Prediction Capability Assessment of Data-Driven Prognostic Methods for Railway Applications

Francesco Di Maio[1], Pietro Turati[2], Enrico Zio[3]

[1,3] *Energy Department, Politecnico di Milano, Via La Masa 34, Milano, 20156, Italy*
francesco.dimaio@polimi.it
enrico.zio@polimi.it

[2,3] *Chair System Science and the Energy Challenge, Fondation Electricité de France (EDF), CentraleSupélec, Laboratoire Genie Industriel Université Paris Saclay, Grande Voie des Vignes, 92290 Chatenay-Malabry, France*
pietro.turati@centralesupelec.fr
enrico.zio@centralesupelec.fr

[1,3] *Aramis Srl, Via Pergolesi 5, Milano, Italy*

ABSTRACT

In the development of Prognostics and Health Management (PHM) for industrial applications, the question of which predictive method to use is fundamental. The choice is typically driven by the data and/or the physics-based models available, and the cost-benefit considerations related to PHM implementation, wherein prediction capability plays an important role. By prediction capability of a prognostic method we refer to its ability to provide trustable predictions of the Remaining Useful Life (RUL) of a component or system, with the characteristics required by the given application. A set of Prognostic Performance Indicators (PPIs) is used to guide the choice of the method to be implemented. These PPIs measure different characteristics of a prognostic method and need to be aggregated to enable a final choice of prognostic method, based on its overall performance. We propose an aggregation strategy to identify the prognostic method with the best compromise performance on all PPIs. The strategy is exemplified on a case study with real data taken from industry, whose structure is general and, therefore, applicable to railway industry.

1. INTRODUCTION

In Prognostic and Health Management (PHM), the prediction of the Remaining Useful Life (RUL) of a component or system is of paramount importance for maintenance strategy definition as Condition Based Maintenance (CBM) (Engel, Gilmartin, Bongort, & Hess, 2000; Jardine, Lin & Banjevic, 2006; Vachtsevanos, Lewis, Roemer, Hess & Wu, 2006; Kan, Tan & Mathew, 2015) or Preventive Maintenance (PM) (Zio & Compare, 2013; Lee, Wu, Zhao, Ghaffari, Liao & Siegel, 2014; Compare & Zio, 2014).

Prognostic methods for RUL estimation can be classified into model-based methods (Luo, Namburu, Pattipati, Qiao, Kawamoto, & Chigusa, 2003; Vichare & Pecht, 2006; Luo, Pattipati, Qiao, & Chigusa, 2008; Pecht & Jaai, 2010) and data driven methods (Schwabacher, 2005; Zio, 2009; Si, Wang, Hu & Zhou, 2011; Tsui, Chen, Zhou, Hai & Wang, 2015; Zhang, Si, Hu & Kong, 2015).

In practice, the decision makers' attitude can be either prone or averse to the risk of relying on a predicted RUL to plan maintenance services. Undoubtedly, such attitude is heavily influenced by the prediction capability of the prognostic method used. For instance, a prognostic method with high prediction capability might make the decision maker risk-prone, because he/she feels that he/she can trust the RUL predictions provided by the method and, thus, he/she is willing to take the risk of using them to plan predictive maintenance. On the other hand, if the prediction capability of the prognostic method is not sufficient, the decision maker might be risk-averse towards using the RUL predictions to support any maintenance decision. The prediction capability of a prognostic method is, thus, an important information for deciding which predictive method should be selected for the development of PHM in a given industrial application

Prediction capability can be seen as the capability of a method of guaranteeing good *prediction quality*, e.g., accurate and precise RUL estimates as well as the trust a decision maker can put on a specific prognostic method result before implementing it on his/her particular application (Zeng, Di Maio, Zio & Kang, 2016). Figure 1 reports a scheme of all the factors influencing the prediction capability of a method. Trustworthiness can, indeed, play an important

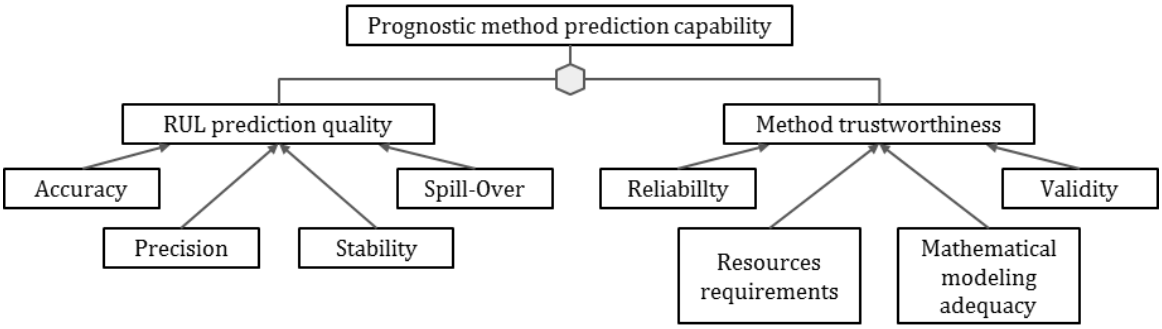

Figure 1. Block diagram describing the prediction capability.

role in the selection process of the prediction method to be used for an application (Paulk, 1993; Herbsleb, Zubrow, Goldenson, Hayes & Paulk, 1997; Farrell & Gallagher, 2015). For example, if two prediction methods, named A and B, provide RUL estimates of the same quality, but method A has been applied in several fields and in many similar applications, while method B has been used only few times and just for a particular application, then, the decision maker is likely to choose method A, since he/she feels that he/she can trust it more. Thus, trustworthiness assesses in a formal way, what typically the expert knowledge is: information about previous applications in the same industrial context, successful applications in different fields, the resource requirements (e.g., data and computational cost), the mathematical adequacy (e.g., the capability of dealing with linear and nonlinear problems), etc. (Zeng et al., 2016).

On the other side, it would be useless to consider only the trustworthiness of a method if it is not capable of providing good RUL estimates, whose quality can be assessed with Prognostic Performance Indicators (PPI) (Saxena et al., 2008; Saxena, Celaya, Saha, Saha & Goebel, 2010; Saxena, Sankararaman & Goebel, 2014). Each PPI aims at quantitatively assessing a specific characteristic of the prediction, such as: *i)* accuracy, *ii)* precision, *iii)* stability and *iv)* spill-over (Walther & Moore, 2005).

