM.

Interpreting the results. Performance in one or more data quality areas or in several scorecard metrics should not be interpreted as a go/no-go threshold, but rather the first step in experimentation and model building. The specific contributions of this work are formalization of a process of preparing data for building prognostics models which accounts for data quality challenges commonly found in the field, with some recommendations for specific data quality measures for scenarios which are concerned with predicting lifetime regarding major removal or replacement events of critical components.

The paper is organized as follows. Section 2 provides background information on data and data quality considerations for prognostics model development and reviews related literature on which the approach was built from. Section 3 summarizes a full pipeline for evaluating data for prognostics modeling suitability, with a series of data quality checks and decision points. Section 4 introduces the case study and data quality measures specific to the case study. Section 5 provides examples of the data quality scorecard process for the case study. Section 6 contains conclusions, remarks and future work.

2. BACKGROUND

The goal of a PHM initiative is to incorporate asset condition data with other relevant information sources, such as available labor resources, parts demand and maintenance histories to predict the need for maintenance actions on components prior to failure. With enough advanced notice, such predictions can lead to more accurate demand forecasts driving changes in supply posture, provisioning of labor resources, prevention of catastrophic failure or other actions that ultimately increase up-time and reduce cost of asset operations.

Many organizations have years of sensor data, such as process control data or diagnostics data, but have not evaluated its use towards developing lifetime prediction models (Nguyen et al., 2019; Corrêa et al., 2022; Kwon, Hodkiewicz, Fan, Shibutani, & Pecht, 2016). There are also organizations who may already have PHM programs in place, but are interested in updating their modeling approaches or for using the data and models in other areas of their organization beyond the maintenance department, such as fleet-level demand forecasting of spare parts and inventory. Evaluating the suitability of already available data for such purposes is a bottom-up approach.

From a formal, top-down perspective, the industry accepted best practice for evaluating the implementation of a PHM program on existing systems is to perform some sort of maintenance strategy evaluation process, such as Reliability-Centered Maintenance (RCM). In RCM, the risk-cost trade-offs are considered in the context of the intended purpose of the asset or system. Information from the data provides an opportunity to augment the classic RCM process, which can help for prioritization (Baker, Nixon, Banks, Reichard, & Castelle, 2020). From a business or operations perspective, it does not make sense to apply PHM as a maintenance strategy for every asset, component or failure mode. In fact, there are many assets or components where it makes much more sense to have fixed interval replacement or run to failure strategies (Goebel et al., 2017; Gulati & Smith, 2021; Atamuradov, Medjaher, Dersin, Lamoureux, & Zerhouni, 2017). Guidelines for requirements specifications for a prognostics initiative which integrates safety, reliability, cost and real-time viability are found in (Saxena et al., 2010; Goebel et al., 2017; Walker & Kapadia, 2009).

2.1. Prognostics Modeling for Equipment Life Prediction

Prognostics modeling, as defined by techniques for equipment life prediction such as Remaining Useful Life (RUL), is one component of a full PHM system. Equipment life predictions are used to inform predictive maintenance (PdM) or condition-based maintenance (CBM) programs aimed at predicting and preventing incipient failures on critical assets. Fleet-level decision making also consume equipment life predictions such as for informing demand forecasts for stockroom inventory, determining optimal fixed intervals in preventative maintenance tasks, and for various life cycle costing analyses such as analyses based on reliability, availability and maintainability of assets.

For PdM/CBM programs, if the available data is not suitable for equipment life predictions, other approaches such as fault detection or diagnostics can still be deployed effectively for early failure detection (Atamuradov et al., 2017; Elattar, Elminir, & Riad, 2016). In contrast, if the available data is not suitable for equipment life predictions for fleet-level decision making, other population-level statistical approaches such as Weibull analysis or Cox Proportional Hazards model can be deployed effectively for making fleet-level predictions (Coble & Hines, 2011). Prognostic models, individualized to an asset through incorporation of asset condition data as well as historical and expected operating conditions, may be useful for both the execution of a predictive maintenance strategy and for refined fidelity in fleet-level decision making.

The major components of the PHM modeling life cycle framework are data operations, model operations and communications. Data operations is concerned with data acquisition, such as how data is collected (such as manually taking measurements or use of sensors) as well as data storage, aggregation and pre-processing considerations. Model operations is concerned with model development, training, validation and storage capabilities of PHM models. Communications refers to processes around initiating action based on the model.

2.2. Data Sources for Prognostics Modeling

Asset condition measurements are usually in the form of time-series data, which can be collected automatically from sensors or manually such as from regularly occurring spot readings. Information measured by sensors includes condition monitoring data, time series data for process measurements and diagnostic readings. Diagnostic variables are typically boolean or categorical, such as an indicator at a certain time when an alarm triggers. This paper is concerned with sensor data, referring to time series data collected by sensors either for measuring asset condition or for operations or process which may possibly also be used for condition assessment. Special considerations are often required for handling sensor data, which is larger in volume than most other data sources due to higher sampling rates and possibly high number of sensor variables.

Transactional data often contains information providing *context* around events regarding an asset or component such as dates and details around failure mode occurrence and maintenance activities. For prognostics model development, the nature of the required contextual data will depend on the specific application desired. Examples of contextual data sources include maintenance work orders, logbooks, financial reporting and operator entered information.

Aging or usage data contains information on the aging of the equipment or component and is used in prognostics models to map between calendar time and equipment operating time. Such information is not typically uniformly captured across different asset categorizations and needs to be determined case by case. For example, for vehicular assets such as aircraft or ground vehicles, usage data may include vehicle trip start and end logs or mileage. In manufacturing operations, where asset usage is scheduled and often continuous, usage data may be extracted from data sources containing stoppage or scheduling information.

2.3. Quality of Sensor Data

Data quality has been defined as data which is fit for a purpose (Hodkiewicz, Kelly, Sikorska, & Gouws, 2006). Quality data for the development of prognostic models requires sensor data which is available and labeled with the needed contextual information. In practice, data for training prognostics models is often unavailable, unlabeled and not organized. Failed sensors, sensors out of calibration or faulty data collection and processing systems may lead to distorted or missing data. Environmental conditions, aging and degradation may also affect accuracy of sensor data. The redundancy from multiple sensors on a system may address some of these challenges, but introduce new challenges such as handling misaligned timestamps and highly correlated data. Consistently labeled data is required for training data-driven models, but often sensor readings are unlabeled. When records are labeled, class imbalance is a common challenge (Omri, Al Masry, Mairot, Giampiccolo, & Zerhouni, 2021; Dangut, Skaf, & Jennions, 2021; Zhang et al., 2019).

