Boosting field data using synthetic SCADA datasets for wind turbine condition monitoring

Milani, Ali Eftekhari; Zappalá, Donatella; Castellani, Francesco; Watson, Simon

DOI
10.1088/1742-6596/2767/3/032033

Publication date
2024

Document Version
Final published version

Published in
Journal of Physics: Conference Series

Boosting field data using synthetic SCADA datasets for wind turbine condition monitoring

Boosting field data using synthetic SCADA datasets for wind turbine condition monitoring

Ali Eftekhari Milani[1], Donatella Zappalá[1], Francesco Castellani[2], Simon Watson[1]

[1]TU Delft, Kluyverweg 1, Delft, 2629 HS, Netherlands
[2]University of Perugia, Department of Engineering, Via Duranti, 06125, Perugia, Italy

E-mail: a.eftekharimilani@tudelft.nl, d.zappala@tudelft.nl, francesco.castellani@unipg.it, s.j.watson@tudelft.nl

Abstract. State-of-the-art Deep Learning (DL) methods based on Supervisory Control and Data Acquisition (SCADA) system data for the detection and prognosis of wind turbine faults require large amounts of failure data for successful training and generalisation, which are generally not available. This limitation prevents benefiting from the superior performance of these methods, especially in SCADA-based failure prognosis. Data augmentation approaches have been proposed in the literature for generating failure data instances within a SCADA sequence to reduce the imbalance between healthy and faulty state data points, which is relevant to fault detection tasks. However, the successful implementation of DL-based failure prognosis methods requires the availability of multiple run-to-failure SCADA sequences. This paper proposes a data-driven method for generating synthetic run-to-failure SCADA sequences with custom operational and environmental conditions and progression of degradation. An Artificial Neural Network (ANN) is trained with signals that represent these factors to reconstruct the SCADA signals. Then, it is used to generate synthetic SCADA datasets based on data available from a wind turbine that experienced a gearbox failure. Synthetic data sets generated are evaluated on the basis of the similarity of their signal distributions, the temporal dynamics within each signal, and the temporal dynamics among different SCADA signals with those in similar field datasets. The results show that the generated synthetic datasets are consistent with their field counterparts, with a comparatively lower diversity in their dynamic behaviour in time.

1. Introduction

State-of-the-art Deep Learning (DL) methods for SCADA-based fault detection and prognosis in wind turbines require a large amount of high-quality faulty-state training data for satisfactory performance [1], which is usually not available. In SCADA datasets of wind turbines experiencing failures, the time periods related to faulty operation are generally significantly shorter in length than those of healthy operation. During fault detection, this imbalance introduces a bias towards the majority (healthy) class, resulting in poor detection performance. Furthermore, a specific component usually fails very few times in a wind farm, leading to an insufficient number of run-to-failure sequences, which can negatively impact the performance of failure prognosis methods. Another reason for this limitation can be the reluctance of wind farm operators to share potentially sensitive failure data [1].

The problem of class imbalance in classification is usually tackled by data augmentation, i.e. oversampling of the minority class. Data augmentation for SCADA-based wind turbine fault detection has been well studied in the literature, such as in [1, 2]. However, there is a significant research gap regarding the generation of synthetic SCADA datasets, which can be a viable solution to mitigate the lack of sufficient run-to-failure SCADA sequences.

An approach to generating synthetic SCADA signals is to use physics-based and hybrid (combining both physics-based and data-driven) models of wind turbine components. For example, in [3], a digital twin based on a hybrid model of a wind turbine drivetrain is developed and used to generate synthetic stator winding temperature signals. The generator failure is modelled as a heat exchanger, simulating elevated stator winding temperatures, and the generated signals are used for fault detection. However, such models always introduce some simplifications in the description of the actual component behaviour. Furthermore, they have limited capacity to model component degradation and failure. As a result, they are unable to generate realistic run-to-failure sequences of multiple SCADA signals useful for prognostic purposes.

Data-driven methods for generating synthetic time series have been proposed in other fields. In [4], a method based on Generative Adversarial Networks (GAN) is developed to generate synthetic time series for smart grid applications. In [5], a survey of time series data generation methods in the field of the Internet of Things is presented. In [6], a GAN-based method is developed to generate synthetic financial time series. However, to the best of the authors' knowledge, data-driven methods for creating synthetic SCADA datasets to mitigate the lack of sufficient run-to-failure data sequences have not been proposed yet. The methods already proposed mainly aim to learn the underlying distribution and dynamic behaviour in time of a known dataset and to sample new time series from the learnt distribution, generating synthetic time series that closely follow the behaviour and distribution of the original data [7]. Therefore, based on very few field datasets with failure, they are not capable of generating a diverse set of synthetic run-to-failure SCADA data sequences. In this paper, a data-driven method is developed to address this limitation. This method allows the customisation of the operational, environmental, and degradation behaviours under which the synthetic SCADA datasets are generated.

The rest of the paper is organised as follows: Section 2 introduces the developed method and the metrics used for its validation. In section 3, the method is applied to the SCADA dataset of a wind turbine with a gearbox failure, and the consistency of the generated synthetic datasets with field datasets is evaluated in terms of the introduced metrics. Section 4 draws conclusions regarding the validity of the method and the ongoing work for assessing its effectiveness in boosting condition monitoring performance.

