Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets

Emmanuel Ramasso[1] and Abhinav Saxena[2]

[1] *FEMTO-ST Institute, Dep. AS2M/DMA, UMR CNRS 6174 - UFC / ENSMM / UTBM, 25000 Besançon, France*
emmanuel.ramasso@femto-st.fr

[2] *SGT Inc., NASA Ames Research Center, Intelligent Systems Division, Moffett Field, CA, 94035-1000, USA*
abhinav.saxena@nasa.gov

ABSTRACT

Six years and more than seventy publications later this paper looks back and analyzes the development of prognostic algorithms using C-MAPSS datasets generated and disseminated by the prognostic center of excellence at NASA Ames Research Center. Among those datasets are five run-to-failure C-MAPSS datasets that have been popular due to various characteristics applicable to prognostics. The C-MAPSS datasets pose several challenges that are inherent to general prognostics applications. In particular, management of high variability due to sensor noise, effects of operating conditions, and presence of multiple simultaneous fault modes are some factors that have great impact on the generalization capabilities of prognostics algorithms. More than seventy publications have used the C-MAPSS datasets for developing data-driven prognostic algorithms. However, in the absence of performance benchmarking results and due to common misunderstandings in interpreting the relationships between these datasets, it has been difficult for the users to suitably compare their results. In addition to identifying differentiating characteristics in these datasets, this paper also provides performance results for the PHM'08 data challenge wining entries to serve as performance baseline. This paper summarizes various prognostic modeling efforts that used C-MAPSS datasets and provides guidelines and references to further usage of these datasets in a manner that allows clear and consistent comparison between different approaches.

1. INTRODUCTION

Run-to-failure datasets from a turbofan engine simulation model were first published by NASA's Prognostics Center of Excellence (PCoE) in 2008. This dataset was originally used in a data challenge competition in PHM'08 conference, where PHM researchers were invited to develop prognostic methods as part of the competition. The competition was well received with a good set of winning methods. These data were further made available following the competition through NASA PCoE data repository (Saxena & Goebel, 2008) so that researchers can use them to build, test, benchmark, and compare data-driven prognostic methods. Since then, these datasets have been widely used by researchers around the world and results published in over 70 publications. This paper reviews these approaches, published in the last five years and analyzes them to understand why some approaches worked better than others, how did researchers use these datasets to compare their methods, and what were the difficulties faced so necessary improvements can be made to these datasets to make them more useful.

1.1. Background

Prognostics has gained significant attention for its promise to improve systems health management through advance warnings about performance degradation and impending failures. Predicting with confidence, however, has posed its own challenges due to various uncertainties involved in the process. Several government and industry supported programs have helped push the thrust in prognostics technology development all round the globe. The science of prognostics has fairly matured and the general understanding of health prediction problem and its applications has greatly improved in the past decade. Both data-driven and physics based methods have been shown to possess unique advantages that are specific to application contexts. However, until very recently, a common bottleneck in development of data-driven methods was the lack of availability of run-to-failure data sets. In most cases real-world data contain fault signatures for a growing fault at various severity levels but no or little data capture fault evolution all the way through failure. Procuring actual system fault progression data is typically time consuming and expensive. Fielded systems are, most of the time, not properly instrumented for collection of relevant data or are unable to distribute such data due to proprietary constraints. The lack of common data sets, which researchers can use to compare their approaches, has been an impediment to progress in the field of prognostics. To tackle this problem, several datasets have been published on a prognostics data repository (Saxena & Goebel, 2008), which have been used by many researchers. Among these datasets are five datasets from a turbofan engine simulation model - C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) (Frederick, DeCastro, & Litt, 2007). By simulating a variety of operational conditions and injecting faults of varying degree degradation, datasets were generated for prognostics development (Saxena, Goebel, Simon, & Eklund, 2008a). One of the first datasets was used for a prognostics data challenge at the PHM'08 conference. A subsequent set was then released later with varying degrees of complexity. These datasets have since been used very widely in publications for benchmarking prognostics algorithms.

1.2. Motivation

The turbofan degradation datasets have received over seven thousand unique downloads in the last five years and yielded about seventy publications with various algorithms. However, results obtained by those algorithms can generally not be compared due to confusions and inconsistencies in how these datasets have been interpreted and used. Therefore, this paper intends to analyze various approaches that researchers have taken to implement prognostics using turbofan datasets. Some unique characteristics of these datasets are also identified that led to use certain methods more often than others. Specifically, various differences among these datasets are pointed out. A commentary is provided on how these approaches fared compared to the winners of the data challenge. Furthermore, this paper also attempts to clear several issues so that researchers, in the future, can take these factors into account in comparing their approaches with the benchmarks.

The paper is organised as follows. In Section 2, the C-MAPSS datasets are presented. Section 3 is dedicated to the literature review. Section 4 presents a taxonomy of prognostics approaches for C-MAPSS datasets. Finally, Section 5 provides some guidelines to give a hand to future users in developing new prognostic algorithms applied to these datasets and in facilitating algorithms benchmarking.

2. TURBOFAN SIMULATION DATASETS

C-MAPSS is a tool, coded in the MATLAB-Simulink ® environment for simulating engine model of the 90,000 lb thrust class (Frederick et al., 2007). Using a number of editable input parameters, it is possible to specify operational profile, closed-loop controllers, environmental conditions (various altitudes and temperatures), etc. Additionally, there are provisions to modify some efficiency parameters to simulate various degradations in different sections of the engine system.

2.1. Datasets characteristics

Using this simulation environment, five datasets were generated. By creating a custom code wrapper, as described in (Saxena, Goebel, et al., 2008a), selected fault injection parameters were varied to simulate continuous degradation trends. Data from various parts of the system were collected to record effects of degradations on sensor measurements and provide time series exhibiting degradation behaviors in multiple units. These datasets possess unique characteristics that make them very useful and suitable for developing prognostic algorithms:

1. Data represent a multi-dimensional response from a complex non-linear system from a high fidelity simulation that very closely models a real system.
2. These simulations incorporated high levels of noise introduced at various stages to accommodate the nature of variability generally encountered.
3. The effects of faults are masked due to operational conditions, which is yet another common trait of most operational systems.
4. Data from plenty of units is provided to allow algorithms to extract trends and build associations for learning system behavior useful for predicting RULs.

These datasets were geared towards data-driven approaches where very little or no system information was made available to PHM developers.

As described in detail in Section 3, the analysis of the publications using these datasets shows that many researchers have tried to make comparisons between results obtained from these similar yet different datasets. This section briefly describes and distinguishes the five datasets and explains why it may or may not be appropriate to make such comparisons.

Table 1 summarizes the five datasets. The fundamental difference between these datasets is attributed to the number of simultaneous fault modes and the operational conditions simulated in these experiments. Datasets $\#1$ through $\#4$ incorporate an increasing level of complexity and may be used to incrementally learn the effects of faults and operational conditions. Furthermore, what sets these four datasets apart from the challenge datasets is the availability of ground truth to measure performance. Datasets $1-4$ consist of a *training set* that users can use to train their algorithms and a *test set* to test the algorithms. The ground truth RUL values for the test set are also given to assess prediction errors and compute any metrics for comparison purposes. Results between these datasets may not always be comparable as these data simulate different levels of complexity, unless a universal generalized model is available that regards datasets $1-3$ as special cases of dataset $\#4$.

Table 1. Description of the five turbofan degradation datasets available from NASA repository.

Datasets		$\#$Fault Modes	$\#$Conditions	$\#$Train Units	$\#$Test Units
Turbofan data from NASA repository	$\#1$	1	1	100	100
	$\#2$	1	6	260	259
	$\#3$	2	1	100	100
	$\#4$	2	6	249	248
PHM2008 Data Challenge	$\#5T$	1	6	218	218
	$\#5V$	1	6	218	435

The PHM challenge datasets are designed in a slightly different way and divided into three parts. Dataset $\#5T$ contains a *train set* and *test set* just like for datasets $1-4$ except with one difference. The ground truth RULs for the test set are not revealed. The challenge participants were asked to upload their results (only once per day) to receive a score based on an asymmetrical scoring function (Saxena, Goebel, et al., 2008a). Users can still get their results evaluated using the same scoring function by uploading their results on the repository page, but otherwise it is not possible to compute any other metric on the results in absence of ground truth to allow error computation. The third part of the challenge set is dataset $\#5V$, the final *validation set* that was used to rank the challenge participants, where they were allowed only once chance to submit their results. The challenge since then is still continuing and a participant may submit final results (only once) for evaluation per instructions posted with the dataset on the NASA repository (Saxena & Goebel, 2008).