However, despite the large number of PPIs available from the literature, it is hard for a practitioner to decide which PPI must be taken into account for the prediction quality assessment. As common in the performance assessment, a single indicator usually assesses a single property of the predictive method (e.g., its accuracy or its precision): in (Micea, Ungurean, Cârstoiu & Groza, 2011) the authors consider only the accuracy to assess the performance of a prognostic method specifically designed for battery management; in (Peng, Liu, Saxena & Goebel, 2015), the authors resort to a single PPI for the prognostic quality assessment of a Bayesian inference framework, whereas their ensemble catches the overall performance: Tobon-Mejia, Medjaher, Zerhouni and Tripot (2012) consider a large set of PPIs for assessing the prognostic quality of a mixture of Gaussian hidden Markov models and Xian, Long, Li and Wang (2014) employ a set of PPIs for assessing a particle filter-based prognostic method.

However, the selection of the most proper combination of PPIs to be used in support of the comparison of different prediction methods in different applications is still an open issue of PHM.

In this paper, we propose two techniques for aggregating PPIs, in order to have a synthetic measure of prediction quality. This measure can be used by the decision maker in the choice of the most apt prognostic method for the application under consideration. A case study from the industry is considered to show how these techniques can be also used in railway applications.

The rest of the paper is structured in this way: in Section 2, a general description of the RUL prediction quality problem is outlined; in Section 3, the proposed aggregation techniques are described; in Section 4, the case study taken from the industry is presented and, finally, in Section 5, some conclusions and possible future perspectives are drawn.

2. RUL PREDICTION QUALITY

Both model-based and data-driven prognostic methods process a number of signals and information collected from sensors placed on the system under analysis, in order to predict the RUL (Engel et al. 2000; Vachtsevanos et al. 2006; Vichare & Pecht, 2006; Zio, 2009; Pecht & Jaai, 2010; Si et al. 2011; Tsui et al. 2015; Zhang et al. 2015). However, the performance of the prediction can largely vary according to the type of signals and data used. Expert knowledge plays a key role, at this stage, in understanding if the best prediction can be directly achieved by a direct elaboration of the raw signals or if a pre-processing of the raw signals is needed to extract features to be used as input for the prediction model.

The quality of a RUL prediction is here assessed by measuring four main characteristics which include: *i)* accuracy, *ii)* precision, *iii)* stability and *iv)* spill-over (Walther et Moore, 2005; Saxena et al., 2008; Johnson et al., 2011; Saxena et al., 2014). In a few words:

- accuracy PPIs quantify the closeness between the model output and the true value: a very accurate model will have an estimated RUL very close to the true one.
- Precision PPIs measure the statistical variability of the predicted value: in practice, they quantify the uncertainty in the estimate and, consequently, the degree to which a repetition of the prognosis will lead to the same results.
- Stability PPIs quantify the sensitivity of the predicted values with respect to the available inputs in order to evaluate the robustness of the estimate and the capability of the method of estimating the correct value while approaching the end of life of the system (convergence), i.e. while the prediction horizon decreases.
- Spill-Over PPIs measure the effects of varying the set of inputs of the predictive model: in practice, what happens if one or more of the inputs is not available.

In literature and industrial applications, a large number of PPIs have been defined. In Figure 2, a large, even though not exhaustive, selection is listed. In Appendix, the PPIs of Figure 2 are formulated in such a way that they take the maximum value 1 when the method performs well, and lower values when the method performance is not optimal.

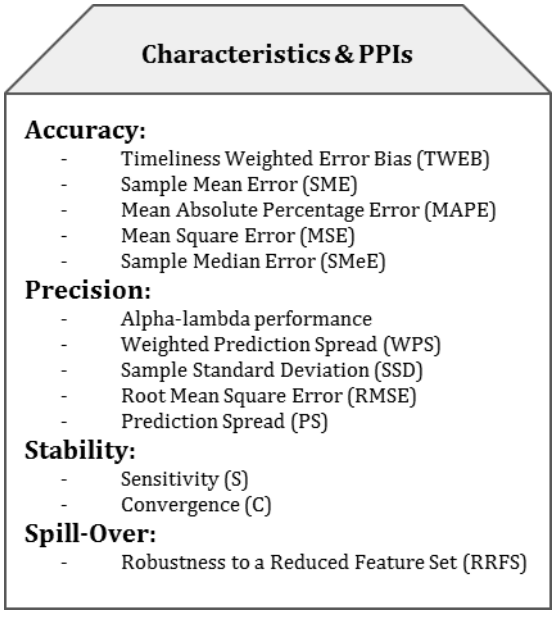

Figure 2. Selection of PPIs, grouped according to their characteristics (see Appendix for their definitions).

3. PROGNOSTIC PERFORMANCE INDICATOR AGGREGATION

In this Section, we present two strategies to manage and get benefit from the large number of PPIs available in the literature, to enable a robust final choice of the prognostic method based on its overall performance. Indeed, instead of a priori selecting an ensemble of PPIs as guiding criterion for the assessment of the prediction quality (Micea et al., 2011; Tobon-Mejia et al., 2012; Xian et al. 2014; Peng et al., 2015), the strategies here proposed give a synthetic quantitative assessment of the overall performance. In both cases, the results of the aggregation should give an intuitive and clear suggestion to the analyst regarding which prognostic method to choose in terms of prediction quality, for the application of interest. The two strategies, which differ in the implementation, complexity and time demand, are described in detail in the following two subsections.

3.1. Weighted Average Strategy

In the Weighted Average Strategy (WAS), the values of the different PPIs are averaged according to specific weights. Hereafter, we refer to the list of PPIs reported in Figure 2 and in the Appendix, where, for each characteristic (accuracy, precision, etc.) the rank reflects the trust that experts have with respect to the PPIs: the higher the rank of a PPI, the more the experts trust the information conveyed by that PPI. Nonetheless, different rankings can be employed.

The overall prognostic quality of a method can be quantified as in Eq. (1):

$$WAS = \frac{1}{\sum_{j=1}^{N_C}\sum_{p=1}^{N_j} w_p^j} \sum_{j=1}^{N_C}\sum_{p=1}^{N_j} w_p^j \cdot PPI_p^j \qquad (1)$$

where the subscript and the superscript indicate the position p of the PPI in the rank of the characteristic j, respectively, with $p = 1, \dots, N_j$ and $j = 1, \dots, N_C$, where N_j is the number of PPIs for the characteristic j and N_C is the number of characteristics considered in the prognostic quality assessment. The weight w_p^j corresponds to the p^{th} PPI of the j^{th} characteristic and is computed as in Eq. (2):

$$w_p^j = 1 - \frac{p-1}{N_j} \qquad (2)$$

For example, the weight of the Timeliness Weighted Error Bias (PPI A.1 in Appendix, for details) is 1, as it is ranked first in the accuracy list:

$$w_1^1 = 1 - \frac{p-1}{N_1} = 1 - \frac{1-1}{5} = 1$$

whereas the weight of the Sample Mean Error (PPI A.2 in Appendix) is 0.8, as it is the second out of 5 PPIs in the accuracy list:

$$w_2^1 = 1 - \frac{p-1}{N_1} = 1 - \frac{2-1}{5} = 0.8$$

Following this procedure, it is possible to compute the weights for all PPIs of the accuracy, precision and sensitivity characteristics, whatever is the number of PPIs considered.