Data quality measurement and improvement frameworks have been proposed and implemented for transactional maintenance data (Hodkiewicz et al., 2006; Lukens, Naik, Saetia, & Hu, 2019). In order to measure (and improve) the quality of the data, it is important to have defined goals and purposes for the data. In this paper, we assume that the goals are for the development and training of prognostics models for predictive maintenance or for fleet management. Once goals are defined, data quality can be measured and data identified as “sufficiently good” can be used. For data which is identified as insufficient, data quality can be improved through improving historical data if possible while implementing best practices for improving future data. Analytical tools can often be employed to improve historical data, such as imputation for handling missing values. For improving the quality of future sensor data, the data quality measures help identify and guide what data is needed and how it needs to improve. Data quality measures may also help identify requirements for where to install sensors and data collection resources (thus informing an RCM process) or specifying calibration tolerances.

2.4. Related Literature

There are published frameworks providing guidance for assessing the suitability of sensor data for building PHM models. Chen, Zhu and Lee developed a methodology for both evaluating and improving the quality of training data for training fault detection and diagnostic models (Chen, 2012; Chen, Zhu, & Lee, 2013). They established procedures and quantitative metrics for assessing acquired training data in its suitability for classification modeling for fault detection and proposed an efficient approach for filtering out outliers and inefficient features from the data. Jia, Zhao, Di, Yang and Lee proposed an iterative framework for assessing the suitability of sensor data for PHM modeling tasks specifically for PdM/CBM applications, specifically fault detection, diagnostic and prognosis (Jia, Zhao, Di, Yang, & Lee, 2017). This framework includes recommended metrics for evaluating the suitability of the data; detectibility if the data supports fault detection, diagnosibility if the data supports diagnostics and trendability if the data supports prognostics models. Omri, Masry, Mairot, Giampiccolo and Zerhouni (Omri et al., 2021) added to this framework through proposing data quality metrics for the fault detection task considering data quality dimensions of data volume, data accuracy and data completeness.

Coble and Hines proposed suitability metrics for prognostics models based on three key qualities for prognostics parameters: monotonicity, prognosability and trendability (Coble & Hines, 2009). Calculation of these metrics helps not only identify the suitability of data, but also helps in fitting a prognostic parameters through methods such as regression. Data reduction, feature selection and the above frameworks for assessing the suitability of the data for PHM modeling assume the data is already "clean" (labeled, in equipment operating time and with few or no missing values).

In many field cases, there is often a gap between raw data and data in a form for evaluating the suitability for different PHM modeling tasks. Recently, published works with specific recommendations for how to handle and prepare raw sensor data for PHM analysis have emerged. Griffiths, Corrêa, Hodkiewicz and Polpo recommend best practices in the alignment of time series variables (Griffiths, Corrêa, Hodkiewicz, & Polpo, 2022), addressing challenges when sensors have different sampling rates. Griffiths et al. (2022) also make recommendations for integrating contextual data sources, such as linking contextual data sources to time series data for labeling the data while also mitigating the large storage needs of streaming data. Cofre-Martel, Lopez Droguett and Modarres developed a step-by-step guideline for processing sensor data for PHM modeling with emphasis on handling the high volume of data and incorporation of expert knowledge from field engineers (Cofre-Martel, Lopez Droguett, & Modarres, 2021). They stress the importance of reproducible and consistent processes for data pre-processing and show the impact of pre-processing decisions on final PHM models. Addressing the challenge of labeling sensor data, Corrêa, Polpo, Small, Srikanth, Hollins and Hodkiewicz proposed a data-driven approach for labeling process data using contextual data sources (Corrêa et al., 2022).

Data reduction and feature extraction and selection are important data preparation steps for prognostics model building which are well-covered in the literature. Data reduction typically includes identifying highly correlated variables in order to discard redundant variables as well as variables with low variability (Eg: close to constant value) which may not act strongly as explanatory variables (Cofre-Martel et al., 2021; Nguyen et al., 2019; Griffiths et al., 2022). Feature extraction and selection may involve preparing data ranging from calculating features such as lag times to employing analytics for identifying significant explanatory variables. Feature extraction techniques are covered in many places as a key area for data pre-processing in a PHM model development pipeline (Atamuradov et al., 2017; Elattar et al., 2016), and have different considerations when deep learning models are introduced (Fink et al., 2020).

3. FULL DATA OPERATIONS PIPELINE FOR PROGNOSTICS SUITABILITY

Data Operations, the first stage in the PHM modeling life cycle, consists of components related to data acquisition and processing data for model development. For field data, this also includes identification if the available data quality is sufficient for model development and any analytical approaches for measuring and improving data quality. Figure 1 shows the full data operations framework from input data sources to outputs suitable for model operations, which includes the development, training, testing, validation and storage of models. The dashed box illustrates the portion of data operations covered in detail by this paper. The following sections contain brief summaries and reviews of the components of the individual steps.

3.1. Data Survey and Data Model Assessment

The first step in preparing raw sensor data for analysis is the Data Survey. Key metadata documented during this step in-

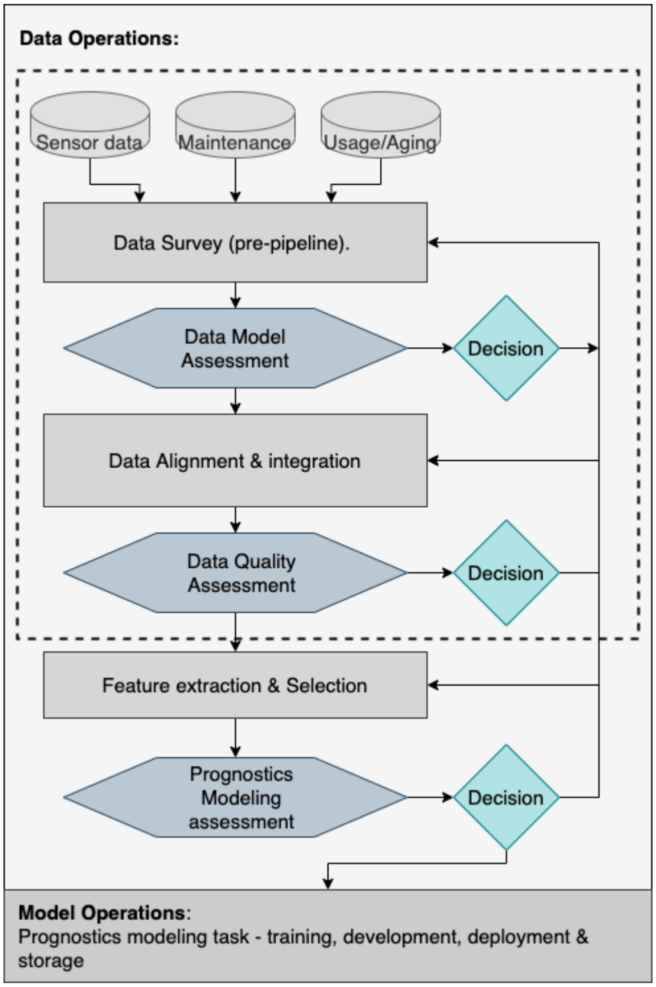

Data Operations:
Sensor data
Maintenance
Usage/Aging
Data Survey (pre-pipeline).
Data Model Assessment
Decision
Data Alignment & integration
Data Quality Assessment
Decision
Feature extraction & Selection
Prognostics Modeling assessment
Decision
Model Operations:
Prognostics modeling task - training, development, deployment & storage

Figure 1. Data operations pipeline for identifying, collecting and assessing the feasibility and suitability of sensor data for prognostics modeling. The portion covered by this paper in detail is in the dashed lines.

clude scope of the asset coverage, component coverage, calendar period covered, and numbers of variables and observations. The authors have found that simply documenting and comparing date ranges between disparate data sources can uncover issues often encountered downstream. In creating the data survey, we also document relevant field names and data types which assists in identifying true time series sensor variables, binary indicators, categorical indicators, and identifying which to keep for the analysis.