2.2. Establishing Performance Baseline

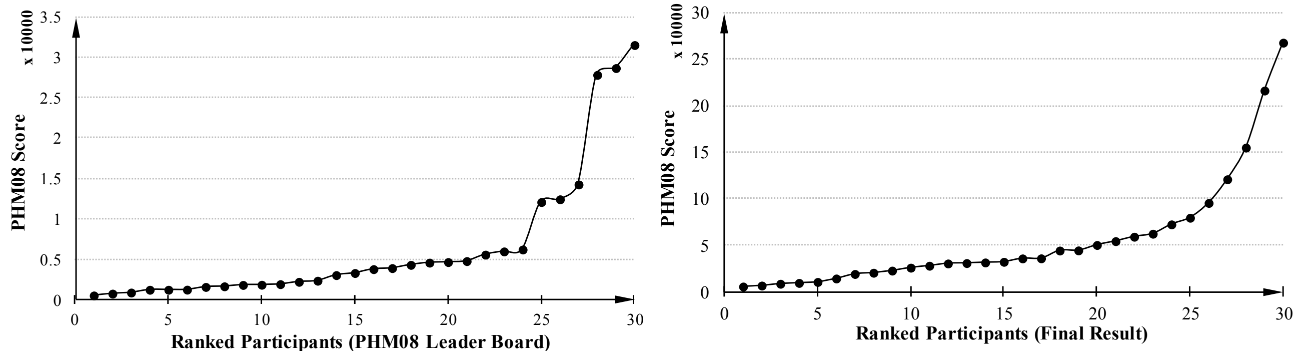
For performance comparison purposes, data challenge winning entries can be regarded as the benchmark performance with scores published on the challenge website in 2008. Since that webpage was taken down in subsequent years, these scores are not available except as reported in the published publications from the winners. Therefore, those scores are included here for reference in future. It is, therefore, realized that a direct comparison with the winners has not been possible and researchers have inconsistently compared performance between these different datasets. To alleviate that problem, this paper provides a number of error-based metrics listed in (Saxena, Celaya, et al., 2008) computed for top ten entries to formally establish a performance benchmark for datasets $\#5T$ and $\#5V$. Since there is no common record of results from datasets $1-4$, no such benchmark can be easily established and can only be partially collected from the published publications. It is however, expected that the performance obtained on dataset $\#2$ may be comparable to that from $\#5T$ due to similarity in fault mode and operational conditions.

Figures 1(a) and 1(b) graphically present the performance of top thirty scores from the challenge datasets. The scores were calculated using the asymmetric scoring function described in (Saxena, Goebel, Simon, & Eklund, 2008b). Similarly a number of error based performance metrics were computed for the participant entries and are presented in Tables 2 and 3 for the test and validation sets respectively. While interpreting these results following must be kept in mind:

- Scores obtained on dataset $\#5V$ are expectedly poorer in general than those obtained on dataset $\#5T$. This is due to two reasons - (1) PHM score metric is an aggregate over all units and is not normalized by number of units (note that $\#5V$ has almost twice the number of units in $\#5V$), and (2) the RUL statistics were intentionally changed for dataset $\#5V$ to check against overfitting or the methods that impose thresholds to avoid overpredictions that are penalized more severely by the scoring function.
- Owing to the above two reasons the top ranking algorithms on $\#5T$ did not necessarily perform equally well on $\#5V$. Therefore, it must not be interpreted that the top ranked algorithm in Figure 1(a) is also the top ranking one in Figure 1(b). At the same time both tables should be regarded as separate benchmarks for the two datasets.
- For the sake of a common ground, the ranks in Tables 2 and 3 are assigned based on PHM'08 scoring function. As observed, that poorer ranking algorithms may perform better when evaluated on other metrics. This illustrates a key point that metrics must be chosen based on what is important to achieve the goals in a specific application and that there may not be a universal metric for performance.
- These tables do not include the prognostic metrics such as those proposed in (Saxena, Celaya, Saha, Saha, & Goebel, 2010). The primary reason being that for these datasets predictions are made for multiple units and only once, which is a different scenario than for continuous prediction for prognostic metrics to be applicable.

3. C-MAPSS DATASET LITERATURE REVIEW

To analyze various approaches that have been used to solve C-MAPSS dataset problem, all the publications that either cite these datasets or the citation recommended by the repository were collected through standard web search. The search results returned over seventy publications which were then



(a) Dataset $\#5T$

(b) Dataset $\#5V$

Figure 1. Top thirty scores from PHM'08 data Challenge (datasets $\#5T$ and $\#5V$) computed using competition scoring function.

Table 2. Comprehensive set of metrics as computed from PHM'08 Challenge leader board based on test set (dataset $\#5T$). Ranking in column 1 is established based on PHM Scores.

Rank on $\#5T$	Competition Score	MSE	FPR (%)	FNR (%)	MAPE (%)	MAE	Corr. Score	Std. Dev.	MAD	MdAD
1	512.12	152.71	56.35	43.65	15.81	8.67	0.96	0.64	8.68	5.69
2	740.31	224.79	57.73	38.12	18.92	10.77	0.94	0.76	10.72	7.00
3	873.36	265.62	53.59	28.18	19.19	11.47	0.93	0.81	11.75	8.00
4	1218.43	269.68	51.93	44.20	20.15	11.87	0.93	0.85	12.05	8.00
5	1218.76	331.30	50.55	49.45	33.14	13.81	0.91	0.95	14.03	10.87
6	1232.27	334.52	42.27	57.73	32.90	14.14	0.91	0.96	14.28	10.37
7	1568.98	394.46	50.55	47.24	36.75	15.37	0.89	1.03	15.48	12.00
8	1645.77	330.02	47.24	50.28	30.00	13.47	0.91	0.95	13.59	10.00
9	1816.60	359.97	50.28	49.17	26.47	13.82	0.90	0.99	14.07	9.75
10	1839.06	377.01	45.86	53.59	27.72	14.31	0.89	1.02	14.43	9.10

preprocessed to identify overlapping efforts by same authors or the publications that only cite the dataset but perceivably did not use them for algorithm development. This resulted in forty unique publications that were then considered for review and analysis for this work. For the sake of readability, each of these publications were assigned a unique ID to use in various tables summarizing the results presented in this section. This mapping between publication and IDs is presented in Table 16. Furthermore, to keep the paper length short, a detailed review analysis of each of the forty publications is not included but only the summarized findings. However, the three winning methods from PHM'08 challenge are reviewed and summarized here to provide an appreciation of the approaches that were used by the top scorers. These publications also reported a lot of observations and other relevant information about the data with several insights that may be useful in future developments.

3.1. The Three Winning Methods

Similarity Based Approach (T. Wang, Yu, Siegel, & Lee, 2008) (Publications ID 1 and 7) $\rightarrow$ *Approach:* The authors' winning approach implemented a similarity-based prognostics algorithm. The method included several data analysis techniques to preprocess the data such as PCA and kernel smoothing (a detailed version can be found in the PhD dissertation (T. Wang, 2010)). These data were then divided into six bins corresponding to the six operating conditions by using a K-means clustering applied on channels $3, 4$, and 5 (these columns represent operating conditions)[1]. Sensors $7, 8, 9, 12, 16, 17$, and 20 were selected manually as relevant features because they exhibit continuous and consistent trends suitable for regression and generalization. The first 5% and the last 95% of each training instance (trajectory) were considered as healthy and failure data respectively. For each operating regime, the data were used to estimate the parameters of an exponential regression model describing the evolution of the health indicator (HI) from the healthy state to the failure state. These regression models were then used to estimate HI given a test trajectory for all operating regimes individually, and then fused together to get one HI using specific fusion rules. A threshold on maximum RUL estimates was applied to decrease the risk of being penalized by late predictions. Being part of the competition this approach was applied on datasets $\#4, \#5T, \#5V$, and most of the results were published in resulting publications.