It must be noticed that the Spill-Over PPI (SO.1 in Appendix) can be either computationally expensive when a large number of features feeds the method, since it requires evaluating the PPIs for several subsets of features, or trivial when only one feature is to be fed to the method. For this reason, Spill-Over should specifically be treated either by making an average over the different subsets of features (if the computational cost permits it), or by excluding it form the WAS evaluation due to its too large computational demand.

In synthesis, the overall performance is the weighted average of all PPIs provided by the WAS and gives an indication of the average weighted performance for all the characteristics: as all PPIs have ranges between $-\infty/0$ and 1, a good performance will give a value close to 1 and a relatively bad performance would give a value lower than 1.

It is worth mentioning that, since the result of the WAS strategy depends on the values assigned to weights w_p^j, different rules for the weights definition can be designed in order to prefer some characteristic rather than others.

3.2. In-Depth Quality Control Strategy

The In-Depth Quality Control Strategy (IDQCS) bases the overall performance assessment on weighting the PPIs on their spill-over performance. The rationale is that in real applications, where it is not rare to have many sensors monitoring the same system, it is necessary (or at least desirable) that the prognostic method keeps the same or a similar prediction quality, even when some of the signals are out of order. Thus, the spill-over indicator SO.1 in Appendix, which assesses the variation of the PPIs when reducing the number of features fed to the method, plays a pivotal role within the IDQCS aggregation strategy.

The strategy whose flowchart is given in Figure 3, proceeds as follows: 1) for the first characteristic (i.e. accuracy), the spill-over SO.1 is evaluated for all PPIs which are, then, sorted in descending order from the one having the highest value of SO.1 (i.e., the one whose value is the least affected by a feature reduction) to the one having the lowest value (i.e., the one whose value is the most affected); 2) following the rank, each PPI is evaluated and its value compared with the corresponding acceptance threshold (see Table 1): if the PPI is above the threshold, the PPI is saved as representative of the associated characteristic and the strategy returns at step 1) for the following characteristic, otherwise the algorithm is interrupted since the prognostic quality of the method under test is deemed *a priori* unacceptable with respect to the minimum requirements for the characteristic under analysis. The values for the acceptance thresholds used in this paper are reported in Table 1; they have been chosen according to expert judgment. However, more or less restrictive values can be used, depending on the application requirements.

Once a PPI has been selected for each characteristic, then, they are aggregated as in Eq. (3):

$$IDQCS = \frac{1}{3}\sum_{j=1}^{N_C-1} w_j \cdot PPI_j \qquad (3)$$

where w_j and PPI_j are the weight and the selected PPI for the j^{th} characteristic, respectively. For the sake of clarity $j = 1$ refers to the accuracy, $j = 2$ to the precision and $j = 3$ to the stability.

Spill-over influences IDQCS not only during the sorting process but also in the weights definition, as follows:

$$w_j = SO.1\left(PPI_j\right) \qquad (4)$$

where w_j represents the spill-over indicator of the PPI selected for the j^{th} characteristic, so that PPIs having a low spill-over indicator are penalized in the weighting process.

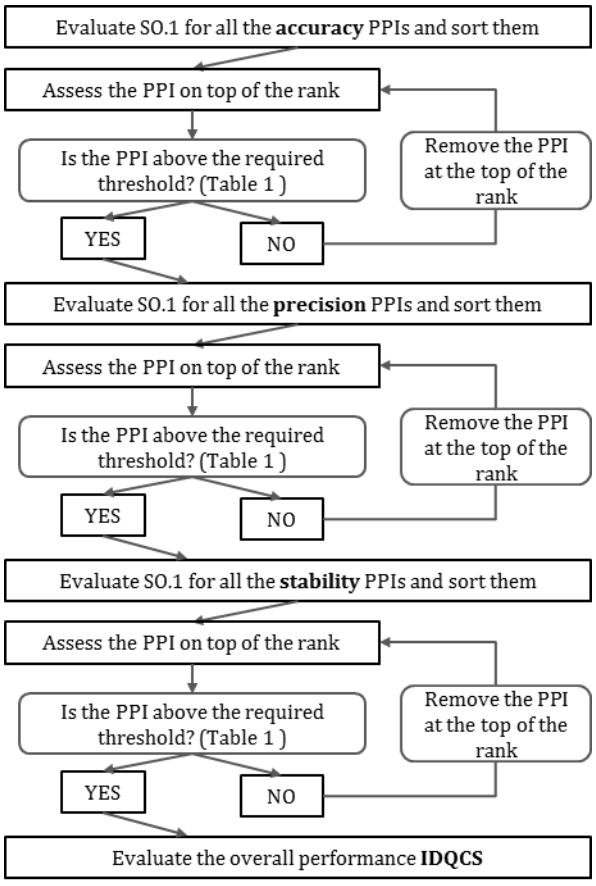

Figure 3. Flowchart of the In-Depth Quality Control Strategy.

As in the WAS strategy, the spill-over indicator can be assessed as the average over the different subsets of features. However, differently from the WAS, in the IDQCS, the spill-over characteristic, which involves the most computational demanding indicator, is not evaluated directly for all the PPIs, but in an iterative way. Firstly, the spill-over for the accuracy PPIs is evaluated; then, only if the prognostic method has already satisfied the accuracy requirements, the spill-over for the precision PPIs is evaluated. Once both accuracy and precision requirements have been satisfied, the algorithm proceeds to the evaluation of the spill-over for the stability PPIs. This iterative procedure allows evaluating the spill-over indicator only if some minimum requirements have been already achieved.

Table 1. List of acceptance thresholds.

PPI	Acceptance Treshold	PPI	Acceptance Treshold	PPI	Acceptance Treshold
A.1	0.75	P.1	0.8	S.1	0.8
A.2	0.8	P.2	0.8	S.2	0.3
A.3	0.75	P.3	0.8	-	-
A.4	0.8	P.4	0.75	-	-
A.5	0.8	P.5	0.75	-	-

4. CASE STUDY

The case study under analysis mimics the behavior of a component of the railway sector and concerns a mono-dimensional signal. Available data include 9 independent run-to-failure trajectories of a physical quantity monitored on the component. In must be mentioned that the real data are not specific to a railway component, but taken from another industrial sector with similar characteristics. From Figure 4, it can be seen that the trajectories have: different initial conditions, high noise/signal ratio and different signal values at failure (the unit of measure has been removed for confidentiality). For these reasons, to improve the tractability of the signals for prediction purposes, the raw data have been preprocessed to extract a relevant feature to be fed to the prognostic method.

