Journal of Quality in Maintenance Engineering

19

Synthetic maintenance data generation for industrial assets based on historic statistical distribution using pseudo-random algorithm

Sebastian Diaz Vivas
Center for Optimization and Applied Probability (COPA), Universidad de Los Andes, Bogotá, Colombia
Rocco Tarantino Alvarado and Sandra Aranguren Zambrano
Facultad de Ingenierías, Arquitectura y Diseño Industrial, Universidad de Pamplona, Pamplona, Colombia, and
Alejandra Tabares Pozos
Center for Optimization and Applied Probability (COPA), Universidad de Los Andes, Bogotá, Colombia

Received 19 March 2025
Revised 27 September 2025
Accepted 5 December 2025

Abstract

Purpose – The article aims to address the challenge of partial or complete absence of maintenance data records for industrial assets by generating synthetic maintenance data under a high-quality maintenance data structure established in the framework of International Organization for Standardization (ISO) 14224:2016. The preceding contributes to maintenance engineering, a strategy to obtain meaningful synthetic data in maintenance management analysis without exposing industrial assets to failures that may lead to undesired consequences.

Design/methodology/approach – The research was conducted under an experimental study aimed at generating synthetic maintenance data from historical statistical distributions of industrial assets. For experimental purposes, based on the criticality of the studied process context, the research was carried out on a centrifugal pump, with its primary data source from the Offshore Reliability Data Handbook (OREDA), from which the four failure modes with the highest failure rate and the non-maintainable components related to the failure rate by probability were selected. The data were processed using Python 3.10.12, using a methodology of standardizing the data structure, for which a pseudo-code was established.

Findings – The article addresses the generation of synthetic maintenance data using historical statistical distributions from the OREDA. Two sets of synthetic data were obtained for a centrifugal pump, with the second set maintaining originality by defining the maximum failure rate as the mean of the global failure rate based on accurate data, demonstrated with an error of 1.96%. This approach allows for objective decision-making when forecasting different scenarios, as the synthetic data set acquires its dynamics dependent on the statistical distribution of the failure rate by failure modes, evidenced by the error in the standard deviation.

Originality/value – The article focuses on generating synthetic maintenance data by developing an algorithm based on internationally recognized statistical distributions aligned with the international standards of ISO 14224:2016. This approach aims to create a synthetic maintenance dataset with maintenance records from which maintenance variables and indicators can be derived. These derived insights enable maintenance optimization through data-driven decision-making feedback loops.

Keywords Synthetic maintenance data, Missing data, Data quality

Paper type Research article

© Sebastian Diaz Vivas, Rocco Tarantino Alvarado, Sandra Aranguren Zambrano and Alejandra Tabares Pozos. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at Link to the terms of the CC BY 4.0 licence.

Journal of Quality in Maintenance Engineering
Vol. 32 No. 5, 2026
pp. 19-33
Emerald Publishing Limited
e-ISSN: 1758-7832
p-ISSN: 1355-2511
DOI 10.1108/JQME-03-2025-0020

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

1. Introduction

Technological evolution has transformed maintenance management into a critical function for industrial competitiveness, leading companies to prioritize strategies that maximize asset performance and operational profitability (Mora, 2009). In this context, data has emerged as a pivotal industrial asset, enabling pivotal insights and informed decision-making through diverse analytical strategies (Merkt, 2019; Bekar *et al*., 2020; Bousdekis *et al*., 2021; Filz *et al*., 2021; Sajid *et al*., 2021; Abbate *et al*., 2022; Cui *et al*., 2022). These data-driven approaches facilitate the iterative optimization of maintenance management via continuous feedback mechanisms (Ciliberti *et al*., 2019; Tarantino, 2021; Diaz *et al*., 2023).

However, a significant obstacle to implementing these advanced analytics is the frequent scarcity or complete absence of high-quality maintenance data records. This often forces a reliance on established, yet inherently limited, maintenance techniques like TPM, RCM, and FMEA (Mora, 2009; International Organization for Standardization, 2016). While foundational, such approaches fall short of the comprehensive, data-driven engineering analysis required for holistic asset management – which is necessary to optimize strategies without compromising assets, production, or safety (Jones, 1995; SAE, International, 2009). This challenge is critical in high-consequence environments where fault impacts and data integration are vital (Hannam, 1997). A further complication involves ensuring data quality and reliability, which demands a well-planned acquisition process that prioritizes essential assets and protects their integrity (Díaz, 2023; Diaz *et al*., 2023; International Organization for Standardization, 2016).

To address this, ISO 14224:2016 outlines a structured framework for data acquisition. Figure 1 illustrates the planning procedure, which involves steps such as determining data acquisition processes, verifying and researching data sources, defining the maintenance data to acquire, setting time, population, and operation parameters, ensuring uniformity in failure definition and classification, and training staff. Subsequently, Figure 2 depicts the acquisition procedure, which includes accessing consequential data, interpreting data, transferring data to a database, and evaluating and analyzing gained insights.

