machines

MDPI

Article

Data Reduction in Proportional Hazards Models Applied to Reliability Prediction of Centrifugal Pumps

Marc Vila Forteza [1,2,*], Diego Galar [1,*], Uday Kumar [1] and Kai Goebel [1,3]

[1] Division of Operation and Maintenance Engineering, Luleå University of Technology, 97187 Luleå, Sweden; uday.kumar@ltu.se (U.K.); kai.goebel@fragumglobal.com (K.G.)
[2] Petronor, 48550 Muskiz, Spain
[3] Fragum Global, LLC, Mountain View, CA 90012, USA
* Correspondence: marc.vila.forteza@associated.ltu.se (M.V.F.); diego.galar@ltu.se (D.G.)

Abstract: This paper presents the use of proportional hazards regression models for predicting the Mean Time Between Failures (MTBF) of centrifugal pumps in the oil and gas industry. To that end, a dataset collected over 8 years including both design and operational variables from 675 pumps in an oil refinery was used to fit statistical models. Parametric and non-parametric transformations and restricted cubic splines were used to fit the covariates, thereby relaxing linearity assumptions and potentiating predictors with strong nonlinear effects on the outcome. Standard Principal Component Analysis (PCA) and sparse robust PCA methods were used for data reduction to simplify the fitted models and minimize overfitting. Models fitted with sparse robust PCA on non-parametrically transformed variables using an additive variance stabilizing (AVAS) method are suggested for further investigation. The complexity of the fitted models was reduced by 85% while at the same time providing for a more robust model as indicated by an improvement of the calibration slope from 0.830 to 0.936 with an essentially stable Akaike information criterion (AIC) (0.34% increase).

Keywords: centrifugal pumps; MTBF; API standard; reliability prediction; proportional hazards model; data reduction

check for updates

Academic Editors: Phuc Do, Cristiano Cavalcante and Elias Strangas

Received: 20 January 2025
Revised: 14 February 2025
Accepted: 26 February 2025
Published: 7 March 2025

Citation: Vila Forteza, M.; Galar, D.; Kumar, U.; Goebel, K. Data Reduction in Proportional Hazards Models Applied to Reliability Prediction of Centrifugal Pumps. *Machines* **2025**, *13*, 215. https://doi.org/10.3390/machines13030215

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Centrifugal pumps are the most commonly used rotating equipment in oil refineries. They handle hazardous fluids that are flammable and may affect the environment and human health. These pumps are robust, durable, and capable of operating reliably under extreme conditions without specific pressure or temperature limits [1] (p.8). Considering their essential role, refineries implement strict maintenance schedules to prevent breakdowns, which could lead to disruption in the production process or even industrial accidents.

Failures and unwanted incidents are being reduced with the development of new field sensors and prediction algorithms in the oil and gas industry [2], boosting the use of complex fault diagnosis and remaining useful life (RUL) models. These models help to improve the safety and reliability of industrial plants by using asset design features, online operational data, and vibration data obtained from the distributed control system (DCS) of the plant [3]. For a list of acronyms used in this paper, refer to Appendix A.

Several studies have applied various modeling and prediction techniques to detect component failures and unwanted operating conditions and to determine the RUL of centrifugal pumps. Physical-based models combined with Support Vector Machines (SVM) have been used to detect unbalance, bearing issues, and misalignment using vibration

data and process signals [4]. Similarly, Artificial Neural Networks (ANN) were applied by Adraoui et al. [5] to predict pump degradation using flow data, while Cubillo et al. [6] analyzed roller bearing degradation with empirical design parameters and temperature data using Physical-based models (PbM). Advanced statistical techniques, such as Eigenvector analysis and Principal Component Analysis (PCA), have been used by Zhang et al. [7] to assess mechanical conditions through vibration and flow data. Machine learning methods like Support Vector Machines (SVM) [8] and Neural Networks with fuzzy modeling [9] have been extensively employed to detect bearing failures using vibration data. For seal degradation, a Variance Gamma Process (VGP) has been proposed by Fouladirad et al. [10], using leakage rate data. In the context of deep learning and machine learning, Convolutional Neural Networks (CNN) [11,12], Deep Neural Networks (DNN) [13], Recurrent Neural Networks (RNN) [14], and Bayesian approaches such as k-Nearest Neighbours (k-NN) combined with Bayesian filtering [15] have been applied for bearing failure detection, degradation, and other failures using vibration data, acoustic images, and pump operating parameters. Additionally, probabilistic methods like Continuous Time Bayesian Networks (CTBN) and Fault Tree Analysis (FTA) have been used by Forrester et al. [16] to address bent shafts, rubbing, and sealing issues for failure rate prediction. Machine learning techniques such as the Gaussian Mixed Model (GMM) and Particle Filter (PF) method have been proposed by Wang et al. [17] for adaptive prognosis under varying operating conditions, including vibration data and pump discharge pressure. The Mixture Weibull Proportional Hazard Model (MWPHM) has been employed by Zhang et al. [18] to predict thrust bearing and sealing ring wear based on vibration data and discharge pressure. A Relevance Vector Machine (RVM) approach has also been applied for oil-sand pump prognostics, particularly in detecting impeller damage using vibration data [19]. Finally, a combined PCA and GMM approach has been used by Cao et al. [20] to detect impeller blade cracks with vibration data and expert knowledge. As a reference, in Figure 1, the main components of an overhung centrifugal pump are illustrated.

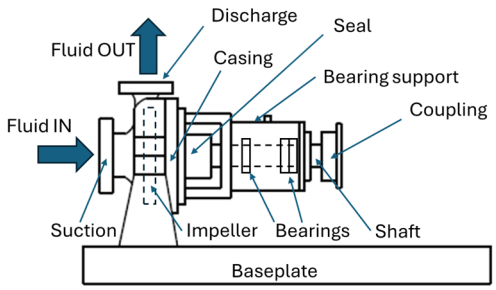


Figure 1. Main component parts of an overhung centrifugal pump.

Irrespective of whether the desired information is the residual life [21,22] of the system or the lifetime of its components, these methods require in situ sensor data to allow for the observation of relevant symptoms of root causes of failures and to estimate RUL. In many cases, when such data are not available, lifetime estimation can be accomplished with a customized estimation of Mean Time Between Failures (MTBF), a widely used key performance indicator (KPI) to assess reliability in the industry.

Having an accurate MTBF estimation is beneficial both when an asset is commissioned and when it is returned to service after repairs. This may help in calculating the expected operating costs, planning maintenance activities, setting KPIs in line with the actual fleet, and improving the design of specific assets with short lifetimes and high economic or safety impact [3]. However, MTBF targets are often set without considering the specific operating conditions and design characteristics of the pumps. In practice, these factors significantly influence the achievable time between failures. Consequently, the accuracy of the actual

MTBF is generally not well known. Operators try to improve it by using empirical MTBF data from the OREDA database or from specialized books [23] and journals, but still, it is not straightforward to extrapolate the findings to specific industrial plants. In this paper, we suggest that it may be possible to create more accurate MTBF targets by using predictive models leveraging field data.

Several methodologies have been proposed in the past to predict the MTBF of centrifugal pumps. For example, Bevilacqua et al. [24] suggested the use of regression trees to generate rules using historical data and operating conditions from an oil refinery dataset. Braglia et al. [25] proposed a stepwise multivariate data classification technique to classify assets' MTBF and identify the parameters that explain most of the reliability. Braglia et al. [26] later presented a framework based on an ensemble learning model to classify assets by their MTBF, using the same dataset as before [25]. They developed these approaches from the perspective of a classification problem by clustering the predicted MTBF into different groups based on the MTBF values. Bevilacqua et al. [27] suggested a multi-layer perceptron-based artificial neural network (MLP-ANN) to evaluate the expected failure rates of centrifugal pumps. Orrù et al. [28] analyzed MLP and SVM techniques for early fault prediction of a centrifugal pump. In more recent work, Sudadiyo [29] proposed the use of a non-homogeneous Poisson process (NHPP) model to obtain an estimate of the MTBF of a centrifugal pump of a research nuclear reactor.

We suggest here to revisit the Cox Proportional Hazard Model (PHM) to model and predict MTBF. In 1972, David Cox introduced the PHM to describe how various factors influence mortality or failure rate at a specific time. This statistical approach was actually first used in biomedicine to study the survival of cancer patients, but it became popular in other disciplines, such as industrial reliability [30]. Note that the acronym PHM is overloaded in the reliability community, as it means both Proportional Hazards Model and Prognostics and Health Management. We will use PHM solely in the context of the Proportional Hazards Model for the remainder of this paper.

Since Jardine et al. [31] proposed a Weibull PHM to model aircraft engine and marine gas turbine failure data together with condition monitoring data, the use of PHM has been further investigated for industrial applications. Sharma et al. [32] reviewed the literature on the use of traditional imperfect maintenance (TIM) models and PHM in maintenance from 1965 to 2020. Chaoqun et al. [30] recently published an additional review focused on the use of the PHM in prognostics applied to industrial systems.

The PHM can employ data on the asset condition and failure and maintenance history [33] and has the characteristics of universality, flexibility, and simplicity. It can also deal effectively with censored data. It turns out that these attributes match our available dataset, which includes an 8-year maintenance and vibration history, mechanical and hydraulic features, and design process conditions of 675 centrifugal pumps in an oil refinery.