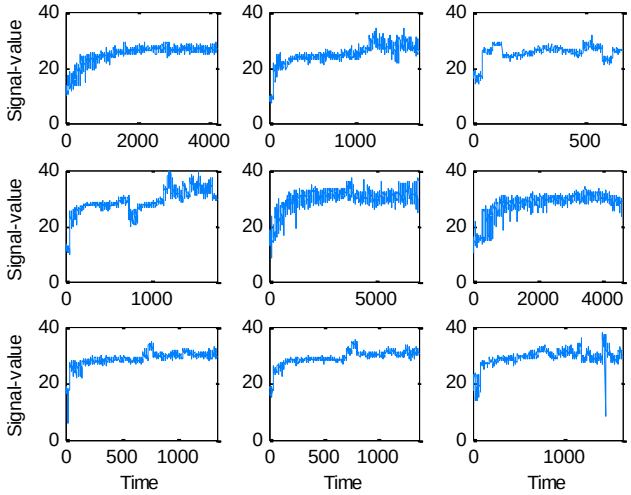

Figure 4. Run-to-failure trajectories available.

Empirical Mode Decomposition (Huang & Wu, 2008) has been employed to extract one feature reflecting the degradation process of the component. The trajectories are reported in Figure 5 (the different shades of colors will turn to be useful in what follows). Two main trends can be observed: *i*) an initial fast degradation followed by low degradation until reaching failure (top-left trajectory); *ii*) an initial fast degradation leading to a relative short plateau followed by a slower degradation until reaching failure (middle-right trajectory).

Three prognostic methods have been chosen as candidates for providing RUL predictions based on a literature review

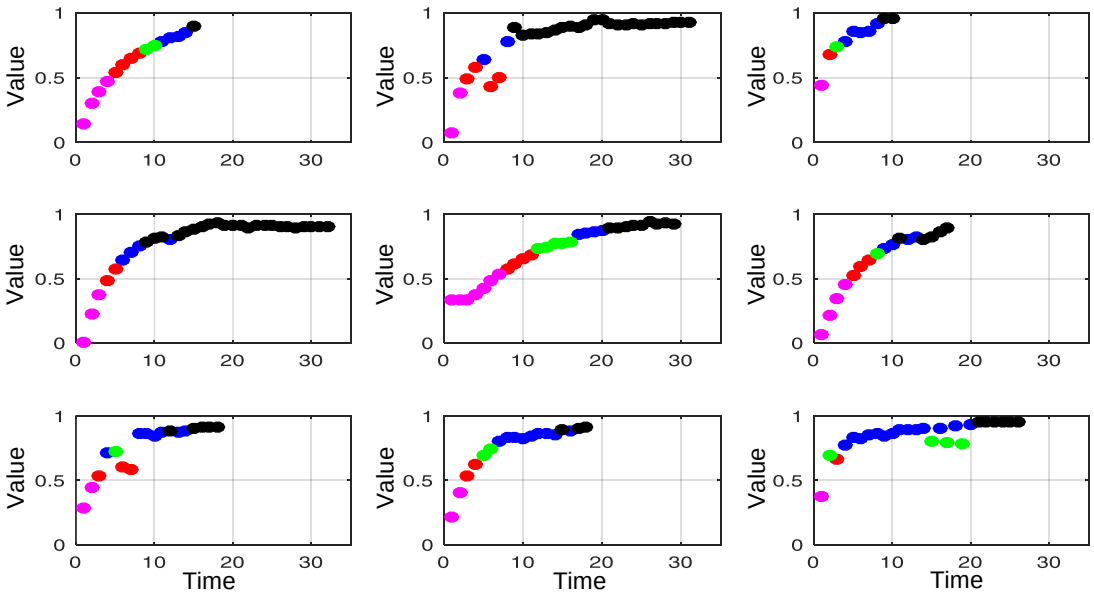

Figure 5. Trajectories where the 5 degradation states of the HSMM are shown with different shade of colors.

tackling similar problems: Fuzzy Similarity (FS), feed-forward Artificial Neural Networks (ANN) and Hidden Semi Markov Model (HSMM).

For the Fuzzy Similarity (Mendel, 1995; Zio & Di Maio, 2010; Di Maio & Zio, 2013), the basic idea is to consider each trajectory available as an historical dataset. Whenever a new trajectory is monitored, it is compared with those in the historical set through a fuzzy similarity measure and it is matched with the most similar. The RUL prediction for the new trajectories is, thus, the weighted average of the most similar historical trajectories. To avoid identifying the trajectory under analysis as the one similar to itself, a leave-one-out, cross-validation procedure has been employed (Efron, 1983; Polikar, 2007).
For the ANN, a two-layer feed-forward network with sigmoid hidden neurons and linear output neurons is trained with the Levenberg-Marquardt backpropagation algorithm (Mahamad, Saon, & Hiyama, 2010; Sikorska, Hodkiewicz & Ma 2011; Yu & Wilamowski 2011; Kan, Tan & Mathew, 2015). The input considered is the value of the feature and the output is the corresponding RUL. The assessment process of the ANN has followed the same cross-validation strategy as for the Fuzzy Similarity.
For the HSMM (Bechhoefer, Bernhard, He, & Banerjee, 2006; Dong & He, 2007; Yu, 2010), the degradation trajectories have been discretized in 5 states on the basis of 4 transition thresholds, which are represented by the different shades of color in Figure 5.The threshold values has been identified by means of a non-supervised clustering algorithm. The 5th state represents the component failure. In this light, the RUL is estimated by means of the time of first entrance in state 5. It must be noticed that 2 out of the 9 degradation processes jump directly from the degradation state 2 (grey) to the degradation state 4 (black) without passing through the degradation state 3 (light grey). This strongly affects the parameter estimation and, therefore, could affect the RUL prediction performance, obtained by Monte Carlo simulation of the chain.

4.1. PPIs Calculation

Table 2 lists the PPIs that are, then, aggregated in the following Subsection 4.2. All PPIs have been computed following the definitions provided in Appendix. However, some specific settings concerning some PPIs are reported in what follows.