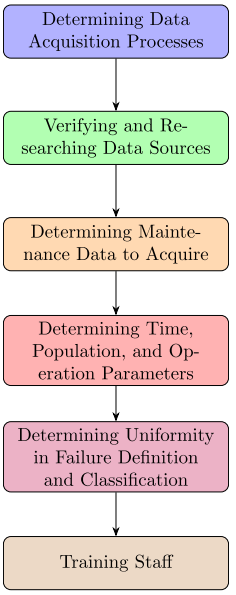

Figure 1. Planning procedure for acquiring quality data according to ISO 14224:2016

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

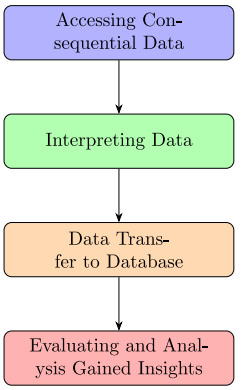

Figure 2. Procedure for acquiring quality data according to ISO 14224:2016

The escalating value of data as a convertible asset for business intelligence underscores the need to overcome inherent data issues, which can be categorized into concerns regarding privacy release and sensitivity, data bias and variance, the need for increased data robustness, and limited or zero data availability (Cole *et al*., 2015; Mannino and Abouzied, 2019; Dankar and Ibrahim, 2021; Jordon *et al*., 2022). Synthetic data emerges as a viable solution to these challenges, defined as artificially generated data that acquires the statistical and phenomenological properties of a real dataset (Dankar and Ibrahim, 2021; Jordon *et al*., 2022). Methodological advancements highlight that synthetic datasets vary in fitness for purpose, necessitating tailored evaluation protocols and practical metrics to enhance reproducibility (Lautrup *et al*., 2024; Giuffrè and Shung, 2023).

Synthetic data offers significant benefits, including privacy protection through the absence of personally identifiable information, improved accessibility by creating larger datasets, flexibility to replicate specific statistical characteristics, and the potential to enhance original data quality (El Emam *et al*., 2020). Various generation techniques exist, such as those based on Bayesian networks, copulas, parametric fitting (e.g. Monte Carlo methods), and non-parametric trees (Mannino and Abouzied, 2019; Dankar and Ibrahim, 2021; El Emam *et al*., 2020; Li *et al*., 2020; Jordon *et al*., 2022; Okagbue *et al*., 2020).

Within reliable maintenance engineering, synthetic data generation employs diverse strategies, including machine learning, such as generative adversarial networks (GANs) and mathematical functions, maintaining statistical fidelity (Lakshmanan *et al*., 2023; Martínez-Heredia and Ventura, 2025).. Contemporary research trends encompass digital twin frameworks for physics-informed data streams and Bayesian methods for uncertainty-aware reliability estimates, providing a context where a parameterized pseudo-random generator offers a reproducible baseline for transparent benchmarking (Zio and Miqueles, 2024; Liu *et al*., 2024; Pan *et al*., 2024; Zheng *et al*., 2024). To address the challenge of absent maintenance records, this work develops a pseudo-random algorithm grounded in the high-quality data attributes prescribed by ISO 14224:2016 (Diaz *et al*., 2023; Díaz, 2023) and leveraging statistical distributions from the OREDA database (SINTEF, 2009), where failure rates follow a gamma distribution. This methodology ensures the generation of valid and practically useful synthetic maintenance data for industrial assets.

Applied to centrifugal pumps in an operational chemical engineering setting —specifically, the ethanol-water separation process at the University of Pamplona — our approach generated two distinct datasets ($\mathbb{X}_1$ and $\mathbb{X}_2$). Dataset $\mathbb{X}_2$ demonstrated strong alignment with OREDA's mean failure rate (exhibiting only a 1.96% error), validating its statistical fidelity. Despite higher deviations in standard deviation and maximum values, both datasets yield realistic

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

maintenance records suitable for engineering analysis and decision-making without jeopardizing actual assets or operations.

2. Methodology

The generation of synthetic maintenance data requires a robust statistical foundation and a structured methodology. This study utilizes the OREDA database as its primary source for failure mode distributions, providing historically validated reliability data from major petrochemical companies (SINTEF, 2009). OREDA's structure organizes failure data by asset taxonomy, with statistical distributions categorized by failure mode severity (critical, degraded and incipient) and including key metrics such as failure rates, active repair hours, and man-hours.

