Synthetic data generation for condition monitoring of railway switches*

Miguel del Álamo[1], Judith Heusel[1], and Daniela Narezo Guzmán[2][0000−0001−9748−1354]

[1] German Aerospace Center (DLR), Lilienthalplatz 7, 38108 Braunschweig, Germany
[2] German Aerospace Center (DLR), Rutherfordstr. 2, 12489 Berlin, Germany
{miguel.delalamoruiz, judith.heusel, daniela.narezoguzman}@dlr.de

Abstract. The application of AI methods to industry requires a large amount of training data that covers all situations appearing in practice. It is often a challenge to collect a sufficient amount of such data. An alternative is to artificially generate realistic data based on training examples. In this paper we present a method for generating the electric current time series produced by railway switch engines during switch-blades repositioning. In practice, this electrical signal is monitored and can be used to detect unusual behaviour associated to switch faults. The generation method requires a sample of real curves and exploits their systematic temperature dependence to reduce their dimensionality. This is done by extracting the effect of temperature on specific parameters, which are then re-sampled and used to generate new curves. The model is analyzed in different practice-relevant scenarios and shows potential for improving condition monitoring methods.

Keywords: Railway switches · Data generation · Anomaly detection.

1 Introduction

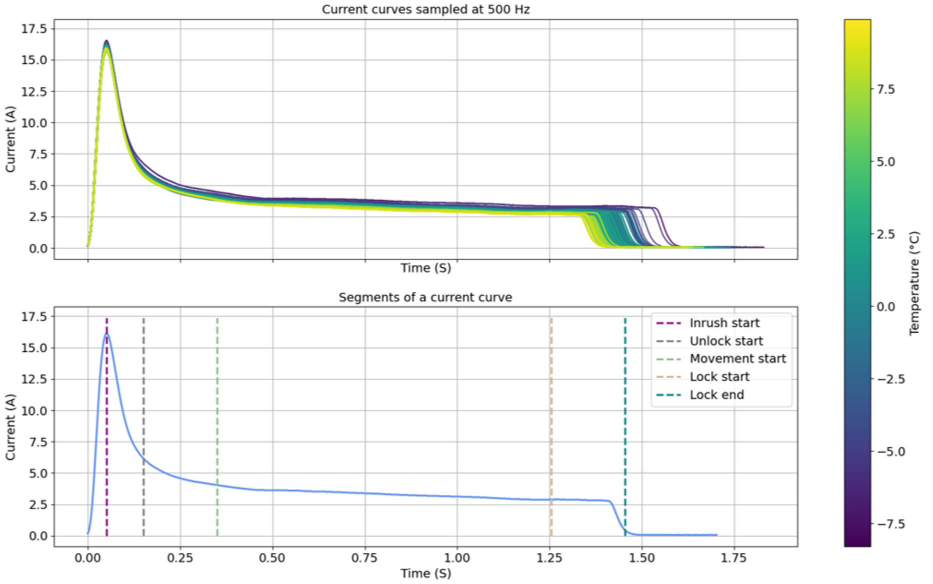
1.1 Motivation: anomaly detection for railway switches

Railway switches are an essential element of the railway infrastructure: they are the crossing nodes which allow trains to change tracks. Their condition is safety relevant, since switch faults can lead to accidents and derailments. In addition, switch malfunctions have a negative impact on the infrastructure availability and reputation. Finally, switches are subject to regular maintenance, renewal and repairs, which makes them cost-intensive assets. All of this makes condition-based and predictive maintenance a desirable goal. In order to achieve it, continuous monitoring and automatic assessment of the switch condition is required.

* This project has received funding from the Shift2Rail Joint Undertaking (JU) under grant agreement No 881574. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Shift2Rail JU members other than the Union. The authors thank Strukton Rail and Wolfgang Riedler for their support on this work.

Strukton Rail (SR), a Dutch railway infrastructure maintenance operator, monitors the condition of thousands of switches in the Netherlands using the in-house developed system POSS®[†]. Every time the switch blades are repositioned, POSS® collects the electric current at the point machine together with the air temperature at the relay house to which the switch is connected. The electric current at the point machine is a measure of the power needed by the engine to move the switch blades from their start to end position. We refer to the current measured during a single repositioning as current curve. The majority of known switch defects has an influence on the shape of these current curves [1,2].