To optimize the accuracy of PHMs, it is important to reduce the number of covariates or simplify them [34]. In this paper, we apply the modeling process of the Cox PHM to our dataset, with a special emphasis on data reduction.

This paper is organised as follows. Section 2 briefly presents the formulation of the PHM and explains the reliability and maintenance concepts and how the outcome variable (MTBF) is calculated. It includes a description of the data reduction methods used in this research. Section 3 details the methodology proposed to reduce the dimension of Cox PHMs applied to machinery reliability. The methodology is applied to the above-mentioned dataset consisting of 675 centrifugal pumps; the results are detailed in Section 4, followed by a discussion of the important issues of the research in Section 5. The conclusions and future research proposals are included in the final sections.

2. Brief Overview of MTBF, PHM, and Data Reduction

2.1. Overview of MTBF Concept

Several metrics are available to help industries measure and monitor uptime, downtime, and the efficiency with which they address reliability and maintenance issues. Some of the most used indices of this type are MTBF, MTTR (Mean Time To Repair), and MTTF (Mean Time To Failure). MTBF is a KPI frequently used in the oil and gas industry and is referenced in ISO 14224:2016 standard [35]. It represents the average time an asset is expected to operate between two consecutive failures and falls under the durability of the physical assets category. Conceptually, MTBF applies to assets that can be repaired; hence, the use of 'time between failures'. It can be expressed in time units like hours, days, or months, depending on the lifetime of the asset. It can be calculated for an individual asset or a population using Equations (1) and (2), respectively.

$$\mathrm{MTBF_{asset}} = \frac{\text{Operating time recorded}}{\text{Number of Failures}} \tag{1}$$

$$\mathrm{MTBF_{population}} = \frac{\text{Operating time recorded} * \text{Number of assets}}{\text{Number of Failures}} \tag{2}$$

The MTBF of an asset is directly linked to its reliability R(t), which is the probability that a component, equipment, or system will operate as required under specific conditions for a defined time period (t). For an asset with a constant hazard rate (λ), the reliability distribution is exponential and is represented by Equation (3), the inverse of MTBF, as shown in Equation (4).

$$\mathrm{MTBF_{asset}} = \frac{1}{\lambda} \tag{3}$$

$$\mathrm{R(t)} = \mathrm{e}^{-\mathrm{MTBF_{asset}} * \mathrm{t}} \tag{4}$$

These relationships can be used when (1) the probability of failure per unit of time remains the same, regardless of the age or usage of the component or system, and (2) the time between failures is exponentially distributed, which implies that each failure is independent of the last and λ remains unchanged.

However, in many real-world situations, the hazard rate λ is not constant and varies with time, leading to a time-dependent hazard rate (λ(t)). For example, λ(t) may increase as the asset ages or decrease if early failures are resolved (representing infant mortality). In these cases, the time between failures does not follow an exponential distribution. When the hazard rate is time-dependent, the concept of MTBF is no longer the simple inverse of the hazard rate because λ changes over time. In addition, the reliability function R(t) must be integrated using λ(t):

$$\mathrm{R(t)} = \mathrm{e}^{-\int_0^{\mathrm{t}} \lambda(\mathrm{u})\mathrm{du}} \tag{5}$$

When λ(t) changes over time, the Weibull distribution is often used to model the failure time because it is flexible enough to approximate different patterns of hazard rate behavior by adjusting parameters to model how λ(t) changes. The Weibull distribution includes parameters that allow for it to represent an increasing, decreasing, or constant hazard rate, depending on the shape and scale parameters. These parameters are (1) the shape parameter β and (2) the scale parameter η. The behavior of the hazard rate over time is determined by β, how quickly failures are likely to occur is modeled by η, and their relationship with the hazard rate is defined by Equation (6).

$$\lambda(\mathrm{t}) = \frac{\beta}{\eta}\left(\frac{\mathrm{t}}{\eta}\right)^{\beta-1} \tag{6}$$

2.2. Brief Overview of Cox PHM

The Cox Proportional Hazards Model (PHM) is a regression model that relates event timing to covariates through the hazard rate h(t | X).

$$h(t|X) = h_0(t) \cdot e^{\beta X} \quad (7)$$

Here, $h_0(t)$ is the baseline hazard function and β represents covariate effect coefficients. While $h_0(t)$ is typically unspecified, it can follow various distributions like Weibull, log-normal distributions, generalized gamma distributions, or spline hazard models [36] (pp. 423–424). Though determining appropriate parametric models is challenging [37], the Cox PHM's robustness provides reliable results even with suboptimal model choices [38] (p. 110).

The model assumes multiplicative covariate effects on $h_0(t)$, maintaining constant hazard ratios between items with different covariate values. However, time-dependent covariates require specialized approaches [39,40]. The model features log-linear covariate effects, meaning that the hazard logarithm is a linear function of covariates.

Key assumptions include constant covariate impact over time, independent survival times, and the accommodation of non-informative right-censoring (when subjects do not experience the event by the study's end).

2.3. Data Reduction and Variable Selection in PHM

Statistical models excel at prognostic predictions when properly developed but can perform poorly when assumptions are violated or when models are used inappropriately [34]. One of the problems that can arise is overfitting, caused by using too many variables, which can be mitigated through variable selection and data reduction techniques.

Variable selection helps when there is insufficient domain knowledge to pre-specify model variables [36] (p. 67). Common techniques applied to PHM [41–46] include all subset selection, stepwise selection, ridge regression, lasso, least angle regression, and tree-based methods. These approaches range from exhaustive searches using AIC criteria to regularization methods like ridge and lasso regression, which handle multicollinearity and improve prediction accuracy.

Data reduction transforms datasets through various methods, including literature reviews, removing narrow-distribution variables, eliminating frequently missing predictors, and statistical techniques like principal component regression [36]. Principal Component Analysis (PCA), widely used in machinery reliability [7,15,47–49], transforms variables into uncorrelated principal components that maximize variance. Standard PCA creates linear combinations of original variables, while its variants [50–53] address issues like non-linearity and sparsity. Sparse PCA [52], used in this paper, improves interpretability by introducing sparsity into principal components through penalty functions, allowing only subset variables to contribute to each component.

3. Methodology and Analytical Approach

This section outlines the proposed method and its associated analytical approach. As seen in the flow chart in Figure 2, the steps comprise asset analysis, data collection, data cleaning and filtering, statistical analysis, fitting Cox PHMs, calibration, validation, parametric reduction, PCA, and component selection. For the List of RStudio version 2024.12.1 packages used for used in this paper refer to Appendix B

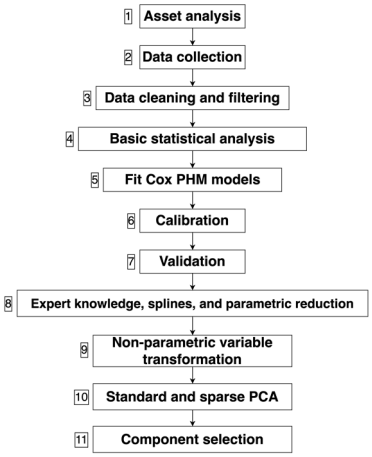


Figure 2. Steps of the proposed methodology for data reduction.

These steps are described in more detail in the following section.

Process Steps

1. Asset analysis

The first step is to thoroughly analyze the type of asset to be studied and to correctly select the variables that may have an influence on its useful life. Other than the asset type, it is also important to select a group of assets with equivalent components and comparable MTBR. For example, reciprocating compressors and centrifugal pumps or steam turbines with centrifugal compressors cannot easily be grouped together for analysis, as they have different operating principles and different components.

2. Data collection

The required data for fitting the model may be gathered from technical manuals and asset data sheets and from maintenance and reliability data history. This can be performed either automatically or manually but may require intensive manual work to ascertain that the information collected is correct. With respect to sample size, we chose the number of predictors $p > m/15$ [36] (p. 73) to minimize overfitting, where m is the number of failures registered for the time period analyzed.

Generally, the MTBF of each item of the fleet of assets to be analyzed is found in the records of the reliability and maintenance KPIs. This variable is the output of the future model, that is, the variable to be predicted.

Relevant information for modeling the MTBF of mechanical assets could be operating conditions such as pressures, temperatures, fluid characteristics, vibrations, etc. Also, hydraulic design features (e.g., suction-specific speed, number of impeller vanes, etc.), mechanical design characteristics, age, and maintenance historian variables can also be considered.

3. Data cleaning and filtering

The third step involves the removal of incomplete observations. Similarly, entries with infrequent or extreme working conditions were removed. Note that the units of measurement for the different variables must be unified so that they are consistent between predictors, and observations can be compared.

Although step four (described next) could help in this task, it is strongly recommended that the decision to delete items be made by a specialist with specific knowledge of the assets to be modeled, in particular, if very large datasets are not encountered, which was the case here.

4. Basic statistical analysis

When working with both categorical and quantitative variables, the preliminary statistical analyses will be different for each case. For categorical variables, the number of possible categories, their frequency, and their proportion will be calculated and plotted to better interpret their distribution in the dataset. Descriptive statistics should be obtained for continuous variables using computing means, quartiles, and standard deviations.