2. Method

This section first describes the method developed for generating synthetic SCADA datasets. Then, the metrics and the approach to evaluate the consistency between the synthetic and field datasets are discussed.

2.1. Synthetic SCADA dataset generation

Sensor signals of degrading industrial components are affected by different factors, including the operational and environmental conditions and the level of degradation [8]. Likewise, a wind turbine SCADA dataset consists of operation- and environment-dependant signals, such as the rotor speed and the wind speed, and signals which depend on those along with the level of degradation, such as gearbox bearing temperature. Any such signal S_i can be expressed as:

$$S_i = \{s_{i,t} | t \in [0, T]\} = \mathcal{F}_i(\boldsymbol{O}, \boldsymbol{E}, D) \tag{1}$$

where $\boldsymbol{O} = \{O_j | j \in [1, C_O]\} = \{o_{j,t} | j \in [1, C_O], t \in [0, T]\}$ and $\boldsymbol{E} = \{E_k | k \in [1, C_E]\} = \{e_{k,t} | k \in [1, C_E], t \in [0, T]\}$ denote the operation- and environment-dependant signals with C_O and C_E indicating the number of such signals in the SCADA dataset, $D = \{d(t)\}$ is the degradation factor which is not directly measured, $\mathcal{F}_i$ is a function that describes the relationship between these factors and S_i, t indicates time, and T is the length of the dataset. The SCADA dataset $\boldsymbol{S}$ can, then, be denoted as:

$$\boldsymbol{S} = \{S_i | i \in [1, C_S]\} \cup \{\boldsymbol{O}, \boldsymbol{E}\} \tag{2}$$

where C_S is the number of signals which depend on the operational and environmental conditions and the level of degradation. Knowing $\mathcal{F}_i$, a synthetic SCADA signal, S_i^s, can be generated by any set of synthetic $\boldsymbol{O}$, $\boldsymbol{E}$, and D.

$$S_i^s = \mathcal{F}_i(\boldsymbol{O}^s, \boldsymbol{E}^s, D^s) \tag{3}$$

where $\boldsymbol{O}^s$, $\boldsymbol{E}^s$, and D^s are synthetic operational, environmental, and degradation factors.

If $\boldsymbol{O}$, $\boldsymbol{E}$, and D are known, a way to learn $\mathcal{F}_i$ is to use a supervised machine learning approach, such as an Artificial Neural Network (ANN). In SCADA datasets, several signals usually exist that represent operational and environmental conditions, such as rotor speed, ambient temperature, and wind speed. However, the level of degradation is not directly measurable by sensors. Since degradation is, to a large extent, an irreversible process, D is expected to be highly monotonic [8]. In [9], a Convolutional Autoencoder (CAE) trained by the Particle Swarm Optimisation (PSO) algorithm has been proposed to extract D from the signals of a degrading component. The CAE receives input as tabular data in Equation (2) and reconstructs them in the output, passing the information through a bottleneck in the middle layer where the degradation factor is extracted. The CAE is trained using the PSO algorithm to minimise the reconstruction error and maximise the monotonicity of the degradation factor D obtained in its middle layer.

The flowchart of the data-driven method proposed to generate synthetic SCADA datasets is shown in Figure 1. The D factor is extracted from a given dataset, $\boldsymbol{S}$, using the CAE-PSO algorithm, and is input, along with the already known $\boldsymbol{O}$ and $\boldsymbol{E}$, to an ANN, that is trained to reconstruct S_i, implicitly modelling $\mathcal{F}_i$ (Figure 1 - Training step). The trained ANNs are then used as a generative model (Generator ANN) to generate synthetic SCADA signals S_i^s in the generation step (Figure 1 - Generation step), where $\boldsymbol{O}$, $\boldsymbol{E}$, and D can be customised or modified to generate synthetic SCADA datasets under desired conditions. For example, $\boldsymbol{O}$ and $\boldsymbol{E}$ can be replaced with those from a different wind turbine and/or a different time frame to replicate the degradation associated with D in those conditions. It is important to note that since $\boldsymbol{O}$ and $\boldsymbol{E}$ are correlated, they must refer to the same time frame and wind turbine. Alternatively, the trajectory of D can be modified to simulate a different degradation history.

2.2. Model evaluation

The problem of evaluating synthetic time series generation methods is an active area of research. However, no consensus has yet been reached among researchers on what characterises high-quality synthetic time series and how to quantify quality [7], and different approaches referring to various metrics and measures have been proposed in the literature [10]. Besides the sensor signal value distributions, a synthetic time series generation method should preserve the temporal dynamics in the original data [11]. Therefore, in this paper, the developed method is evaluated with respect to the generated signal distribution, the temporal dynamics within each generated signal and among the generated signals within a synthetic SCADA dataset.