[1] The partitioning into operating conditions can be directly obtained by using the results obtained in (Richter, 2012)

Table 3. Comprehensive set of metrics as computed from PHM'08 Challenge leader board based on test set (dataset $\#5V$). Ranking in column 1 is established based on PHM Scores.

Rank on $\#5V$	Competition Score	MSE	FPR (%)	FNR (%)	MAPE (%)	MAE	Corr. Score	Std. Dev.	MAD	MdAD
1	5636.06	546.60	64.83	31.72	19.99	16.23	0.93	1.01	16.33	11.00
2	6691.86	560.12	63.68	36.32	17.92	15.38	0.94	1.03	16.29	8.08
3	8637.57	672.17	61.38	23.45	20.72	17.69	0.92	1.09	17.79	11.00
4	9530.35	741.20	58.39	39.54	34.93	20.19	0.90	1.22	20.17	15.00
5	10571.58	764.82	58.85	41.15	32.60	20.05	0.91	1.22	20.41	14.23
6	14275.60	716.65	59.77	37.01	21.61	18.16	0.90	1.17	18.57	11.00
7	19148.88	822.06	56.09	41.84	30.25	20.23	0.88	1.29	20.89	13.00
8	20471.33	1000.06	51.95	48.05	33.63	22.44	0.88	1.42	24.05	14.78
9	22755.85	1078.19	62.53	35.40	39.90	24.51	0.86	1.45	24.08	20.00
10	25921.26	854.57	34.25	64.83	51.38	22.66	0.86	1.36	21.49	16.00

Recurrent Neural Network Approach (Heimes, 2008) (Publication ID 2) → *Approach:* The authors implemented an advanced proprietary algorithm initially developed for modeling complex dynamics of aircraft components (such as turbines) at *BAE Systems*. This algorithm was based on the work of (Feldkamp & Puskorius, 1998) able to solve various problems in adaptation, filtering and classification. The first part of the publication is dedicated to the data exploration phase. A classifier based on MLP was first trained to distinguish between healthy and faulty states yielding an error of 1%. The author then focused on regression methods such as MLP and RNN to cope with truncated instances as it is the case for the data challenge. RNN was preferred to MLP because RNN was able to cope with time-dependent data. The data exploration phase ends with some observations that may be useful for other algorithms: The degradation is made of four phases (steady, knee, acceleration of degradation and failure), and a threshold on maximum RUL estimates is set to 130. The second part of the publication is then focused on the algorithm implemented. The parameters of the RNN were estimated using all sensor data and operating conditions. Gradients were computed by a truncated back-propagation through time algorithm jointly with an extended Kalman filter to refine the weights. An evolutionary approach based on differential evolution was also used to maintain an ensemble of solutions with various and efficient parameterization. The optimization was performed using cross-validation by splitting the training data into distinct training and validation sets.

Multi Layer Perceptron and Kalman Filter Based Approach (Peel, 2008) (Publication ID 3) → *Approach:* The author presented an architecture based on MLP and RBF (implemented on Netlab) including a Kalman filter. The first part of the publication is dedicated to the data exploration phase and in particular effective visualization techniques such as neuroscale mapping. This technique allowed the author to point out the six operating conditions from sensor data and to conclude that these conditions cannot be used alone for predictions. Similarly to (Heimes, 2008), the various degradation levels were also pointed as one important difficulty (due to operating conditions (Saxena, Goebel, et al., 2008b)). The second part of the publication is then dedicated to the algorithm. The data processing included data standardization defined specifically for each operating condition to obtain features in similar scales. The author investigated a tournament heuristic approach to select various sensor subsets by minimizing the error on RUL prediction. The RUL is estimated by an ensemble of RBFs and MLPs with multiple parameterization. The step phase is the Bayes-optimal combination of RUL estimates using a Kalman filter. The Kalman filter also allowed the author to reduce sensor noise and to treat instances as time-series by integrating past information.

The analysis of the collected publications reveals several important observations that are summarized here. First, these publications are binned into various different categories and then analyzed. These categories and corresponding findings are presented in the sequel.

3.2. C-MAPSS Dataset Used

Table 4 identifies specific publications that use one or more of these five datasets. It can be observed that the dataset $\#1$ was the most used one (55%), followed by the test set ($\#5T$) from the PHM'08 challenge (35%), whereas rest of the other datasets are relatively under utilized. Three publications report generating their own datasets using the C-MAPSS simulator.

The heavy usage of the dataset $\#1$ ($\approx 70\%$) compared to all other datasets among the four from the NASA Repository may be attributed to its simplicity compared to the rest (see Figure 2), whereas high usage of dataset $\#5T$ is attributed to the PHM'08 challenge, where several teams had already used these data extensively, thereby gaining significant familiarity with the dataset as well as a preference due to availability of corresponding benchmark performance from the challenge leader board.

Several publications mentioned in Table 4 have used only the training datasets that have complete (run-to-failure) trajectories. Using data with complete trajectories gives access to

Table 4. List of publications for each dataset.

Datasets		Publication ID	Rate
Turbofan data from NASA repository	#1	5, 6, 10, 13, 14, 15, 19, 20, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 36, 37, 38, 40	22/40
	#2	13, 22, 34, 40	4/40
	#3	34, 40	2/40
	#4	7, 34, 40	3/40
PHM'08 Data challenge	#5T	1, 2, 3, 4, 8, 12, 16, 17, 21, 29, 30, 34, 35, 40	14/40
	#5V	1, 2, 3, 40	4/40
Simulator	OWN	9, 11, 39	3/40
Other	-	18	1/40

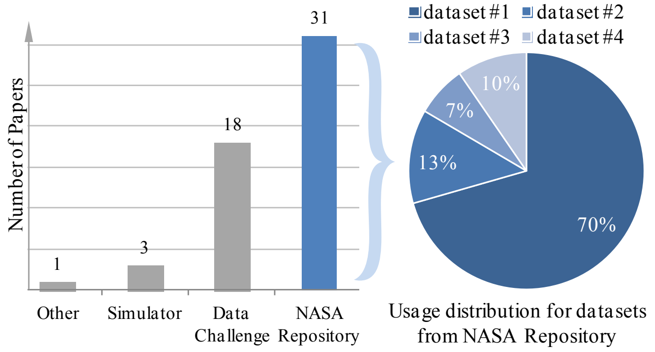


Figure 2. Relative usage of various C-MAPSS datasets in literature.

the true End-of-Life (EOL) to compute RUL from any time point in a degradation trajectory which could be used to generate a larger set of training data. This approach is also relevant to estimating RULs at different time points and allows the usage of prognostics metrics (Saxena, Celaya, et al., 2008) such as prognostics Horizon, $\alpha - \lambda$ metric, or the convergence measure. However, in true learning sense the algorithm, once trained, should be tested on unseen data for proper validation, as was required for the PHM'08 challenge datasets. Table 5 shows that 11 different publications used the "full" training/testing datasets, meaning the training dataset for estimating the parameters and the full testing dataset for performance evaluation.

Table 5. List of publications using only *full* training/testing datasets.

Datasets		Publication ID	Rate
Turbofan dataset from NASA repository	#1	20, 27, 28, 40	5/40
	#2	40	1/40
	#3	40	1/40
	#4	40	1/40
PHM'08 Data challenge	#5T	1, 2, 3, 4, 16, 21, 40	7/40
	#5V	1, 2, 3, 40	4/40

3.3. Target Problem to Solve

As normally expected, there is a wide variety of approaches taken in interpreting the datasets, formulating the problem and modeling the system to solve it. However, contrary to expectations, a significant number of publications have utilized these datasets for "fault detection" purpose by considering "multi-class classification" or "clustering" rather than prognostics. Table 6 identifies and distinguishes between publications that focus on detection versus prognostics.

By posing a multi-class classification problem various publications attempt to solve mainly three types of problems:

- Supervised classification: The training dataset is labeled (known classes for each feature vector);
- Unsupervised classification: The classes are not known apriori and data are not labeled;
- Partially supervised classification: Some classes are precisely known, others are unknown or are attached with a confidence value to express belief in that class.