Several different statistical indexes can be used to quantify discrimination ability (e.g., R^2, model χ^2, Somers' D_{xy}, Spearman's ρ, the area under ROC curve) [36] (p. 92). For our purposes, we will use Spearman's ρ and Somers' D_{xy} rank indices corrected by the d.f. The first can measure how well each predictor and the MTBF can be associated using a monotonic function. The second can be used to rank the correlation between predictors and the outcome by measuring the difference between concordance and discordance probabilities. When $D_{xy} = 0$, a model makes random predictions, but if $D_{xy} = 1$, the predictions are perfectly discriminating. More information about the importance of concordant and discordant pairs in Cox PHM can be found in the literature [36,54].

The next step is a redundancy analysis of the variables. The main goal is to provide insights about variables that do not add new information. The findings will be useful in the formal data reduction stage to remove unnecessary predictors. The algorithm transforms each continuous predictor into an RCS, and each categorical predictor is transformed into a dummy ordinal variable. Predictors are then predicted one by one from all remaining predictors using flexible parametric additive models. Then the predictor that can be predicted from the remaining set with the highest R^2 is removed. Later, all remaining predictors are predicted from their complement, and so on, until no variable remaining in the list of predictors can be predicted with an R^2 greater than a certain threshold [36] (p. 80).

5. Fit Cox PHMs

Before starting to fit the models, the survival curves of the assets stratified by the categorical variables must be compared using a log-rank test. It should include (1) the chi-square statistic under the assumption that there is no difference in survival between the groups, (2) the degrees of freedom (d.f.) of each variable, and (3) the *p*-value derived from the chi-square statistic to test the null hypothesis. This latter test helps to compare the overall survival times between different groups; groups can later be stratified in the models if necessary. After the log-rank test, it is time to fit the Cox PHMs.

The first approach is to fit the models using all the predictors available in the dataset, as the full models containing all the variables normally predict the most accurately on new data. Although confidence limits and statistical tests have the desired properties, these models are not very parsimonious [36] (p. 119). A better approach is to include only those predictors that meaningfully contribute to the explanatory power, thus leading to more robust, interpretable, and generalizable models. Two common statistical methods that help to assess model parsimony are AIC [55] and Bayesian information criterion (BIC) [56]. Both penalize the addition of unnecessary predictors, helping to identify models that balance fit and simplicity. As discussed by Vrieze [57], AIC and BIC have different advantages in model selection depending on the context. We suggest using AIC to assess the model complexity. AIC is more focused on predictive accuracy, while BIC aims to identify the true and simpler model more reliably as the sample size grows—this should not be an issue in our case.

Cox PHM regression assumes the log-hazard of the outcome has a linear relationship with the baseline hazard rate of continuous predictors. However, relationships between variables are not always linear, except in certain cases [36] (p. 18). Thus, it is important to

begin by fitting two Cox models: one assuming a linear relationship between predictors and the outcome, and the other using a more flexible approach to capture non-linear effects.

The second approach is to use RCS [58] to approximate the behavior of continuous predictors in regression models. The method is widely used in epidemiology research, as it allows for flexible modeling of the relationship between the predictors and outcome, making the method suitable for capturing non-linear relationships and complex data patterns. It also ensures that the resulting function is smooth, which is important for maintaining interpretability and stability in predictions. In addition, it helps control the complexity of the model by choosing the number of knots (points where the spline changes direction) and providing visual representations of relationships between predictors and outcomes [36].

6. Calibration

The fitted model is calibrated in the sixth step. This step assesses how well the model's predicted survival probabilities agree with the observed survival probabilities over time. A well-calibrated model provides reliable estimates of survival probability, reducing the risk of overfitting the data and making the model more generalizable.

Harrell [36] advocates using the flexible adaptive hazard regression approach proposed by Kooperberg et al. [59] to estimate the calibration curves for survival models. This approach does not rely on assumptions of linearity or proportional hazards. By applying hazard regression, the relationship between predicted survival probabilities and observed outcomes can be assessed, resulting in a calibration curve. Bootstrap resampling can be used to reduce bias in the estimates and adjust for overfitting, thereby providing a more accurate projection of the calibration performance of the model in future data [36] (p. 506).

7. Validation

The validation process for a regression model is critical to ensure that the model is both accurate and generalizable to the unseen data. The main causes of failure to validate a model are overfitting, changes in measurement methods/definition of categorical variables, and major changes in subject inclusion criteria [36] (p. 109).

Validation can be external or internal. While external validation evaluates the performance of the model using an independent dataset, internal validation uses the dataset from which it was developed. External validation can be performed either by testing the model on a completely new dataset or by applying it in different contexts; both are common approaches. Key methods to obtain datasets for internal validation are train-test subsets, cross-validation, and bootstrapping. The first divides the dataset into training and testing subsets, the second splits the data into multiple folds, and the third resamples the data with replacements.

Because of the difficulty of obtaining datasets of different machine fleets and using all the information of the complete dataset to fit the models, we suggest the use of bootstrap internal validation for MTBF modeling. Bootstrapping can validate the procedure used to fit the original model and provide accuracy indices adjusted for the likelihood of overfitting. The model trained on a bootstrap sample is evaluated against the original dataset. Since the original dataset contains observations the model may not have seen (due to resampling), this evaluation measures how well the model generalizes to unseen data. The difference in performance between super-overfitting, where the model is evaluated on the same bootstrap sample used for training, and regular overfitting (evaluating the bootstrap-trained model on the original dataset) reflects the extent of overfitting. A significant performance drop compared to the super-overfitting scenario indicates potential overfitting. A small difference suggests minimal overfitting, as the model performs similarly on both datasets. This difference is called optimism [36] (p. 114) and the average optimism across

all bootstrap samples can be calculated and subtracted from the performance indices of the original model, resulting in bias-corrected estimations.

8. Expert knowledge, RCS, and parametric reduction

The data reduction begins in this step and has two main cornerstones: the first is the use of expert knowledge and the second is the minimization of d.f. to model the behavior of the predictors versus the outcome.

The number of covariates may be reduced by relying on the expert knowledge of machinery reliability specialists. For example, the number of variables with fluid characteristics can be limited to those that are more significant to the type of asset under study. New variables can also be created as a combination of the original ones to reduce d.f. The collinearities and redundant variables detected in step 4 are verified by expert knowledge and the available literature, excluding those that do not offer valuable information for the predicting models. Finally, the variables to reconsider depend on the available data and the asset being studied.

Data reduction can be further addressed by analyzing the effect of each predictor on survival time. As detailed by Harrell [36] (p. 467), the first step is to visually analyze the shape of the effect of each predictor on survival time, and the second is to compute the relative LR (log-likelihood ratio) χ^2 statistics, penalized for d.f. The first step detects non-linearities and the possible parameterizations of the variables versus the outcome depending on the shape of the curve. This process can lead to changes in some approximations made with RCS to simple parametric equations, thus reducing the d.f. used for the variable to one. In the second step, the number of knots used for each RCS approximation is adjusted depending on the importance of the variables. In general, more d.f. can be spent on the important variables, i.e., those contributing a higher relative LR χ^2 in the complete model.

9. Variable transformation

In multivariate regression, transforming variables on one or both sides of the equation (Y and X) can be useful for a number of reasons: to achieve linearity using parametric or polynomial transformations; to reduce skewness, making the data more normally distributed and improving model performance; to stabilize the variance of the outcome or predictor variables; to make coefficients easier to interpret; to reveal hidden interaction effects that are otherwise missed with raw data, enhancing model complexity; and to improve the fit, minimizing residuals and improving model performance metrics.

The literature proposes different techniques of variable transformation for both continuous and categorical variables. The Maximum Total Variance (MTV) method proposed by Young, Takane, and de Leeuw [60] seeks to maximize the total variance explained by a set of transformed variables. By maximizing variance, the MTV method ensures that the reduced dimensions or components retain as much information from the original data as possible. This technique is often implemented using Alternating Least Squares (ALS) for efficient computation. The Maximum Generalized Variance (MGV) method proposed by Sarle [61,62] aims to maximize the determinant of the covariance matrix of optimally scaled variables. This method is also implemented using ALS to optimize the transformations iteratively. Kunfeld describes an approach using MTV and MGV to transform predictors, supporting options like monotonic splines and standard cubic splines in ref. [63].

A more flexible transforming method is Alternating Conditional Expectation (ACE), developed by Breiman and Friedman [64]. This non-parametric, iterative method is designed to find optimal transformations for predictors and the response variable in regression analysis. Lastly, the Additive Variance Stabilizing (AVAS) method developed by Tibshirani [65] is designed to stabilize variance in response variables by adding a constant. This approach

ensures that g(Y), the transformed version of Y, is monotonic, and the fitting criterion focuses on maximizing R^2 while ensuring that the transformation of Y leads to residuals with nearly constant variance [36] (p. 391).

We apply different transformations only to the raw X variables and fit the models with them, conducting calibration and validation procedures again (steps 6 and 7). The results will provide a basis for comparing the key metrics and, if signs of overfitting are detected, the process can move to the following step.

10. Standard and sparse PCA on raw and transformed variables

The next step in the data reduction process is to apply standard PCA and sparse PCA to the raw and previously transformed sets of variables. As explained in Section 2, both are dimensionality reduction techniques that convert the original variables into a new set of uncorrelated variables, referred to as principal components, to capture the maximum variance of the data.

11. Selection of principal components

The key idea of this step is to select the minimum set of principal components that will maximize the performance of the model and the explained variance of the original dataset. This should be performed for each of the previously transformed datasets with both PCAs.

A Cox PHM should be fitted with each number of principal components from 1 to the maximum for each set. In addition, the AIC should be computed for each model to measure its performance. At this point, it should be noted that the principal components are ordered by the eigenvalues in descending order. Therefore, the fitted model for each number of components may not be the one that minimizes the AIC.

