Creating Models for Predictive Maintenance of Field Equipment in the Oil Industry Using Simulation Based Uncertainty Modelling

Raul Ramirez-Velarde1[0000-0001-7186-1914], Laura Hervert-Escobar[0000-0003-2465-7106] and Neil Hernandez-Gress[0000-0003-0966-5685]

Tecnologico de Monterrey. Eugenio Garza Sada 2501 Sur, Col. Tecnológico, Monterrey, N. L., Mexico, 64849

rramirez@tec.mx

Abstract. Determining what causes field equipment malfunction and predicting when those malfunctions will occur can save large amounts of money for corporations that are capital-intensive. To avert equipment downtime, field equipment maintenance departments must be adequately resourced. Herein, we demonstrate the efficacy of machine learning to determine time between failure, repair time (equipment downtime) and repair cost. Additionally, a mean value analysis is carried out to determine the maintenance department capacity. Uncertainty is modelled using statistical analysis and simulation.

Keywords: predictive maintenance, machine learning, heavy-tail simulation.

1 Introduction

Predictive equipment maintenance is one of the most important areas in industries that are heavily reliant on capital, as all other processes depend on the correct performance of its "clients" (equipment). It is well-known that over time, the performance of field equipment decreases, and their failure rate increases, so it is logical and a good strategy to diagnose areas of opportunity for minimizing repair costs, equipment down-time, production delays, and equipment failure frequency. In this paper, we will analyse maintenance work orders data for field equipment maintenance provided by an oil extraction corporation corresponding to upstream gas extraction operations to create predictive models that can allow companies to reduce costs and improve efficiency. However, data and the models developed with it bring inherent uncertainty both epistemically and ontologically. Uncertainty modelling techniques include stochastic simulation, chance-constrained models, Markov processes, stochastic optimization, Bayesian models, evidence theory, fuzzy theory, information-gap theory, and statistics and probability. In this paper, we present and study how uncertainty is modelled using statistics and probability, and stochastic simulation. We carry out a thorough analysis of the company's maintenance area by applying unsupervised and supervised data science analysis to its work orders database, which consists of more than 55 variables and over 500K records.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Specifically, the research goals are:

1. Determine how Time between Failures (TBF), repair costs, and repair duration can be characterized and identify which factors influence these variables. Create statistical and mathematical models to make predictions.
2. Develop models to predict how repair wait time and the number of repair reports waiting to be serviced can change with changes in the organization, characteristics, and structure of the company's maintenance areas.
3. Provide a framework for uncertainty modelling through simulation.

2 Related Work

The research presented in this paper falls under the predictive maintenance field, which is an area that has gained increased attention in the context of equipment maintenance systems. Predictive maintenance is defined by [1] as "regular monitoring of the actual mechanical condition, operating efficiency, and other indicators of the operating condition of machine-trains to ensure the maximum interval between repairs as well as to minimize the number and cost of unscheduled outages caused by machine-train failures". Due to its importance for the efficiency of field operations, many studies have performed predictive maintenance. In [2], a comprehensive review of the state of the art of equipment maintenance systems is presented, and it is concluded that predictive maintenance technology is a growing research field with fast-increasing contributions from different application areas, such as mechanical, chemical, energy, automation, etc. Predictive maintenance proceeds by creating machine behaviour models. One example is shown in [3], where a bi-level optimization model is presented, which addresses the need to balance an optimization procedure locally (level one) and globally (level two) while minimizing system average interruption frequency. In [4], an application of predictive maintenance using a model-based analysis is presented, to enhance the accuracy of performance diagnostics. Different models under different operating conditions are created and then performance is recalculated and compared for efficiency. Another model-based approach is shown in [5].

Recently, the body of knowledge related to machine learning, big data, artificial intelligence, and cloud computing has been combined with predictive maintenance, resulting in the emergence of Cyber-Physical Systems (CPS). An extensive literature review of CPS as of 2020 is presented in [6], which concludes that the main keywords are internet of things, industry 4.0, predictive maintenance, machine learning, artificial intelligence, cloud computing, and big data. In [7], a comprehensive review of machine learning algorithms applied to tool wear selection is provided, with adversarial neural networks and random forests being identified as the best models for tool wear prediction. A comparison of machine learning algorithms used to reliably estimate performance and detect anomalies in a representative combined cycle power plant is found in [8]. Here, unlike in [7], the prediction is carried out using multivariable models with features such as temperature, humidity, and pressure. Early-stage malfunctions are detected using anomaly estimation via unsupervised machine learning algorithms and

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

principal component analysis for dimension reduction. A definition of Cyber-Physical Systems and its relation to Industry 4.0 is presented in [9], and the concepts of big data, cloud computing, and machine learning and their applications to industry 4.0 are reviewed. Big data and cloud computing are mentioned in [10] in which a systematic architecture is proposed. An application using neural networks and deep learning to improve maintenance support and wear prediction of field equipment is shown in [11].