Publications 1, 7, 10, 20, 24, 27, 32 use classification for preprocessing steps towards solving a prognostics problem. Specifically, unsupervised classification algorithms are used in publications 1, 7 to segment the dataset into the six operating conditions. For reference, detailed information about various simulated operating conditions in C-MAPSS is described in (Richter, 2012), which can also be used to label these datasets. Supervised and unsupervised classification algorithms are also used in publications 6, 10, 20, 27, 32 to assign a degradation level according to sensor measurements. The sequence of discrete failure degradation stages is indeed relevant for the estimation of the current health state and its prediction (Kim, 2010).

Health assessment, anomaly detection (seen as a 1-class classification problem) or fault identification are tackled in publications 6, 11, 12, 13, 26, 31, 35 using supervised classification methods, and partially supervised classification techniques in publications 12, 27, 33. For these approaches, a known target (or a degradation level) is required to evaluate the classification rate. For instance, four degradation levels were defined for labeling data in publications 6, 10, 27, 33: normal degradation (class 1), knee corresponding to a noticeable degradation (class 2 viewed as a transition between class 1 and 3), accelerated degradation (class 3) and failure (class 4). Some hand-made segmentations have been provided by some authors (publication 13). Using these segmented data (clusters) as proxy to ground truth, some level of classification performance can be evaluated for comparison purposes. These classification methods span all three classes of learning as mentioned above and corresponding publications are summarized in Table 7.

Similar to several classification approaches used, many approaches were employed for solving the prognostics problem for predicting RULs. In order to give due attention to the analysis of prognostic methods, a discussion is presented separately in Section 4.

Table 6. List of publications focused on detection and prediction.

Purpose	Publication ID	Rate
Detection	6, 10, 11, 12, 13, 31, 33, 34, 35, 37	10/40
Prediction	1, 2, 3, 4, 5, 7, 8, 9, 10, 13, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 32, 34, 36, 37, 38, 39, 40	31/40
Other	15, 18, 26	3/40

Table 7. Methods and references for state/fault classification.

Classification method	Publication ID
Supervised	6,10,11,12,13,26,31,35
Unsupervised	1,7,11,20,24,32
Partially supervised	27,33

3.4. Method for Treatment of Uncertainty

Given the inherent nature of datasets that include several noise factors and lack of specific information on the effects of operational conditions it is important for algorithms to model and account for uncertainty in the system. Different publications have dealt with uncertainty at various stages of processing as described below:

1. **Signal processing step** such as noise filtering using a Kalman filter as in publications 2, 3, 20, Gaussian kernel smoothing in publications 1, 7, and functional principal component analysis in publication 15.
2. **Feature extraction/selection step** such as using principal component analysis and other variants of it as suggested in publications 1, 7, 13, grey-correlation in publication 22, relevance of features for prediction in publication 23, and a greedy search with results visualization in publication 40.
3. **Health estimation step** such as based on operating conditions assessment to normalize/factor out the effects of operating conditions as proposed in publications 1, 7, 21, 40 and using non-linear regression.
4. **Classification step** where uncertainty modeling plays a role on data labeling using noisy and imprecise degradation levels as shown in publications 12, 27, 33, or on the inference of a sequence of degradation levels such as using Markov Models or multi-models as in publications 6, 10, 24, 32, 34.
5. **Prediction step** such as gradually incorporating prior knowledge during estimation in presence of noise as proposed in publications 4, 14, 16, 17, 19, 21, 30, in determining failure thresholds as in publications 10, 27, 32 or in representing health indicator such as in publication 40 to be used in prediction.
6. **Information fusion step** by merging multiple RUL estimates through Bayesian updating as pointed in publications 4, 21 or in similarity-based matching as in publications 1, 27, 40.

A variety of different uncertainty representation theories are found to be used. Table 8 classifies different publications according to the theory of uncertainty treatment used in corresponding analysis (Klir & Wierman, 1999). As shown in the table, the probability theory is the most popular one (65%) followed by set-membership approaches (in particular fuzzy-sets with 15%), Dempster-Shafer's theory of belief functions (13%), and other measures (such as polygon area and Choquet integral).

Table 8. Methods for uncertainty management used on C-MAPSS datasets.

Theories	Publication ID	Rate
Probability theory	1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 26, 28, 29, 30, 31, 32, 33, 34, 35	26/40
Set-membership	10, 14, 23, 25, 36, 39	6/40
Belief functions	6, 10, 24, 27, 33	5/40
Other measures	10, 40	2/40

3.5. Methods used for Performance Evaluation

Table 9 summarizes the performance measures that have been used for prognostics-oriented publications. A taxonomy of performance measures for RUL estimation was proposed in (Saxena, Celaya, et al., 2008; Saxena et al., 2010), where different categories were presented: accuracy-based, precision-based, robustness-based, trajectory-based, computational performance and cost/benefit measures, as well as some measures dedicated specifically to prognostics (PHM metrics). Since this problem involves predictions on multiple units, it is expected that the majority of publications would use error-based accuracy and precision metrics. Metric like the Mean Squared Error (MSE) has been used in two different ways-for the estimation of the goodness of fit between a predicted and a real signal, and as an accuracy-based metric to aggregate errors in RUL estimation. Only the publications that fall under latter category are included in the table. The table clearly shows that accuracy-based measures were most widely used, in particular the scoring function from PHM'08 challenge, which also weighs accuracy by timeliness of predictions. Broader usage of this metric is also explained by the fact that this is the only metric for which scores from data challenge were available and can be used as benchmark to compare with any new development. However, one may also compute additional measures if using only the training datasets where full trajectories are available. In that case, approaches like leave-one-out validation become applicable where all training instances but one are used for training each time and the remaining ones can be used for performance evaluation. Then the average of the performance measures is computed from all the runs. Publication 27 presents this approach for dataset $\#1$ and a cross-validation procedure for dataset $\#5T$ is used in publication 21. Note that publications 19, 20, 32 provide the only RULs estimates for all testing instances (without computing any metrics) and publications 10, 27 present distribution of errors.

Table 9. Performance measures used in prognostics-oriented publications applied on C-MAPSS.

Categories	Measures	Publication ID	Rate
Accuracy	PHM'08 Score	1, 2, 4, 5, 8, 16, 21, 29, 30, 40	10/40
	FPR, FNR	8, 10, 27, 40	4/40
	MSE	3, 8, 15, 17, 29, 40	6/40
	MAPE	4, 23, 28, 32, 34, 39, 40	7/40
	MAE	5, 13, 38, 40	4/40
Precision	ME	25,28,32,39	4/40
	MAD	25	1/40
Prognostics	PH	7, 22	2/40
	$\alpha - \lambda$	7, 22	2/40
	RA	7, 22, 34	3/40
	CV	7, 22, 34	3/40
	AB	34	1/40

4. PROGNOSTIC APPROACHES

As shown previously, C-MAPSS datasets have been used for the development and benchmarking of various detection and prognostics approaches. This section focuses specifically on the prognostic approaches which can be divided into three broad categories as described in the sequel.

4.1. Category 1: Using functional mappings between set of inputs and RUL

Methods in this category (see Table 10) first transform the training data (trajectories) into a multidimensional feature space and use the RULs to label the feature vectors. Then, using supervised learning methods, a mapping between feature vectors and RULs is developed. Methods within this category are mostly based on Neural Networks with various architectures. Different sensor channels were used to generate corresponding features. However, it was observed that the approaches yielding good performance also included a feature selection step through advanced parameter optimization such as using genetic algorithm and Kalman filtering as described in publications 2, 3 that ranked 2d and 3rd respectively in the competition.