To address this issue, we suggest using the algorithm implemented by Wen et al. [66]. It systematically evaluates all possible combinations of predictors to determine the best-fitting model based on a specified criterion (e.g., AIC). It computes the best function and its AIC for each number of components (from 1 to maximum) and extracts the best combination of components using the AIC criteria. In the next step, it is possible to plot the obtained AIC versus each number of components for each PCA set to compare the results together with the full models. Note that the AIC will normally decrease sequentially as the number of components increases. The selection can be performed by comparing the AIC achieved by each model for each number of components, thus detecting the best balance between these two features; the minimum is better for both. Finally, the explained variance for each set of principal components should be evaluated to determine if the selected groups of principal components sufficiently capture the variance of the original dataset.

4. Results

The methodology outlined in Section 3 was applied to field data comprising 675 pumps at an oil refinery. The plant has a fleet of 3200 rotating machines, of which around 1200 are centrifugal process pumps, representing 38% of the total rotating machinery fleet. All of them are designed under the American Petroleum Institute standard (API std) 610 from the 5th to 11th editions according to Table 1. The dataset includes pumps of different types, including OH1, OH2, OH3, OH5, OH6, BB1, BB2, BB3, and BB5, which are installed outdoors and are not immersed in liquid. This means that while they may represent different models and sizes, they are functionally similar enough to be included in the data analysis. That said, the pumps are deployed in different parts of the plant and work under a wide range of operating conditions (see Table 2), which presents some challenges for the analysis.

Table 1. Pump classification type identification. Source: API 610 12th ed.

Pump Code	Pump Type	Orientation
OH1	Overhung, flexibly coupled	Horizontal, foot-mounted
OH2	Overhung, flexibly coupled	Horizontal, centerline-supported
OH3	Overhung, flexibly coupled	Vertical in-line, with bearing bracket
OH4	Overhung, rigidly coupled	Vertical in-line, rigid coupling
OH5	Overhung, close-coupled	Vertical in-line, close-coupled
OH6	Overhung, close-coupled	High-speed, integrally geared
BB1-A	Between-bearings, single or two stage	Axially split, foot-mounted
BB1-B	Between-bearings, single or two stage	Axially split, near-centerline-mounted
BB2	Between-bearings, single or two stage	Radially split, centerline-supported
BB3	Between-bearings, multistage	Axially split, near-centerline-supported
BB4	Between-bearings, multistage	Radially split, single-casing
BB5	Between-bearings, multistage	Radially split, double-casing
VS1	Vertically suspended	Single-casing, discharge through column
VS2	Vertically suspended	Single-casing, discharge through column
VS3	Vertically suspended	Single-casing, discharge through column
VS4	Vertically suspended	Separate discharge pipe, line shaft
VS5	Vertically suspended	Separate discharge pipe, cantilever shaft
VS6	Vertically suspended	Double-casing, radially split
VS7	Vertically suspended	Double-casing, radially split

Table 2. Distribution and number of centrifugal pumps installed in the refinery and selected for the dataset by production area.

Production Area	Refinery		Dataset	
	% Pumps	Qty of Pumps	% Pumps	Qty of Pumps
Atmospheric distillation area	41.3%	503	44.8%	303
Conversion area	26.3%	320	35.8%	241
Fuel reduction area	10.8%	132	12.2%	82
Tanks and dock	21.6%	263	7.3%	49

Twenty-nine potential predictors were selected in the first step of this study, which are listed in Table 3. They are the most commonly used data for selecting centrifugal pumps in the industry and are available in the API 610 data sheets and maintenance records. Note that the operating conditions are not acquired and processed in the model in real time. Instead, a condensed operating condition state, which is specified in the pump's data sheet, is used for each specific pump. They were considered to have an impact on the MTBF of centrifugal pumps according to the available literature [3,23,24] and based on the expert knowledge of the research team. The data were divided into six groups: operating conditions (including vibration values measured with non-intrusive piezoelectric accelerometers), hydraulics, mechanics, sealing, age, and maintenance history (see Table 3). This allows us to check for interactions between groups of variables as opposed to individual predictors, which is a more challenging task.

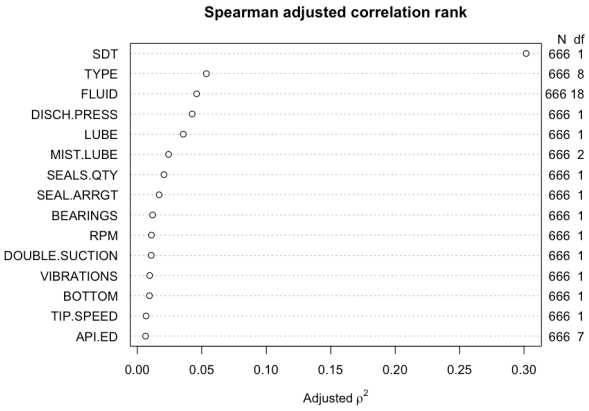
Table 3. Set of potential predictors considered in the research.

Operating Conditions	Hydraulics	Mechanical	Sealing	Age	Maintenance Historian
Fluid type	Double suction	rpm	Seal arrgt.	API 610 ed.	Lube workorders
ISO 10816-7 [67] vib zone	Tip speed	Power	Seal type		Ordinary workorders
Bottom pump	Diameter ratio	Bearing type	API 682 plan		
Flow ratio	Efficiency	Lube type	Number seals		
NPSH margin	N_{ss}				
Relative density	Ns				
Dynamic viscosity	N_{ss} ratio				
Vapor pressure					
Discharge pressure					
Fluid temperature					
Vibration level					

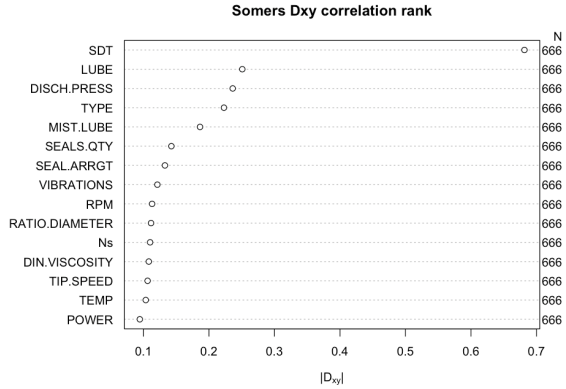
Note that advancements in impeller design and computational tools have led to a re-evaluation of the allowable N_{ss} working limit. Consequently, the achievable N_{ss} may vary depending on the year of the pump's design, which is important for evaluating the N_{ss} effect on the MTBF [68]. This was considered by including the variable N_{ss} ratio (achievable N_{ss} vs. actual N_{ss}) in the dataset.

The units of each covariate were carefully reviewed and standardized as needed to ensure consistency across all observations. Nine assets were removed from the dataset because they work in extreme conditions and fell outside the research scope. Next, data were pre-processed and cleaned to remove outliers and detect possible collinearities between the predictors.

Following step 4 outlined in Section 3, Spearman's and Somers' D_{xy} rank correlations were calculated to identify potential correlation issues between the variables. The Spearman index ρ^2 was adjusted to the d.f. of each covariate, assigning the continuous variables 1 d.f. while categorical variables use their possible categories minus 1 d.f. The rankings of the first 15 covariates with higher values are shown in Figure 3a,b. The most correlated variable in both ranks is SDT, which corresponds to the number of ordinary work orders, as indicated in Table 3.



(**a**)



(**b**)

Figure 3. Correlation ranks of the first 15 variables with higher values: (**a**) Spearman adjusted correlation rank; (**b**) Somers' D_{xy} correlation rank.

Next, a redundancy analysis was performed to check which variables were explained by the others and that could therefore be candidates for reducing the dimension of the models. Redundancies with a fixed threshold R^2 of 0.9 were identified using the method proposed by Harrell [36] (p. 80). After this calculation, the variables that could be con-

sidered redundant were the following: number of mechanical seals, power, relative fluid density, and pump efficiency.

4.1. Full Models and Necessity of Data Reduction

After data preparation and preliminary analyses, the first two Cox PHMs were fitted using all the covariates. One model was fitted considering the continuous covariates had a linear effect on the survival time. The second model was fitted using an RCS approximation with a maximum of 5 knots [36] (p. 28).

In the next step, the ability of the model to make unbiased estimates of outcome was ascertained. That is, the calibration curves using the method described in step 6 of Section 3 were used along with computing random predictions for the 1500-day survival probability, employing Efron's bootstrap method with 1000 resamples of 666 observations.

Finally, to validate the fitted models, the validation method described in step 7 of Section 3 was carried out using the resampled datasets used in the calibration. This method allows us to estimate the future performance of the model with other datasets as well as to evaluate the likelihood of significant data overfitting. If the calibration slope [36] (p. 75) falls below 0.9, there may be a lack of calibration of the model on new data [69]. The model using the RCS approximation achieves better validation results for indices such as D_{xy} and R^2; however, it exhibits a higher risk of overfitting compared to the linear model approximation, as reflected in a lower calibration slope. In any event, the slopes of both the linear and splined models were well below 0.9, indicating the need for either shrunken estimators or data reduction.

The results of the complete tests and indices of the two fitted models are summarized in Table 4.

Table 4. Summary of linear and spline models; full, domain-reduced linear and domain-reduced splines with adjusted number of knots and parametrization of covariates.