Within the Shift2Rail project In2Smart2, the German Aerospace Center and SR are developing and validating methods for anomaly detection and diagnosis of switch defects [3]. The goal of these efforts is to support maintenance engineers at SR's Control Center in identifying faulty switches. The project comprises electromechanical switches of type NSE (Nederlandsche Spoorwegen Elektrisch). Current curves of these switches have typically a similar, yet switch- and repositioning direction dependent shape. However, they all are characterised by a systematic temperature dependence (see Figure 1, [4]). In addition, the curves can be split into segments which roughly correspond to the phases of the blades repositioning (inrush current, unlocking, blades movement and locking, see Figure 1).


Current curves sampled at 500 Hz

Segments of a current curve

Fig. 1. Upper row: temperature dependence of switch current curves (sampled at 500 Hz). Bottom row: segments of a current curve. The inrush, unlock and lock segments have pre-defined lengths. The curve start is defined as the global maximum's position.

[†] POSS®: Preventief Onderhoud - en Storingsdiagnosesysteem Strukton, http://www.POSSinfo.com

Anomaly detection (AD) approaches applied to current curves can help to identify switch degradation at an early stage as well as sudden failures. Some AD methods (e.g. [4]) employ parameters derived from current curves and their segments (throughout the paper we use the word "parameter" to denote current curve quantities such as length, mean, standard deviation, maximum, kurtosis, etc). Such models are trained with parameters derived from a set of historical curves for which the switch is assumed to behave normally. The set of historical curves is required to represent typical temperatures found in all seasons.

The validation of AD models is challenging due to the lack of annotated data. The amount of labelled current curves is very low and does not cover a large enough period of time. In this context, the generation of synthetic current curves can help to validate and compare AD models, and to test and improve AD algorithms as well as the considered curve parameters. In addition, it can help to make the current implementation of the AD model more robust by generating realistic data for sparse temperature bins and applicable even when there is few historical data to train the model. This is the case e.g. briefly after a new switch is installed or after maintenance actions on switches, which can strongly modify the current curve typical shape, making it necessary to retrain the AD model. This paper focuses on the generation of synthetic current curves which imitate the normal behaviour of a switch, especially by capturing the temperature dependent variation of the current curves.

1.2 Challenge: sampling from a complex high-dimensional distribution

The task of generating new current curves based on real ones can be reformulated as sampling from a high-dimensional distribution. This is a well-known and extremely difficult problem in modern data science (see e.g. [5]).

In our application, temperature is found to be responsible for the main variation in current curves, see Figure 1. Thus temperature can be used for dimension-reduction. Further, the effect of temperature on the current curves is well captured by three parameters: the curve length, its maximal height, and its median height in the movement segment. In other words, if we take any two curves at different temperatures and manipulate them such that these three parameters match, then they look approximately equal. Two caveats are due here: first, switch condition may change or degrade over time, so this holds for curves measured not too far apart from each other; second, we have tested this phenomenon by taking into account curves from four switches of type NSE. Since switches can vary a lot from one type to another, we do not claim that these three parameters plus temperature are enough to characterize all switch types.

With these observations, the task of sampling from the distribution of curves can be reduced to sampling from their temperature distribution, and then for each temperature, sampling from their parameter distributions, which are all one-dimensional distributions. Finally, the sampled parameters are imposed into an "ideal curve", see Section 2. Altogether, this method allows to sample from a complex high-dimensional distribution. Since it is quite flexible, the method can potentially be used in other applications involving electromechanical components.