An important part of the research presented in this paper is modelling uncertainty. There is extensive bibliography regarding uncertainty management. Uncertainty is usually modelled using statistical analysis, probability theory, and Bayesian methods, stochastic optimization, constrained models, fuzzy theory, information-gap theory, Monte Carlo simulation, and discrete event simulation based on statistical characterization of critical variables. An important reference for modelling uncertainty using probability theory is found in [12] which gives the mathematical background for uncertainty diffusion and draws heavily on inferential statistics to model systematic errors. Another important reference is [13], which contains a wide range of applications of uncertainty modelling, from probability and statistics principles to applications in economics, art, psychology, and sciences. Another important reference is [14], which is a compendium of mathematical tools for engineering, ranging from decision making to optimization, to probability applications and curve fitting, and certainly, modelling uncertainty. In chapters 7, uses of probability theory, Monte Carlo simulation, chance constrained models, Markov processes, and stochastic optimization are reviewed. [15] is a specific survey of modern techniques for uncertainty modelling. In this paper, uses of probability theory and probability theory-derived techniques such as Monte Carlo methods, Bayesian methods, and evidence theory are reviewed. Also, fuzzy theory and Information-Gap theory are reviewed. The mathematical background for uncertainty propagation found in [12] is applied in [16] for modelling uncertainty in linear regression, which shows how difficult modelling uncertainty can be even for the simplest of machine learning algorithms. An important discussion of the differences between uncertainty and risk can be found in [17]. In this paper, it is argued that a change in risk is a change in the spread of the probability spectrum conserving the mean, whereas uncertainty is a change in the probability curves that can change the mean. It is also argued that some agents might be fond of risk but averse to uncertainty, whereas some agents might accept high uncertainty if the risk is low. High-risk almost always indicates high gains but high uncertainty might indicate the contrary. An example of the use of standard statistics such as standard deviation to model uncertainty can be found in [18], whereas in [19] bootstrapping is used to run several simulations taking traditional standard deviation statistics to create confidence intervals for biomanufacturing processes. An interesting approach is carried out in [20], where uncertainty is managed by using a genetic algorithm to improve a simulation carried out with a mixed integer linear programming model. Simulation is again utilized, and uncertainty is modelled using Bayesian averaging for modelling watersheds and stream flows in [21], whereas in [22], Monte Carlo simulation is thoroughly studied and compared against conventional uncertainty estimation methods such as uncertainty diffusion mathematical models. The paper indicates when it is appropriate to use Monte Carlo simulation.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

3 Methodology

The following methodology for analytics and model creation, called **Predictive Variance Association**, was followed for three types of analysis: Time between failures (TBF), repair cost, and repair duration. We are given a matrix X with data in which the rows are samples (m samples) with numerical values, and the columns are variables (n variables). The explanatory variables are separated into matrix E (o variables), and the dependent variables (p variables) or variables of interest, are separated into matrix G. Naturally, $n = o + p$. Additionally, from PCA [23], samples scores and variables loadings, dimension reduction and clustering can also be carried out. See Figure 1, which shows the process.

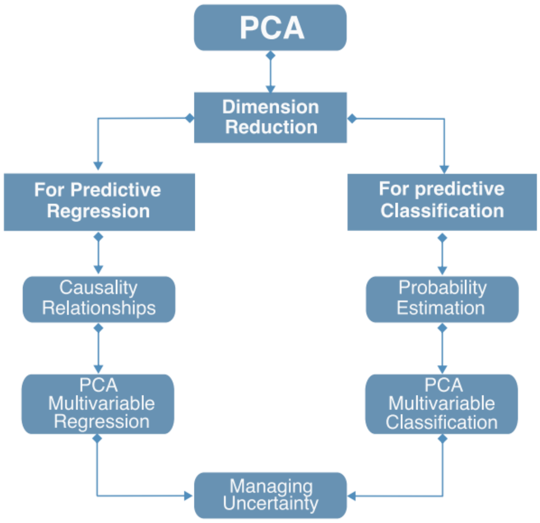

Fig. 1. Predictive Factor Association Steps.