The adaptation to new situations can be difficult with such approaches as it would be necessary to adapt model parameters if the process evolves as opposed to keep using the static mapping. Some further developments have considered these issues, for instance, Echo State Networks and Extreme Learning Machine are known to reduce the learning time and complexity (Jaeger & Haas, 2004; Huang, Zhu, & Siew, 2004; Siqueira, Boccato, Attux, & Lyra, 2012; Butcher, Verstraeten, Schrauwen, Day, & Haycock, 2013), while evolving models (Angelov, Filev, & Kasabov, 2010) are well-suited for adaptive real-time learning. However, these later approaches were evaluated only on a few instances of C-MAPSS training datasets in publications 24, 25, 32. Therefore, it was difficult to evaluate whether reducing learning time, making models gray-boxes, or even using evolving models are efficient strategies to perform well on C-MAPSS datasets. More experiments using the full training/testing datasets may be useful to evaluate these approaches better.

Table 10. Category 1 methods using a mapping learned between a subset of sensor measurements as inputs and RUL as output.

Methods	Publication ID
RNN, EKF	2
MLP, RBF, KF, Ensemble	3
MLP	8
ANN	9
ESN	20
Fuzzy rules, genetic algorithm	36
MLP, adaboost	38

4.2. Category 2: Functional mapping between health index (HI) and RUL

Methods listed in Table 11 are based on the estimation of two mapping functions: One maps sensor measurements to a health index (1-D variable) for each training unit based on sensor measurements; The second mapping links health index values to the RUL. These approaches construct a library of degradation models. Inference of the RUL for a given test instance includes using the library as prior knowledge to update the parameters of the model corresponding to the new test instance. Updating can be done using Bayes rule as proposed in publication 4 or other model averaging or ensemble techniques designed to take into account the uncertainty inherent to the model selection process (Raftery, Gneiting, Balabdaoui, & Polakowski, 2003).

Table 12 lists some other approaches that use approximation functions to represent the evolution of individual sensor measurement through time. Given a test instance as many predictions are made as the number of sensors. These predictions are then used in a classifier that assigns a class label related to identified degradation level. Some of these approaches also update classifier parameters with new measurements using some Bayesian updating rules as mentioned previously. These methods were however applied only on dataset $\#1$ in which sensors depict clear monotonic trends.

Methods in the category 2 using health index to RUL mapping are flexible since new instances are easily incorporated into the library. The main drawbacks include the generalization capability of updating techniques (with performance esti-

Table 11. Type 2 methods using health index as input and RUL as output.

Methods	Publication ID
Quadratic fit, Bayesian updating	4
Logistic regression	5
Kernel regression, RVM	7
RVM	16
Gamma process	17
Linear, Bayesian updating	19
RVM, SVM, RNN, Exponential and quadratic fit, Bayesian updating	21
Exponential fit	28
Wiener process	29
Copula	30
HMM, LS-SVR	34

Table 12. Category 2 methods based on individual sensor modeling and classification.

Methods	Publication ID
exTS, supervised classification	10
SVR	13
exTS, ARX	14
ANN, ANFIS	23
Piece-wise linear (multi-models)	24
exTS	25
ELM, unsupervised classification	32

mation on full training/testing datasets). The performance for RUL prediction depends on both mapping functions which are excellent research topics. The approaches for health index (degradation) estimation and modeling are of great importance, not only for C-MAPSS datasets but more generally for PHM. The method used in publications 1, 7 was used the most in estimation of the health index (the first mapping function), with some variants proposed in publications 21, 17, 40. The health index in C-MAPSS datasets is a temporal hidden variable and strongly occluded by operating conditions which have to be inferred from noisy sensor measurements. The estimation of the first mapping function is thus challenging and advanced machine learning and pattern recognition tools have to be considered. Since the operating conditions represent usage load profile as pointed out in publications 9, 18, their sequence is important. Therefore, the estimation of the health index can be improved by taking into account the *sequence* of operating conditions.

The second mapping function generally takes the form of a regression model (Table 11) for which new methods for uncertainty estimation and propagation can be developed (Table 8). Nonlinear methods demonstrated better performance since the degradations of the units demonstrate changing trends, i.e. gradual and slow, sudden and steep, or regular as shown in publications 1, 2, 3. Approaches based on sequences of multiple states (Moghaddass & Zuo, 2014) could be considered to cope with slope change in the health index. Moreover, adaptation to complex noise appears to be a critical issue since generally, in practice, noise is not simply additive and Gaussian but multimodal and dependent on operating conditions (Saxena, Goebel, et al., 2008b). Finally, new model averaging algorithms can be a relevant source of improvement as illustrated in publications 1, 2, 3, 4, 7, 21, 38, 40, for example by considering information fusion tools and ensemble techniques (Kuncheva, Bezdek, & Duin, 2001; Francois, Grandvalet, Denoeux, & Roger, 2003; Kuncheva, 2004; Monteith, Carroll, Seppi, & Martinez, 2011; Hu, Youn, Wang, & Yoon, 2012).

4.3. Category 3: Similarity-based matching

In these methods (Table 13), historical instances of the system (sensor measurements trajectories labeled with known failure times) are used to create a library. For a given test instance similarity with instances in the library is evaluated generating a set of Remaining Useful Life (RUL) estimates that are eventually aggregated using different methods. Compared to category 2 methods, these methods do not make use of training trajectory abstraction into features, but trajectory data (possibly filtered) are themselves stored. Similarity is computed in the sensor space as in publication 27 or using health indices as in publications 1, 7, 17, 21, 40.

As mentioned in publications 1, 7, the test instance and the training instance may take different time in reaching a particular degradation level from the initial healthy state. Therefore, similarity-based matching must accommodate this difference in the early phases of degradation curves. In publication 40, this problem was tackled by assuming a constant initial wear for all instances yielding an offset on health indices. Efficient similarity measures are also necessary to cope with noise and degradation paths. For instance, in publications 1, 7 three different similarity measures were used, and in publication 40, computational geometry tools were used for instance representation and similarity evaluation.

Table 13. Category 3 methods using similarity-based matching.

Methods	Publication ID
HI-based 3 similarity measures and kernel smoothing	1, 7
Similar to 1 and 7 using 1 similarity measure	22
Feature-based similarity, 1 similarity measure, ensemble, degradation levels classification	27
HI-based similarity, polygon coverage similarity, ensemble	40

An advantage of approaches in this category is that new instances can be easily incorporated. Moreover, similarity-based matching approaches have demonstrated good generalization capability on all C-MAPSS datasets as shown in publications 1, 7, 40 despite a high level of noise, multiple simultaneous fault modes, and a number of operating conditions.

This category of algorithms are relatively easily parallelized to reduce computational times needed for inference.

Similarity-based approaches are generally said to be sensitive to the training dataset. However, to our knowledge, this sensitivity has not been studied for C-MAPSS datasets. The important number of training and testing instances in C-MAPSS should help to draw interesting conclusions about the behavior of an algorithm and its ability to map features to RUL with respect to both the quantity and the quality of data (Gouriveau, Ramasso, & Zerhouni, 2013). An interesting direction should be to consider the performances of the similarity-based approaches using multiple metrics and considering various operating conditions and fault modes. This study may help future end-users of prognostics algorithms to select the appropriate one according to the possibility of gathering sufficient and representative data.

5. SOME GUIDELINES TO USING C-MAPSS DATASETS

Another contribution from this paper is through summarizing some guidelines in using C-MAPSS datasets that may help future users to understand and utilize these datasets better. It summarizes information gathered from the literature review and authors' own experiences, which in many cases go beyond the documentation provided along with the datasets. Specifically, it offers some general processing steps and lists relevant publications that describe implementation of these preprocessing steps that could be useful in developing a prognostic algorithm (Figure 3).

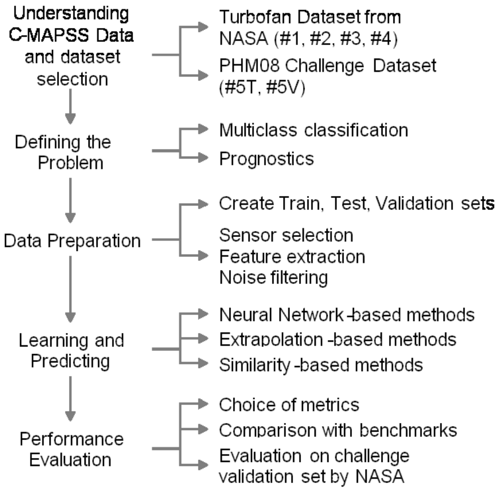

Understanding C-MAPSS Data and dataset selection
Turbofan Dataset from NASA (#1, #2, #3, #4)
PHM08 Challenge Dataset (#5T, #5V)
Defining the Problem
Multiclass classification
Prognostics
Data Preparation
Create Train, Test, Validation sets
Sensor selection
Feature extraction
Noise filtering
Learning and Predicting
Neural Network-based methods
Extrapolation-based methods
Similarity-based methods
Performance Evaluation
Choice of metrics
Comparison with benchmarks
Evaluation on challenge validation set by NASA

Figure 3. Guidelines to Using C-MAPSS Datasets.