2 Generation method and hyperparameters

2.1 Method description

Our current curve generation method is based on the sampling methodology discussed in Section 1.2, and thus on the effect of temperature on specific parameters (i.e., maximum, length, and median of the movement segment). The underlying assumption is that if the switch is in perfect condition, the shape of the curve does not change and only the three aforementioned parameters vary in dependence on the temperature. In practice, sampling and resolution have an additional impact on the resulting current curves. Due to these assumptions, the curve manipulation for generating new current curves of a given switch is as follows (see Figure 2 for a schematic workflow visualization): an input curve, which is assumed to be ideal and to represent the current condition of the switch, is chosen per repositioning direction - e.g., the respective first curves of the time series - and serves as model pattern for the synthetically generated ones. Then, the three named parameters are randomly sampled from a temperature-dependent distribution which is learned from real data; details are given in Section 2.2. Subsequently the model curve is stretched to have the target length. The resulting curve is then multiplied by a spatially varying scaling function that sets the maximum and the median of the movement segment to the target values.

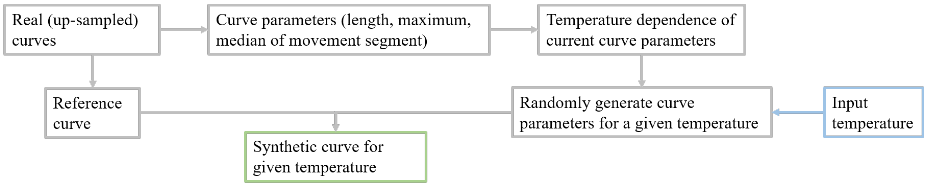

Real (up-sampled) curves
Curve parameters (length, maximum, median of movement segment)
Temperature dependence of current curve parameters
Reference curve
Randomly generate curve parameters for a given temperature
Input temperature
Synthetic curve for given temperature

Fig. 2. Description of the process of generating synthetic current curves.

POSS® usually samples current curves at 50 Hz. This sampling rate has a non-negligible effect on the current curves, as it induces discretization issues that add undesired variation and may ruin the generated curves. A way to mitigate this problem is to work with current curves sampled at a higher rate (we employ 500 Hz), which are available only for a few switches. Alternatively the data are up-sampled from 50 to 500 Hz using a section-wise quadratic and linear interpolation, see [3] for details.

2.2 Hyperparameters

In this section we discuss several hyperparameters and variants of our method. As shown in Section 3, these variants can have a significant effect on the performance of the simulation results in terms of e.g. distribution similarity of real and synthetic current curve parameters. Moreover, certain variants are specially suitable for particular applications.

Linear vs empirical sampling One key ingredient of the method is the way in which we sample new parameters for a certain temperature given real data. We present two variants here:

1) Empirical method: the new parameter is generated by randomly drawing from the distribution of real parameters in a temperature bin containing the target temperature.
2) Linear method: first we perform a linear regression of the real parameters against temperature. Then the new parameter is randomly sampled from a normal distribution whose mean is given by the linear prediction at that temperature, and whose variance is estimated from real data.

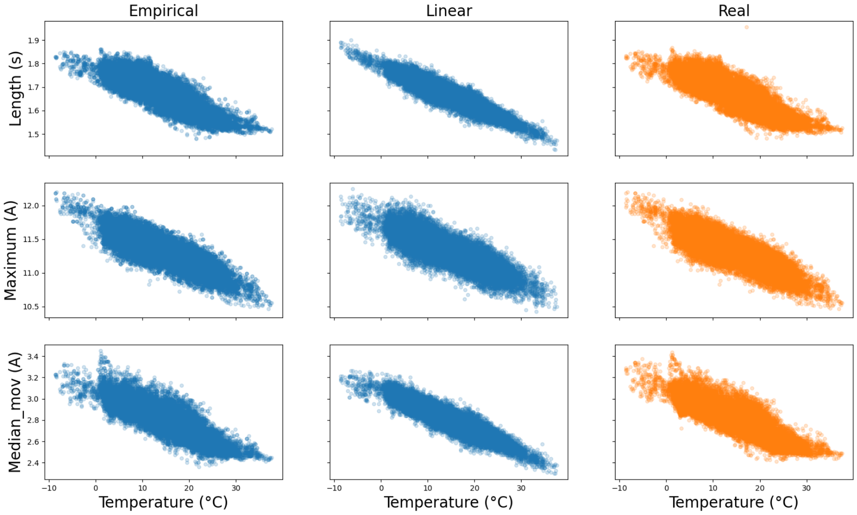


Fig. 3. Temperature dependence of three parameters (length, maximum and median of movement segment) for empirical method (left), linear method (middle), and real data (right).