Description	Index	Full Model Linear	Full Model RCS	Domain Linear	Domain RCS	Domain Red. RCS and Param.
Model Tests	LR X^2	623.65	873.07	542.98	623.65	792.30
	d.f.	60	103	35	75	56
Discrimination Indices	$R^2_{df,666}$	0.608	0.731	0.558	0.697	0.696
	D_{xy}	−0.691	−0.768	−0.665	−0.757	−0.760
	$R^2_{df,501}$	0.675	0.785	0.637	0.762	0.770
Predictive Discrimination	Concordance	0.846	0.883	0.832	0.878	0.880
	AIC	5469.44	5299.74	5500.11	5328.58	5292.79
	AIC X^2 scale	503.65	673.35	472.98	644.51	680.30
Validation	LR R^2	0.5355	0.6377	0.5036	0.5987	0.6958
	D_{xy}	−0.6552	−0.7152	−0.6369	−0.7071	−0.7596
	Calibration Slope	0.8302	0.7215	0.8701	0.7378	0.8144
Calibration	Mean |error|	0.0768	0.1006	0.0497	0.1031	0.1110
	0.9 quantile |error|	0.1275	0.2390	0.0732	0.2190	0.2410

As the full models were prone to overfitting, it was necessary to reduce the d.f. of the covariates. Several strategies were successively employed to find a balance between the use of the potential predictors in the regression analysis and the sample size of our dataset.

The first data reduction considered the domain knowledge of the research team and the results of the redundancy analysis. This reduction applies to all future models regardless of the method used in later variable transformations. Table 5 shows which variables were removed from the list of potential predictors or that had their d.f. reduced. The first linear

and spline full models had 60 and 103 d.f., respectively, and with the first data reduction, we obtained models of 35 and 75 d.f., respectively.

Table 5. Data reduction based on expert domain knowledge and redundancies.

Predictor	Reduction	d.f. Saved		Justification
		Linear	Spline	
API edition	Change from categorical to continuous variable (manufacturing year).	6	6	Reduce d.f. by using a continuous variable instead of a categorical.
Pump type	Reduction categories and include lubrication information.	1	1	Reduce d.f. and group less frequent categories.
Fluid type	Group similar categories.	12	12	Reduce d.f. and modeling issues with less frequent categories.
Bearings	Remove covariate.	1	1	Keep consistency in high-speed pumps.
ISO 10816-7 vib. zone	Remove covariate.	3	3	Redundant with global vibration level.
Seal type	Include variable pressurized in seal type cat. Increase 1 d.f.	−1	−1	Modeling issues with pressurized covariate.
Seals quantity	Remove covariate.	1	1	Redundant information, explained by pump type covariate.
Pressurized	Remove pressurized covariate.	1	1	Add the covariate information in seal type predictor. Modeling issues with this predictor.
Relative density	Remove covariate.	1	4	It is explained by vapor pressure, viscosity, fluid and temperature.
N_{ss}	Remove and change into a different predictor (stable).	1	4	Change the predictor to stable parameter.
Stable	Add a predictor that contains more information than N_{ss}.	−1	−4	To include information lost by removing N_{ss}.
TOTAL		**25**	**28**	

Next, the non-linear effects were evaluated for each predictor and their Chi-square (χ^2) statistics, the associated d.f., and *p*-values were computed to estimate their significance. The relative likelihood ratio χ^2 for the non-linear components was 249.42, compared to 873.07 for the overall model, indicating the presence of non-linearity in the covariates. Consequently, models were fitted using RCS approximations.

The subsequent dimension reduction was carried out by reducing the number of knots used for some continuous variables and by parameterizing other covariates using simple equations. This approximation was performed by looking at the shape of the relative hazard curves versus the covariate inputs and quantifying the change in the partial effect of each covariate after each reduction. The partial effects were evaluated separately for each predictor by computing the likelihood ratio χ^2 statistic. As can be seen in Table 6, it was possible to improve the explained likelihood for most of the covariates, and for some others, the reduction was not significant.

After these two reductions, the validation parameters of the spline models were still well below the desirable values to minimize overfitting. Therefore, we applied PCA and variable transformations as described in Section 3.

Table 6. Data reduction from RCS adjustment of used knots and parameterization.

Predictor	Reduction	d.f. Saved	Change of LR X^2
Number of Workorders	Parametrized log (Workorders + 1).	3	+33.00
Fluid temperature	Reduce number of knots.	1	+18.42
Discharge pressure	Change from spline to linear.	3	−5.78
Speed (rpm)	Parametrized from linear to sqrt (rpm).	1	+0.79
Power	Reduce number of knots.	1	−1.25
Overall vibration	Reduce number of knots.	1	−3.35
Flow ratio	Reduce number of knots.	2	+0.70
NPSH margin	Change from spline to linear.	2	+2.02
Vapor pressure	Reduce number of knots and convergence issues.	1	+1.99
Tip speed	Reduce number of knots.	1	+5.00
Ratio diameter	Reduce number of knots.	1	−5.93
Suction stability	Reduce number of knots.	1	−2.49
Number of Lube Workorders	Reduce number of knots.	1	+2.60
TOTAL		**19**	**+45.72**

4.2. PCA Reduction

The PCA reduction was applied considering four transformed sets of covariates. The main objective was to reduce the dimension of the final models.

The first variable transformation based on expert knowledge, variable parameterization, and reduced dimension of the RCS had already been performed. The other three proposed transformations were the following: the application of the ACE method to the raw variables, the application of RCS on these transformed variables, and the application of the AVAS transformation. These transformations and the relevant literature are described in Section 3.

Next, standard PCA and sparse PCA were applied to the eight sets of transformed covariates, thus computing eight new sets of features. The goal in using PCA was to reduce the dimension of the data to maximize the performance of the model and to minimize the risk of overfitting (by using the minimum number of components).

Once the PCA process was performed, eight models were developed using both standard PCA and Sparse PCA reduction on raw splined covariates, nonparametric transformed raw covariates, nonparametric transformed splined covariates, and AVAS transformed covariates. Figures 4 and 5 display the number of components on the x-axis and the corresponding AIC values for each model fitted using standard PCA and sparse PCA, respectively, on the y-axis. It is observed that as the number of components increases, the AIC decreases, indicating an improvement in the model's predictive capacity. However, there is a region where no further reduction in AIC is achieved despite adding new predictors to the model.

As can be seen in Figure 4, the raw splines PCA model reaches the lowest AIC score and one can observe that the score eventually increases after more than 40 components are added. Similar results are obtained with PCA transformed splines. On the other hand, both non-parametric and AVAS transformed variables have a limit of around 20 components because these transformations use fewer d.f. and do not model the non-linearities. As a consequence, the models have limited prediction capacity and the minimum achievable AICs are higher than those obtained with spline models. Despite this, the PCA AVAS model

achieved a slightly lower AIC value compared to the PCA raw splines and transformed splines models when using 21 components (d.f.), with a good balance of model complexity and model performance.

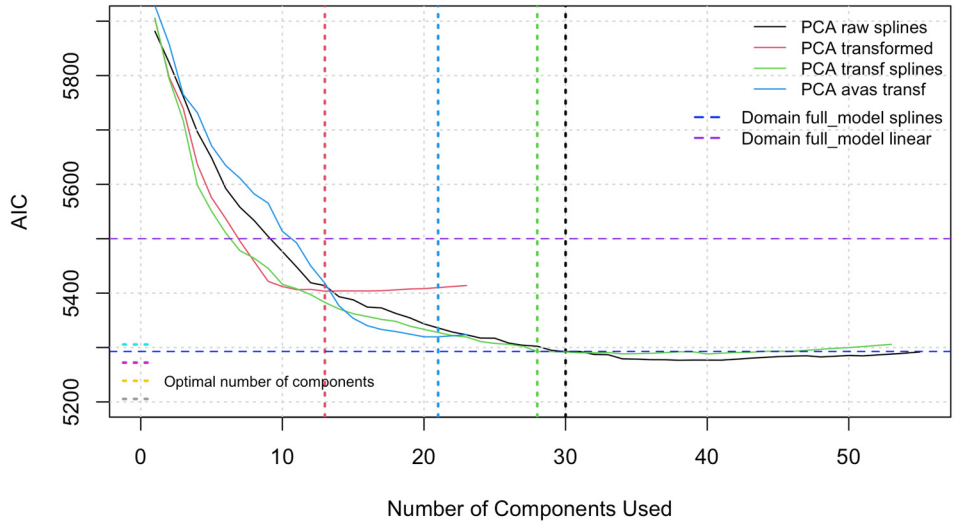


Figure 4. AIC of Cox models fitted with progressively more principal components, obtained with standard PCA, compared to the obtained AIC of full models, linear and spline.