Based on the analysis presented in (Section 3), five general data processing and algorithmic steps are considered:

[Step 1:] Understanding C-MAPSS datasets – Comprehensive background information on turbofan engines and C-MAPSS datasets is well presented in three publications, (Saxena, Goebel, et al., 2008b), (Richter, 2012), and (T. Wang, 2010). More details about the hierarchical decomposition of the simulated system into critical components can also be found in (Frederick et al., 2007; Abbas, 2010), which provides valuable domain knowledge. These publications do not focus on the physics-of-failure of turbofan engines but describe generation of these datasets and various practical aspects when using C-MAPSS datasets for prognostics. These include description of sensors measurements, illustrations of operating conditions, impact of fault modes, etc., which can play an important role in improving data-driven prognostics algorithms as well. Going from dataset $\#1$ to $\#4$ represents varying degrees of complexity and, therefore, it is recommended to use them in that order to incrementally develop methods to accommodating individual complexity one by one. The challenge datasets fall somewhere in the middle as far as complexity level goes but suffer from availability of ground truth information for a quicker feedback during algorithm development. Therefore, these datasets may be used as validation examples and should be compared to other approaches using benchmarks presented in Section 2.2.

[Step 2:] Defining the problem – Given the nature of these datasets several types of problems can be defined. As mentioned in Section 3.3 in addition to prediction, a multi-class classification problem can be defined for a multidimensional feature space. However, the intent behind these data was to promote prognostics algorithm development. Since these data consist of multiple trajectories, the problem to predict the RUL for all trajectories can be constructed just as the one posed in the data challenge. However, one could also define the problem at a higher granularity by modeling the degradation for each trajectory individually and predict RUL at multiple time instances, which would be more of a condition based prognostics context.

[Step 3:] Data preparation – After a dataset (turbofan or data challenge) is selected, it is suggested to split the original training dataset into two subsets: a training dataset for model parameter estimation (learning) and a testing dataset to test the learned model 7 (see for example publications 21, 40). For the datasets $\#1-4$ corresponding RUL vectors are provided for the test sets so users can validate their algorithms. However, for the challenge datasets, the evaluations can only be obtained by uploading the RUL to the data repository website. Therefore, it may be desirable to split the training set itself for training, test, and validation purposes during algorithm development. The next step is to downselect sensors to reduce problem dimensionality. Some data exploration and preparation approaches for the data challenge (datasets $\#5T$ and $\#5V$) are well described in publications 1, 2 and 7. Some "heuristic rules" to avoid over-predictions are also presented in publication 40 and applied on all five C-MAPSS datasets. Some of the better performing methods are based on a PCA such as in publication 1, and other sensor selection procedures such as in publications 2, 3 and 40. From the survey, it was noted that the most commonly selected subset of sensors was $7, 8, 9, 12, 16, 17, 20$ (as it was also initially suggested in publication 1). Additional sensors may also be considered, similar to the approach proposed in publication 40 where a total of 511 combinations were studied for each dataset for an exhaustive evaluation.

[Step 4:] Learning and Predicting – This step forms the core of the prediction problem. As described in Section 3 a variety of learning approaches can be employed to learn various mappings between the sensor data and system health to compute RUL. Some of these methods try to learn RUL as a function of sensor data (system state) or features thereof, others estimate a health index first. Each of the trajectory can be modeled into a degradation process to predict when they cross the zero health threshold using regression methods. Approaches based on health index computation can be applied to all datasets. The approach proposed in publications 1, 7 is the simplest to implement. To deal with normalization (or alternatively segmentation) of data by operating conditions one could use a clustering approach as suggested by the authors above, or one may directly use the parameters described in publication 18 to validate the performance of segmentation. Some variants for health indicator estimation can also be picked from publications 21 and 40.

[Step 5:] Performance evaluation – Once a learned model results in to satisfactory results on the testing set aside by partitioning the training data, one may use the actual test dataset provided with the datasets. After further tuning, especially for datasets ($\#5T$ and $\#5V$), a final validation can be done by submitting the results to the NASA repository. Before uploading the final submission, the generalization capability should be ensured by computing using several performance metrics as discussed in Section 2.2. Some benchmarks have been provided in Section 2.2 using metrics that aggregate prediction performance from multiple units. While the exact numbers would not match, the performance is expected to be in the similar range for results obtained from turbofan datasets that have access to RUL. For comparison purposes, the scores obtained in previous works on full C-MAPSS datasets are summarized in Table 14 for the PHM'08 data challenge and Table 15 for the turbofan datasets. Note that here using the full trajectory data it is possible to compute prognostics metrics as presented in (Saxena, Celaya, et al., 2008; Saxena et al., 2010) as the actual EOL is known apriori. This allows testing the critical time aspect of a prediction in addition to accuracy and precision measures.

6. CONCLUSION

As observed from published PHM literature the most widely used datasets for data-driven prognostics come from the C-MAPSS turbofan simulator from among the other openly available prognostic datasets. Guided by this observation, a survey of approaches developed using these datasets (since 2008) was carried out with the purpose of understanding the current state-of-the-art and assess how these datasets have helped in development of prognostic algorithms. However, it was noticed that due to several factors these datasets did not get used as intended and any meaningful comparison between approaches was not trivial. Specifically, following observations were made and this paper tries to alleviate some of these factors to improve usage of these datasets as originally intended:

- Several thousand downloads and 70 papers referring to C-MAPSS were found in the published literature. The ratio suggests that a vast majority of those who downloaded did not get to utilize these data to the point of publishing the results in a publication. Therefore, some guidance has been provided to help in understanding these datasets and how a prognostics problem may be set up

Table 14. Performance of approaches on the PHM'08 datasets (testing dataset $\#5T$ and final validation dataset $\#5V$) after 2008 (published work).

Algorithm & Publication ID	$\#5T$	$\#5V$
RULCLIPPER in 40	752	11572
SBL in 16	1139	-
DW in 21	1334	-
OW in 21	1349	-
MLP in 8	1540	-
AW in 21	1863	-
SVM-SBI in 21	2047	-
RVM-SBI in 21	2230	-
EXP-SBI in 21	2282	-
GPM3 in 4	2500	-
RNN in 21	4390	-
REG2 in 8	6877	-
GPM2B in 4	19200	-
GPM2v in 4	20600	-
GPM1 in 4	22500	-
QUAD in 21	53846	-

Table 15. Performance of approaches on the full training/testing turbofan datasets.

ID	Measures	$\#1$	$\#2$	$\#3$	$\#4$
20	MSE	3969	-	-	-
	MSE	44100	-	-	-
27	Accuracy (%)	53	-	-	-
	FPR (%)	66	-	-	-
	FNR (%)	34	-	-	-
28	MAPE (%)	9	-	-	-
40	PHM08 Score	216	2796	317	3132
	Accuracy (%)	67	46	59	45
	FPR (%)	56	51	66	49
	FNR (%)	44	49	34	51
	MAPE (%)	20	32	23	34
	MAE	10	17	12	18
	MSE	176	524	256	592

in few different ways. Furthermore, a description of all five C-MAPSS datasets is provided identifying their distinguishing characteristics and clearing up some misunderstandings as identified from the survey.