Both methods have their advantages and shortcomings. The empirical method performs well when a lot of data is available, and poorly for few training curves. Additionally, the empirical method will naturally mimic the empirical distribution of the real data, hence making the generated curves arguably more realistic (Figure 4). On the other hand, the linear method tends to overregularize, thus yielding a different feature distribution than the real data (see Figure 3). However, through this implicit regularization, the linear method is able to extrapolate from few curves to a new temperature range, as discussed in Section 3.2. This makes the linear method attractive when few training curves are available. In addition, when outliers are already present in the parameter training data, the empirical method will reproduce them and abnormal curves may be contained in the synthetic data, too. This problem can be solved by filtering out statistical outliers before using the parameters for simulation. In contrast, the linear sampling method makes the assumption that the parameters used for curve manipulation follow a linear relationship to temperature and produces ideal distributions and curves. Note that a certain variation is naturally caused by measurement uncertainty related to the fact that the temperature measured at the relay house is only a proxy of the asset temperature.

Updating the reference curve Synthetic curves can be generated by only using one reference curve per repositioning direction and a few samples of the real parameters length, maximum and median of the movement phase. Sometimes the typical shape of the current curves from a given switch can vary over time without this change being a critical development (e.g. slowly developing changes in track geometry), especially when looking at a long time-span or at a frequently used switch. When the objective is to replicate the current curves and related distributions for such a time-span, the synthetic curves benefit from an update of the reference curve after e.g. some fixed time interval. In practice, an update should be performed by a switch maintenance analyst when the reference curve begins to differ from real curves that are deemed normal.

Local vs global In practice, the temperature dependence of real current curves can develop over time. This is, the input parameters belonging to a narrower time span exhibit a lower variance per temperature bin than the whole parameter set and the slope and intercept of the linear regressions of the parameters in dependence to temperature slowly develop over time (see Figure 7). This leads to the artifact that a randomly generated parameter set which used the whole available time span as training ("global") possesses a similar parameter distribution as the real data, but the parameter time series varies too much. Alternatively, linear regressions and variance estimations can be fitted for chronological data subsets ("local"), as is done in Section 3.1. This improves the parameter time series while the whole parameter distribution is still well represented.

3 Applications

In this section we illustrate the performance of our simulation machinery in two scenarios: 1. we generate data following the same distribution as in the given training data. And 2. we extrapolate from the training data to generate current curves at unseen temperatures. The performance of the simulation is evaluated in temperature windows of width 2°C using different quality measures.

3.1 Scenario 1: replicate the observed distribution

We consider a two years long sequence of current curves corresponding to the blades movement in one direction of a switch (about 22.000 current curves). The challenge is to generate a new sequence of synthetic curves covering the same time interval. We present the results from the linear and the empirical methods as discussed in Section 2 in Figure 3.

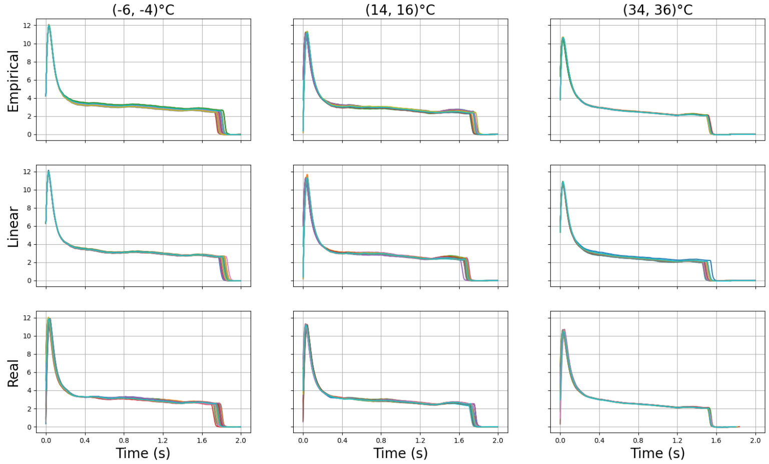


Fig. 4. Current curves generated by empirical (top) and linear (middle) methods, compared to real curves (bottom), in temperature range $(-6, -4)$°C (left), $(14, 16)$°C (middle) and $(34, 36)$°C (right). The vertical axes show electric current measured in Ampere. In each temperature window we display 20 randomly sampled curves.