For this research, the critical risk state was selected, with failure rates following a gamma distribution as confirmed by Kolmogorov-Smirnov and χ^2 goodness-of-fit tests (ibid.). The probability density function is given by:

$$f(x) = \frac{\frac{\theta}{\hat{\sigma}}^{\left(\frac{\theta^2}{\hat{\sigma}}\right)}}{\Gamma\left(\frac{\theta^2}{\hat{\sigma}}\right)} x^{\left(\frac{\theta^2}{\hat{\sigma}}-1\right)} e^{\left(-\frac{\theta}{\hat{\sigma}}x\right)}, \tag{1}$$

and its cumulative distribution function by:

$$F(x) = \frac{\gamma\left(\frac{\theta^2}{\hat{\sigma}}, \frac{\theta}{\hat{\sigma}}x\right)}{\Gamma\left(\frac{\theta^2}{\hat{\sigma}}\right)}, \tag{2}$$

where θ^* represents the mean failure rate and $\hat{\sigma}$ the standard deviation.

The methodology also incorporates probability distributions for failure modes relative to non-maintainable components, ensuring the total probability sums to 100% as expressed by:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} X_{ji} = 100\%, \tag{3}$$

where j denotes failure modes, i non-maintainable components, and X_{ji} the occurrence probability.

The synthetic data generation workflow, depicted in Figure 3, is implemented through Algorithm 1. This algorithm details the modular steps from statistical initialization to dataset cleaning, ensuring reproducibility and transparency.

Algorithm 1. Synthetic Data Generation for Maintenance Asset Records

1: **Function** SyntheticDataGeneration()

2: **Input:** Failure rate distribution parameters, component costs, maintenance parameters

3: **Output:** Synthetic maintenance dataset including Failure Mode, Component, TBF, Maintenance Cost, etc.

4: # Step 1: Initialization

5: Define constants and distributions (e.g. mean and standard deviation for failure rates)

6: Initialize data structures for storing generated records

7: # Step 2: Data Sampling

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

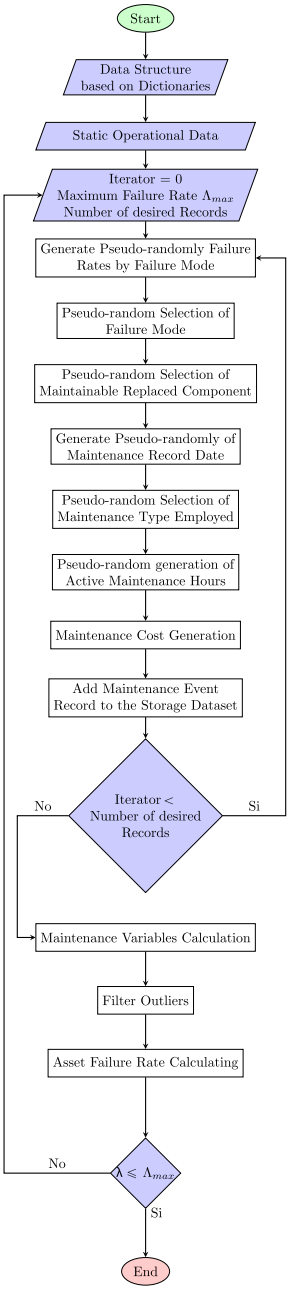

Figure 3. Methodology for synthetic maintenance data generation

8: Generate random values $[r_1, r_2, r_3, r_4, r_5]$ from $N(0, 1)$

9: # Step 3: Failure Mode and Component Selection

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

10: Sample Failure Mode and Component using probabilistic sampling with r_1

11: Sample Failure Description using probabilistic sampling with r_2

12: # Step 4: Failure Rate and TBF

13: Compute parameters $\beta = \frac{\mu}{\sigma}$ and $\alpha = \beta \cdot \mu$ based on failure rate distributions

14: Calculate Failure Rate $\lambda = F_X^{-1}(r_3, \beta, \alpha)$

15: Calculate $TBF = -\ln\left(\frac{r_4}{\lambda}\right)$

16: # Step 5: Dataset Filtering and Adjustments

17: **while** $\frac{1}{\lambda_{\max}} \leq TBF \leq \lambda_{\min}$ **do**

18: Filter dataset records based on *TBF* constraints

19: **end while**

20: Select subset n_2 from n_1 records

21: # Step 6: Calculate Maintenance Costs

22: Calculate Failure Date = Start Operation Date + TBF

23: Sort the dataset by Failure Date in ascending order

24: Calculate ManHours = $F_X^{-1}(r_5, \mu, \sigma)$ based on manhour distribution

25: Compute Maintenance Cost = Component Cost[Component] + (ManHours × ManHour Cost)

26: Update *TBF* = Current Failure Date - Last Failure Date (in hours)

27: Calculate Downtime = ManHours + (1 + Average Administrative Time Percentage)

28: # Step 7: Repeat and Clean Dataset

29: **while** Additional Filtering Needed **do**

30: Reapply filters to the dataset

31: **end while**

32: Clean dataset to finalize output

33: **Return** Synthetic maintenance dataset

Initialization: The process begins with defining the key statistical distributions and parameters, such as the mean and variance for failure rates and component costs. These parameters will shape the random variables generated later in the process. Next, data structures are initialized to store records that will eventually contain attributes such as failure modes, components, Time Between Failures (TBF), and maintenance costs.