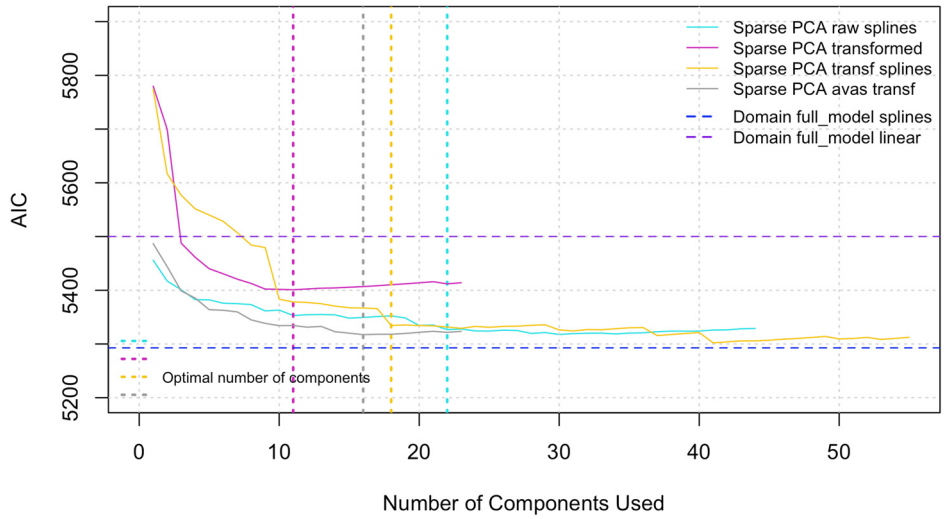


Figure 5. AIC of Cox models fitted with progressively more principal components, obtained with sparse PCA, compared to the obtained AIC of full models, linear and spline.

The results of Sparse PCA models are depicted in Figure 5. It can be observed that similar AIC results to those obtained with standard PCA are achieved using fewer d.f. Consequently, the optimal number of components was determined for each model. The AIC for the sparse PCA AVAS model settled for a low score of AIC = 5317 when 16 components were added to the model. This represents a good trade-off between model complexity and model performance and was chosen for further research. Similar AIC results were achieved with spline models using 22 components or more.

The relevant indices of both standard and sparse PCA optimal models were computed, as shown in Table 7.

Table 7. Summary of PCA models. Sparse PCA raw covariates (linear). Domain full model with RCS. Nonparametric transformed raw covariates. Transformed splined covariates. AVAS transformed covariates.

Description	Index	PCA Dom Red RCS & Param.	PCA Transf.	PCA Transf. RCS	PCA AVAS	Sparse PCA Raw RCS	Sparse PCA Transf.	Sparse PCA Transf. RCS	Sparse PCA AVAS
Model Tests	LR X^2	740.76	595.51	736.40	695.50	690.35	594.28	674.62	687.74
	d.f.	30	13	28	21	22	11	18	16
	Explained Var.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Discrimination Indices	R^2	0.671	0.591	0.669	0.648	0.645	0.590	0.637	0.644
	D_{xy}	−0.742	−0.706	−0.740	−0.744	−0.740	−0.706	−0.736	−0.744
	$R^2_{df,501}$	0.758	0.687	0.757	0.740	0.737	0.688	0.730	0.738
Predictive Discrimination	Concordance	0.871	0.853	0.870	0.872	0.870	0.853	0.867	0.872
	AIC	5292.33	5403.58	5292.70	5319.60	5326.74	5400.8	5334.48	5317.34
	AIC X^2 scale	680.76	569.51	704.40	653.49	646.35	572.27	638.61	655.74
Validation	LR R^2	0.633	0.568	0.633	0.618	0.617	0.566	0.613	0.620
	D_{xy}	−0.722	−0.697	−0.722	−0.730	−0.725	−0.697	−0.724	−0.734
	Calibration Slope	0.886	0.937	0.898	0.920	0.923	0.934	0.933	0.936
Calibration	Mean |err|	0.1278	0.0343	0.0740	0.0102	0.0077	0.0325	0.0116	0.0154
	0.9 quantile |err|	0.2583	0.0647	0.1581	0.0269	0.0182	0.0667	0.0264	0.0339

5. Discussion

The Cox PHM was applied to predict the MTBF of centrifugal pumps, where challenges with numerous covariates were encountered, which can be linked to overfitting and interpretation difficulties. While stepwise methods are commonly implemented, these were avoided due to statistical estimation concerns [36,58].

The complete dataset was utilized rather than being split into training and testing sets to maximize information. Two solutions were implemented to address the observed non-linear effects: restricted cubic splines (RCS) with simple parametrizations and nonparametric additive regression transformations. Degrees of freedom were increased by RCS, while interpretability was sacrificed by nonparametric variable transformations.

Unsupervised variable selection was applied to both transformed datasets, whereby established Cox model philosophies were followed [34,36,70]. Although regular PCA was attempted, better results were achieved through robust sparse PCA despite its limited documentation in the machinery reliability literature.

Model quality was evaluated using an AIC index, where variables were ordered using Wen et al.'s primal dual active set algorithm [66] instead of standard PCA ordering. Variables were selected by balancing AIC scores and variable count. Performance was assessed through Efron's bootstrap algorithm, whereby it was found that similar AIC values could be achieved by reduced degrees of freedom models while overfitting was minimized.

The effectiveness of these methods discussed here was demonstrated in managing numerous covariates and collinearity in estimating MTBF.

6. Conclusions and Future Work

This research demonstrated how data reduction techniques can be applied for MTBF estimation, achieving an 85% reduction in model complexity with equivalent predictive performance. At the same time, the model robustness was improved, as indicated by an increase in the calibration slope from 0.830 to 0.936. These findings confirm that MTBF can be predicted using the Cox PHM. In addition, the following observations were made:

- Full linear and spline models showed calibration slopes of 0.830 and 0.722, respectively. Because these slopes are below 0.9, excessive overfitting would be expected, requiring the need for data reduction.
- Strong non-linear components in the full model made it necessary to transform the covariates to relax the linearity assumptions of the regression (which would have resulted in poor model fit).
- The models applying sparse robust PCA obtained results similar to those fitted with the standard PCA method but using fewer d.f.
- The preferred model was fitted using principal components obtained from sparse robust PCA, applied to X variables transformed with the AVAS algorithm. The resulting AIC was 5317.34 with a calibration slope of 0.936 for the prediction of MTBF, indicating a superior result.
- The dimension reduction achieved with the final model was 16 d.f., down from 103 d.f. in the full model with RCS with a corresponding AIC increase of 0.34%.

In addition, preliminary correlation ranks of potential predictors with MTBF are provided but the findings suggest that further research is needed to obtain meaningful insights. Furthermore, the impact of each predictor on the reliability of the analyzed pumps can be assessed. Specifically, the following topics are suggested for future work:

- Determine the most important variables and check how they impact the predicted MTBF.
- Rank the importance and prediction ability of the raw covariates and principal components of the full and reduced models.
- Examine the assumptions of the Cox PHM to identify potential issues and evaluate their impact on the final models.
- Assess which variables make some pumps behave differently from others, focusing on various groups of covariates: operating conditions, hydraulic design, mechanical design, age, sealing, and maintenance history.
- Repeated failures of the assets were modeled under perfect repair conditions, which implies that the machine has the same lifetime distribution and the same rate function as a new one [71] after repair. This assumption might be challenged.
- Check for variable interactions and their importance in the prediction ability of the model.
- Models were fitted considering that the effect of the covariates remains constant over time. Future work could take a time-dependent approach using the same variables for centrifugal pumps.

Author Contributions: Conceptualization, M.V.F., D.G. and U.K.; methodology, M.V.F. and D.G.; software, M.V.F.; validation, M.V.F. and D.G.; formal analysis, M.V.F.; investigation, M.V.F.; resources, M.V.F., D.G. and U.K.; data curation, M.V.F.; writing—original draft preparation, M.V.F.; writing—review and editing, K.G.; visualization, M.V.F.; supervision, D.G.; project administration, D.G. and U.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of the dataset. The dataset for this study is not publicly available due to third-party ownership.

Conflicts of Interest: Author Marc Vila Forteza was employed by the company Petronor. Author Kai Goebel was employed by the company Fragum Global. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. List of Acronyms and Symbols

For a list of acronyms and symbols used in this paper, refer to Tables A1 and A2.

Table A1. List of acronyms used in this paper.

Acronym	Explanation
ACE	Alternating Conditional Expectation
AIC	Akaike information criterion
ALS	Alternating Least Squares
ANN	Artificial Neural Networks
API	American Petroleum Institute
AVAS	Additive Variance Stabilizing
BIC	Bayesian information criterion
CNN	Convolutional Neural Networks
CTBN	Continuous Time Bayesian Networks
d.f.	Degrees of freedom
DCS	Distributed control system
DNN	Deep Neural Networks
FTA	Failure Tree Analysis
GMM	Gaussian Mixed Model
ISO	Internation Organization for Standardization
k-NN	k-Nearest Neighbours
KPI	Key Performance Indicator
LARS	Least angle regression
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Log-Likelihood ratio
MGV	Maximum Generalized Variance
MTBF	Mean Time Between Failures
MTTR	Mean Time To Repair
MTV	Maximum Total Variance
MWPHM	Mixture Weibull Proportional Hazards Model
NHPP	Non-homogeneous Poisson process
NPSH	Net Positive Suction Head
N_s	Specific speed
N_{ss}	Suction specific speed
OLS	Ordinary Least Squares
OREDA	Offshore and Onshore Reliability Data
PbM	Physical-based Models
PCA	Principal Component Analysis
PF	Particle Filter
PHM	Proportional Hazards Model
RCS	Restricted cubic splines
RMV	Relevance Vector Machine
ROC	Receiver-Operating Characteristic Curve
rpm	Revolutions per minute
RNN	Recurrent Neural Networks
RUL	Remaining useful life
SAS	Statistical Analysis System
SVM	Support Vector Machine
TIM	Traditional imperfect maintenance
VGP	Variance Gamma Process

Table A2. List of symbols used in this paper.