In a nutshell, we observe that the distribution of parameters is more realistic for the empirical method. We verify this in Figure 4, where synthetic curves generated by the empirical and linear methods are compared with real curves in three temperature windows. Here, the linear method is seen to overregularize sometimes (see middle temperature window), while the empirical method yields visually correct results. This mirrors the intuitive recommendation to use the empirical method when plenty of training samples are available.

In Figure 5 we compare the distributions (histograms) of several parameters obtained from real and synthetic curves generated by the empirical sampling method. Beyond the visual similarity, the p-value for the Kolmogorov-Smirnov test between the two distributions (see Chapter 14.2 in [6]) is shown, and indicates that the distributions are indeed similar in a statistically rigorous sense. Similar yet slightly worse results are obtained for the linear method, as expected.

We compute another quality measure based on the Hausdorff distance. This is a well-known metric in mathematics that can be used to compare the geometry of quite general objects and sets. The Hausdorff metric provides a measure of distributional similarity. It does so by comparing individual points with an underlying metric, and then aggregating the individual distances into a global quantity (details can be found in Chapter 4 in [7]). Here, we want to measure the Hausdorff distance between the real and synthetic data. If we measured the distance directly, we would just get one number with no reference of whether it is big or small. We circumvent this problem by a bootstrap-type argument [8],

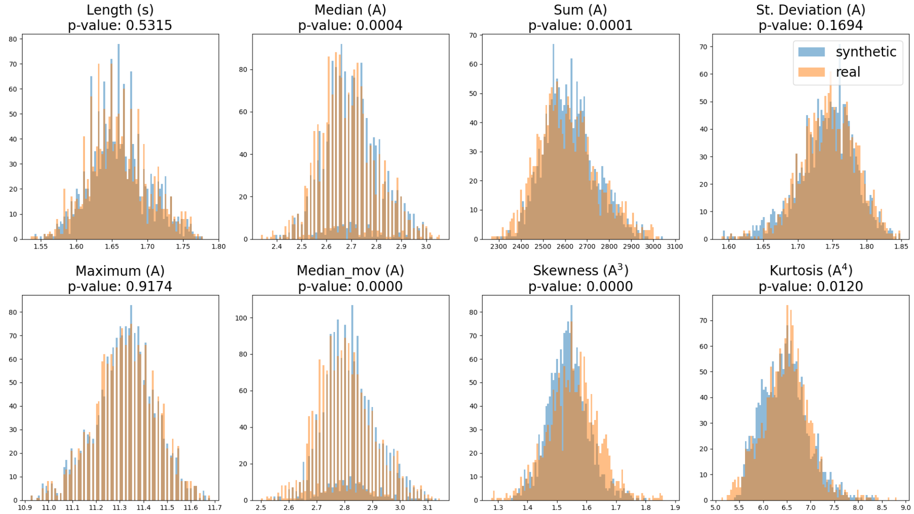


Fig. 5. Histograms of parameters from curves generated by the empirical method (blue) and from real curves (orange). The curves belong to the temperature range $(14, 16)$°C.

that is by measuring the Hausdorff distance between several randomly sampled subsets of our sets, and then comparing the distances. Specifically, we randomly sample two subsets and measure their distance, and we do so in three different fashions: sampling both subsets from the real data (R), both from the synthetic data (S), and one from each set (RS). Each of these three distributions of distances can be plotted as a histogram, as done in Figure 6. If the distances from the RS category are much larger than from the R and S categories, it means that the distances *between* the real and synthetic sets are larger than *within* those sets. In that case, we interpret the synthetic data to be significantly different from the real data.

On the other hand, if all three distributions of distances were identical, it would mean that it does not make a difference to sample from the real or from the synthetic set of curves, concluding that their distributions are similar.