2.2.1. Signal distribution. The signal distributions of a synthetic SCADA dataset generated from a given set of operational, environmental, and degradation conditions should be congruent

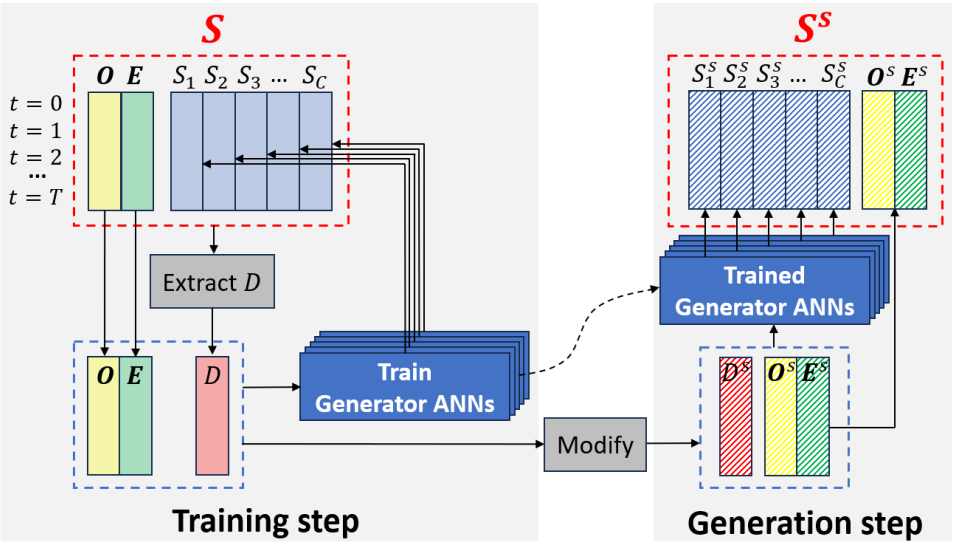

Figure 1. Synthetic SCADA data generation method. D denotes the degradation trend, $\boldsymbol{O}$ and $\boldsymbol{E}$ the operation- and the environment-dependant signals respectively, and $\boldsymbol{S}$ the SCADA dataset.

with those of a field SCADA dataset referring to similar conditions. For example, in the case of a synthetic SCADA dataset including a gearbox failure mode known to cause elevated temperatures [12], an increase in the level of degradation (D) in a given period should translate into an increase in the gearbox-related temperature signals in the same period. Furthermore, the increase in the temperature signals should be proportional to the increase in the degradation level. The former assertion is assessed by applying the two-sample Kolmogorov–Smirnov (KS) test [13] and by verifying that the null hypothesis that the signals related to the lower and higher D come from identical underlying distributions is rejected. The p-value is calculated, which indicates the probability that the two sets of samples have identical underlying distributions. A p-value less than 0.05 typically indicates that the null hypothesis is rejected. The latter analysis is performed by measuring the dissimilarity between signal distributions corresponding to different quantities of increase in D, which is measured by the Wasserstein Distance (WD) [14]. This metric is calculated by associating a hypothetical mass to distributions, finding the optimal way to transport it from one distribution to the other, and calculating the required work to do that. The WD is calculated for each S_i and S_i^s pair (WD_i) and then the Average WD (AWD) between $\boldsymbol{S}$ and $\boldsymbol{S}^s$ is calculated as the mean WD_i for $i = 1 \dots C$.

2.2.2. Intra-signal temporal dynamics. In addition to exhibiting congruent distributions, the synthetic and field signals corresponding to similar conditions should display similar dynamics in time [11]. This property is measured by the Auto-Correlation Function (f_{AC}) [6]. The auto-correlation of a signal S_i with a time lag of τ, $\mathcal{C}(S_i, \tau)$, is measured by the Pearson correlation coefficient between $S_{i,0:T-\tau} = \{s_{i,t} | t \in [0, T-\tau]\}$ and $S_{i,\tau:T} = \{s_{i,t} | t \in [\tau, T]\}$ where T is the length of S_i.

$$\mathcal{C}(S_i, \tau) = Corr(S_{i,0:T-\tau}, S_{i,\tau:T}) \tag{4}$$

Then, the f_{AC} function is defined as:

$$f_{\mathrm{AC}}(S_i) := \{\mathcal{C}(S_i, \tau) | \tau \in [1, T-1]\} \tag{5}$$

which returns a vector of auto-correlations at lags of 1 through $T-1$. The dispersion among the f_{AC} vectors in field datasets should ideally be replicated in the synthetic datasets referring to similar conditions, indicating similar diversity in intra-signal temporal dynamics. The autocorrelation score among the field and synthetic signals, $s_{AC}(S_i, S_i^s)$, is defined as the Euclidean norm of the difference between the mean $f_{\mathrm{AC}}(S_i)$ among field datasets and the mean $f_{\mathrm{AC}}(S_i^s)$ among synthetic datasets.

$$s_{AC}(S_i, S_i^s) := ||\overline{f_{\mathrm{AC}}(S_i)} - \overline{f_{\mathrm{AC}}(S_i^s)}||_2 \tag{6}$$

The lower $s_{AC}(S_i, S_i^s)$ is, the higher the similarity of temporal dynamics in S_i and S_i^s.