During the data survey we consider data size and its implications on storage and computation as well. We audit the size of each data source on disk before and after any compression or optimization to better understand burden on hardware and compute resources moving forward. Size of the data often places important requirements on algorithm chosen during modeling phase of PHM, and the goal is to establish those requirements as early as possible.

We conclude the Data Survey step with creation of Data Model Assessment, an important outcome aimed at answering the following question: *does the way in which information is stored support alignment?* In particular, we apply the following criteria for each of the three data categories:

- The data exists and is human readable.
- The data is structured and machine readable.
- Join key/s exist.
- Join key/s are unambiguous.

Furthermore, we impose additional criteria on the maintenance dataset.

- Maintenance action is captured and human readable.
- Maintenance action is codified.
- Failure cause is captured and human readable.
- Failure cause is codified.

The above criteria can be evaluated as a stoplight chart with green (criterion completely satisfied), yellow (criterion partially satisfied), or red (criterion not satisfied) color coding.

3.2. Data Alignment and Integration of Contextual Data Sources

Once the data scope and model have satisfied the criteria above and deemed viable, the next step in the data preparation process is aligning the three major data categories: sensor, maintenance, and operating time. First the different sensor variables are aligned with each other, which may include truncating or normalizing data and timestamps to adjust misaligned sensor readings (Griffiths et al., 2022). Once the sensor variables are aligned, integration with the contextual data is performed. A general approach is to map the desired information from the other data sources to time-series data, such as creating and populating an array or column containing a representation of the contextual information of interest at specified time points. Equipment aging may be represented in time series form where each point contains the component's age at the time stamp. The desired labeling information is often an event such as a detected fault, failure mode or time interval indicating the period between an installation event and the subsequent removal. The occurrence of significant events can be represented as a discrete event or as a calculated value such as an aggregation of observed events during a specified time window or measuring the time leading up to the occurrence (Simon & Schoenhof, 2021; Alam, Jalali, Ghosh, Farahat, & Gupta, 2021; Griffiths et al., 2022). The end result of data alignment and integration is a transformed time-series data which is aligned in time and labeled.

3.3. Data Quality Assessment

The data quality assessment measures if the aligned and integrated data is suitable for use in prognostics modeling. Data quality measures include specific checks around data completeness, data volume and accuracy. Completeness is the most straightforward of the data quality dimensions, measuring how much data is present and how much is missing. Other data quality dimensions address the quality of the data which is not missing. Once the sensor data is aligned and integrated with contextual data sources, data completeness can be evaluated with respect to partially missing data.

For instance, if after aligning and integrating the data, ideally each time point contains a label representing the occurrence of a significant event. Data quality can be measured to evaluate if, when and how often sensor data is measured but knowledge about an event is unknown, or if there is an event occurrence but no supporting sensor data. For the data which is not missing, other data quality measures such as detecting outliers or noisy data may be measured. Data volume measures can be assessed to see how much data falls in each class, assessing class imbalance possibilities and seeing if enough data is in the minority class (Omri et al., 2021).

As equipment usage information is needed for prognostics model building for conversions between calendar time and operating time, sensor data missing while an asset was in operation should be accounted for as such data may impact prognostic model performance. How and how much such information is missing may have various degrees of impact. This paper is focused primarily on recommending specific methodology, and more specific details on the methodology are in Section 4.2.

If the quality of the sensor data is deemed sufficient from this step (or specific areas where the data is sufficient is identified), the data is ready for the next step - data reduction and feature selection. For data that may have missing areas, the process can assist in identifying if imputation may be possible and what imputation approach is appropriate. For example, if sensor values are missing intermittently, a different imputation approach may be used than if sensor values are missing for an entire component over a significant period of time. Additionally, the decision points from this step can also identify gaps in asset condition coverage, helping to inform a decision process of where to install new sensors and where to use already collected data.

3.4. Data Reduction, Feature Extraction and Selection and Model Suitability Assessments

In practice, the next step in the prognostics modeling pipeline are data reduction, feature extraction and selection, which were reviewed in Section 2.4 and assumes by this point the data is sufficiently clean and labeled. The tests for PHM model suitability (such as trend-ability) are applicable at this stage to inform the decision process. For PdM/CBM programs, if the data is determined unsuitable for prognostic modeling, the data may still be suitable for diagnostics or fault detection algorithms. If the goal is fleet level insights and the data is determined unsuitable for prognostics modeling, the data may be suitable for reliability-based approaches such as Cox Proportional Hazards modeling or Weibull Analysis (Coble & Hines, 2011; Nguyen et al., 2019).

4. ILLUSTRATIVE CASE STUDY FOR DATA QUALITY SCORECARD

Specific details on how data is labeled and evaluated for prognostic modelings are specific to the organizational goals from modeling. For this reason, in this section we define a case study scenario which will drive specific details presented for the rest of the paper, but will be generic enough that the desired prognostics models may be useful in both CBM/PdM programs and for fleet-level decisions using equipment lifetime predictions for demand forecasting. In this scenario, the event (or response) of interest are component removals which lead to a major repair and/or replacement event. Specifically, a removal event which causes “supply chain engagement”. We define a maintenance interval as the time unit between a component install and removal (component lifetime) and treat the membership of a sensor reading to its maintenance interval as a “label”. The duration of the interval, or RUL, is the historical lifetime value.

The case study is intended to provide an example of the different steps of the pipeline, with emphasis on the data quality assessment after the sensor data is integrated with the maintenance and usage data sources. Specifically, the main focus of the case study will be showing how gaps in the data between the sensor readings and the equipment usage data are identified, characterized and presented in a scorecard for decision making towards determining modeling adequacy. The case study assumes some sort of vehicular asset which has irregular usage and operations in the form of vehicle starts and stops, where a ”sufficiently good” mapping between equipment operating time and calendar time is a strong requirement for prognostic model performance.

We develop synthetic data to illustrate the data quality scorecard assessment process based on the widely used open source C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) data set, which is simulated data for turbo-fan engine degradation (Saxena et al., 2008). The C-MAPSS data in its downloaded state is clean and suitable for prognostics modeling. To illustrate the case study, we apply a “backwards mapping” process to the C-MAPSS data to synthetically make the data more resemble a “raw” data source. For simplicity, the most basic C-MAPSS data subset (FD001)

Table 1. Representative sample of vehicle trip data used for creating calendar times based on vehicle usage

Trip ID	Trip start time	Trip end time	Duration (hrs)	Cum. usage (hrs)
1	2022-05-03 19:55	2022-5-03 21:44	1.9	1.9
2	2022-05-04 02:16	2022-5-04 05:24	3.2	5.1
3	2022-05-17 04:40	2022-5-17 06:35	2.0	7.1
4	2022-05-18 00:00	2022-5-18 03:00	3.0	10.1

is used which has 100 units, one fault mode and one operating condition.