Sampling Random Values: Once initialization is complete, random values are generated to simulate various aspects of the maintenance process. Values such as r_1, r_2, r_3, r_4, and r_5 are drawn from a standard normal distribution, $N(0, 1)$, which will later be used in probabilistic sampling.

Failure Mode and Component Selection: Using the generated random values, failure modes and components are selected for each record. The value r_1 is applied within cumulative probability distributions for failure modes and components, allowing for probabilistic sampling. Similarly, r_2 is used to choose a failure description based on the cumulative probabilities associated with each failure mode.

Compute Failure Rate and TBF: After selecting the failure mode and component, the next step is to compute the failure rate (λ) and TBF for each record. Parameters β and α are

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

calculated based on the mean and variance of the failure rate distributions. Using these parameters along with r_3, the failure rate λ is determined by inverting the cumulative distribution function, $F_X^{-1}(r_3, \beta, \alpha)$. The TBF is then calculated as $TBF = -\ln\left(\frac{r_4}{\lambda}\right)$.

Dataset Filtering and Adjustments: With preliminary data generated, the dataset undergoes filtering based on the TBF values to ensure that records meet realistic operational limits. Only TBF values that fall within a specified range (between $\frac{1}{\lambda_{\max}}$ and $\lambda_{\min}$) are kept. This step ensures that the synthetic data aligns with practical operational expectations.

Calculate Maintenance Costs and Downtime: The maintenance cost and downtime associated with each failure are then calculated. The failure date is determined by adding TBF to the start operation date. Man-hours are computed using distribution-based calculations. Maintenance cost is calculated by combining the component cost with the product of man-hours and the man-hour cost rate. Downtime is calculated by adding man-hours and an additional administrative time percentage.

Final Filtering and Dataset Cleaning: After calculating costs and downtime, the dataset undergoes a final filtering and cleaning stage. Additional filters are reapplied if necessary to ensure that all records meet the predefined conditions. The dataset is then organized and prepared for output.

This structured process yields a synthetic maintenance dataset with realistic attributes, including varied failure modes, components, times between failures, costs, and downtimes, based on statistically grounded parameters and distributions.

3. Case study and results

3.1 Operational context

The selected study field for the research was the chemical engineering laboratory at the University of Pamplona, whereby the ethanol-water mixture separation subprocess was selected, focusing on a plate column with liquid feed, direct steam injection, and the possibility of having the top product as vapor or liquid through a condenser that can be used as total or partial condenser. The column feed can come from a feed tank or the rectification column (see Figure 4).

Based on an analysis conducted during the research, the assets with high criticality involved in the subprocess are the centrifugal pumps $P-400$ and $P-405$, located in the P&ID of Figure 4.

Similarly, four failure modes were determined for the study that are common in the mentioned assets: External Leak (ELU), Abnormal Instrument Reading (AIR), Vibration (VIB), and Breakdown (BRD) [1].

For this reason, a dataset of synthetic maintenance data for centrifugal pumps with four failure modes was generated based on the methodology for generating synthetic maintenance data and the constructed programming algorithms.

3.2 Statistical information characterization

The OREDA provided statistical information based on statistical and percentage distributions per failure mode and non-maintainable components (SINTEF, 2009, pp.138-145).

To satisfy (2) so that the data resembles actual behavior, it is determined, from the statistical distribution of the overall failure rate found in the OREDA during critical phases of the asset, (4). It is worth noting that the values found are failure rates per million hours (10^6 h).

$$\lambda \in [\Lambda_{\min}; \Lambda_{\max}]: \quad \begin{aligned} \Lambda_{\min} &= 3\times10^{-4}, \\ \Lambda_{\max} &= 136.71, \\ \mu &= 28.08, \\ \sigma &= 56.95. \end{aligned} \tag{4}$$