2.2.3. Inter-signal temporal dynamics. A SCADA dataset consists of multiple signals with different levels of correlation. Therefore, other than the temporal dynamics within each signal, the signals in a synthetic dataset and a field dataset corresponding to similar conditions should display similar cross-correlations. To assess this property, a measure called the Cross-Correlation Function (f_{CC}) is defined based on [15].

$$f_{\mathrm{CC}}(\boldsymbol{S}) := \{Corr(S_i, S_j) | i \in [2, C], j \in [1, i]\} \tag{7}$$

where $Corr(S_i, S_j)$ is the cross-correlation between S_i and S_j defined by their Pearson correlation coefficient, and C is the number of signals in the SCADA dataset. In [15], the whole cross-correlation matrix is used for measuring inter-signal temporal dynamics, where the values in the higher and lower triangles are identical, and the main diagonal is equal to 1. Equation 7 returns a vector with the lower triangle elements of the cross-correlation matrix to eliminate redundant values. The dispersion in f_{CC} vectors of field datasets should ideally be replicated in the synthetic datasets, indicating similar diversity in inter-signal temporal dynamics.

3. Results

In this section, the proposed method is used to generate synthetic datasets based on a wind turbine with a gearbox failure. The synthetic datasets are compared with field datasets corresponding to similar conditions, considering the three properties introduced in the previous section.

3.1. Dataset

The dataset used in this paper contains SCADA data from nine $2MW$ wind turbines (WT1–9) with $100m$ diameter, $80m$ hub height, and three-stage gearboxes with one planetary and two parallel stages. One wind turbine (WT8) experienced a failure in the gearbox planetary stage. Signals are recorded in 10-minute time frames from 2017-01-01 until 2022-08-01, and the gearbox failure occurred around 2022-02-23. Among the available data, the 14 signals listed in Table 1 are selected in this work as they refer to the turbine operational and environmental conditions and the gearbox operation.

The data pre-processing includes omitting values outside their acceptable physical ranges and those corresponding to non-operational conditions of the turbines. The former includes gearbox-related temperatures lower than the ambient temperature, zero gearbox oil pressures, and negative rotor speeds, which are caused by sensor errors and constitute around 5% of the total data points. The latter refers to data points corresponding to negative produced power values which occur when the turbine is idling. They constitute around 20% of the total data points. Hence, the pre-processing step eliminates around 25% of the original data. Then, the pre-processed signals are resampled in 6-hour time frames. A larger resampling window leads to a lower number of data points in a given time period and, hence, a lower computational

Table 1. Selected SCADA signals.

Gearbox-related signals	Operational signals	Environmental signals
(1) Gearbox oil temperature	(8) Produced power	(12) Wind speed
(2) Gearbox bearing 1 temperature	(9) Rotor RPM	(13)Ambient temperature
(3) Gearbox bearing 2 temperature	(10) Rotor RPM Max.	(14) Nacelle temperature
(4) Gearbox bearing 3 temperature	(11) Blades pitch angle	
(5) Gearbox oil inlet temperature		
(6) Gearbox oil pressure after filter		
(7) Gearbox oil pressure before filter		

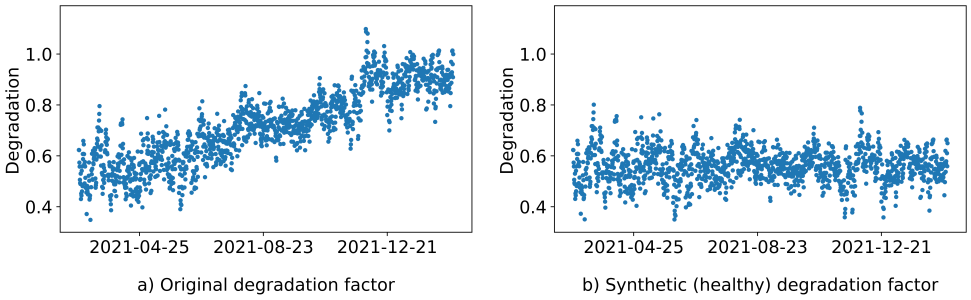

Figure 2. (a) degradation factor extracted from WT8 field data and (b) synthetic degradation factor mimicking a healthy WT8.

burden, especially during the extraction of the degradation factor. Furthermore, it reduces the ratio of missing data points, which is crucial for maintaining the signal continuity for the analysis of temporal dynamics. However, it can affect the quality and accuracy of the extracted degradation factor and the suitability of the datasets for failure prognosis tasks. A window of 6 hours, reducing the number of data points in 1 year from around 52000 to around 1400 and the percentage of the missing data from 25% to 3.6%, has been identified as a good trade-off among these conflicting factors.

3.2. Extraction of the degradation factor

Since WT8 includes a gearbox failure, it is used as a reference for the generation of synthetic SCADA datasets. The period of 1 year leading to the gearbox failure is selected as the run-to-failure SCADA dataset, and according to the method developed in [9], its degradation factor, shown in Figure 2 (a), is extracted.