While preparing the data for the case study, some assumptions regarding the physical system were made. While the time unit on the C-MAPSS data (one engine cycle) is supposed to represent a single vehicle trip, we assume that each time point is a reading in time, distributed over different cycles. As a result, the resulting calendar time window and number of vehicle trips over the time window do not reflect the intended C-MAPSS time scale of engine fault initiation and degradation. We made the simple assumption that a unit number in the C-MAPSS dataset corresponds to one engine which is treated as a component on a vehicular asset.

Asset operation time intervals were created in the form of vehicle trip start and stop data for vehicular assets expressed in calendar time. The authors developed this usage data based on generating trip times and durations in statistically similar ways as heuristically observed on real vehicular data. A sample of the generated operating times are shown in Table 1. For each observed sensor value, a calendar dates is generated at a spacing of five minute intervals. A plot of one of the variables in the C-MAPSS dataset is plotted in Figure 2 to give a visual comparison between equipment operating time and calendar time. Gaps in the calendar time plot correspond to periods where the component is not operating (for a vehicular asset, when the engine is off). In the field, longitudinal plots of raw sensor data for vehicular assets resemble the plot in Figure 2(b) and need to be converted to operating time for model building.

The maintenance interval, signifying the time between component install and removal, are simulated using the first time series calendar date as the component install date and the last time series calendar date as the last date as a suspension. For simplicity, the age of the asset and the age of the component are both zero at the start of the interval (component installation). In reality, both the asset and component may have ages at installation which need to be incorporated into the equipment lifetime calculations. It is important to note that in field data, there is nearly always a significant amount of work preparing maintenance intervals from maintenance data.

Since the original data is of "perfect" data quality, we would expect that the backwards mapped "raw" version will return perfect scores at every step of the way. In the case study examples, the scores from the perfect data are reported as the baseline (Scenario 1) and two additional scenarios are created for comparison. Scenario 2 is the "promising data" scenario, where data quality issues exist, but there is promise for prognostic modeling possibilities. Scenario 3 is the "insufficient data" scenario, where the data quality is so poor, it is insufficient for prognostics modeling. The data quality measures supplied in the case study are synthetic, but have values which are simulated based on actual observations from the authors' experiences.

4.1. Alignment and Integration Methodology

For the case study, preparing data for use in prognostics algorithms requires three sources: sensor readings, maintenance intervals, and operating time. Figure 3 shows how the three data sources are plotted together over a single maintenance interval, which is shown as the thick gray bar. The blue tick marks on the top row show the regions where condition indicator data observed, and the gray tick marks on the second row show the regions where "flight"/vehicle trip cycle data is observed.

The outcomes of the alignment step are the appropriate labels (membership to maintenance interval and interval duration) and an age value at each measurement time. Formally, the first step is to align the different sensor readings to common time points. For the case study, we assume this is completed (observe the C-MAPSS data comes aligned) and focus on alignment and integration between sensor data and contextual sources. The sensor readings are aligned against the maintenance intervals in calendar time through labeling each sensor reading with the corresponding maintenance interval. This may involve adding columns to the sensor data with the interval information, such as unique ID, install and removal date and duration. The next step involves the mapping between calendar time and operating time, using the operations data set. At this step, the age of the component can be inferred at every measurement time point. An example of the fused data is shown in Table 2.

4.2. Data Quality Scorecard Methodology

The proposed methodology for evaluating the data quality of aligned and integrated sensor data for prognostics modeling is detailed in this section. For the purposes of this case study we assume that maintenance data is always recorded (install and removal events are in the data), remarking that maintenance data quality is a topic by itself. Table 3 summarizes different observed challenges observed when comparing coverage of the different data sources. Note that this table does not contain an exhaustive list, but rather suggests different data quality issues which could be observed and could be used as a starting point.