Regarding PPIs for precision: for P_λ^α, the parameters have been set to $\alpha = 0.2$ and $\lambda = 0.5$, that means that a RUL prediction made at time t has a positive contribution to P_λ^α if the percentage error α at half of the predicted $RUL^*(t)$ (i.e., $\lambda = 0.5$) is lower than 20% (a specific discussion on the selection of those parameters which can strongly affect the results is out of the scope of this paper and is treated in (Saxena, Roychoudhury, Celaya, Saha, Saha, & Goebel, 2012)); for PS, which in principle can be assessed for each accuracy PPI, we evaluate it only for $TWEB$ both for the sake of readability of the results and because in the authors opinion, it is the most trustable one. In addition, $TWEB$ gets a good score and, therefore, it is worth evaluating the corresponding precision.

Regarding PPIs for stability: we have been focusing on the Convergence C_M only. Moreover, even though C_M can be assessed for all the accuracy and precision PPIs, we have decided, coherently with the choice made for PS, to calculate it only for $TWEB$ and P_λ^α.

Finally, since the prognostic method has only one feature as input, Spill-Over PPIs cannot be quantified, as it makes no sense to talk about Spill-Over effects for a one-dimensional signal.

Table 2. PPIs for the three selected methods.

Accuracy	*Fuzzy Similarity*	*ANN*	*HSMM*
$A.1\ TWEB$	0.98	0.94	0.11
$A.2\ SME$	0.37	0.56	-9.47
$A.3\ MAPE$	0.85	0.63	-1.44
$A.4\ MSE$	-5.14	-7.56	-143.79
$A.5\ SMeE$	0.98	0.57	-10.65
Precision			
$P.1\ P_\lambda^\alpha$	0.61	0.34	0.02
$P.2\ WPS$	0.97	0.94	0.67
$P.3\ SSD$	-0.71	-1.11	-4.16
$P.4\ RMSE$	-0.28	-1.42	-10.20
$P.5\ PS_{TWEB}$	0.98	0.97	0.14
Stability			
$S.2\ C_{TWEB}$	0.37	0.35	0.25
$S.2\ C_{P_\lambda^\alpha}$	0.32	0.32	0.22

4.2. PPIs Aggregation

Both the WAS and the $IDQCS$ strategies have been employed for comparison.

Regarding WAS, all the necessary information is reported in Table 2 and the results listed in Table 4: the Fuzzy Similarity has an overall performance better than ANN, whereas HSMM obtains a very bad score that would discourage its employment in the case under analysis.

Regarding $IDQCS$, since the Spill-Over cannot be evaluated, we have resorted to the same ranking proposed in Figure 2, which reflects the authors preference with respect to the different PPIs. Table 3 reports the PPIs that have been selected according to $IDQCS$ for each one of the methods and for each characteristic. The column regarding the HSMM has not been reported since the IDQCS has not found any accuracy PPIs satisfying the acceptance threshold and, thus, the aggregation strategy has been interrupted. For both Fuzzy Similarity and ANN, the same PPIs have been selected for all the characteristics. In particular, coherently with the choice made for the accuracy and precision PPIs, the stability PPI

has been evaluated as the average of the Convergence PPIs (i.e., C_{TWEB} and C_{WPS}). Furthermore, it must be noticed that during the aggregation process of $IDQCS$ the weights w_j have all been set equal to 1 due to the impossibility in evaluating the Spill-Over indicators.

Table 3. Indicators selected by the $IDQCS$ for the three selected methods.

Metric	***Fuzzy Similarity***	***ANN***
Accuracy	$TWEB = 0.98$	$TWEB = 0.89$
Precision	$WPS = 0.97$	$WPS = 0.94$
Stability	$\frac{C_{TWEB} + C_{WPS}}{2} = 0.40$	$\frac{C_{TWEB} + C_{WPS}}{2} = 0.37$

From the results reported in Table 4, it can be seen that both WAS and IDQCS suggest the Fuzzy Similarity as best candidate method to be utilized in the case under analysis. However, while WAS shows large evidence to discriminate between Fuzzy Similarity and ANN, the evidence is strongly reduced in the IDQCS evaluation.

Table 4. Aggregated performances for the three selected methods.

Strategy	***Fuzzy Similarity***	***ANN***	***HSMM***
WAS	0.20	-0.08	-9.81
IDQCS	0.58	0.56	*NaN*

To conclude, we report in Figure 6 the RUL estimates obtained by the prognostic methods analyzed for all the 9 trajectories available, so that it can be confirmed that the aggregation strategies proposed really select the best prognostic method. At a first sight, it can be noticed that the HSMM provides poor RUL predictions with a general trend to overestimating the real RUL. On the contrary, the difference between the Fuzzy Similarity and the ANN is not so evident.

As a last remark, it is worth mentioning the weak performance of the methods for the bottom-right trajectory (except close to the end of life): the aggregated scores take values far from the optimal score, suggesting that more run-to-failure trajectories should be collected to improve the reliability of the overall prediction quality and to support the decision on the method selection.

5. CONCLUSIONS

In the design process of PHM for industrial applications, the selection of an adequate predictive method is of paramount importance. Recent works have underlined that in industrial application two main drivers for an effective selection are: *i*) the trustworthiness of the method, which can be qualitatively quantified with a maturity assessment; *ii*) the prediction quality of the method, i.e. its capacity of providing accurate, precise and robust RUL for the application under analysis.

This paper aims at underlying and drawing attention on the necessity of providing industrial decision makers with more synthetic and easy to interpret performance indicators for the prognostic quality assessment. In this light, we have shown that the aggregation of the PPIs, which are typically involved in the prediction quality assessment, is a viable way. Aggregation strategies provide a synthetic result capable of intuitively guiding the selection of the method with highest overall prediction quality. In particular, IDQCS allows not only ranking the different prognostic methods, but also verifying if a minimum overall prediction quality has been