3.3. Generation of the synthetic SCADA datasets

Generator ANNs are trained to map the operational and environmental factors and the extracted degradation factor of the WT8 run-to-failure SCADA dataset (input) to each gearbox-related signal. Therefore, 7 generator ANNs are developed for the 7 gearbox-related signals listed in Table 1. In this study, the rotor speed of the wind turbine is used as the operational factor, and the wind speed and ambient temperature are used as the environmental factors. Therefore, the input of the generator ANN includes these signals along with the degradation factor. Including additional operation- and environment-related signals reported in Table 1 did not affect the ANN performance significantly while increasing the computational burden. Once trained, these ANNs can generate synthetic SCADA signals corresponding to modifications of any of their input signals. During training, 20% of the training data points are randomly selected as a validation set to monitor the training process in terms of convergence and overfitting. Each

ANN contains six layers (including the output layer), each with a "ReLU" activation function, and the number of neurons per layer is 8, 8, 8, 12, 12, and 1. The network architecture is obtained using a constructive trial-and-error approach where the number of layers and neurons gradually increases to obtain a trade-off between performance, measured by the prediction Mean Squared Error, and computational burden.

The extracted degradation factor (D) in Figure 2(a) describes the gradual WT8 gearbox degradation with the failure occurring at the final time instance. The remaining eight wind turbines are healthy in the considered time frame. Therefore, to enable a comparison with the available field datasets, the generated synthetic signals refer to two specific degradation factors, one describing the known run-to-failure WT8 behaviour and the other a synthetic healthy turbine. In this latter case, the synthetic degradation factor (D^s) is built by mimicking a constant degradation equal to the initial WT8 level within the analysed 1-year period. This is achieved by applying the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise algorithm [16] to decompose D into its trend, D_{trend}, and several other intrinsic mode functions (IMFs), $D_{IMF,i}$.

$$D = D_{trend} + \sum_{i=1}^{N_{IMF}} D_{IMF,i} \tag{8}$$

where N_{IMF} is the number of IMFs other than the trend. D^s, shown in Figure 2(b), is then obtained by replacing the increasing trend (D_{trend}) with a constant one (D^s_{trend}) whose value is equal to the initial value of D_{trend}.

$$D^s_{trend,t} = D_{trend,t}|_{t=0} \tag{9}$$

$$D^s_{trend} = \{D^s_{trend,t} | t \in [0, T]\} \tag{10}$$

$$D^s = D^s_{trend} + \sum_{i=1}^{N_{IMF}} D_{IMF,i} \tag{11}$$

By inputting either D or D^s along with the three operational and environmental signals related to any wind turbine within any 1-year time frame, two sets of synthetic healthy $\boldsymbol{S}_i^{s,h}, i = 1, \ldots, N_{s,h}$ and synthetic failed $\boldsymbol{S}_i^{s,f}, i = 1, \ldots, N_{s,f}$ datasets can be built, where $N_{s,h}$ and $N_{s,f}$ are the number of generated datasets. In this work, one healthy and one failed synthetic SCADA dataset is generated from each wind turbine's operational and environmental signals in each one-year time frame. Hence $N_{s,h}$=9 and $N_{s,f}$=9.

3.4. Analysis of the signal distributions

In addition to an increase in gearbox-related temperatures, the degradation and failure in WT8 led to a reduction in the gearbox oil pressure. Post-mortem analysis uncovered an accumulation of wear particles in the oil, clogging up the oil filter and leading to a drop in the oil pressure. The replacement of D with D^s while generating the synthetic signals should reverse these changes. To evaluate if the developed method is able to mimic this behaviour, WT8 data in the year leading to the gearbox failure and its equivalent synthetic healthy SCADA dataset are analysed and compared (Figure 3-top), where only the degradation factor is modified as shown in Figure 3-bottom. Sections A, B, and C indicate the first, middle, and last 3 months of the data, respectively. KS tests performed on sections A and C result in average p-values of around 0.35 and 10^{-6}. This indicates that the null hypothesis, i.e., the two datasets have identical underlying signal distributions, can be rejected with higher certainty for the last three months, where the degradation factors have higher divergence. However, in the first three months, given the strong similarities between the two degradation behaviours, the signals must have similar underlying distributions, which is confirmed by the result of the KS test.