- Among the 70 papers, only a few actually used the testing datasets for evaluating their methods. A mix of different datasets and the metrics used to evaluate performance was observed from the survey. This made it difficult to compare performance between different reported methods in a consistent manner. Therefore, a better explanation of differences in these datasets and providing the top thirty scores from challenge datasets should help future users in comparing their methods against a benchmark in a more consistent manner. Furthermore, it is also suggested how results from datasets that are not from the challenge could be compared against this benchmark established on the challenge set. In addition to the timeliness, several other metrics have been computed from the PHM'08 Challenge leader board and presented in this paper to help in comparisons.
- The survey reveals usage of various prognostics approaches that can be divided into three main categories. These approaches are briefly described with potential areas for further improvement. The survey also demonstrated that C-MAPSS datasets can be used for developing and testing methods for several intermediate steps in prognostics such as sensor selection, health indicator estimation, operating conditions modeling in addition to fault estimation and prediction.

With the analysis presented in this paper and references to a variety of approaches employed, this paper hopes to establish public knowledge that can be used by future users in prognostic algorithm development and aid in fulfilling the underlying intent of data repository to facilitate algorithm benchmarking and further development.

NOMENCLATURE

PHM	Prognostics and Health Management
RUL	Remaining Useful Life
CMAPSS	Commercial Modular Aero-Propulsion System Simulation
HI	Health index
MLP	MultiLayer Perceptron
ANN	Artificial neural network
RNN	Recurrent neural network
RBF	Radial basis function
ESN	Echo state network
ELM	Extreme learning machine
EKF	Extended Kalman filter
KF	Kalman filter
SVR	Support vector regression
LS-SVR	Least squared support vector regression
exTS	Evolving extended Takagi-Sugeno system
ARX	Autoregressive exogeneous model
ANFIS	Adaptive neuro fuzzy inference system
RVM	Relevance vector machine
HMM	Hidden Markov model
PCA	Principal components analysis
MSE	Mean squared error
MAPE	Mean absolute percentage error
MAE	Mean absolute error
ME	Mean error
PH	Prediction horizon
AP	Acceptable predictions (rate)
$\alpha - \lambda$	Accuracy at specific times
RA	Relative accuracy
CV	Convergence
AB	Average bias
FPR	False positive rate
FNR	False negative rate

ACKNOWLEDGMENT

This work has been partly carried out within the Laboratory of Excellence ACTION through the program "Investments for the future" managed by the French National Agency for Research (ANR-11-LABX-01-01) and with partial support from NASA's System-wide Safety and Assurance Technologies (SSAT) Project under ARMD/Aviation Safety program. The authors also thank the anonymous reviewers for their valuable comments.

REFERENCES

Abbas, M. (2010). *System level health assessment of complex engineered processes*. Unpublished doctoral dissertation, Georgia Institute of Technology.

Al-Salah, T., Zein-Sabatto, S., & Bodruzzaman, M. (2012). Decision fusion software system for turbine engine fault diagnostics. In *Southeastcon, 2012 proceedings of ieee* (p. 1-6).

Angelov, P., Filev, D., & Kasabov, N. (2010). *Evolving intelligent systems: Methodology and applications* (J. Willey & Sons, Eds.). IEEE Press Series on Computational Intelligence.

Butcher, J. B., Verstraeten, D., Schrauwen, B., Day, C. R., & Haycock, P. W. (2013). Reservoir computing and extreme learning machines for non-linear time-series data analysis. *Neural Network*, *38*, 76–89.

Coble, J. (2010). *Merging data sources to predict remaining useful life - an automated method to identify prognostic parameters*. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Coble, J., & Hines, J. (2008). Prognostic algorithm categorization with phm challenge application. In *Ieee int. conf. on prognostics and health management.*

Coble, J., & Hines, W. (2011). Applying the general path model to estimation of remaining useful life. *International Journal of Prognostics and Health Management*, *2*, 1-13.

El-Koujok, M., Gouriveau, R., & Zerhouni, N. (2011). Reducing arbitrary choices in model building for prognostics: An approach by applying parsimony principle on an evolving neuro-fuzzy system. *Microelectronics Reliability*, *51*(2), 310 - 320.

Feldkamp, L., & Puskorius, G. (1998, Nov). A signal processing framework based on dynamic neural networks with application to problems in adaptation, filtering, and classification. *Proceedings of the IEEE*, *86*(11), 2259-2277.

Francois, J., Grandvalet, Y., Denoeux, T., & Roger, J.-M. (2003). Resample and combine: An approach to improving uncertainty representation in evidential pattern classification. *Information Fusion*, *4*(2), 75-85.

Frederick, D., DeCastro, J., & Litt, J. (2007). *User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS)* (Tech. Rep.). Cleveland, Ohio 44135, USA: National Aeronautics and Space Administration (NASA), Glenn Research Center.

Gouriveau, R., Ramasso, E., & Zerhouni, N. (2013). Strategies to face imbalanced and unlabelled data in PHM applications. *Chemical Engineering Transactions*, *33*, 115-120.

Gouriveau, R., & Zerhouni, N. (2012). Connexionist-systems-based long term prediction approaches for prognostics. *IEEE Trans. on Reliability*, *61*, 909-920.

Heimes, F. (2008). Recurrent neural networks for remaining useful life estimation. In *Ieee int. conf. on prognostics and health management.*

Hu, C., Youn, B., Wang, P., & Yoon, J. (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. *Reliability Engineering and System Safety*, *103*, 120 - 135.

Huang, G., Zhu, Q., & Siew, C. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. In *International joint conference on neural networks.*

Ishibashi, R., & Nascimento Junior, C. (2013). GFRBS-PHM: A genetic fuzzy rule-based system for phm with improved interpretability. In *Ieee conference on prognostics and health management (phm)* (p. 1-7).

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, *304*(5667), 78-80.

Javed, K., Gouriveau, R., Zemouri, R., & Zerhouni, N. (2012). Features selection procedure for prognostics: An approach based on predictability. In *8th ifac symposium on fault detection, supervision and safety of technical processes* (Vol. 8, p. 25-30).

Javed, K., Gouriveau, R., & Zerhouni, N. (2013). Novel failure prognostics approach with dynamic thresholds for machine degradation. In *Ieee industrial electronics conference.*

Jianzhong, S., Hongfu, Z., Haibin, Y., & Pecht, M. (2010). Study of ensemble learning-based fusion prognostics. In *Prognostics and health management conference, 2010. phm '10.* (p. 1-7).

Kim, H.-E. (2010). *Machine prognostics using health state probability estimation.* Unpublished doctoral dissertation, School of engineering systems, Faculty of built environmental engineering, Queensland university of technology.

Klir, G., & Wierman, M. (1999). Uncertainty-based information. elements of generalized information theory. In (chap. Studies in fuzzyness and soft computing). Physica-Verlag.

Kuncheva, L. (2004). Classifier ensembles for changing environments. In *Int. workshop on multiple classifier systems* (Vol. 3077, p. 1-15).

Kuncheva, L., Bezdek, J., & Duin, R. (2001). Decision templates for multiple classifier fusion. *Pattern Recognition*, *34*, 299-314.

Li, X., Qian, J., & Wang, G. (2013). Fault prognostic based on hybrid method of state judgment and regression. *Advances in Mechanical Engineering*, *2013*(149562), 1-10.

Liao, H., & Sun, J. (2011). Nonparametric and semi-parametric sensor recovery in multichannel condition monitoring systems. *IEEE Transactions on Automation Science and Engineering*, *8*(4), 744-753.

Lin, Y., Chen, M., & Zhou, D. (2013). Online probabilistic operational safety assessment of multi-mode engineering systems using Bayesian methods. *Reliability Engineering & System Safety*, *119*(0), 150 - 157.

Liu, K., Gebraeel, N. Z., & Shi, J. (2013). A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. *IEEE Trans. on Automation Science and Engineering*.

Moghaddass, R., & Zuo, M. (2014). An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process. *Reliability Engineering and System Safety*, *124*, 92–104.

Monteith, K., Carroll, J., Seppi, K., & Martinez, T. (2011, July). Turning bayesian model averaging into bayesian model combination. In *International joint conference on neural networks* (p. 2657-2663). doi: 10.1109/IJCNN.2011.6033566

Peel, L. (2008). Data driven prognostics using a Kalman filter ensemble of neural network models. In *Int. conf. on prognostics and health management.*

Peng, Y., Wang, H., Wang, J., Liu, D., & Peng, X. (2012). A modified echo state network based remaining useful life estimation approach. In *Ieee phm conference.*

Peng, Y., Xu, Y., Liu, D., & Peng, X. (2012). Sensor selection with grey correlation analysis for remaining useful life evaluation. In *Annual conference of the phm society.*

Raftery, A., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2003). Using bayesian model averaging to calibrate forecast ensembles. *American Meteorological Society*, *133*(5), 1155-1174.