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

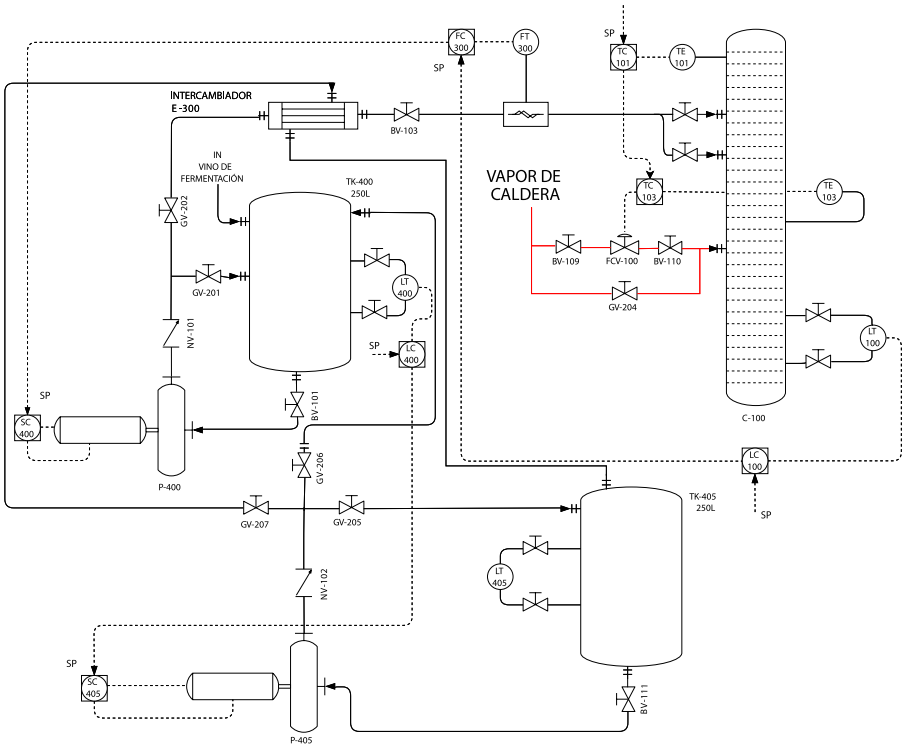

Figure 4. Process P&ID

Under the table structure, these distributions were characterized, allowing for the definition of the data structures proposed in the algorithms presented in this article.

In Table 1, the statistical distribution of failure rates about the selected failure modes for the asset is observed.

Table 2 shows the probability of a failure mode caused by a non-maintainable asset component.

Based on a market analysis, the cost associated with replacing each non-maintainable component of the asset has been defined in Colombian pesos (COP). The prices of the components are expressed in monetary units (see Table 3).

Finally, Table 4 defines statistical information regarding the active maintenance hours required per failure mode of the asset.

Table 1. Statistical distribution of failure rate by failure mode

Failure mode	min	μ	max	σ
BRD	0.51	2.45	5.59	1.63
VIB	0.51	3.82	9.68	3.0
ELU	0.0	3.24	11.62	15.60
AIR	0.0	0.31	1.72	0.89

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

Table 2. Occurrence probability of failure mode per non-maintainable component

Failure mode	Non-maintainable component	Failure probability [0; 1]
BRD	Bearing	0
BRD	Cabling and junction boxes	0
BRD	Casing	0.41
BRD	Control unit	0
...	...	...
AIR	Wiring	0.89
AIR	Valves	0

Table 3. Associated costs per non-maintainable component

Component	Unit price (COP)
Bearing	150,000.00
Cabling and junction boxes	100,000.00
Casing	50,000.00
Control unit	300,000.00
...	...
Wiring	30,000.00
Valves	40,000.00

Table 4. Active maintenance hours

Failure mode	μ	max
BRD	15	30
VIB	27	77
ELU	15	45
AIR	44	44

3.3 Synthetic maintenance dataset obtained

Two synthetic maintenance datasets were generated for the centrifugal pump under study based on the four identified failure modes and the provided statistical information. The datasets include maintenance records with relevant and non-redundant maintenance data, as required by ISO 14224:2016. These two data sets will be called $\mathbb{X}_1$ and $\mathbb{X}_2$, respectively. The online Supplementary Material includes Dataset $\mathbb{X}_1$ and Dataset $\mathbb{X}_2$.

These two data sets were generated with different statistical properties. Data set $\mathbb{X}_1$ behaves in such a way that it satisfies the ranges of the observed set in (4), while data set $\mathbb{X}_2$ follows the behavior observed in (5). Despite the latter, data set $\mathbb{X}_2$ retains statistical characteristics that can be considered as actual for synthetic data, considering that it also complies with (4).

$$\lambda \in [\Lambda_{\min}; \Lambda_{\max}]: \quad \begin{aligned} \Lambda_{\min} &= 3\times10^{-4}, \\ \Lambda_{\max} &= \mu, \\ \mu &= 28.08, \\ \sigma &= 56.95. \end{aligned} \tag{5}$$

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

During the code execution, technological limitations were encountered that made it impossible to increase the expected number of maintenance records (K) without affecting the expected statistical properties, thus obtaining the data sets described by (6) and (7).

$$\mathbb{X}_1 = \left\{x_{i,j} |\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, K;\ 2 \leq K \leq 140\right\} \tag{6}$$

$$\mathbb{X}_2 = \left\{x_{i,j} |\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, K;\ 2 \leq K \leq 30\right\} \tag{7}$$

When the variable filtering process is applied to the maintenance data sets as seen in (6) and (7), the synthetic maintenance data sets with similar statistical and business characteristics are obtained (see 2, 4, 8, and 9).

$$\mathbb{H}_1 = \left\{x_{i,j} |\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m;\ m \leq K\right\} \tag{8}$$

$$\mathbb{H}_2 = \left\{x_{i,j} |\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m;\ m \leq K\right\} \tag{9}$$

Table 5 presents a representative subset of the records generated for the dataset $\mathbb{X}_1$. This subset illustrates different maintenance records, whether corrective or preventive, with all the corresponding data for each record.