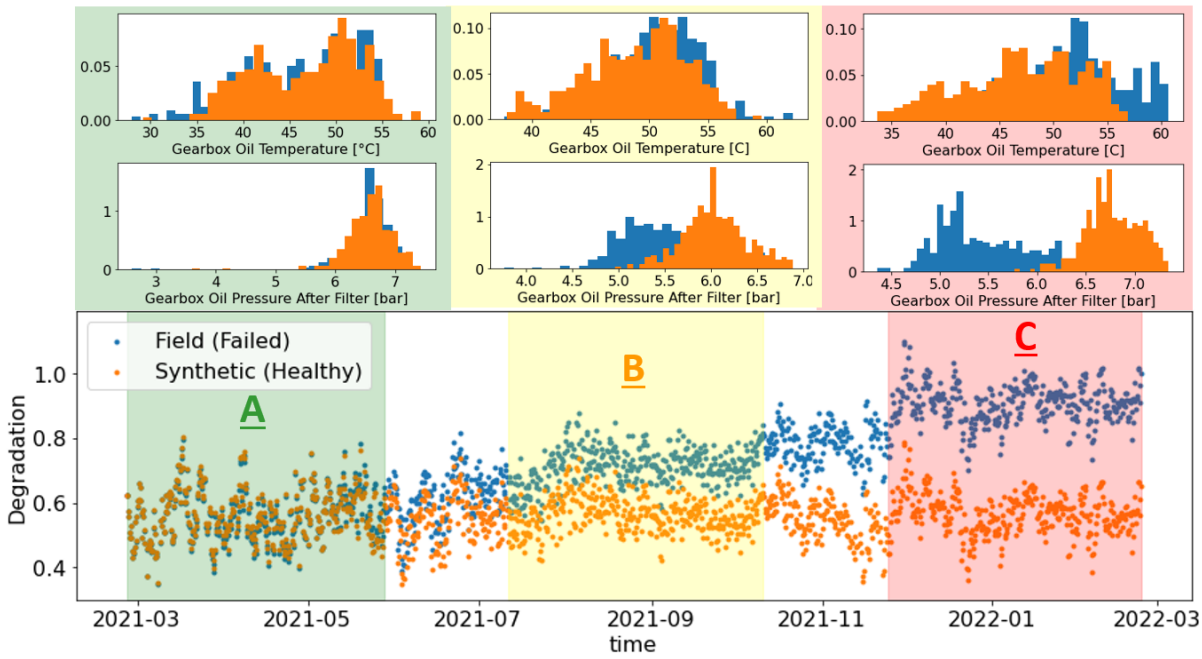

Figure 3. WT 8 (failed) SCADA dataset and its corresponding synthetic healthy SCADA dataset. Top: distributions of the gearbox oil temperature and oil pressure after filter in the first, middle, and last 3 months. Bottom: degradation factors.

The level of divergence in the gearbox-related signal distributions for the field (failed) and synthetic healthy SCADA datasets is expected to correspond to the level of divergence between their degradation factors, D and D^s, i.e. these signals must show increasing divergence as we move closer to the failure. This can be observed in Figure 3-top showing the distributions of two gearbox-related signals, the oil temperature and the oil pressure after the filter, in sections A, B, and C of the analysed 1-year period. When approaching the failure point, in the field dataset with failure, as expected, the gearbox oil temperature shifts to higher values and the gearbox oil pressure shifts to lower values, while the same behaviour is not observed in the healthy synthetic dataset. This leads to a clear divergence between the two signal distributions. The divergence among all the gearbox-related signals in these two datasets is measured with the *AWD* metric, whose increasing values of 0.33, 1.28, and 3.22 for sections A, B, and C, respectively, confirm the trend shown in Figure 3-top. This increasingly shifting trend between the two datasets is also clearly visible after applying Principal Component Analysis to reduce the dimensionality of the gearbox-related signals from 7 to 2 for visualisation purposes, as shown in Figure 4.

The *AWD* metric is calculated between the nine field datasets and the 18 synthetic datasets generated from their operational and environmental signals, along with either the healthy or failed D, in the time frame of three months leading to failure in WT8. The results reported in Table 2 show that the signal distributions of the synthetic healthy datasets are more similar to field healthy datasets than the field failed dataset and vice versa, demonstrating the above-mentioned shift in signals in all the generated synthetic datasets.

3.5. Analysis of the intra-signal temporal dynamics

To assess the similarity of intra-signal temporal dynamics in field and synthetic datasets, the f_{AC} in gearbox-related signals $(f_{\mathrm{AC}}(S_i), i = 1, \ldots, 7)$ is analysed to measure how their autocorrelation varies with the lag. It was observed that all temperature signals display similar f_{AC} behaviours,

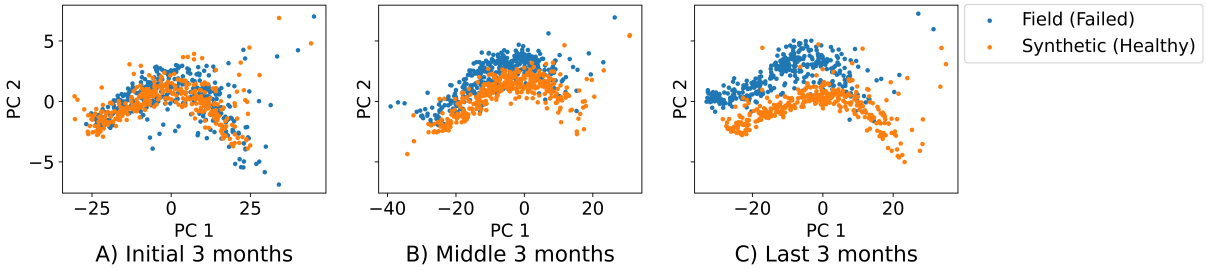

Figure 4. First vs. second principal components of the field (failed) and synthetic healthy datasets in the first, middle, and last three months.

Table 2. *AWD* metric between each pair of dataset groups.

	Field healthy	Field failed
Synthetic healthy	1.44	3.05
Synthetic failed	3.77	1.67

Table 3. Mean s_{AC} metric of pressure signals for each pair of dataset groups.