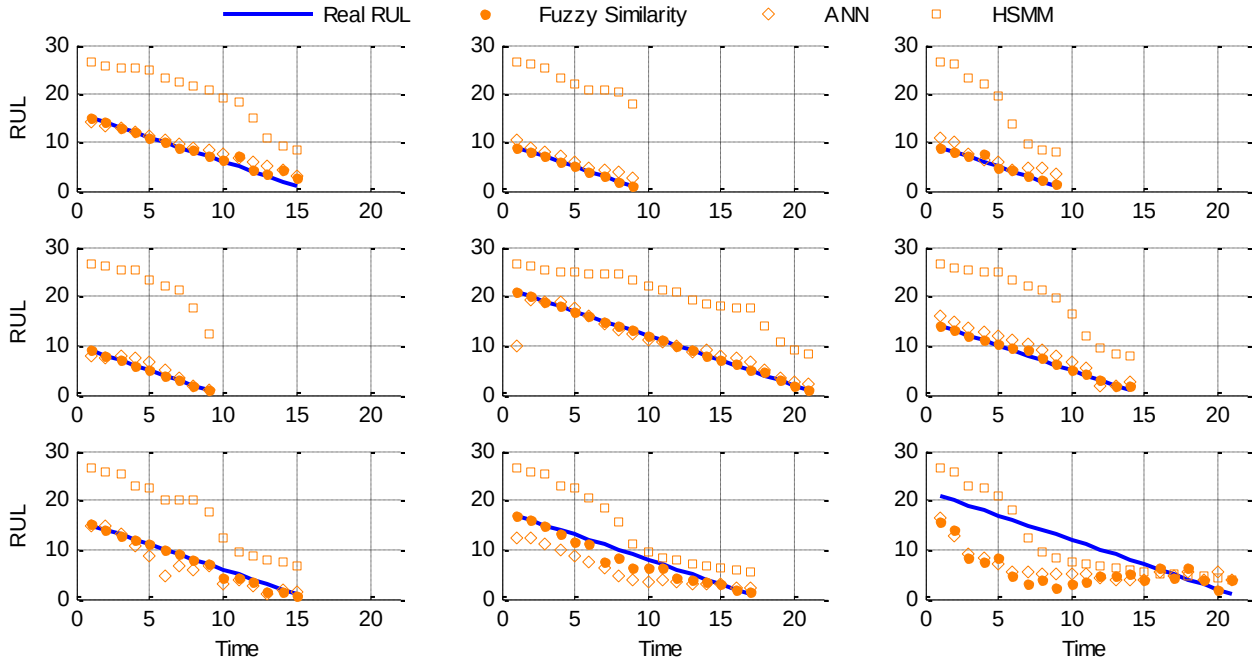

Figure 6. Estimated (light marks) and real (bold dark line) RUL for each component analyzed

achieved by the prognostic methods considered, eventually giving a warning to the analyst. The convenience of the proposed strategies has been tested on a case study related to real industrial data.

It must be mentioned that the results of the proposed strategies are dependent on some design parameters which are here set according to authors' expertise and whose optimal design needs further research. However, the main goal of the paper was to propose and show that an aggregation strategy can provide the decision maker with a valid and synthetic performance index.

ACKNOWLEDGEMENT

This research has been carried out within the study contact "Elaboration of a maturity assessment of data-driven methods in railway industry" with ALSTOM TRANSPORT S.A., Paris, France.

REFERENCES

Bechhoefer, E., Bernhard, A., He, D., & Banerjee, P. (2006). Use of Hidden Semi-Markov Models in the Prognostics of Shaft Failure. In *AHS International 62nd Annual Forum Proceedings*, vol. 2, May 9-11, Phoenix, AZ.

Compare, M., & Zio, E. (2014). Predictive Maintenance by Risk Sensitive Particle Filtering. *Reliability, IEEE Transactions on*,vol *63*(1), pp. 134-143.

Di Maio, F., & Zio, E. (2013). Failure prognostics by a data-driven similarity-based approach. *International Journal of Reliability, Quality and Safety Engineering*, vol. *20* (2), pp. 1-17, doi: 10.1142/S0218539313500010.

Dong, M., & He, D. (2007). A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing*, vol. *21*(5), pp. 2248-2266.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, vol. *78*(382), pp. 316-331.

Engel, S. J., Gilmartin, B. J., Bongort, K., & Hess, A. (2000). Prognostics, the real issues involved with predicting life remaining. In *Aerospace Conference Proceedings, 2000 IEEE*, vol. 6, pp. 457-469.

Farrell, M., & Gallagher, R. (2015). The valuation implications of enterprise risk management maturity. *Journal of Risk and Insurance*, vol. *82*(3), pp. 625-657.

Herbsleb, J., Zubrow, D., Goldenson, D., Hayes, W., & Paulk, M. (1997). Software quality and the capability maturity model. *Communications of the ACM*, vol. *40*(6), pp. 30-40.

Huang, N. E., & Wu, Z. (2008). A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics*, vol. *46*(2), doi: 10.1029/2007RG000228.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, vol *20*(7), pp. 1483-1510.

Johnson, S. B., Gormley, T., Kessler, S., Mott, C., Patterson-Hine, A., Reichard, K., & Scandura Jr, P. (Eds.). (2011). *System health management: with aerospace applications*. John Wiley & Sons.

Kan, M. S., Tan, A. C., & Mathew, J. (2015). A review on prognostic techniques for non-stationary and non-linear rotating systems. *Mechanical Systems and Signal Processing*, vol. 62,pp. 1-20.

Kan, M. S., Tan, A. C., & Mathew, J. (2015). A review on prognostic techniques for non-stationary and non-linear rotating systems. *Mechanical Systems and Signal Processing*, vol. *62*, pp. 1-20.

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, vol *42*(1),pp. 314-334.

Luo, J., Namburu, M., Pattipati, K., Qiao, L., Kawamoto, M., & Chigusa, S. (2003). Model-based prognostic techniques [maintenance applications]. In *AUTOTESTCON 2003. IEEE Systems Readiness Technology Conference. Proceedings*, pp. 330-340.

Luo, J., Pattipati, K. R., Qiao, L., & Chigusa, S. (2008). Model-based prognostic techniques applied to a suspension system. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. *38*(5), pp. 1156-1168.

Mahamad, A. K., Saon, S., & Hiyama, T. (2010). Predicting remaining useful life of rotating machinery based artificial neural network. *Computers & Mathematics with Applications*, vol.*60*(4), pp. 1078-1087.

Mendel, J. M. (1995). Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, vol. *83*(3), pp. 345-377.

Micea, M. V., Ungurean, L., Cârstoiu, G. N., & Groza, V. (2011). Online state-of-health assessment for battery management systems. *Instrumentation and Measurement, IEEE Transactions on*, vol. *60*(6), pp. 1997-2006.

Paulk, M. (1993). *Capability maturity model for software*. John Wiley & Sons, Inc.

Pecht, M., & Jaai, R. (2010). A prognostics and health management roadmap for information and electronics-rich systems. *Microelectronics Reliability*, vol. *50*(3), pp. 317-323.

Peng, T., Liu, Y., Saxena, A., & Goebel, K. (2015). In-situ fatigue life prognosis for composite laminates based on stiffness degradation. *Composite Structures,132*, 155-165.

Polikar, R. (2007). Bootstrap-Inspired Techniques in Computation Intelligence. *IEEE Signal Processing Magazine*, vol. *4*(24), pp. 59-72.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. In *Prognostics and health management, 2008. PHM 2008. international conference on* (pp. 1-17). IEEE, October, Denver, CO, USA.