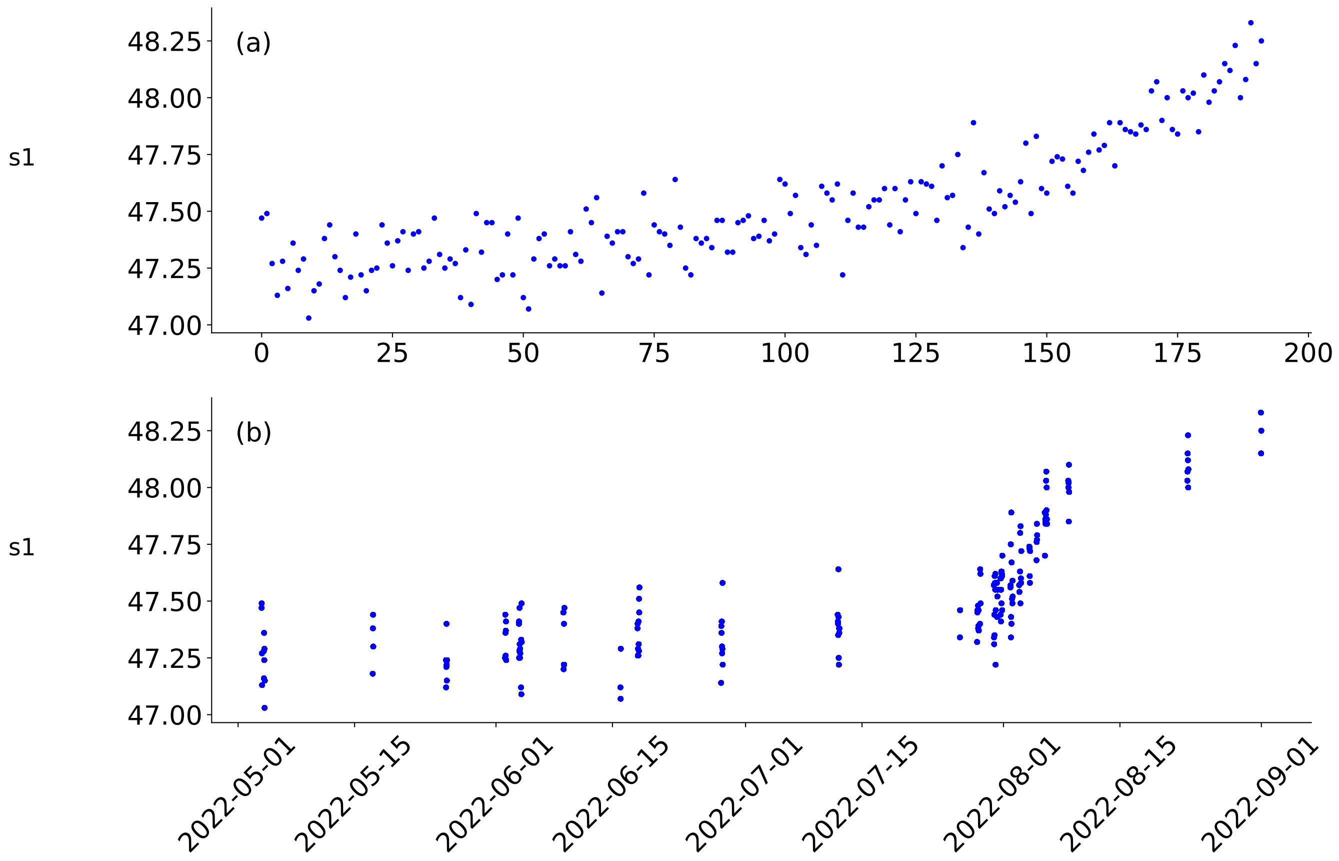


Figure 2. One variable (s1) from the C-MAPSS data (unit 1) plotted in equipment operating units and calendar units. (a) Sensor readings plotting in equipment operating units. Each point here is a "cycle" in the original data. (b) After "backwards" transformation, showing the values read in calendar time. The large gaps correspond to periods when the component is not running.

The main possible data issues reported for data quality identification are organized into categories which are used to categorize each maintenance interval. These data quality categories are: no operating data, incomplete operating data (observed gaps), low component aging, no sensor data and measurement ambiguity between healthy and unhealthy readings. Measurement ambiguity may occur when there are time series readings at the time of a significant event, such as install or removal. In this case, it is ambiguous if the readings correspond to a newly installed or about to be removed part and may lead to training models on mislabeled data. Knowing where this occurs and how often helps guide decisions made in model development and training (decisions such as if to remove data or consult a subject matter expert). When there is low component aging observed, such as a very small usage time, the recommended action may be to inspect the data and determine special logic such as remove the data or try models with or without the data.

In cases when there is no sensor data, it is unlikely that sensor-based prognostics can be built, but there is need to further investigate the issue. Depending on the scope and nature of where the sensor data is missing, it may or may not be possible to use data imputation and use the coverage measures to help inform what type of imputation. For example, sensor data may be missing for an asset, but data from a similar asset with similar operating conditions can be used to contextually help impute missing values. Additionally, for fleet management insights, such information can be used to justify using reliability-based approaches in the modeling initiative.

The other major issues are around missing or incomplete operating data. In both cases, challenges may occur when mapping between calendar and equipment operating time. It may be possible to infer component ages at different reading times, but it is important to be mindful of the uncertainty that this practice may introduce into the prognostics model, both in training and for making predictions. There is a lot of ambiguity in when and how these cases occur, particularly when the gaps are incomplete. For this reason, the data quality scorecard approach further assists in drilling down to characterizing these gaps.

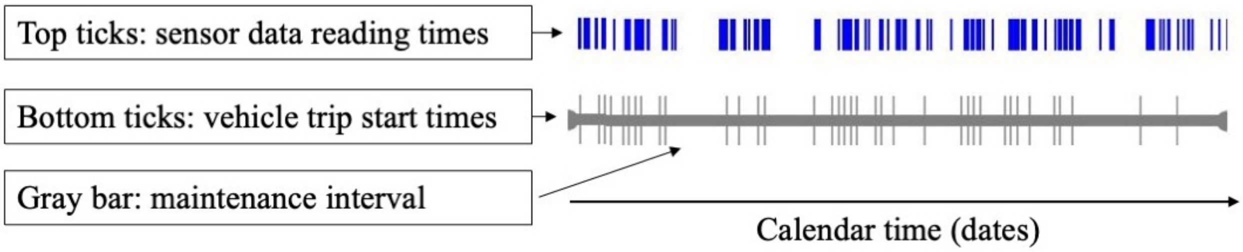

Top ticks: sensor data reading times
Bottom ticks: vehicle trip start times
Gray bar: maintenance interval
Calendar time (dates)

Figure 3. Legend for reading data alignment plots. Three data sources are plotted together over calendar time (x-axis): sensor data, where occurrences are shown as ticks in the top row; usage data, where vehicle trips are shown as ticks in the bottom over; and maintenance intervals, which are shown as thick gray bars over the vehicle trips.

Table 2. Example of aligned data with integration of contextual sources. For the case study, the key information is the maintenance interval where each reading has membership (represented by Interval ID, install date and duration columns) and usage/operating window (represented by Trip ID). The Age column is calculated from the integration and denotes the operating age of the component at each time reading and is necessary for prognostics model development.

Time	Sensor 1	Interval ID	Install date	Duration	Trip ID	Age
2022-5-03 01:00	641.82	1	2022-5-01	31	1	0
2022-5-03 01:05	642.15	1	2022-5-01	31	1	5
2022-5-03 01:10	642.35	1	2022-5-01	31	1	10
2022-5-04 03:45	642.35	1	2022-5-01	31	2	15
2022-5-04 03:50	642.37	1	2022-5-01	31	2	20

Each interval is classified based on consistency and completeness of data within the range of the interval. As modelers, we are particularly interested in the size of the gaps and the location of the gaps of sensor readings within each maintenance interval. Gaps in sensor data in the beginning or end of the interval represent missing positive identification of either healthy or unhealthy measurements respectively. Gaps in the middle of the maintenance interval could inhibit the ability to model degradation from healthy to unhealthy behavior if it is observable for specific failure modes. Small gaps may possible be overcome by other modeling techniques such as imputation whereas large gaps may force throwing away maintenance intervals in the dataset.

Table 4 lists the categories each maintenance interval is placed along with an image of an example. Observe the data shown in the table has been mapped from calendar time to equipment operating time at the best of the ability of the available data, so that any gaps shown represent missing sensor readings at periods when the asset was in operation. The table shows the categories in increasing order of severity, where Type A represents the "cleanest" and D is the "dirtiest." The first characterization, A.1, is the type of data desired for prognostics modeling, while the other extreme, D.2, is where the asset is in operation but there are no measurements. The "B" category corresponds to small gaps and the "C" category is large gaps, and the cut-off between small and large can be user specified. The different qualitative characterizations are missing data at the beginning or end of the interval (*.2) and missing data in the middle of the interval (*.1).

Application of this measurement process involves characterizing every maintenance interval in scope of the data as falling in to one of these categories. The result is a high level scorecard summarizing the data quality.

5. Results of Data Quality Scorecard Applied to Case Study

The below sections walk through the various components of the data quality scorecard process and illustrate how the different measurements could look in the three case study scenarios.

5.1. Data Survey and Data Model Assessment

The "stoplight chart" summarizing the results of the Data Model Assessment (Section 3.1) is shown in Table 5, applied to the three scenarios described in section 4. The first scenario, which depicts the situation where perfect data is available and underlying data model does not introduce any ambiguity, is colored as completely green.

Scenario 2, the "Promising data" scenario, has several areas where the Data Model Assessment falls short of a perfect score but does not entirely halt progress on development of

Table 3. Summary of different data quality challenges which arise from integrating sensor data with maintenance and usage data. We focus further on characterizing the different gaps in operating data for better recommendations on the suitability of a dataset.