3.1 Predictive Analytics for Continuous Variables by Regression

This process is accomplished when the objective is to predict the future value of a continuous variable of interest:

1. **First PCA**. Carry out PCA for matrix X and find matrices $X = PDQ^t$ [23] [24].
2. Let $F = XQ_{1 \ldots k}$ be the **principal component scores**, where $k < o$. Let $L = Q\sqrt{D}$ be the loadings matrix. The square cosine table is generated by squaring L.
3. Use square cosines and a clustering algorithm over the columns of F to determine:
   a. **Collinearity and dimension reduction**. Explanatory variables that are grouped together, that is, have a high cosine in the same component column or are clustered in the circle correlation plot are collinear and thus can eliminate those columns from matrix X. Repeat PCA, and then go to step 1.
   b. **Explain components**. Explanatory variables with the highest squared cosine in a particular column of L can be used to associate a component to a particular

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

variable of interest, thus elucidating the meaning of the component through the variable of interest. Explanatory variables with high squared cosine in the same column can also be associated with that variable of interest.

c. **Causality relationships**. If explanatory variables are grouped with variables of interest, then that is evidence of a causality relationship.
d. The **square cosine** is the square of the cosine of the angle α between the vectors of variable loadings, i.e., the correlation. If variables are close, the cosine will tend to one. If variables are separated, the cosine will approach zero. If $\rho^2(x,y) = cos^2(\alpha) > 0.5$ then $\cos(\alpha) > \mp 0.7071$, thus associate variables with squared cosine greater than 0.5.

4. **Generate PCA based synthetic attributes.** Eliminate variables of interest leaving only matrix E. Carry out PCA over E.
5. **Curve Fitting**. Carry our curve fitting between PCA derived explanatory variables and variables of interest clustered together in step 3.b using the new sample scores from step 4.
6. **Evaluate Curve Fitting**. Carry out curve fitting. Discard all R^2<0.5. This indicates which explanatory variables influence the most the variable of interest. Then $\widehat{G_\iota} = f_{ij}(F_j) = f_{ij}(q_1X_1, q_2X_2, \ldots, q_oX_o)$ where the coefficients q_i are determined by matrix Q.
7. **New samples** called X' can give predictions on variables, for $F' = X'Q$ and $\widehat{G_\iota'} = f_{ij}(F_j')$.

3.2 Categorical Variable Prediction by Multi-label classification

This process is carried out when the objective is to predict categorical variables. The procedure is similar to the one for continuous variables, except that the prediction is accomplished with a classification algorithm instead of a curve fitting algorithm:

1. Same as 3.1
2. Same as 3.1
3. Same as 3.1
4. Same as 3.1
5. **Apply classification algorithm**. Since the relationship between variables of interest and components is not always evident in classification problems, a search algorithm is carried out to determine the best components to use to predict a particular categorical variable of interest.
   a. **Match** F_i row by row with variable of interest j (assumed to be categorical with 0 or 1 as value), where $i = 1 \ \ldots \ o$ and $j = 1 \ldots p$.
   b. **Estimate class probabilities**. This can be carried out using different procedures, such as window-based probability estimation shown in [25]. Other machine learning classification algorithms are also used. Once this is completed, calculate performance measures. Each Fi is a linear combination of explanatory variables that potentially has enough information to give good predictions for the variables of interest. Then $\widehat{G_\iota} = f_{ij}(F_j) = f_{ij}(q_1X_1, q_2X_2, \ldots, q_oX_o)$ where the coefficients q_i are determined by matrix Q. If $\widehat{G_\iota} > 0.5$ then assume a result of 1, and 0 otherwise.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

6. **New samples** called X' can give predictions on variables, for $F' = X'Q$ and $\widehat{G_i'} = f_{ij}(F_j')$. If $\widehat{G_i} > 0.5$ then assume a result of 1, and 0 otherwise.

4 Results and Discussion

Throughout this document, an analysis of the company's maintenance area is done by applying data analysis to its Work Orders database, which uses the standard **ISO 14224:2016** and consists of 55 variables and 1,429,919 observations. After cleaning up of invalid values and some critical missing entries, and keeping records from 2015 to 2018, the database was reduced to 590,604 entries. Additionally, when talking about corrective jobs, we refer to jobs of JobType labeled "corrective", consisting, after value clean-up, of 89,480 elements (rows).