Ramasso, E. (2009). Contribution of belief functions to hidden Markov models with an application to fault diagnosis. In *Machine learning for signal processing.*

Ramasso, E. (2014a). Investigating computational geometry for failure prognostics. *Int. Journal on Prognostics and Health Management*. (submitted)

Ramasso, E. (2014b). Investigating computational geometry for failure prognostics in presence of imprecise health indicator: Results and comparisons on c-mapss datasets. In *European conf. on prognostics and health management.*

Ramasso, E., & Denoeux, T. (2013). Making use of partial knowledge about hidden states in hidden Markov models: an approach based on belief functions. *IEEE Transactions on Fuzzy Systems*, *10.1109/TFUZZ.2013.2259496*.

Ramasso, E., & Gouriveau, R. (2010). Prognostics in switching systems: Evidential Markovian classification of real-time neuro-fuzzy predictions. In *Ieee prognostics and health management conference.*

Ramasso, E., & Gouriveau, R. (2013). RUL estimation by classification of predictions: an approach based on a neuro-fuzzy system and theory of belief functions. *IEEE Transactions on Reliability*, *Accepted*.

Ramasso, E., Rombaut, M., & Zerhouni, N. (2013). Joint prediction of observations and states in time-series based on belief functions. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, *43*, 37-50.

Riad, A., Elminir, H., & Elattar, H. (2010). Evaluation of neural networks in the subject of prognostics as compared to linear regression model. *International Journal of Engineering & Technology*, *10*, 52-58.

Richter, H. (2012). Engine models and simulation tools. In *Advanced control of turbofan engines* (p. 19-33). Springer New York.

Sarkar, S., Jin, X., & Ray, A. (2011). Data-driven fault detection in aircraft engines with noisy sensor measurements. *Journal of Engineering for Gas Turbines and Power*, *133*, 081602.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, W. (2008). Metrics for evaluating performance of prognostic techniques. In *Int. conf. on prognostics and health management.*

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*.

Saxena, A., & Goebel, K. (2008). C-mapss data set. *NASA Ames Prognostics Data Repository*.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008a). Damage propagation modeling for aircraft engine run-to-failure simulation. In *Prognostics and health management, 2008. phm 2008. international conference on* (pp. 1–9).

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008b). Damage propagation modeling for aircraft engine run-to-failure simulation. In *Int. conf. on prognostics and health management* (p. 1-9). Denver, CO, USA.

Serir, L., Ramasso, E., & Zerhouni, N. (2012). An evidential evolving multimodeling approach for systems behavior prediction. In *Annual conference of the phm society.*

Siegel, D. (2009). *Evaluation of health assessment techniques for rotating machinery*. Unpublished master's thesis, Division of Research and Advanced Studies of the University of Cincinnati.

Siqueira, H., Boccato, L., Attux, R., & Lyra, C. (2012). Echo state networks and extreme learning machines: A comparative study on seasonal streamflow series prediction. In *Neural information processing* (Vol. 7664, p. 491-500). Springer Berlin Heidelberg.

Son, K. L., Fouladirad, M., & Barros, A. (2012). Remaining useful life estimation on the non-homogenous gamma with noise deterioration based on gibbs filtering : A case study. In *Ieee int. conf. on prognostics and health management.*

Son, K. L., Fouladirad, M., Barros, A., Levrat, E., & Iung, B. (2013). Remaining useful life estimation based on stochastic deterioration models: A comparative study. *Reliability Engineering and System Safety*, *112*, 165 - 175.

Sun, J., Zuo, H., Wang, W., & Pecht, M. (2012). Application of a state space modeling technique to system prognostics based on a health index for condition-based maintenance. *Mechanical Systems and Signal Processing*, *28*, 585 - 596.

Tamilselvan, P., & Wang, P. (2013). Failure diagnosis using deep belief learning based health state classification. *Reliability Engineering and System Safety*, *115*(0), 124 - 135.

Wang, P., Youn, B., & Hu, C. (2012). A generic probabilistic framework for structural health prognostics and uncertainty management. *Mechanical Systems and Signal Processing*, *28*, 622 - 637.

Wang, T. (2010). *Trajectory similarity based prediction for remaining useful life estimation*. Unpublished doctoral dissertation, University of Cincinnati.

Wang, T., Yu, J., Siegel, D., & Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *Int. conf. on prognostics and health management* (p. 1-6).

Xi, Z., Jing, R., Wang, P., & Hu, C. (2013). A copula-based sampling method for data-driven prognostics and health management. In *Asme 2013 international design engineering technical conferences and computers and information in engineering conference.*

Xue, Y., Williams, D., & Qiu, H. (2011). Classification with imperfect labels for fault prediction. In *Proceedings of the first international workshop on data mining for service and maintenance* (pp. 12–16). ACM.

Yu, J. (2013). A nonlinear probabilistic method and contribution analysis for machine condition monitoring. *Mechanical Systems and Signal Processing*, *37*, 293-314.

Zein-Sabatto, S., Bodruzzaman, J., & Mikhail, M. (2013). Statistical approach to online prognostics of turbine engine components. In *Southeastcon, 2013 proceedings of ieee* (p. 1-6).

Zhao, D., P., R. G., & Willett. (2011). Comparison of data reduction techniques based on SVM classifier and SVR performance. In *Proc. spie, signal and data processing of small targets* (Vol. 8137, p. 1-15).

APPENDIX

All references were mapped to numeric identifiers to be used in survey and analysis results for better readability. This mapping is provided in the Table 16 below.

Table 16. References to ID mapping.

Reference	Publication ID
(T. Wang et al., 2008)	1
(Heimes, 2008)	2
(Peel, 2008)	3
(Coble & Hines, 2008) (Coble, 2010) (Coble & Hines, 2011)	4
(Siegel, 2009)	5
(Ramasso, 2009)	6
(T. Wang, 2010)	7
(Riad, Elminir, & Elattar, 2010)	8
(Abbas, 2010)	9
(Ramasso & Gouriveau, 2010) (Ramasso & Gouriveau, 2013)	10
(Sarkar, Jin, & Ray, 2011)	11
(Xue, Williams, & Qiu, 2011)	12
(Zhao, P., & Willett, 2011)	13
(El-Koujok, Gouriveau, & Zerhouni, 2011)	14
(Liao & Sun, 2011)	15
(P. Wang, Youn, & Hu, 2012)	16
(Son, Fouladirad, & Barros, 2012)	17
(Richter, 2012)	18
(Sun, Zuo, Wang, & Pecht, 2012)	19
(Peng, Wang, Wang, Liu, & Peng, 2012)	20
(Hu et al., 2012)	21
(Peng, Xu, Liu, & Peng, 2012)	22
(Javed, Gouriveau, Zemouri, & Zerhouni, 2012)	23
(Serir, Ramasso, & Zerhouni, 2012)	24
(Gouriveau & Zerhouni, 2012)	25
(Yu, 2013)	26
(Ramasso, Rombaut, & Zerhouni, 2013)	27
(Liu, Gebraeel, & Shi, 2013)	28
(Son, Fouladirad, Barros, Levrat, & Iung, 2013)	29
(Xi, Jing, Wang, & Hu, 2013)	30
(Lin, Chen, & Zhou, 2013)	31
(Javed, Gouriveau, & Zerhouni, 2013)	32
(Ramasso & Denoeux, 2013)	33
(Li, Qian, & Wang, 2013)	34
(Tamilselvan & Wang, 2013)	35
(Ishibashi & Nascimento Junior, 2013)	36
(Gouriveau et al., 2013)	37
(Jianzhong, Hongfu, Haibin, & Pecht, 2010)	38
(Zein-Sabatto, Bodruzzaman, & Mikhail, 2013) (Al-Salah, Zein-Sabatto, & Bodruzzaman, 2012)	39
(Ramasso, 2014b) (Ramasso, 2014a)	40