Figure 6 includes the distribution of distances for both real and synthetic empirical data for two different underlying metrics (L^2-norm and Wasserstein distance). Both metrics indicate similarity between distributions in the three central temperature windows, but not in the extreme temperatures.

3.2 Scenario 2: extrapolate to unseen temperatures

We consider a short sequence of curves in a limited temperature range, and our task is to generate a longer sequence with temperatures outside this range and time span. Specifically, the input consists of a three month long sequence (comprising winter and spring), thus the parameters span a limited temperature range (from -2.7 to 18.3°C). The task consists of extrapolating the observed temperature dependence to unseen temperatures.

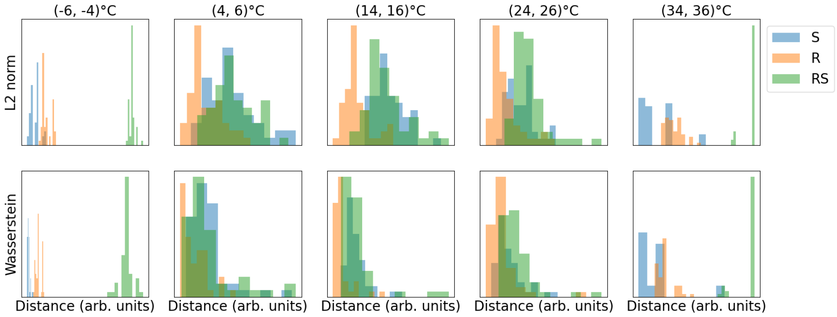


Fig. 6. Histograms of Hausdorff distances for subsets of the empirical synthetic data (blue), for the real data (orange) and between them (green), computed for different temperature windows and underlying metrics: L^2-norm and Wasserstein distance.

As discussed in Section 2, the linear method is able to extrapolate to unseen temperature ranges while considering only few data points.

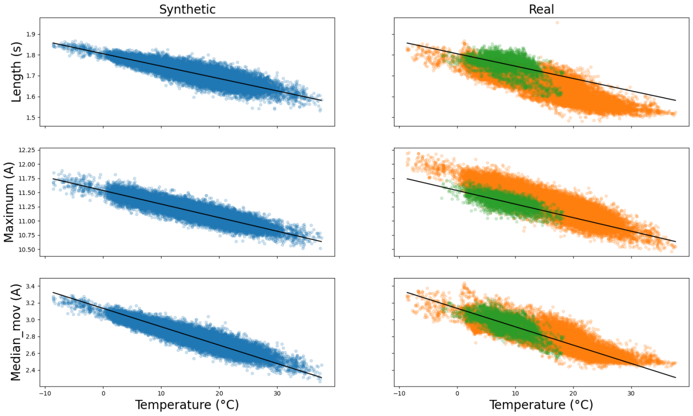


Fig. 7. Curve parameters as function of temperature for synthetic (left) and real (right) curves. The linear regression (black line) is computed with a limited amount of training curves (green dots) belonging to a small time interval and amounting to 15% of the real curves.

Figure 7 shows a linear regression performed on the three input parameters derived from real curves (green points) and the generated synthetic parameters (blue points). We identify that the final distribution of the synthetic parameters is quite different from the real distribution, specially for extreme temperatures, but also regarding its overall shape. Still, the synthetic curves generated with those parameters are visually good, as shown in Figure 8 for three temperature windows. However, we also see that some parameters of the synthetic curves are different from those of the real curves. This is due to the fact that the method