4.1 Dataset Description

New variables were created by establishing thresholds for the variables of interest which created classes. For Duration, the threshold was **7 days** for all jobs, which was 95% percentile and **6 days for corrective jobs**, which was also 95% percentile. A Duration of less than the threshold was labelled as 0. A Duration above the threshold was labelled as 1. For **Total cost**, the threshold was established at **$6,520.00** which is 99% percentile (above is 1, below is 0) and **$6,700.00 for corrective jobs** (99% percentile). For **time between reports** for the same equipment, the threshold was set to **40 days** for both general and corrective reports (63% percentile general, 42% corrective, above is 1, below is 0).

Many of the variables had to be recoded as they were originally coded as labels or text. Statistical Analysis was conducted. Table 1 summarizes statistics of all reports and Table 2 summarizes statistics of only corrective reports. The “Time between Reports same Equipment” and “Time between Failures same Equipment” were calculated by searching the unique equipment ID and subtracting the report dates, giving a time span in days. The statistics given are for data corresponding to years 2015 through 2018.

Table 1. Statistical summary for all reports

Variable	Mean	Std Dev	80% per	90% per	95% per	99%per
Duration	2.7929	13.1885	0	1	7.3	26.4300
Total Cost	397.67	8,675.53	0	0	0	6,520.00
Time bet Reports	0.00214	0.046213	0	0	0	0.0010
TBR Same Equip	91.7292	165.6745	133.94	322.59	443.33	799.88

Table 2. Statistical summary for corrective reports

Variable	Mean	Std Dev	80% per	90% per	95% per	99%per
Duration	1.9790	12.6575	0	1	6	16.65
Total Cost	724.52	17,677.36	0	0	0	16,700.00
Time bet Reports	0.0141	0.1213	0	0	0	1.0000
TBR Same Equip	124.0433	168.1949	213.72	343.63	473.54	800.41

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Time Between Failures

We analysed time between reports of any kind for any equipment. That is, the time between global consecutive reports. Data shows that there is always more than one report per day of any kind. The information given for time between reports does not have enough granularity to determine probability distribution as data was given in days, but actual arrival rate is minutes.

Analysis for time between reports of any kind (corrective, verification, update, scheduled maintenance, etc.) for a particular equipment indicates that reports for any given individual equipment follow an exponential distribution which will allows to establish probability bands on the time between reports for the same equipment.

As for corrective reports, analysis of time between any failures of any equipment indicates that corrective reports arrive at a frequency of several reports per day. The event of less than one report a day is very rare.

Analysis pf time between corrective reports, that is, time between failures for a particular equipment indicates that reports for any given individual equipment follow an exponential distribution which will allows to establish probability bands on the time between failures for the same equipment.

Duration

A variable called "Duration" was also analyzed. Analysis of report duration for all reports showed a very long tail, indicating that most reports take a day or less to be completed, but a significant number take more than 6 days. Of those, around 14,000 take 7 days, more than 16,000 take 26 days, and a very small fraction take more than 1,000 days.

For corrective reports, repair duration has the same characteristics. 91% of reports are resolved within a day, while a small percentage can take up to 1,000 days.

Total Cost

Analysis of Total Cost for all reports shows a very long tail. 57% of all report have zero cost, whereas 99.59% of all reports have a cost of $6,524.00 or less. As for corrective reports, 41.51% report a total cost of zero and 99.76% report a total cost of $16,758.00 or less.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

5 Principal Component Analysis

The next step is to analyse the relationship between our main variables of interest, Duration, Time between Reports/Failures and Cost and the other variables by carrying out principal component analysis.

5.1 Duration

Principal component analysis was also carried out. Correlation plots and square cosine lead us to conclude that **Equipment Type**, **Equipment Criticality**, **Area**, **Material Cost**, **Labor Cost**, **IsAffectingProduction** and **WordOrder** are the variables that correlate the most to duration, and these can be used to predict repair duration. Duration is associated mostly to the second principal component.

5.2 Repair Cost

Principal component analysis was also carried out to determine the factors affecting repair cost. According to the square cosines table, most of the variation of the total cost is captured by the first two principal components. We determined that the most important relations for the cost are the material cost and labour cost, which we know are directly related (total cost being the sum of labour and material). Thus, it was inferred that other factors which affect the cost are the level of **equipment criticality, WO-Type, TradeGroup, IsAffectingProduction and Job Type**.

5.3 Time Between Reports/Failures

The results of PCA for time between reports indicate that there are no variables that influence the time between reports, and that there are very few variables that influence the time between reports for the same equipment, mainly: **IsAffectingProduction, Equipmentcode, EquipmentRollupCode, ActualDuration and LabortCost**. But the influence is limited.