One hundred and seventeen (117) filtered maintenance synthetic data records were obtained for dataset $\mathbb{X}_1$. Meanwhile, twenty-six (26) filtered maintenance synthetic data records were obtained for dataset $\mathbb{X}_2$.

The statistical comparison between the synthetic datasets and the OREDA benchmark is summarized in Table 6, which presents key failure rate metrics. Subsequently, the Time Between Failures (TBF) distributions are visualized in Figures 5–7.

Figure 5 depicts the gamma distribution of TBF derived from OREDA data, serving as the reference for synthetic data generation. Figure 6 shows the TBF distribution for dataset $\mathbb{X}_1$, which exhibits a higher mean failure rate and distinct spread compared to OREDA. Figure 7 presents the TBF distribution for dataset $\mathbb{X}_2$, which aligns closely with OREDA's mean.

As observed in Table 6, dataset $\mathbb{X}_2$ demonstrates better alignment with OREDA's mean failure rate. The percentage errors for the mean ($\%_\mu$), standard deviation ($\%_\sigma$), and maximum value ($\%_{\max}$) are calculated as follows (Montgomery, 2004):

$$\%_\mu = \left|\frac{28.08 - 27.53}{28.08}\right| \cdot 100\% = 1.96\%. \tag{10}$$

$$\%_\sigma = \left|\frac{56.95 - 13.72}{56.95}\right| \cdot 100\% = 75.91\%. \tag{11}$$

$$\%_{\max} = \left|\frac{136.71 - 536.80}{136.71}\right| \cdot 100\% = 292.66\%. \tag{12}$$

4. Conclusion

This paper established a methodology and developed pseudo-code algorithms to generate synthetic maintenance data for industrial assets. By implementing pseudo-random functions within statistical, probabilistic, and exponential-variate distributions based on the OREDA, the approach successfully produces realistic maintenance records.

The results demonstrate the method's practical value. As evidenced in Table 6, the $\mathbb{X}_2$ dataset shows a notably strong performance, with its mean value exhibiting a minimal error of

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

Table 5. Subset of synthetic maintenance data generated for centrifugal pump

	Registration date	Maintenance Type	Replaced component	Failure mode	Active maintenance hours	Costs	UT	DT	TTR	TBF	OT
0	2016–08-27 15:30:11.141296	Corrective	Instrument, pressure	AIR	44.0	405752.0	3273.451392	44.0	44.0	3273.451392	3317.451392
1	2029–10-23 15:34:54.101196	Preventive	Instrument, flow	AIR	44.0	395752.0	1536.894218	44.0	44.0	1536.894218	1580.894218
2	2078–10-21 19:13:21.852802	Corrective	Instrument, flow	AIR	44.0	395752.0	8748.577526	44.0	44.0	8748.577526	8792.577526
3	2085–02-13 11:29:20.306991	Preventive	Instrument, pressure	AIR	44.0	405752.0	168077.471786	44.0	44.0	168077.471786	168121.471786

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

Table 6. Global failure rate (per 10^6 *h*)

Dataset	Min	μ	σ	Max
OREDA	3×10^{-4}	28.08	56.95	136.71
$\mathbb{X}_1$	5.70	129.20	38.08	84388.18
$\mathbb{X}_2$	2.74	27.53	13.72	536.80

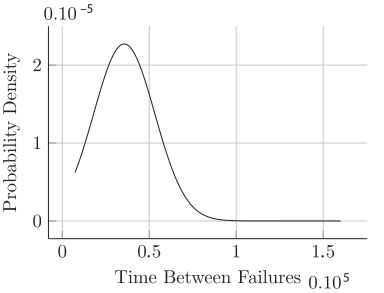

Figure 5. Statistical distribution of TBF for the asset

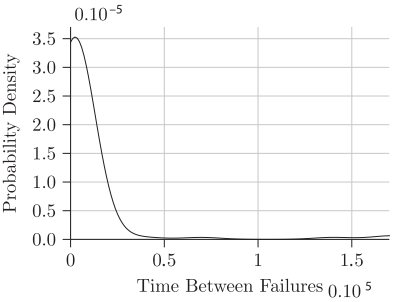

Figure 6. Statistical distribution of TBF for the asset obtained from Synthetic Maintenance Dataset $\mathbb{X}_1$