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*, vol. *1*(1), pp. 4-23.

Saxena, A., Roychoudhury, I., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2012). Requirements Flowdown for Prognostics and Health Management.*Proceedings of the AIAA Infotech@ Aerospace*, Garden Grove,CA.

Saxena, A., Sankararaman, S., & Goebel, K. (2014). Performance evaluation for fleet-based and unit-based prognostic methods. In *Second European conference of the Prognostics and Health Management society*, July 8-10, Nantes, FR.

Schwabacher, M. (2005). A survey of data-driven prognostics. In *Proceedings of the AIAA Infotech@ Aerospace Conference*, pp. 1-5.

Si, X. S., Wang, W., Hu, C. H., & Zhou, D. H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, vol. *213*(1), pp. 1-14.

Sikorska, J. Z., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, vol. *25*(5), pp. 1803-1836.

Tobon-Mejia, D. A., Medjaher, K., Zerhouni, N., & Tripot, G. (2012). A data-driven failure prognostics method based on mixture of gaussians hidden markov models. *Reliability, IEEE Transactions on*, vol. *61*(2), pp. 491-503.

Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., & Wang, W. (2015). Prognostics and health management: A review on data driven approaches. *Mathematical Problems in Engineering*, *2015*.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A. & Wu, B. (2006). Intelligent Fault Diagnosis and Prognosis for Engineering Systems, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470117842.

Vichare, N. M., & Pecht, M. G. (2006). Prognostics and health management of electronics. *Components and Packaging Technologies, IEEE Transactions on*, vol. *29*(1), pp. 222-229.

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, vol. *28*(6), pp. 815-829.

Xian, W., Long, B., Li, M., & Wang, H. (2014). Prognostics of lithium-ion batteries based on the verhulst model, particle swarm optimization and particle filter. *Instrumentation and Measurement, IEEE Transactions on*, vol. *63*(1), pp. 2-17.

Yu, H., & Wilamowski, B. M. (2011). Levenberg–marquardt training. *Industrial electronics handbook*, vol. *5*(12), pp. 1-16.

Yu, S. Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, vol. *174*(2), pp. 215-243.

Zeng, Z., Di Maio, F, Zio, E., & Kang K. (2016). A hierarchical decision making framework for the assessment of the prediction capability of prognostic methods. IEEE Transactions on Reliability. (Submitted).

Zhang, Z., Si, X., Hu, C., & Kong, X. (2015). Degradation modeling–based remaining useful life estimation: A review on approaches for systems with heterogeneity. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 229 (4SI), pp. 343-355, 1748006X15579322.

Zio, E. (2009). Reliability engineering: Old problems and new challenges. *Reliability Engineering & System Safety*, vol. *94*(2), 125-141.

Zio, E., & Compare, M. (2013). Evaluating maintenance policies by quantitative modeling and analysis. *Reliability Engineering & System Safety*, vol. *109*, pp. 53-65.

Zio, E., Di Maio, F. & Stasi, M (2010). A Data-driven Approach for Predicting Failure Scenarios in Nuclear Systems. *Annals of Nuclear Energy*, vol. *37*, pp. 482–491.

APPENDIX

i index for the identification of the unit under test (e.g., the equipment).

N total number of units under test.

t index for the time instant.

T failure time of the unit. Note that each unit has a different T_i value.

EOP End-Of-Prediction: time at which the unit is expected to fail, as predicted by the prognostic model.

$RUL_i^*(t)$ Estimated Remaining Useful Life (RUL) for the unit i, at time index t.

$RUL_i(t)$ Real RUL value for component *i* at time index *t*

$M_i(t)$ PPI calculated at time *t* for the *i*-th unit.

f number of features (i.e., signals) available for the prognostic model.

f' subset of f indicating a reduced number of features.

$\Delta input_i$ variation of the inputs of the *i*-th unit.

ACCURACY PPIs	
A.1 Timeliness Weighted Error Bias (TWEB)	Exploits an average out of the *N* units of a penalized weighted prediction error over the entire lifetime T_i. The penalizing function considered is an exponential function $\varphi(z)$ that penalizes late predictions $(z \geq 0)$ with respect to early predictions $(z < 0)$ .The weighting function $w_i(t)$ is a Gaussian Kernel Function with a mean value set to the lifetime of the unit, T_i, and a standard deviation set to 50% of that lifetime: this places an emphasis on the errors made at the end of lifetime. The optimal value for the TWEB is 1, indicating that the average penalized weighted prediction value is centered on the true RUL. Values smaller than 1 indicate that the predictions dispersion is above, or under, the true RUL.
$TWEB = 1 - \frac{1}{N}\sum_{i=1}^{N} \varphi\left(\sum_{t=1}^{T_i} w_i(t)\frac{RUL_i^*(t) - RUL_i(t)}{T_i}\right)$ $\varphi(z) = \begin{cases} \exp\left(\frac{\lvert z\rvert}{a_1}\right) - 1, & for\ z < 0 \\ \exp\left(\frac{\lvert z\rvert}{a_2}\right) - 1, & for\ z \geq 0 \end{cases}$ $a_1 > a_2 > 0$	
A.2 Sample Mean Error (SME)	Concerns the average among all *N* samples of the mean error during the respective lifetime T_i. The optimal value of the SME is 1, indicating that the average error for *N* unit samples over their whole lifetime T_i is 0, thus it is centered on the true RUL. Low values of SME indicate a greater overall discrepancy between the true RUL and the estimated one.
$SME = 1 - \left\lvert \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t) - RUL_i(t)\right)\right)\right\rvert$	
A.3 Mean Absolute Percentage Error	Exploits the average absolute percentage error of all *N* units throughout their lifetime T_i. In other words, the absolute percentage error takes into account the importance of having more accurate estimates of RUL when the unit approach its failure time.t. The optimum value for MAPE is 1, indicating that the average absolute percentage error for all *N* units during their lifetime T_i is small. A low value tells the user that a discrepancy between the estimated RUL and the true one occurs.
$MAPE = 1 - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left\lvert\frac{RUL_i^*(t) - RUL_i(t)}{RUL_i(t)}\right\rvert\right)$	
A.4 Mean Square Error (MSE)	Takes into account the average for all *N* units of the average quadratic error of the RULs estimated during the lifetime T_i. An optimum value for the MSE is 1, indicating that the estimated RULs are equal to the real ones for all units i.. A low value indicates that, during the lifetime of the *N* components, the errors in the RUL estimates are high.
$MSE = 1 - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t) - RUL_i(t)\right)^2\right)$	
A.5 Sample Median Error (SMeE)	Exploits the absolute value of the median of all mean errors, for all *N* units, over their lifetime T_i. The median is chosen as an indicator, as it can sometimes be more robust than the mean, and more representative of the data if the data is not distributed symmetrically around the mean. An optimum value for SMeE is 1, indicating that the modulus of the median error is zero. A low SMeE indicates that most RUL estimates are wrong.
$SMeE = 1 - \left\lvert \underset{i=1,\ldots,N}{\text{Median}}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t) - RUL_i(t)\right)\right)\right\rvert$	