Also, PCA results show that there are no variables that influence the time between failures, that is, corrective reports, when considering all reports, and that there are very few variables that influence the time between failures for the same equipment. These variables are: **IsAffectingProduction, Equipmentcode, ActualDuration, CauseC and SafetyC**. But the influence is limited.

5.4 Machine Learning Prediction

Prediction models for our variables of interest were developed. These are classification type models, meaning that given a set of features represented as columns in a data matrix, the output is 1 or 0 for each sample; 1 indicates that the sample corresponds to a class and 0 indicates that it does not. In our case, the classes were determined as follows (Tab. 3):

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Table 3. Thresholds use to create the classification classes.

Variable	Threshold (equals to 1)
Repair Duration	$\geq 1\ day$
Total Repair Cost	$\geq \$250.00$
Time Between Failures	$\geq 40\ days$

Repair Duration.

For repair duration machine learning prediction, we used the class "Repair Time > 1 Day". The machine learning algorithm used for repair duration was a multi-layer perceptron classifier (MLP, an artificial neural network). The performance parameters for repair duration prediction using MLP are shown in Table 4. The global accuracy is 0.984, and the F1 obtained for class 1 is 0.99, which are considered excellent results. The likelihood ratio obtained for class 1 is 200, which is considered very good.

Table 4. Repair duration prediction performance parameters using deep learning.

Class	Precision	Recall	F1/score
0	0.86	0.74	0.80
1	0.99	1.00	0.99

Repair Cost

A MLP model was also used to create a model to predict whether the repair cost would be $250.00 dollars or higher. As predictors, the variables determined by PCA to be more closely related to repair cost were used. The results are shown in Table 5. The global accuracy achieved was 0.90, but the likelihood ratio for class 1 was only 1.25, which is not considered good. Additionally, the recall for class 1 was too low at 0.56.

Table 5. Total cost prediction performance parameters using deep learning. Accuracy of 0.90 is considered good, but F1 of 0.643 is considered low. Likelihood ratio of 1.25 in not good.

Class	Precision	Recall	F1-score
0	0.92	0.97	0.94
1	0.76	0.56	0.64

Time between Failures

To approximate a prediction model for time between failures, we used the somewhat loose category of time between failures greater than or equal to 40 days. We used MLP and Logistic Regression classifiers and obtained similar results. The results for the MLP classifier are shown in Table 6. This variable proved to be the hardest to predict. The likelihood ratio of 3.823 for class 1 is considered adequate, but the recall for class 0 is too low at 0.38. The global accuracy achieved is 0.615.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Table 6. Repair duration prediction performance parameters using deep learning. Accuracy obtained is 0.615, considered regular, but F1 of 0.701 not so much. Likelihood ratio obtained is 3.823 which is adequate.

Class	Precision	Recall	F1-score
0	0.58	0.38	0.46
1	0.63	0.79	0.70

5.5 Performance Analysis: Mean Value Analysis

What resources are necessary to service all reports with the correct quality level? Firstly, we will address this question using mean-value analysis; the notation for this follows in Table 7. Secondly, we will use simulation to model variability and uncertainty.

Table 7. Notation for mean value analysis

Variable	**Meaning**	**Units**
X	Throughput. Reports carried out per unit time	$\frac{Job}{day}$
S	Service time. Time required to complete a report	$\frac{day}{job}$
U	Utilization. Fraction of the time a worker is busy	None
P	Parallel workers. Workers working concurrently at any given moment	Workers
R	Residence Time. Total time required to finish a single job from report to conclusion	Days
N	Jobs in System. Jobs waiting to be finished per worker including the active one.	$\frac{Job}{report}$
D	Delay. The time a report will wait before been worked upon	Days

As mentioned in section 4.2, on average there are $X = 467.239$ reports per day, which represents both the report arrival rate λ and the service throughput X, as $\lambda = X$. Thus, the average inter-arrival time per report is 1/467.239=0.00214 days per report, or 0.0514 hours per report (a report every 3.0816 minutes). In that same section, we learned that the average duration time per report is $S = 2.7929$ days. This can only be possible if many personnel are working in parallel in different problems.