	Field healthy	Field failed
Synthetic healthy	2.80	9.77
Synthetic failed	7.28	3.10

as do all pressure signals. Therefore, the f_{AC} vectors of one temperature and one pressure signal, i.e. the gearbox oil temperature and the pressure after the filter, in all field and synthetic datasets and various one-year periods from 2017/1/1 to 2021/12/1 with one-month shifts in the starting dates, are plotted in Figure 5-top which shows a similar range of values for field and synthetic datasets. However, the pressure signals' f_{AC} vectors in synthetic datasets show a slightly lower dispersion at lower lags compared with field datasets, indicating a lower diversity in temporal dynamic behaviour. The high variability of the autocorrelation observed at high lags is the result of the reduction in the length of the correlated sequences. Unlike the pressure signal, the autocorrelation of the temperature signal is generally close to zero, except at very low lags. This might be because temperature signals are highly dynamic and show higher sensitivity to wind turbine operation. In contrast, pressure signals are mostly static and affected by ambient and oil and filter conditions, among other factors. This explains the larger variations in the f_{AC} vectors associated with the pressure signal. Because of this sensitivity to ambient conditions, the two signals are analysed in the fixed one-year time frame of the run-to-failure dataset among the four groups of field/synthetic healthy/failed datasets, plotted in Figure 5-bottom. The pressure signals show different autocorrelation behaviours in the healthy and failed field datasets, closely replicated in the synthetic datasets. Table 3, which reports the s_{AC} metric between field/synthetic healthy/failed dataset groups, averaged among all pressure signals, confirms this observation. The temporal dynamic behaviour of pressure signals in synthetic healthy datasets is more similar to field healthy datasets than the field failed one and vice versa.

3.6. Analysis of the inter-signal temporal dynamics

The f_{CC} vectors, which include 21 elements corresponding to the number of pairs of gearbox-related signals, are plotted in Figure 6 for all the field and synthetic datasets referring to different one-year periods. Indices 1 through 10 and 21 refer to temperature-temperature and pressure-pressure signal pairs, which display high cross-correlations. The other indices

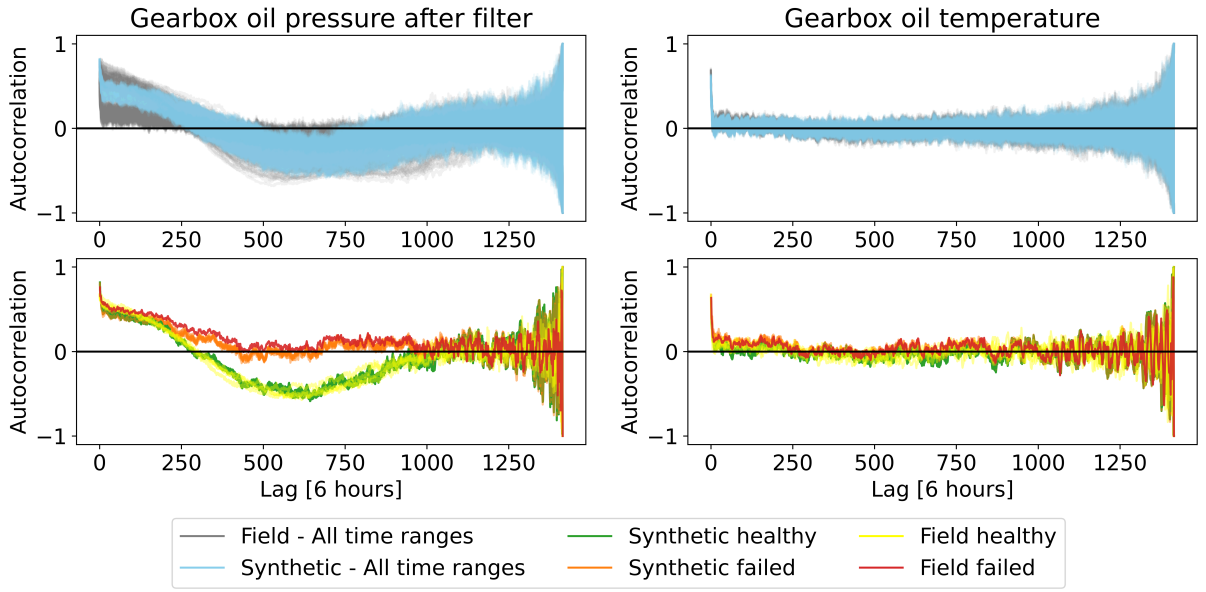

Figure 5. The f_{AC} vectors of gearbox oil pressure after filter and gearbox oil temperature