<table>
<tr><th colspan="2">PRECISION PPIs</th></tr>
<tr><td>P.1 α-λ Performance (P_λ^α)</td><td rowspan="2">Measures the average fraction of points, during the lifetime T_i over all N units, for which the prediction of the RUL estimated at a specific time t before failure is, with α confidence, the true RUL at $t + \lambda(EOP_i - t)$. The points which have an accuracy of prediction of α within a relative time distance λ from the current time t, are considered positive. The optimum value for P_λ^α is 1, indicating that all estimated RULs have still an accuracy at least of α at a relative distance λ from the current prediction time t. Low values indicate that the prediction made at time t is not reliable in the future time window defined by λ. α is the confidence modifier and λ the time window modifier.</td></tr>
<tr><td>$$P_\lambda^\alpha = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i} b(t)\right)$$
$$b(t) = \begin{cases} 1 & if\ (1-\lambda)RUL_i^*(t) \in (1 \pm \alpha)RUL_i\big(t + \lambda(EOP_t - t)\big) \\ 0 & otherwise \end{cases}$$</td></tr>
<tr><td>P.2 Weighted Prediction Spread (WPS)</td><td rowspan="2">Considers the standard deviation of the weighted prediction error during the entire lifetime T_i for all N units . The optimum value for WPS is 1, indicating that all units either share a similar average weighted prediction error or that it is small. A low value of WPS indicates a high dispersion, and, thus, a low precision. The weighting function is the same as A.1</td></tr>
<tr><td>$$WPS = 1 - \sigma_{1,\dots,N}\left(\sum_{t=1}^{T_i} w_i(t) \cdot \frac{\big(RUL_i^*(t) - RUL_i(t)\big)}{T_i}\right)$$</td></tr>
<tr><td>P.3 Sample Standard Deviation (SSD)</td><td rowspan="2">Considers the standard deviation of the average error over the lifetime T_i for all N units. The optimum SSD value is 1, indicating that all errors for all units are closely similar. A low value of SSD indicates that the dispersion of the errors within the N units is high.</td></tr>
<tr><td>$$SSD = 1 - \sqrt{\frac{\sum_{i=1}^{N}(ME_i - SME)^2}{N-1}}$$
$$ME_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\big(RUL_i^*(t) - RUL_i(t)\big)$$</td></tr>
<tr><td>P.4 Root Mean Square Error (RMSE)</td><td rowspan="2">Considers the average out of the N units of the Root Mean Squared Error during the entire lifetime T_i. The optimum value of RMSE is 1, indicating that the error between the estimated RUL and the true RUL is consistent in the model. A low value indicates that the discrepancy between the estimated and the true RUL is inherently stochastic.</td></tr>
<tr><td>$$RMSE = 1 - \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}\Big(RUL_i^*(t) - RUL_i(t)\Big)^2}$$</td></tr>
<tr><td>P.5 Prediction Spread (PS)</td><td rowspan="2">Considers the standard deviation of the Indicator M for all N units. The PS measures how the indicator M varies through all N units. The optimum value of PS is 1, indicating that the standard deviation of the indicator is 0: thus, the indicator is concentrated on one value, reducing the variability of the performance throughout the units. A high value indicates that the indicators behavior varies between units</td></tr>
<tr><td>$$PS = 1 - \sigma_{1,\dots,N}(M_i(T))$$</td></tr>
</table>

<table>
<tr><th colspan="2">STABILITY PPIs</th></tr>
<tr><td>S.1 Sensitivity (S)</td><td rowspan="2">Considers the average over the J inputs of the variation of a generic Mindicator, with respect to the variation of input values in the model, i.e. how sensitive the model is to the input values. The optimum value of S is 1, indicating that, despite a variation in input features, the indicator Mremains unaltered. A high value in S indicates that the variation of the model is significant with respect to the inputs, thus the model is very sensitive to the input data.</td></tr>
<tr><td>$$S = 1 - \frac{1}{J}\sum_{j=1}^{J}\left|\frac{\Delta M}{\Delta input(j)}\right|$$</td></tr>
<tr><td>S.2 Convergence (C_M)</td><td rowspan="2">Measures the ability of the indicator M of improving during time to reach its perfect score of 1. The distance between the origin and the center of mass (x_c, y_c) of the area under the curve of 1 minus the M indicator quantifies its convergence. The optimum value for $C_M$tends to 1 as the value of the indicator tends to 1 and the value of the curve tends to $1 - M = 0$, which is the perfect score. Low values indicate a slow rate of improvement, and thus a slow tendency to reach a perfect value for the indicator M</td></tr>
<tr><td>$$C_M = 1 - \sqrt{\left(x_c - t_p\right)^2 + y_c^2}$$
$$x_c = \frac{\frac{1}{2}\sum_{t=t_p}^{T}((t+1)^2 - t^2)\left(1 - M(t)\right)}{\sum_{t=t_p}^{T}\left((t+1) - t\right)\left(1 - M(t)\right)}$$
$$y_c = \frac{\frac{1}{2}\sum_{t=t_p}^{T}\left((t+1) - t\right)\left(1 - M(t)\right)^2}{\sum_{t=t_p}^{T}\left((t+1) - t\right)\left(1 - M(t)\right)}$$</td></tr>
<tr><th colspan="2">SPILL-OVER PPI</th></tr>
<tr><td>SO.1 Robustness to a Reduced Feature Set ($RRFS_M$)</td><td rowspan="2">Assesses the spill-over effects on a generic PPI M (i.e., A.1, A.2, etc.) due to removing a set of features from a model originally evaluated with f to f' features. A low value of RRFS indicates a PPI with a high sensitivity to the number of features involved and, thus, a non-robust PPI. A high value of RRFS indicates a PPI which tolerates the absence of a set of features and produces a good result anyways. An optimum value for RRFS is 1.</td></tr>
<tr><td>$$RRFS_M(f') = 1 - \left|\frac{M(f) - M(f')}{\max\left(M(f), M(f')\right)}\right|$$</td></tr>
</table>