Thus, by the utilization law:

$$U = \frac{SX}{P}$$

Where P is the number of parallel maintenance workers. If we assume $U = 0.7, 0.8 \; or \; 0.9$ the following numbers of simultaneous maintenance workers P must be employed at project operation time (Tab. 8):

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Table 8. Number of parallel maintenance workers by utilization factor

U	*P*
0.7	1,864
0.8	1,631
0.9	1,450

Let us assume a utilization of $U = 0.8$. By queueing theory:

$$N = \frac{U}{1-U} = \frac{0.8}{1-0.8} = 4\,reports$$

Indicating that each worker has $N = 4$ jobs pending, including the one being serviced now. And by Little's law:

$$R = \frac{N}{\left(\frac{X}{P}\right)} = \frac{4}{\frac{467.239}{1{,}631}} = 13.963\,days$$

Indicating that each report will wait $R = 13.963$ days from creation to conclusion or 5 times the service time 2.7929 days. The wait time, the time a report will remain unattended, would be $D = 13.963 - 2.7929 = 11.1701$ days.

Additionally, we can repeat the procedure to determine the number of workers carrying out corrective repairs. For corrective repairs, we have a duration of 1.9790 days, an actual duration of 1.4961 days, and a time between failures of 0.0141 days per job, giving a demand of 70.922 jobs per day. Again assuming $U = 0.8$, we find:

Using Duration, $X = 70.922$ jobs/day and $S = 1.9790$ days/job (Duration). Assuming again $U = 0.8$ we find that $P = \frac{1.9790(70.922)}{0.8} = 175.4433$ workers doing repairs in parallel, total time $R = \frac{4}{\frac{70.922}{175.4433}} = 9.90$ days (5 times the service time), and wait time $D = 9.90 - 1.9790 = 7.92$ days.

Results are summarized in Table 9.

Table 9. Summary of mean value analysis results. Always U=0.8 and thus N=4 jobs in system per worker including the current one.

S is	**S (Service Time)**	**X (Throughput)**	**R (Total time)**	**P (Parallel Workers)**	**D (Wait to service)**
Duration	2.7929	467.239	13.963	1,631	11.170
Dur Corrective	1.9790	70.922	9.900	175	7.920

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

5.6 Modelling Uncertainty Using Simulation

The mean value analysis carried out in Section 4.5 gives us an idea of the capacity required to provide quality service in field equipment maintenance. To be able to establish different scenarios in which both workloads for repair personnel and available resources can vary substantially, more information is needed. Now, we turn to simulation to establish probability distributions on R, or total system time, and N, the number of jobs pending.

To carry out this simulation, we take the data from Section 4.2 and determine probability distributions for service times (called Duration) and time between failures (TBF). TBF will determine the job arrival rate. We find that service time has a long-tailed probability distribution, which we model using the Pareto probability, since it is the best fit (see Figures 2 and 3), with a shape parameter, α, of 1.4525 and a scale parameter, A, of 1.2344. The information given for time between arrivals does not have enough granularity to determine a probability distribution, so it will be assumed to be exponential, as is the normal in discrete-time simulations (data was given in days, but actual arrival rate is in minutes).

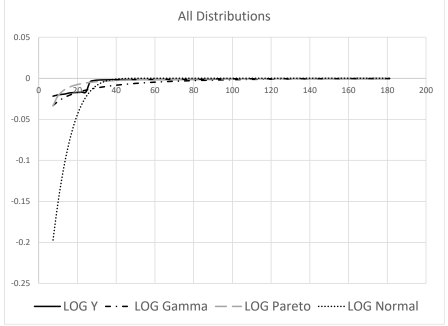

Fig. 2. Fitting different probability distributions to Duration data

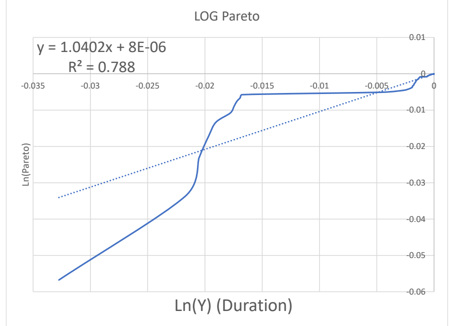

Fig. 3. QQ plot Pareto distribution vs Duration Data

In this simulation, a simplified model was used. The workload was divided between workers. Each worker receives a new job every 5 days, and on average, a worker finishes the job in 3.9623 days. The ratio between arrival rate and service rate is $\rho = 0.7925$. Average simulation results are given in Table 10.