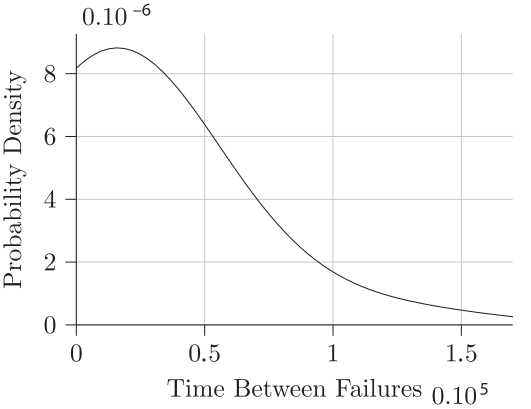

Figure 7. Statistical Distribution of TBF for the asset obtained from Synthetic Maintenance Dataset $\mathbb{X}_2$

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

just 1.96% compared to the OREDA distribution (10). This high degree of accuracy in replicating the central tendency underscores the model's effectiveness. While larger errors were observed for the standard deviation (75.91%) and the maximum global failure rate (292.66%), these are not inherently unfavorable. Instead, they reflect the stochastic nature of the pseudo-random generation across multiple failure modes, which imbues each synthetic dataset with a unique variability from which valuable maintenance engineering insights can be extracted.

Ultimately, this work provides a robust, low-risk tool for simulation and decision-making in asset management. The generated datasets, exemplified by the centrifugal pump category in Table 5, possess statistical and behavioral fidelity to real-world data. This enables organizations to test maintenance strategies, train models, and plan asset lifecycles efficiently and safely—without jeopardizing operational integrity, production, safety, or the environment. Thus, the research bridges a critical gap between theory and practice by leveraging synthetic maintenance data generation as a foundational seed for advanced applications, such as digital twins and machine learning models, which facilitates the incorporation of these and other data-driven tools into decision-making processes to continuously improve maintenance management.

Acknowledgments

This work is supported by University of Pamplona and University of the Andes, Colombia.

Note

1. Although four failure modes were determined, this does not mean that more failure modes cannot be considered.

Supplementary material

The supplementary material for this article can be found online.

References

Abbate, R., Caterino, M., Fera, M. and Caputo, F. (2022), "Maintenance digital twin using vibration data", *Procedia Computer Science*, Vol. 200, pp. 546-555, ISSN: 18770509, doi: 10.1016/j.procs.2022.01.252.

Bekar, E.T., Nyqvist, P. and Skoogh, A. (2020), "An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study", *Advances in Mechanical Engineering*, Vol. 12 No. 5, 168781402091920, ISSN: 1687-8140, doi: 10.1177/1687814020919207.

Bousdekis, A., Lepenioti, K., Apostolou, D. and Mentzas, G. (2021), "A review of data-driven decision making methods for industry 4.0 maintenance applications", *Electronics*, Vol. 10 No. 7, 828, doi: 10.3390/electronics10070828.

Ciliberti, V.A. , Østebø, R., Selvik, J.T. and Alhanati, F.J.S. (2019), "D041S055R003, Optimize safety and profitability by use of the ISO 14224 standard and big data analytics", doi: 10.4043/29634-MS.

Cole, D., Nelson, J. and McDaniel, B. (2015), "Benefits and risks of big data", *SAIS 2015*.

Cui, P.-H., Wang, J.-Q. and Yang, Li (2022), "Data-driven modelling, analysis and improvement of multistage production systems with predictive maintenance and product quality", *International Journal of Production Research*, Vol. 60 No. 22, pp. 6848-6865, ISSN: 0020-7543, doi: 10.1080/00207543.2021.1962558.

Dankar, F.K. and Ibrahim, M. (2021), "Fake it till you make it: guidelines for effective synthetic data generation", *Applied Sciences*, Vol. 11 No. 5, 2158, ISSN: 2076-3417, doi: 10.3390/app11052158.

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

Díaz, S. (2023), Metodología de Análisis de Datos de Mantenimiento en la Industria Aplicando la Norma ISO 14224:2016 mediante el Uso de Ciencia de Datos y Machine Learning en el Contexto del Metamantenimiento, Tesis de Pregrado. Universidad de Pamplona.

Diaz, S., Tarantino, R. and Aranguren, S. (2023), "Metodología de Análisis de Datos aplicado al Metamantenimiento Industrial", *Congreso Internacional de Electrónica y Tecnologías de Avanzada 16*, Universidad de Pamplona.

El Emam, K., Mosquera, L. and Hoptroff, R. (2020), in Hassell, J., Collins, C. and Faucher, C. (Eds), *Practical Synthetic Data Generation*, 1st ed., O'Reilly Media.

Filz, M.-A., Langner, J.E.B., Herrmann, C. and Thiede, S. (2021), "Data-driven failure mode and effect analysis (FMEA) to enhance maintenance planning", *Computers in Industry*, Vol. 129, 103451, ISSN: 01663615, doi: 10.1016/j.compind.2021.103451.

Giuffrè, M. and Shung, D.L. (2023), "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy", *Npj Digital Medicine*, Vol. 6 No. 1, p. 186, doi: 10.1038/s41746-023-00927-3.