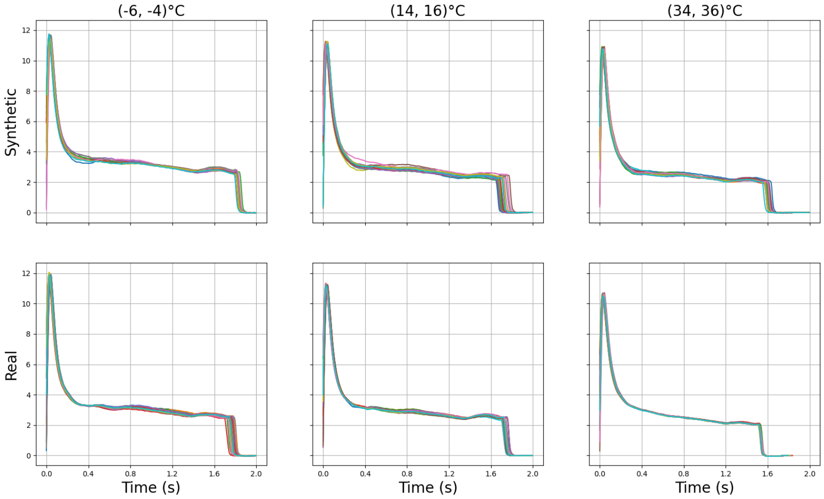


Fig. 8. Synthetic and real current curves from use case 2 for three temperature windows: $(-4, -6)$°C degrees (left), $(14, 16)$°C degrees (middle), and $(34, 36)$°C degrees (right). In each temperature window we display 20 randomly sampled curves.

does not have enough samples for all temperatures to "learn" the distribution correctly (see e.g. parameter length in Figure 7) and that the slope and intercept of the linear regression obtained at the beginning of the time series do not hold for the following year and half (see e.g. parameter maximum).

In Figure 9, we illustrate the real and synthetic parameters as a function of time, together with temperature. We observe that the parameter distributions are generally similar in temperature ranges that span across training temperatures (winter period); especially for the parameter length, the real distribution for higher temperatures outside the training set is underestimated. In addition, performance decreases with time, especially for the parameter maximum. This presumably results from a long-term change in temperature dependence. Nevertheless, the similarity between curves in Figure 8 is specially remarkable in the temperature window $(34, 36)$°C since the synthetic method was not trained with curves in that range, and it is however able to produce quite realistic results. This similarity can be quantified statistically: in this temperature window, the Kolmogorov-Smirnov test applied to the distribution of the maximum, median and length, returns following p-values: $(0.003, 0.012, 0.000)$.

4 Discussion

In this paper we formulate and motivate the problem of generating realistic synthetic current curves. A method for solving that problem is presented, and its performance with respect to several metrics is shown in two scenarios. In the first scenario, we compare the performance of the linear and empirical method, finding that the empirical method is able to match the real distribution better.

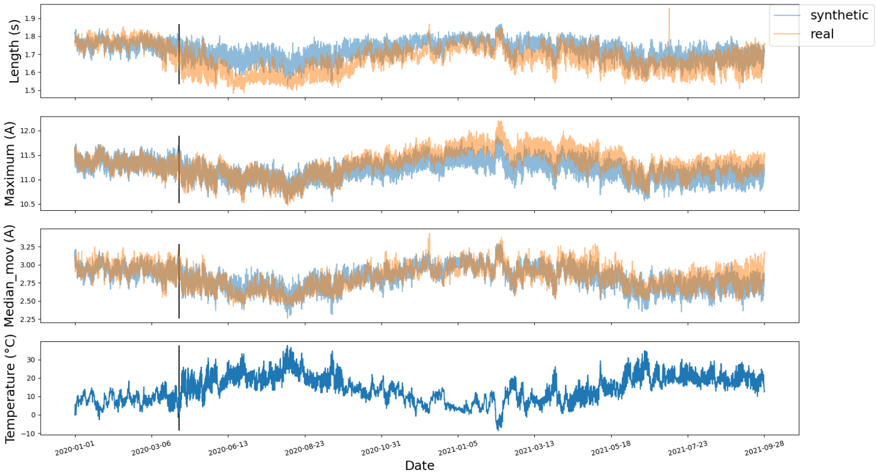


Fig. 9. Time evolution of three parameters: length (top), maximum (middle), and median of movement segment (bottom). The datapoints to the left of the vertical black line are the observed training data, as described in scenario 2 in Section 3.2. The synthetic datapoints are generated with the linear method.

In the second scenario we only employed the linear method, since the empirical method is not suitable for that setting. Here we find that the linear method is able to generate realistic curves in unseen temperature regions, as shown in the third column of Figure 8. In other words, the temperature extrapolation is performed well.