Table 10. Average simulation results

Parameter	Average
U (Utilization)	0.8111
W (R, total wait time)	2,243.58
L (N, pending Jobs on queue)	171.01
Max L (Max N, maximum queue)	1,001

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Max L in Tab. 10 shows that even with a ratio between arrival rate and service rate of $\rho = 0.7925$, which is usually not considered a heavy load, the extreme variability of the service time can create severe bottlenecks. As a result, the probability distribution of the number of jobs in the system has a rather heavy tail, for which the best fit is a Pareto distribution with shape parameter $\alpha = 0.2859$ and scale parameter $A = 1.5$ (see Figs. 4 and 5).

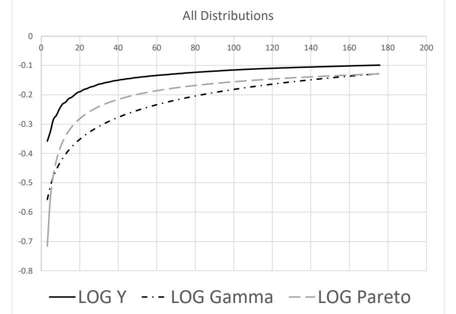

Fig. 4. Fitting different probability distributions to jobs on queue simulation result

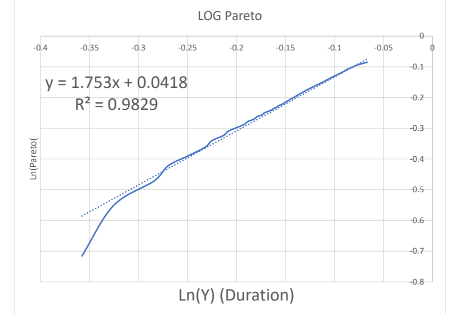

Fig. 5. QQ plot Pareto distribution vs jobs on queue simulation result

Surprisingly, the wait time is not long tailed. The best fit for the total time in the system (wait time R or W) is a Gamma probability distribution (see Figs. 6 and 7).

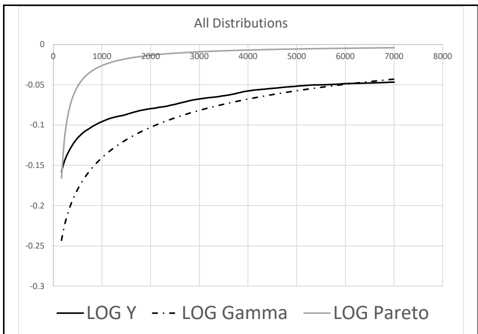

Fig. 6. Fitting different probability distributions to total wait time

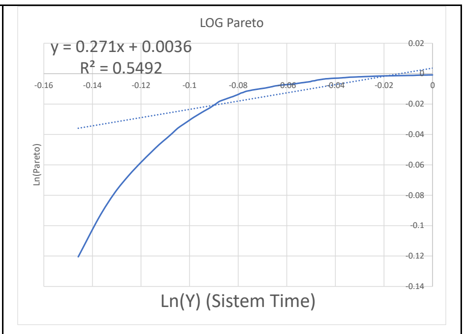

Fig. 7. QQ plot Pareto distribution vs total wait time simulation result

6 Conclusions

Statistical analysis showed that, even though the time between reports or failures for the same equipment follows the well-known exponential decay, the totality of reports duration (all types of reports) and repair cost exhibits a long-tail probability distribution. This is true for both general reports and specific corrective reports, interpreted as equipment failures. This is caused by the fact that the most common value in the three variables is zero.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

Principal component analysis showed that the correlations between the independent variables (50 of them) and the variables of interest, mainly Duration, Total Cost and Time Between Reports/Failures, are not very strong, although not so small that the information is totally random. Principal component analysis provides a roadmap for constructing predictive models.

Following PCA's pointers, predictive models were constructed. The predictions of Total Cost and Time Between Reports/Failures had lower confidence results, but pointed in the right direction. The model for predicting Duration had good predictive performance.

Mean value analysis and simulation confirmed that the job of keeping field equipment is not an easy one, facing severe bottlenecks due to the extreme variability of service times. Nevertheless, the uncertainty models developed can be used to plan for different workload scenarios and resource availability.