Issue	Sensor data	Maint. data	Operating data	Detail	Possible root causes	Modeling implications
No operating data	Yes	Yes	No	Sensor data exists and maintenance records indicate parts are aging. Operating data does not exist.	Missing operating data. Sensors are measured, but certain criteria are not met to record operating data data	Cannot map to calendar data to operating time - either see if aging can be approximated or interval unusable for training.
Gaps in operating data	Yes	Yes	**Gaps**	Component aging calculated from the maintenance data will be higher than the sum of operating hours.	Operating records may be missing, operating recording issues or sensors measured but certain criteria not met	Challenge in mapping calendar data to operating time - introduces high uncertainty to any prognostic modeling
Low component aging	Often missing	Yes	Often missing	Equipment aging between install and removal is very small (eg: less than 10 hours).	Incorrectly installed, infant mortality or controlled exchange	Inclusion (or exclusion) to prognostics model may introduce bias, removal of data may reduce size of training data
No sensor data	No	Yes	Yes	Sensor data does not exist, but operating time indicates component was in operation and aging	Records do not exist, or equipment is not monitored	Cannot build sensor-based prognostics model - need to investigate scope of this issue
Measurement ambiguity	Yes	Yes	Yes	Sensor readings occur at the same time as a significant event, leading to ambiguity on when they are observed	Differences in recording practices between data types	Prognostic algorithm trained on mislabeled data

prognostic models. Two areas where the data partially fulfills the criteria (yellow) are in join key ambiguity. Join key ambiguity could be yellow because times of events are captured with a DATE datatype not TIMESTAMP in both maintenance and operating time data. Events could have occurred anywhere in a 24 hour window, which introduces ambiguity when attempting to align with sensor data, which *does* have a TIMESTAMP datatype.

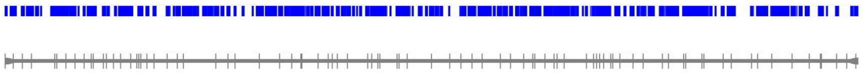
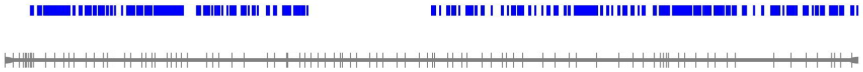
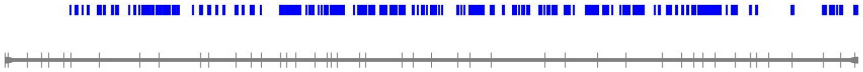
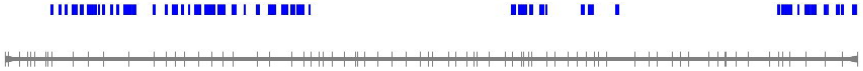
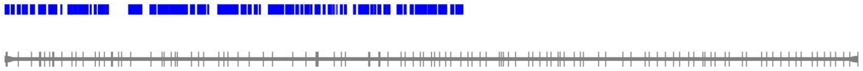
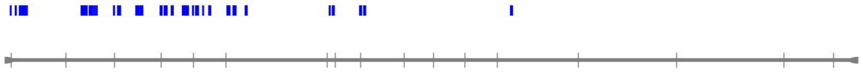
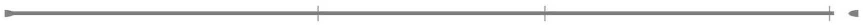
Another aspect where criteria are not met in Scenario 2 (red) is capturing failure cause and maintenance action as codified fields. Instead, these fields could be free-text fields written by the maintenance personnel performing the action. Additional effort will be required to interpret or translate the natural language into something that can be used for modeling.

Scenario 3 represents the most severe case where sensor data is not human readable and thus entirely precludes proceeding with the data alignment step without remediation. In this example, the original sensor data was encrypted by the original equipment manufacturer (OEM) and required significant effort beyond the scope of this paper to decrypt. An additional complication is that failure code information is not available nor could corrective maintenance actions be differentiated from scheduled maintenance.

5.2. Data Quality Scorecard for Sensor Data Coverage

Data Completeness: High Level Assessment. Table 6 shows the total percentage (normalized by maintenance interval) of occurrence of the categories from Table 3 for the three scenarios. Observe that these numbers need not add up to 100 because issues may overlap. For Scenario 3 (insufficient data), we see that 12% of the maintenance intervals

Table 4. Data quality categories based on completeness of sensor data readings over equipment operating time across a maintenance interval (from component installation to removal).

Type	Description	Example
A.1	Sensor data consistent from interval start to end	
B.1	Small gap in sensor data in the middle of interval	
B.2	Small gap in sensor data at beginning or end of interval	
C.1	Large gap in sensor data in middle of interval	
C.2	Large gap in sensor data at beginning or end of interval	
D.1	Sensor data very sparse throughout interval	
D.2	Sensor data does not exist in interval	

have low aging. This may be important to note because there may be insufficient lifetime data volume for model training or data quality issues which need addressing. In this case, the user could drill down to see where and how this happens and make decisions such as whether to make certain exclusions from the training data. Also note that in the insufficient case, 60% of the sensor readings are missing or incomplete over their respective maintenance intervals and 32% are missing or incomplete 2 for the promising case.

Data Quality Scorecard for Missing Operating Data. The amount of sensor data coverage within the maintenance removals for each scenario are reported in Table 7. The number in each cell represent the fraction of total maintenance intervals for that component that fall into that coverage category: A.1 (most pristine, sensor data throughout interval) to D.2 (poorest data quality, e.g. no sensor data within the interval). For the three scenarios, only one sensor variable is shown because we did not assume different sensor variables were missing or had different sampling rates, so all sensors had equal coverage. In reality, understanding the respective coverage across different variables could be hugely informative in selecting which variables potentially use for prognostic model development.

A visualization of the numbers summarized in Table 7 comparing the three scenarios is shown in Figure 4. Such visualizations are powerful for quickly reporting and communicating the results of the data quality analysis.

Using these type of data quality measures, the analyst can gauge the suitability of their data for modeling. In the perfect data scenario, the analyst can begin dimension reduction, feature selection and assessing which prognostic modeling approaches may be suitable given the data. In the insufficient data scenario, the information in the data quality scorecard can be presented to communicate the feasibility of prognostics modeling given the data before attempting to build a model. The promising data scenario is definitely grayer, but the scorecard can give qualitative and quantitative insights towards how to proceed depending on the use case and the nature of the data. In all cases, the results of the data quality scorecard process are not necessarily good or bad, but rather

Table 5. Data Model Assessment, shown as a "stoplight chart", for three scenarios with increasing levels of data *dirtiness*.