Symbol	Meaning	Units (if Applicable)
Λ	Hazard rate	Failures per unit of time
R(t)	Reliability function	Dimensionless (0 to 1)
B	Shape parameter (Weibull)	Dimensionless
H	Scale parameter (Weibull)	Time units (e.g., days)
T	Time	Time units (e.g., days)
X	Matrix of predictors	Variable (depends on context)
B	Regression coefficients (Cox model)	Dimensionless
R^2	Coefficient of determination	Dimensionless (0 to 1)
χ^2	Chi squared	Dimensionless (statistic)
P	Spearman's rank correlation coefficient	Dimensionless (−1 to 1)
D_{xy}	Somers' rank correlation	Dimensionless (−1 to 1)

Appendix B. List of RStudio Version 2024.12.1 Packages Used for Computation

All the algorithms were developed in R statistical programming language using RStudio, an integrated development environment for R. The following list in Table A3 includes the packages that were used to perform the relevant computations of our work.

Table A3. List of RStudio version 2024.12.1 packages used for computation.

Package/Software	Reference
R	R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/ accessed on 24 February 2025
tidyverse	Wickham et al. (2019). Welcome to the tidyverse. Journal of Open-Source Software, 4(43), 1686. doi: 10.21105/joss.01686. https://cran.r-project.org/web/packages/tidyverse/index.html accessed on 24 February 2025
Matrix	Bates et al. (2023). Matrix: Sparse and Dense Matrix Classes and Methods. https://CRAN.R-project.org/package=Matrix accessed on 24 February 2025
survival	Therneau (2023). A Package for Survival Analysis in R. https://CRAN.R-project.org/package=survival accessed on 24 February 2025
rstatix	Kassambara (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://CRAN.R-project.org/package=rstatix accessed on 24 February 2025
survminer	Kassambara & Kosinski (2021). survminer: Drawing Survival Curves using 'ggplot2'. https://CRAN.R-project.org/package=survminer accessed on 24 February 2025
ggcorrplot	Kassambara (2019). ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. https://CRAN.R-project.org/package=ggcorrplot accessed on 24 February 2025
ggplot2	Wickham et al. (2023). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. https://CRAN.R-project.org/package=ggplot2 accessed on 24 February 2025
dplyr	Wickham et al. (2023). dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr accessed on 24 February 2025
MASS	Venables & Ripley (2002). MASS: Modern Applied Statistics with S (4th ed.). Springer.
doBy	Hawthorne & Wesselingh (2016). doBy: Grouping, ordering, and summarizing functions.
glmnet	Friedman et al. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Software, 33(1), 1-22. doi: 10.18637/jss.v033.i01. https://www.jstatsoft.org/article/view/v033i01 accessed on 24 February 2025
rms	Harrell (2021). rms: Regression Modeling Strategies. https://CRAN.R-project.org/package=rms accessed on 24 February 2025
Hmisc	Harrell (2021). Hmisc: Harrell Miscellaneous. https://CRAN.R-project.org/package=Hmisc accessed on 24 February 2025
pcaPP	Lê et al. (2008). pcaPP: Principal component methods: A new approach to principal component analysis. https://CRAN.R-project.org/package=pcaPP accessed on 24 February 2025
splines	R Core Team (2021). splines: Regression Spline Functions. https://CRAN.R-project.org/package=splines accessed on 24 February 2025
acepack	Tyler & Wang (2015). acepack: A Package for Fitting the ACE and AVAS Models. https://CRAN.R-project.org/package=acepack accessed on 24 February 2025
BeSS	Friedman & Popescu (2008). BeSS: Best Subset Selection. https://CRAN.R-project.org/package=BeSS accessed on 24 February 2025

References

1. Nesbitt, B. *Handbook of Pumps and Pumping: Pumping Manual International*, 1st ed.; Elsevier: Amsterdam, The Netherlands, 2006; ISBN 9781856174763.
2. Lu, H.; Guo, L.; Azimi, M.; Huang, K. Oil and Gas 4.0 era: A systematic review and outlook. *Comput. Ind.* **2019**, *111*, 68–90. [CrossRef]
3. Vila Forteza, M.; Pascual, D.; Galar, U.; Kumar, A.K.; Verma. Work-in-progress: Reliability prediction of API centrifugal pumps using survival analysis. In Proceedings of the 19th IMEKO TC10 Conference "Measurement for Diagnostics, Optimisation and Control to Support Sustainability and Resilience", Delft, The Netherlands, 21–22 September 2023. [CrossRef]
4. Vila Forteza, M.; Jimenez Cortadi, A.; Diez Olivan, A.; Seneviratne, D.; Galar Pascual, D. Advanced Prognostics for a Centrifugal Fan and Multistage Centrifugal Pump Using a Hybrid Model. In Proceedings of the 5th International Conference on Maintenance, Condition Monitoring and Diagnostics 2021, Oulu, Finland, 16–17 February 2021. [CrossRef]
5. Adraoui, I.E.; Gziri, H.; Mousrij, A. Prognosis of a degradable hydraulic system: Application on a centrifugal pump. *Int. J. Progn. Health Manag.* **2020**, *11*, 1–11. [CrossRef]
6. Cubillo, A.; Perinpanayagam, S.; Esperon Miguez, M. A review of physics-based models in prognostics: Application to gears and bearings of rotating machinery. *Adv. Mech. Eng.* **2016**, *8*, 1687814016664660. [CrossRef]
7. Zhang, S.; Hodkiewicz, M.; Ma, L.; Mathew, J. Machinery Condition Prognosis Using Multivariate Analysis. In *Engineering Asset Management*; Mathew, J., Kennedy, J., Ma, L., Tan, A., Anderson, D., Eds.; Springer: London, UK, 2006. [CrossRef]

8. Yu, R.; Li, X.; Tao, M.; Ke, Z. Fault Diagnosis of Feedwater Pump in Nuclear Power Plants Using Parameter-Optimized Support Vector Machine. In Proceedings of the 2016 24th International Conference on Nuclear Engineering, Charlotte, NC, USA, 26–30 June 2016; p. V001T03A013. [CrossRef]
9. Zurita Millan, D.; Delgado-Prieto, M.; Saucedo-Dorantes, J.; Cariño-Corrales, J.; Osornio-Rios, R.; Ortega, J.; Romero-Troncoso, R. Vibration Signal Forecasting on Rotating Machinery by means of Signal Decomposition and Neurofuzzy Modeling. *Shock. Vib.* **2016**, *2016*, 2683269. [CrossRef]
10. Fouladirad, M.; Belhaj Salem, M.; Deloux, E. Variance Gamma process as degradation model for prognosis and imperfect maintenance of centrifugal pumps. *Reliab. Eng. Syst. Saf.* **2022**, *223*, 108417. [CrossRef]
11. Souza, R.; Sperandio, N.; Erick, G.; Miranda, U.; Silva, W.; Lepikson, H. Deep learning for diagnosis and classification of faults in industrial rotating machinery. *Comput. Ind. Eng.* **2020**, *153*, 107060. [CrossRef]
12. Kumar, A.; Gandhi, C.; Zhou, Y.; Kumar, R.; Xiang, J. Improved deep convolution neural network (CNN) for the identification of defects in the centrifugal pump using acoustic images. *Appl. Acoust.* **2020**, *167*, 107399; ISSN 0003-682X. [CrossRef]
13. Zhao, L.; Wang, X. A Deep Feature Optimization Fusion Method for Extracting Bearing Degradation Features. *IEEE Access* **2018**, *6*, 19640–19653. [CrossRef]
14. Zhang, Y.; Zhou, T.; Huang, X.; Longchao, C.; Zhou, Q. Fault diagnosis of rotating machinery based on recurrent neural networks. *Measurement* **2020**, *171*, 108774. [CrossRef]
15. Mosallam, A.; Medjaher, K.; Zerhouni, N. Data-driven prognostic method based on Bayesian approaches for direct remaining useful life prediction. *J. Intell. Manuf.* **2014**, *27*, 1037–1048. [CrossRef]
16. Forrester, T.; Harris, M.; Senecal, J.; Sheppard, J. Continuous Time Bayesian Networks in Prognosis and Health Management of Centrifugal Pumps. In Proceedings of the Annual Conference of the PHM Society, Scottsdale, AZ, USA, 23–26 September 2019; p. 11. [CrossRef]
17. Wang, J.; Zhang, L.; Zheng, Y.; Wang, K. Adaptive prognosis of centrifugal pump under variable operating conditions. *Mech. Syst. Signal Process.* **2019**, *131*, 576–591. [CrossRef]
18. Zhang, Q.; Hua, C.; Xu, G. A mixture Weibull proportional hazard model for mechanical system failure prediction utilising lifetime and monitoring data. *Mech. Syst. Signal Process.* **2014**, *43*, 103–112. [CrossRef]
19. Hu, J.; Tse, P. A Relevance Vector Machine-Based Approach with Application to Oil Sand Pump Prognostics. *Sensors* **2013**, *13*, 12663–12686. [CrossRef]
20. Cao, S.; Hu, Z.; Luo, X.; Wang, H. Research on fault diagnosis technology of centrifugal pump blade crack based on PCA and GMM. *Measurement* **2020**, *173*, 108558. [CrossRef]
21. Li, X.; Duan, F.; Mba, D.; Bennett, I. Rotating machine prognostics using system-level models. Lecture Notes in Mechanical Engineering. In *Engineering Asset Management 2016: Proceedings of the 11th World Congress on Engineering Asset Management*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; Volume 3, pp. 123–141. [CrossRef]
22. Kim, S.; Choi, J.H.; Kim, N.H. Challenges and Opportunities of System-Level Prognostics. *Sensors* **2021**, *21*, 7655. [CrossRef] [PubMed] [PubMed Central]
23. Bloch, H.P.; Budris, A.R. *Pump User's Handbook, Life Extension*, 4th ed.; The Fairmont Press Inc.: Atlanta, GA, USA, 2014; p. 103.
24. Bevilacqua, M.; Braglia, M.; Montanari, R. The classification and regression tree approach to pump failure rate analysis. *Reliab. Eng. Syst. Saf.* **2003**, *79*, 59–67. [CrossRef]
25. Braglia, M.; Carmignani, G.; Frosolini, M.; Zammori, F. Data classification and MTBF prediction with a multivariate analysis approach. *Reliab. Eng. Syst. Saf.* **2012**, *97*, 27–35. [CrossRef]
26. Braglia, M.; Castellano, D.; Frosolini, M.; Gabbrielli, R.; Marrazzini, L.; Padellini, L. An ensemble-learning model for failure rate prediction. *Procedia Manuf.* **2020**, *42*, 41–48. [CrossRef]
27. Bevilacqua, M.; Braglia, M.; Frosolini, M.; Montanari, R. Failure rate prediction with artificial neural networks. *J. Qual. Maint. Eng.* **2005**, *11*, 279–294. [CrossRef]
28. Orrù, P.F.; Zoccheddu, A.; Sassu, L.; Mattia, C.; Cozza, R.; Arena, S. Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability* **2020**, *12*, 4776. [CrossRef]
29. Sudadiyo, S. Nonhomogeneous Poisson Process Model for Estimating Mean Time Between Failures of the JE01-AP03 Primary Pump Implemented on the RSG-GAS Reactor. *Nucl. Technol.* **2024**, 1–16. [CrossRef]
30. Chaoqun, D.; Song, L. A Study of Proportional Hazards Models: Its Applications in Prognostics. In *Maintenance Management-Current Challenges, New Developments, and Future Directions*; IntechOpen: London, UK, 2023. [CrossRef]
31. Jardine, A.K.S.; Anderson, P.M.; Mann, D.S. Application of the Weibull proportional hazards model to aircraft and marine engine failure data. *Qual. Reliab. Eng. Int.* **1987**, *3*, 77–82. [CrossRef]
32. Sharma, G.; Sahu, P.K.; Rai, R.N. Imperfect maintenance and proportional hazard models: A literature survey from 1965 to 2020. *Life Cycle Reliab. Saf. Eng.* **2022**, *11*, 87–103. [CrossRef]
33. Gorjian, N.; Sun, Y.; Ma, L.; Yarlagadda, P.; Mittinty, M. Remaining useful life prediction of rotating equipment using covariate-based hazard models–Industry applications. *Aust. J. Mech. Eng.* **2017**, *15*, 36–45. [CrossRef]