7 References

1. R. K. Mobley, An Introduction to Predictive Maintenance, New York City, NY, USA: Elsevier Science, 2002.
2. B. Hoppenstedt, R. Pryss, B. Stelzer, F. Meyer-Brötz, K. Kammere, A. Treß and M. Reichert, "Techniques and emerging trends for state of the art equipment maintenance systems—a bibliometric analysis," Applied Sciences, vol. 8, no. 6, p. 916, 2018.
3. S. A. A. Matin, S. A. Mansouri, M. Bayat, A. R. Jordehi and P. Radmehr, "A multi-objective bi-level optimization framework for dynamic maintenance planning of active distribution networks in the presence of energy storage systems," Journal of Energy Storage, vol. 52, p. 104762, 2022.
4. T. S. Kim, "Model-based performance diagnostics of heavy-duty gas turbines using compressor map adaptation," Applied energy, vol. 212, pp. 1345-1359, 2018.
5. L. Barberá, A. Crespo, P. Viveros and R. Stegmaier, "A case study of GAMM (graphical analysis for maintenance management) in the mining industry," Reliability Engineering & System Safety, vol. 121, p. 113/120, 2014.
6. T. Zonta, C. A. Da Costa, R. d. R. Righi, M. J. de Lima, E. Silveira da Trindade and G. Pyng Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," Computers & Industrial Engineering, vol. 150, p. 106889, 2020.
7. D. C. J. J. T. R. X. G. a. S. K. Wu, "A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests," Journal of Manufacturing Science and Engineering, vol. 139, no. 7, 2017.
8. P. Hundi and S. Rouzbeh , "Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants," Applied Energy, vol. 265, p. 114775, 2020.
9. M. Javaid, A. Haleem, R. P. Singh and R. Suman, "An integrated outlook of Cyber–Physical Systems for Industry 4.0: Topical," Green Technologies and Sustainability, vol. 1, no. 1, p. 100001, 2023.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50

10. J. Lee, H. D. Ardakani, S. Yang and B. Bagheri, "Industrial big data analytics and cyber-physical systems for future maintenance & service innovation," Procedia cirp, vol. 38, pp. 3-7, 2015.
11. Z. Li, J. Hao and C. Gao, "Equipment Maintenance Support Effectiveness Evaluation Based on Improved Generative Adversarial Network and Radial Basis Function Network," Complexity, pp. 1-11, 2021.
12. L. Kirkup and R. B. Frenkel, An introduction to uncertainty in measurement: using the GUM (guide to the expression of uncertainty in measurement), Cambridge University Press, 2006.
13. G. Bammer and M. Smithson, Uncertainty and risk: multidisciplinary perspectives, Routledge, 2012.
14. D. P. Loucks and E. Van Beek, Water resource systems planning and management: An introduction to methods, models, and applications, Springer, 2017.
15. Y. Li, J. Chen and L. Feng, "Dealing with uncertainty: A survey of theories and practices," Transactions on Knowledge and Data Engineering, vol. 25, no. 11, pp. 2463-2482, 2012.
16. K. Klauenberg, S. Martens, A. Bošnjaković, M. G. Cox, A. MH van der Veen and C. Elster, "The GUM perspective on straight-line errors-in-variables regression," Measurement, vol. 187, p. 110340, 2022.
17. C. L. Ilut and S. Martin, "Modeling uncertainty as ambiguity: A review," NATIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge, MA, USA, 2022.
18. G. Papadakos, V. Marinakis, C. Konstas, H. Doukas and A. Papadopoulos, "Managing the uncertainty of the U-value measurement using an auxiliary set along with a thermal camera," Energy and Buildings, vol. 242, p. 110984, 2021.
19. W. Xie, B. L. Nelson, R. R. Barton and K. Wang, "Stochastic Simulation Uncertainty Analysis to Accelerate Flexible Biomanufacturing Process De-," European Journal of Operational Research, p. online preproof, 2023.
20. E. Badakhshan and P. Ball, "A simulation-optimization approach for integrating physical and financial flows in a supply chain under economic uncertainty," Operations Research Perspectives, vol. 10, p. 100270, 2023.
21. M. Moknatian and R. Mukundan, ""Uncertainty analysis of streamflow simulations using multiple objective functions and Bayesian Model Averaging," Journal of Hydrology, vol. 617, p. 128961, 2023.
22. C. E. Papadopoulos and H. Yeung, "Uncertainty estimation and Monte Carlo simulation method," Flow Measurement and Instrumentation, vol. 12, no. 4, pp. 291-298, 2001.
23. I. Jolliffe, Principal component analysis, New York: Springer-Verlang, 2002.
24. G. H. Golub and C. F. Van Loan, Matrix computations, Baltimore: Johns Hopkins University Press, 2012.
25. R. Ramirez-Velarde, L. Hervert-Escobar and N. Hernandez-Gress, "Predictive Analytics with Factor Variance Association," Lecture Notes in Computer Science, vol. 11540, p. 346–359, 2019.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_50