Criterion	Scenario 1: Perfect (%)	Scenario 2: Promising (%)	Scenario 3: Insufficient (%)
Sensor data is human readable	🟩	🟩	🟥
Sensor data is machine readable	🟩	🟩	N/A
Sensor data join key/s exist	🟩	🟩	N/A
Sensor data join key/s unambiguous	🟩	🟩	N/A
Maintenance data is human readable	🟩	🟩	🟩
Maintenance data machine readable	🟩	🟩	🟩
Maintenance data join key/s exist	🟩	🟩	🟩
Maintenance data join key/s unambiguous	🟩	🟨	🟨
Operating time data is human readable	🟩	🟩	🟩
Operating time data machine readable	🟩	🟩	🟩
Operating time join key/s exist	🟩	🟩	🟩
Operating time join key/s unambiguous	🟩	🟨	🟨
Failure cause is human readable	🟩	🟩	🟥
Failure cause is codified	🟩	🟥	🟥
Maintenance action human readable	🟩	🟩	🟥
Maintenance action codified	🟩	🟥	🟥

a formal process for evaluating and better understanding the possibilities of using available data for prognostics modeling.

5.3. Implications

Imbalanced data: When gaps in sensor measurements exist at the start (labeled as healthy) or end (labeled as unhealthy) of maintenance intervals, this introduces imbalance in training samples. Many machine learning models assume balanced data, which causes them to have poor predictive performance for under-sampled classes, often the most interesting. Furthermore, over/under-sampling healthy or unhealthy sensor readings may lead to inaccurate estimate of model performance. For example, a model may be very accurate at predicting healthy components (low false positive rate) at the same time poor at identifying unhealthy components (high false negative rate).

Table 6. Examples of high level sensor data quality measurements comparing the 3 hypothetical data quality scenarios.

Issue	Scenario 1: Perfect (%)	Scenario 2: Promising (%)	Scenario 3: Insufficient (%)
No operating data	0%	1%	4%
Gaps in operating data	0%	32%	60%
Low aging (<10 hours)	0%	3%	12%
No sensor data	0%	2%	19%
Measurement ambiguity	0%	0%	1%

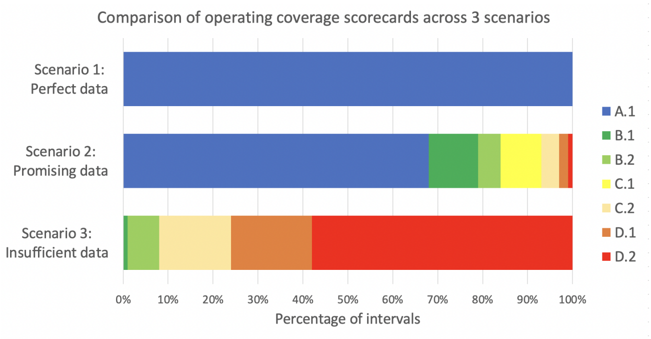


Figure 4. Comparison of operating coverage scorecards across three scenarios

Degradation modeling: For those failure modes that have smooth and predictable transition between healthy and unhealthy (e.g. corrosion and wear), gaps in sensor data lead to step-changes in behavior rather than consistent observations over the degradation period. In this case, the problem may be reduced to anomaly detection rather than prediction of remaining useful life.

Reduced training dataset: At some point, sensor data becomes too sparse, and we are forced to eliminate entire maintenance intervals from our training dataset. This has important modeling implications where less data on which to train can lead to overfitting. Models that overfit on few samples do not generalize well to new observations.

Model-specific implications: In the example of linear regression, we can train a model that gives point estimates of the model parameters (e.g. y-intercept, slope) based on observations, and those point estimates have standard errors that depend on number of observations: fewer samples means higher error.

Performance implications: Lastly, our ability to estimate model accuracy is decreased when we have fewer samples with which to calculate performance metrics. For example, a

Table 7. Example of data quality scorecard for characterizing operating data across the 3 scenarios. Every maintenance interval is classified into a category, with A.1 the most desirable. The number of maintenance intervals in each class are reported as a percentage of the total maintenance intervals.

Type	**Description**	Scenario 1: Perfect **(%)**	Scenario 2: Promising **(%)**	Scenario 3: Insufficient **(%)**
A.1	Sensor data consistent from interval start to end	100%	68%	0%
B.1	Small gap in sensor data in the middle of interval	0%	11%	1%
B.2	Small gap in sensor data at beginning or end of interval	0%	5%	7%
C.1	Large gap in sensor data in middle of interval	0%	9%	0%
C.2	Large gap in sensor data at beginning or end of interval	0%	4%	16%
D.1	Sensor data very sparse throughout interval	0%	2%	18%
D.2	Sensor data does not exist in interval	0%	1%	58%

model that correctly predicts 900 out of 1,000 samples has the same accuracy as a model that predicts 9 out of 10. However, we are much more certain of the estimate in the first case. In the context of a PdM/CBM program, this could be the difference between a viable and inviable maintenance policy.

6. CONCLUSION

A methodology was proposed which can be used for an initial data quality assessment of available sensor data for prognostics model development. A case study concerned with predicting the next major repair or replacement event on a critical component was used to specifically show execution of the methodology and the decision points made before feature selection in a prognostics model building pipeline. The data quality framework provides scope of the suitability of the data for the development of prognostic models and can be used to target data issues for data quality improvement, inform specific data quality improvement strategies, assist the modeler in forming hypotheses, designing experiment and tuning models, and quantify the return on investment for data quality improvement initiatives.

In addition to evaluating existing data for its feasibility for prognostics modeling, there are additional applications. The data quality framework can also be used to inform a data quality improvement initiative, such as identifying challenges in condition measurement sampling. The data quality measurement approaches could be used to inform an RCM process for developing a PHM maintenance strategy by summarizing existing information which could be re-purposed. Comparisons of the different measures from this approach could be to compare sensor data from similar systems with different instrumentation strategies for analysis, such as comparing data from assets with retrofitted sensors against data from assets with health ready components.

While it is known that the quality of the data will impact prognostic model performance, a rigorous process for measuring how much, under what data quality assumptions and managing uncertainty is future work. The synthetic data based on the C-MAPSS dataset is a starting point in this direction, as prognostics modeling approaches and their performance are well documented. Two major categories in how the C-MAPSS data fails to depict real-world data is (1) it is missing many real system characterization features such as maintenance logs, data quality and other diverse sources of factors that add uncertainty to the data, and (2) it is not very large in size. Incorporating considerations around big data into a framework are important too, as in reality, condition indicator dataset is massive. The newer C-MAPSS dataset (N-CMAPSS) is a natural place to expand on creating synthetic data containing data quality challenges (Arias Chao, Kulkarni, Goebel, & Fink, 2021).

Many works focused on PHM analytics are focused on the PdM/CBM use case and not on fleet management applications. This paper has focused on prognostics models and the role of equipment lifetime predictions. There is opportunity for organizations which have existing PdM/CBM programs in place to also use the data and analytics in their organization for fleet management purposes such as demand forecasting. In practice, data, models and information can be in silos across an organization. Restricted access to condition indicator data may not just be across different departments, but also occurs between owner/operators and the OEMs who have proprietary encryption. The challenge of siloed data islands can be larger as decision making uses for the data can have far reaches within different departments in an organization and across the supply chain.

ACKNOWLEDGMENT

The authors wish to acknowledge and thank Nathan Danneman and Thomas Willemain for their constructive feedback on the early drafts.