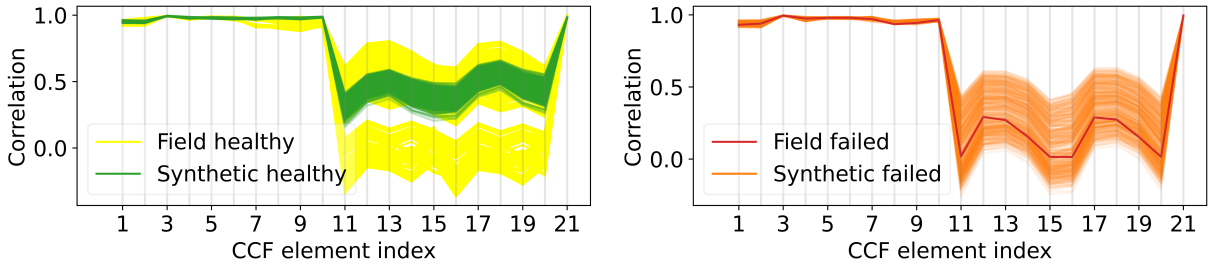

Figure 6. f_{CC} vectors, i.e., cross-correlation (y-axis) for different pairs of gearbox signals (x-axis), in different 1-year periods

refer to pressure-temperature pairs which have lower cross-correlations. The f_{CC} vectors for the synthetic healthy/failed datasets are consistent with the corresponding field observations. However, a noticeably lower dispersion can be seen in synthetic healthy datasets compared with the corresponding field ones, indicating a lower diversity. Conversely, synthetic failed datasets show a relatively larger diversity. Since there is only one field failed dataset available, it is not possible to draw conclusions on the diversity between field and synthetic failed datasets.

4. Conclusions and future work

This paper proposes a data-driven method using an ANN for generating synthetic SCADA datasets with customisable operational, environmental, and degradation conditions. The results show that the signal distributions and temporal dynamic behaviours of the synthetic datasets generated are consistent with field datasets under similar operational, environmental and degradation conditions. However, the lower variability in the signal autocorrelation and cross-correlation indicates that the synthetic datasets, especially in the healthy case, usually feature a slightly lower diversity in temporal dynamic behaviour compared to the field case. The proposed generator ANN is trained with signals related to only one wind turbine gearbox failure, and this limits the possibility of replicating the farm-wide diversity in signal behaviour. With the

availability of more faulty field data, future work will further assess this observation. The robustness of using the proposed method to perform reliable fault detection and prognosis when adequate field SCADA datasets containing multiple failure instances are not available is explored in ongoing research.

5. Acknowledgement

We thank Lucky Wind SpA for providing the SCADA datasets used in this paper.

References

[1] Chatterjee J and Dethlefs N 2021 *Renewable and Sustainable Energy Reviews* **144** 111051 ISSN 1364-0321 URL http://dx.doi.org/10.1016/j.rser.2021.111051

[2] Liu J, Yang G, Li X, Wang Q, He Y and Yang X 2023 *ISA Transactions* **139** 586–605 ISSN 0019-0578 URL http://dx.doi.org/10.1016/j.isatra.2023.03.045

[3] Pujana A, Esteras M, Perea E, Maqueda E and Calvez P 2023 *Energies* **16** 861 ISSN 1996-1073 URL http://dx.doi.org/10.3390/en16020861

[4] Zhang C, Kuppannagari S R, Kannan R and Prasanna V K 2018 *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2018*

[5] Hu C, Sun Z, Li C, Zhang Y and Xing C 2023 *Sensors* **23** 6976 ISSN 1424-8220 URL http://dx.doi.org/10.3390/s23156976

[6] Wiese M, Knobloch R, Korn R and Kretschmer P 2020 *Quantitative Finance* **20** 1419–1440 ISSN 1469-7696 URL http://dx.doi.org/10.1080/14697688.2020.1730426

[7] Leznik M, Lochner A, Wesner S and Domaschka J 2022 *Journal of Systems Research* **2** ISSN 2770-5501 URL http://dx.doi.org/10.5070/SR32159045

[8] Yang Z, Baraldi P and Zio E 2018 Automatic extraction of a health indicator from vibrational data by sparse autoencoders *2018 3rd International Conference on System Reliability and Safety (ICSRS)* (IEEE) URL https://doi.org/10.1109/icsrs.2018.8688720

[9] Eftekhari Milani A, Zappalá D and Watson S J 2024 *Engineering Applications of Artificial Intelligence (under review)*

[10] Stenger M, Lseppich R, Foster I, Kounev S and Bauer A 2023 Evaluation is key: A survey on evaluation measures for synthetic time series URL http://dx.doi.org/10.21203/rs.3.rs-3331381/v1

[11] Yoon J, Jarrett D and van der Schaar M 2019 Time-series generative adversarial networks *Advances in Neural Information Processing Systems*

[12] Salameh J P, Cauet S, Etien E, Sakout A and Rambault L 2018 *Mechanical Systems and Signal Processing* **111** 251–264 ISSN 0888-3270 URL http://dx.doi.org/10.1016/j.ymssp.2018.03.052

[13] Massey F J 1951 *Journal of the American Statistical Association* **46** 68–78 ISSN 1537-274X URL http://dx.doi.org/10.1080/01621459.1951.10500769

[14] Vallender S S 1974 *Theory of Probability and Its Applications* **18** 784–786 ISSN 1095-7219 URL http://dx.doi.org/10.1137/1118101

[15] Wiese M, Bai L, Wood B and Buehler H 2019 *arXiv* URL https://arxiv.org/abs/1911.01700

[16] Colominas M A, Schlotthauer G and Torres M E 2014 *Biomedical Signal Processing and Control* **14** 19–29 ISSN 1746-8094 URL http://dx.doi.org/10.1016/j.bspc.2014.06.009