34. Harrell, F.E., Jr.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions an adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [CrossRef]
35. *ISO 14224:2016*; Petroleum, Petrochemical and Natural Gas Industries—Collection and Exchange of Reliability and Maintenance Data for Equipment. 2016. Available online: https://www.iso.org/standard/64076.html (accessed on 24 February 2025).
36. Harrell, F.E., Jr. *Regression Modeling Strategies*; Springer Series in Statistics; Springer International Publishing: Cham, Switzerland, 2016.
37. Lee, E.T. *Statistical Methods for Survival Data Analysis*; John Wiley & Son: New York, NY, USA, 1982.
38. Kleinbaum, D.G.; Klein, M. *Survival Analysis: A Self-Learning Text*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2012.
39. Jiang, J.; Xiong, Y. Cox models with time-dependent covariates. In *Handbook of Survival Analysis*; CRC Press: Boca Raton, FL, USA, 2011; pp. 205–226.
40. Fisher, L.D.; Lin, D.Y. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* **1999**, *20*, 145–157. [CrossRef]
41. Becker, T. *BSc Report Applied Mathematics: Variable Selection*; Delft University of Technology: Delft, The Netherlands, 2021.
42. Petersson, S.; Sehlstedt, K. Variable Selection Techniques for the Cox Proportional Hazards Model: A Comparative Study. 2018. Available online: https://gupea.ub.gu.se/handle/2077/55936 (accessed on 24 February 2025).
43. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
44. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat. Med.* **1997**, *16*, 385–395. [CrossRef]
45. Fan, J.; Li, R. Variable selection for cox's proportional hazards model and frailty model. *Ann. Stat.* **2002**, *30*, 74–99. [CrossRef]
46. Zhang, H.H.; Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **2007**, *94*, 691–703. [CrossRef]
47. Lin, D.; Banjevic, D.; Jardine, A.K.S. Using principal components in a proportional hazards model with applications in condition-based maintenance. *Reliab. Eng. Syst. Saf.* **2006**, *91*, 59–69. [CrossRef]
48. Bankole-Oye, T.; El-Thalji, I.; Zec, J. Combined principal component analysis and proportional hazard model for optimizing condition-based maintenance. *Mathematics* **2020**, *8*, 1521. [CrossRef]
49. de Carvalho Michalski, M.A.; da Silva, R.F.; de Andrade Melani, A.H.; de Souza, G.F.M. Applying Principal Component Analysis for Multi-parameter Failure Prognosis and Determination of Remaining Useful Life. In Proceedings of the 2021 Annual Reliability and Maintainability Symposium (RAMS), Orlando, FL, USA, 24–27 May 2021; pp. 1–6. [CrossRef]
50. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
51. Liu, W.M.; Chang, C.I. Variants of Principal Components Analysis. In Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–27 July 2007; pp. 1083–1086. [CrossRef]
52. Zou, H.; Hastie, T.; Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **2006**, *15*, 265–286. [CrossRef]
53. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
54. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–220. [CrossRef]
55. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
56. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
57. Vrieze, S.I. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* **2012**, *17*, 228–243. [CrossRef]
58. Harrell, F.E.; Lee, K.L.; Pollock, B.G. Regression models in clinical studies: Determining relationships between predictors and response. *JNCI J. Natl. Cancer Inst.* **1988**, *80*, 1198–1202. [CrossRef]
59. Kooperberg, C.; Stone, C.J.; Truong, Y.K. Hazard regression. *J. Am. Stat. Assoc.* **1995**, *90*, 78–94. [CrossRef]
60. Young, F.W.; Takane, Y.; de Leeuw, J. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika* **1978**, *43*, 279–281. [CrossRef]
61. Sarle, W.S. *SPLIT-CLASS: A Method for Multivariate Categorical Data Analysis*; SAS Institute Technical Report; SAS Institute: Cary, NC, USA, 1995.
62. Kuhfeld, W.F. *Marketing Research Methods in SAS: Experimental Design, Choice, Conjoint, and Graphical Techniques*; SAS Institute Inc.: Cary, NC, USA, 2009; pp. 1267–1268.
63. Kuhfeld, W.F. SAS/STAT® 14.1 User's Guide. The PRINQUAL Procedure. SAS Publishing. 2009. Available online: https://support.sas.com/documentation/onlinedoc/stat/141/prinqual.pdf (accessed on 14 February 2025).
64. Breiman, L.; Friedman, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
65. Tibshirani, R. Estimating transformations for regression via additivity and variance stabilization. *J. Am. Stat. Assoc.* **1988**, *83*, 394–405. [CrossRef]
66. Wen, C.; Zhang, A.; Quan, S.; Wang, X. BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. *J. Stat. Softw.* **2020**, *94*, 1–24. [CrossRef]

67. *ISO 10816-7:2009*; Mechanical Vibration—Evaluation of Machine Vibration by Measurements on Non-Rotating Parts—Part 7: Rotodynamic Pumps for Industrial Applications, Including Measurements on Rotating Shafts. International Organization for Standardization, 2019. Available online: https://www.iso.org/es/contents/data/standard/04/17/41726.html (accessed on 24 February 2025).
68. Bradshaw, S.; Liebner, T.; Cowan, D. Influence of impeller suction specific speed on vibration performance. In Proceedings of the Twenty-Ninth International Pump Users Symposium, Houston, TX, USA, 1–3 October 2013.
69. Pavlou, M.; Ambler, G.; Qu, C.; Seaman, S.R.; White, I.R.; Omar, R.Z. An evaluation of sample size requirements for developing risk prediction models with binary outcomes. *BMC Med Res. Methodol.* **2024**, *24*, 146. [CrossRef]
70. Sauerbrei, W.; Perperoglou, A.; Schmid, M.; Abrahamowicz, M.; Becher, H.; Binder, H.; Dunkler, D.; Harrell, F.E., Jr.; Royston, P.; Georg Heinze for TG2 of the STRATOS initiative. State of the art in selection of variables and functional forms in multivariable analysis—Outstanding issues. *Diagn. Progn. Res.* **2020**, *4*, 3. [CrossRef]
71. De Carlo, F.; Arleo, M.A. Imperfect maintenance models, from theory to practice. In Proceedings of the International Conference on Reliability and Maintenance (ICRM), Buenos Aires, Argentina, 15–19 May 2017; pp. 345–356.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.