NOMENCLATURE

PHM	Prognostics and Health Management
PdM	Predictive Maintenance
CBM	Condition Based Maintenance
RCM	Reliability Centered Maintenance
EDA	Exploratory Data Analysis
OEM	Original Equipment Manufacturer

REFERENCES

Alam, M., Jalali, L., Ghosh, D., Farahat, A., & Gupta, C. (2021). Remaining useful life estimation using event data. In *Annual Conference of the PHM Society* (Vol. 13).

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5.

Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., & Zerhouni, N. (2017). Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation. *International Journal of Prognostics and Health Management*, *8*(3), 1–31.

Baker, W., Nixon, S., Banks, J., Reichard, K., & Castelle, K. (2020). Degrader analysis for diagnostic and predictive capabilities: a demonstration of progress in DoD CBM+ initiatives. *Procedia Computer Science*, *168*, 257–264.

Chen, Y. (2012). *Data quality assessment methodology for improved prognostics modeling* (Unpublished doctoral dissertation). University of Cincinnati.

Chen, Y., Zhu, F., & Lee, J. (2013). Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method. *Computers in industry*, *64*(3), 214–225.

Coble, J., & Hines, J. W. (2009). Identifying optimal prognostic parameters from data: a genetic algorithms approach. In *Annual Conference of the PHM Society* (Vol. 1).

Coble, J., & Hines, J. W. (2011). Applying the general path model to estimation of remaining useful life. *International Journal of Prognostics and Health Management*, *2*(1), 71–82.

Cofre-Martel, S., Lopez Droguett, E., & Modarres, M. (2021). Big machinery data preprocessing methodology for data-driven models in prognostics and health management. *Sensors*, *21*(20), 6841.

Corrêa, D., Polpo, A., Small, M., Srikanth, S., Hollins, K., & Hodkiewicz, M. (2022). Data-driven approach for labelling process plant event data. *International Journal of Prognostics and Health Management*, *13*(1).

Dangut, M. D., Skaf, Z., & Jennions, I. K. (2021). An integrated machine learning model for aircraft components rare failure prognostics with log-based dataset. *ISA transactions*, *113*, 127–139.

Elattar, H. M., Elminir, H. K., & Riad, A. (2016). Prognostics: a literature review. *Complex & Intelligent Systems*, *2*(2), 125–154.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Goebel, K., Daigle, M. J., Saxena, A., Roychoudhury, I., Sankararaman, S., & Celaya, J. R. (2017). *Prognostics: The science of making predictions*.

Griffiths, T., Corrêa, D., Hodkiewicz, M., & Polpo, A. (2022). Managing streamed sensor data for mobile equipment prognostics. *Data-Centric Engineering*, *3*.

Gulati, R., & Smith, R. (2021). *Maintenance and reliability best practices* (3rd ed.). Industrial Press Inc.

Hodkiewicz, M., Kelly, P., Sikorska, J., & Gouws, L. (2006). A framework to assess data quality for reliability variables. In *Engineering Asset Management* (pp. 137–147). Springer.

Jia, X., Zhao, M., Di, Y., Yang, Q., & Lee, J. (2017). Assessment of data suitability for machine prognosis using maximum mean discrepancy. *IEEE transactions on industrial electronics*, *65*(7), 5872–5881.

Kwon, D., Hodkiewicz, M. R., Fan, J., Shibutani, T., & Pecht, M. G. (2016). IoT-based prognostics and systems health management for industrial applications. *IEEE Access*, *4*, 3659–3670.

Lukens, S., Naik, M., Saetia, K., & Hu, X. (2019). Best practices framework for improving maintenance data quality to enable asset performance analytics. In *Annual conference of the phm society* (Vol. 11).

Nguyen, D., Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., & Bäck, T. (2019). A review: Prognostics and health management in automotive and aerospace. *International Journal of Prognostics and Health Management*, *10*(2), 35.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., & Zerhouni, N. (2021). Towards an adapted PHM approach: Data quality requirements methodology for fault detection applications. *Computers in Industry*, *127*, 103414.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management* (pp. 1–9).

Saxena, A., Roychoudhury, I., Celaya, J., Saha, S., Saha, B., & Goebel, K. (2010). Requirements specification for prognostics performance-an overview. *AIAA Infotech@ Aerospace 2010*, 3398.

Simon, H., & Schoenhof, S. (2021). Enhancing the diagnostic performance of condition based maintenance through the fusion of sensor with maintenance data. In *Annual*

Conference of the PHM Society.

Walker, M., & Kapadia, R. (2009). Integrated design of online health and prognostics management. In *Annual Conference of the PHM Society* (Vol. 1).

Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., & Wei, M. (2019). A review on deep learning applications in prognostics and health management. *IEEE Access*, *7*, 162415–162438.

BIOGRAPHIES

Sarah Lukens is a Data Scientist at LMI. Her interests are focused on data-driven modeling for reliability applications by combining modern data science techniques with current industry performance data. This work involves analyzing asset maintenance data and creating statistical models that support asset performance management (APM) work processes using components from natural language processing, machine learning, and reliability engineering. Sarah completed her Ph.D. in mathematics in 2010 from Tulane University with focus on scientific computing and numerical analysis. Sarah is a Certified Maintenance and Reliability Professional (CMRP).

Damon Rousis is a Data Scientist at LMI with extensive experience in applying analytical techniques to predictive maintenance problems. In particular, Damon is focused on design and productization of machine learning algorithms using sensor data for real-time decision making. Damon received his Ph.D. in Aerospace Engineering from Georgia Tech in 2011 focusing on complex system design and computer simulation.

Dominic Thomas is a Data Scientist at LMI with experience developing machine learning models and statistical analysis for the Department of Defense. Dominic's main focus has been on developing predictive maintenance approaches from historical data better inform maintenance decision making. He graduated from American University with a Master of Science in Analytics.

Travis Baer is a Data Scientist at LMI with over 10 years experience in statistical methods in reliability analysis. He graduated from Georgia Institute of Technology in 2010 with a Master of Science in Operations Research. His work includes software design to automate reliability data transformations, modeling, validation and forecasting, as well as general purpose data management to support other forecasting simulations.

Michael Lujan is a Senior Data Scientist at LMI. He has a demonstrated history of working on solutions to highly complex problems for the Department of Defense. Michael is skilled in statistical modeling, simulations, and optimization algorithms. Michael obtained his Doctor of Philosophy (Ph.D.) in Nuclear Physics from The George Washington University, where he is also an adjunct Professor of Physics.

Marshall Smith has 20+ years of technical experience in the development and successful delivery of predictive analytic solutions focused on capital intensive and mission critical assets. He is an expert in modeling and simulation solutions supporting reliability, availability, and maintainability analysis, asset life cycle management analysis, and asset performance management and is adept in reliability/survival analysis, Reliability-Centered Maintenance (RCM), Condition-Based Maintenance Plus (CBM+) analytics, statistical process control, decision support analysis, and generalized linear modeling.