Hannam, R. (1997), *Computer Integrated Manufacturing: From Concepts to Realisation*, 1st ed., Addison Wesley Longman, Harlow.

International Organization for Standardization (2016), ISO 14224:2016.

Jones, R.B. (1995), *Risk-Based Management: A Reliability-Centered Approach*, Gulf Publishing, Houston, TX.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. and Weller, A. (2022), *Synthetic Data - What, Why and How?*, The Alan Turing Institute.

Lakshmanan, K., Tessicini, F., Gil, A.J. and Auricchio, F. (2023), "A fault prognosis strategy for an external gear pump using machine learning algorithms and synthetic data generation methods", *Applied Mathematical Modelling*, Vol. 123, pp. 348-372, ISSN: 0307904X, doi: 10.1016/j.apm.2023.07.001.

Lautrup, A.D., Hyrup, T., Zimek, A. and Schneider-Kamp, P. (2024), "Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data", *ACM Computing Surveys*, Vol. 57 No. 4, pp. 1-38, doi: 10.1145/3704437.

Li, Z., Yue, Z. and Fu, J. (2020), "SynC: a copula based framework for generating synthetic data from aggregated sources". In: IEEE, pp. 571-578. ISBN: 978-1-7281-9012-9, doi: 10.1109/ICDMW51313.2020.00082.

Liu, Y., Feng, J., Lu, J. and Zhou, S. (2024), "A review of digital twin capabilities, technologies, and applications based on the maturity model", *Advanced Engineering Informatics*, Vol. 62, 102592, doi: 10.1016/j.aei.2024.102592.

Mannino, M. and Abouzied, A. (2019), "Is this real? Generating synthetic data that looks real". In: ACM, pp. 549-561. ISBN: 9781450368162, doi: 10.1145/3332165.3347866.

Martínez-Heredia, A.M. and Ventura, S. (2025), "Weak supervision: a survey on predictive maintenance", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 15 No. 2, e70022, doi: 10.1002/widm.70022.

Merkt, O. (2019), "On the use of predictive models for improving the quality of industrial maintenance: an analytical literature review of maintenance strategies", Vol. 18, pp. 693-704, doi: 10.15439/2019F101.

Montgomery, D. (2004), *Diseño y Análisis de Experimentos*, 2nd ed., Limusa Wiley, Mexico D.F.

Mora, L. (2009), *Mantenimiento. Planeación, Ejecución Y Control*, 1st ed., Alfaomega Grupo Editor, Mexico D.F.

Okagbue, H., Adamu, M.O. and Anake, T.A. (2020), "Approximations for the inverse cumulative distribution function of the gamma distribution used in wireless communication", *Heliyon*, Vol. 6 No. 11, e05523, ISSN: 24058440, doi: 10.1016/j.heliyon.2020.e05523.

Pan, J., Sun, B., Wu, Z., Yi, Z., Feng, Q., Ren, Y. and Wang, Z. (2024), "Probabilistic remaining useful life prediction without lifetime labels: a Bayesian deep learning and stochastic process fusion method", *Reliability Engineering and System Safety*, Vol. 250, 110313, doi: 10.1016/j.ress.2024.110313.

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026

SAE, International (2009), SAE, international JA1011, evaluation criteria for reliability-centered maintenance RCM, Standard. SAE International.

Sajid, S., Haleem, A., Bahl, S., Javaid, M., Goyal, T. and Mittal, M. (2021), "Data science applications for predictive maintenance and materials science in context to Industry 4.0", *Materials Today: Proceedings*, Vol. 45, pp. 4898-4905, ISSN: 22147853, doi: 10.1016/j.matpr.2021.01.357.

SINTEF (2009), *OREDA: Offshore Reliability Data Handbook*, 5th ed., Vol. 1, OREDA Participants, Trondheim, 978-82-14-04830-8.

Tarantino, R. (2021), Metamantenimiento: Una Propuesta para Incrementar la Confiabilidad en los Procesos Industriales.

Zheng, X. , Yao, W., Xu, Y. and Wang, N. (2024), "Algorithms for Bayesian network modeling and reliability inference of complex multistate systems with common cause failure", In: *Reliability Engineering and System Safety*, Vol. 241, 109663, doi: 10.1016/j.ress.2023.109663.

Zio, E. and Miqueles, L. (2024), "Digital Twins in safety analysis, risk assessment and emergency management", *Reliability Engineering and System Safety*, Vol. 246, 110040, doi: 10.1016/j.ress.2024.110040.

Corresponding author
Sebastian Diaz Vivas can be contacted at: s.diazv2@uniandes.edu.co

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Downloaded from http://ftp.nowpublishers.com/jqme/article-pdf/32/5/19/11210960/jqme-03-2025-0020en.pdf by guest on 11 June 2026