Overall, we developed a method that can generate data from a complex distribution. Here we want to stress that, even though our method is applied to a very specific type of data (current curves with temperature dependence), the idea behind the method may be applied to other types of data, e.g. current curves with a similar structure from other applications. This opens up interesting possibilities for a general anomaly detection methodology in electromechanical systems. Since our method successfully generates synthetic current curves, it can assist anomaly detection models. Specifically, the method can be used to enlarge the training data of an anomaly detection model in order to cover rare situations (such as extreme temperatures) or even unseen conditions (as in Section 3.2 above). This can make anomaly detection models more robust.

It is an arguably difficult task to compare two sets of complex objects, such as current curves. We choose to split this task into two parts and consider several similarity measures. On one hand, we compare *individual* synthetic and real curves, and test whether they are similar (as in Figure 4). On the other hand, we compare the two sets of curves *as a whole* by looking either at their statistical distributions (as in Figure 5) or at their geometry (in terms of the Hausdorff metric, as in Figure 6). One further similarity measure that could be used is an anomaly detection algorithm that is trained in the real data and then applied to the synthetic data. In that setting, the percentage of found anomalies ("false

positives") would provide a good error measure.
There are further extensions of our work that we want to discuss. First, other interesting use cases can be considered. This includes the validation of anomaly detection methods, but also the retraining of an anomaly detection method after the switch conditions have changed (either by degradation or due to maintenance actions). Second, other generation methods can be used. We have presented what we call a parametric method, since the curves are reduced to a set of parameters that are modelled. Alternatively, one could use *nonparametric* methods, where the curves are generated as a whole. We have explored this idea using Generative Adversarial Networks (GANs) with good preliminary results, which offer an interesting and flexible alternative, although at the price of requiring more training samples. Another idea is to use dictionary learning to extract more accurate parameters, and then perform linear regression against temperature. Preliminary results show this to be a promising approach, as it automatically determines the dimensionality reduction to be performed, which can be useful when analyzing different switch types. And third, here we presented the generation of "normal" curves by modelling their temperature variation. In a future paper, we will discuss how to also generate synthetic abnormal curves related to different fault types and degrees of anomaly. These abnormal curves can be used to train anomaly detection methods in a better way.

References

1. García Márquez, F.P., Lewis, R.W., Tobias, A.M., Roberts, C.: Life cycle costs for railway condition monitoring. In Transportation Research Part E: Logistics and Transportation Review 44, pp 1175–1187 (2008)
2. Innotrack (2009) Deliverable 3.3.2, Available sensors for railway environments for condition monitoring, https://www.charmec.chalmers.se/innotrack/deliverables/sp3/d332-f3p-available_sensors.pdf. Last accessed 3 Jun 2022.
3. Narezo Guzmán, D., Heusel, J., Weik, N., Reetz, S., Buursma, D., van den Berg, A., Schrijver, G., Neumann, T., van den Broek, S., Groos, J.C.: Towards the automation of anomaly detection and integrated fault identification for railway switches in a real operational environment. In: Proceedings World Congress on Railway Research, 2022 (in press)
4. Narezo Guzmán, D., Hadzic, E., Baasch, B., Heusel, J., Neumann, T., Schrijver, G., Buursma, D., Groos, J. C.: Anomaly Detection and Forecasting Methods Applied to Point Machine Monitoring Data for Prevention of Railway Switch Failures. In: Ball, A. et al. (eds) Advances in Asset Management and Condition Monitoring, vol 166, pp 307–318. Springer International Publishing, Cham (2020)
5. Vono, M., Dobigeon, N., Chainais, P.: High-dimensional Gaussian sampling: a review and a unifying approach based on a stochastic proximal point algorithm. SIAM Review **64**(1), 3–56 (2022)
6. Lehmann, E. L., Romano, J. P.: Testing statistical hypotheses. Vol. 3. Springer Science & Business Media, 2005.
7. Rockafellar, R. T., Wets, R. J.-B.: Variational analysis. Vol. 317. Springer Science & Business Media, 2009.
8. Romano, J. P.: A bootstrap revival of some nonparametric distance tests. Journal of the American Statistical Association **83**(403), 698–708 (1988)