Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P3542-cd

Multivariate Simulation of Product Fleets based on Usage Data: Case Study on Light Electric Vehicles

Georgios Ioannou

Chair of Reliability and Risk Analytics, University of Wuppertal, Germany.
E-mail: ioannou@uni-wuppertal.de

Semih Severengiz

Sustainable Technologies Laboratory, Bochum University of Applied Sciences, Germany.
E-mail: semih.severengiz@hs-bochum.de

Stefan Bracke

Chair of Reliability and Risk Analytics, University of Wuppertal, Germany. E-mail: bracke@uni-wuppertal.de

As the requirements for technically complex products and their functionality increase, product complexity continues to rise. At the same time, development times and costs must be reduced to ensure that technical products remain marketable. This leads to an increase in possible damage causalities and potential field failures. This applies in particular to the development of new markets, such as electromobility in the light vehicle sector, known as light electric vehicles (LEVs). The battery systems installed in these vehicles harbor a comparatively high risk of function- and safety-critical failures. These present companies with new challenges when operating product fleets in the private and commercial sectors due to the high level of safety and reliability required. To reduce the risk and increase the reliability of LEVs in the field, analysing product data from the field is highly relevant in order to be able to predict the remaining useful life. One way of supporting the reliability analysis process during the utilisation phase of the product life cycle is to simulate and prognose the further use of a product or a product fleet. The existing and simulated usage data can then be used for forecasts regarding the remaining useful life of products. This paper presents the results of a feasibility study in which a concept for the multivariate simulation of product fleets based on usage data from the field is applied in the context of LEVs. Field data from a Kumpan electric 54 e-moped, which was recorded over a period of several days, serves as the data basis. The available data was analysed and used for the multivariate simulation. Finally, the simulation results were compared with the original data and a conclusion was drawn.

Keywords: Multivariate simulation, Product Fleet, Usage Data, Light Electric Vehicle, Multivariate Analysis.

1. Introduction

The development of technically complex products is associated with increasing requirements and functionalities, which means that product complexity is constantly on the rise. At the same time, development times and costs must be continually reduced in order to make the products marketable. This increases the risk of possible causes of damage, potential field failures and the associated complaints and recalls campaigns. Product development and market support for technically complex products therefore face major challenges in terms of safety and reliability. This is particularly true for the development of new markets where, for example, it is not possible to draw on the knowledge and technologies of previous product generations, such as electromobility in the field of light vehicles (e.g., e-scooters or electric bicycles), also known as Light Electric Vehicles (LEVs).

As the sustainable transport transition towards electromobility progresses, the importance of LEVs in urban areas for the commercial and private sectors is also increasing. The global market share of LEVs is forecast to increase from USD 88 billion in 2023 to USD 225 billion in 2033 (Towards Automotive, 2024; Precedence Research, 2024). For example, the battery system installed in LEVs carries a relatively high risk of functional and safety-critical failures and, in the worst case, can lead to thermal runaway and the burning of the vehicle. The Europe-wide RAPEX system (Rapid Exchange of Information System / Safety Gate) has already been used in the past to recall e-scooters from several manufacturers due to the risk of injury, burns and fire (LEVA-EU, 2023). These present companies with new challenges when operating product fleets of LEVs in the private and commercial sectors due to the high level of safety and reliability required.

In order to counteract the risk of failures in the field, it is advisable to include information from the field in the analysis of the LEVs. For example, multivariate analysis of the remaining useful life of individual vehicles or entire vehicle fleets can be carried out taking into account various lifespan variables from the field. The stochastic simulation of usage data from the field can be used decisively here in this process. By simulating the vehicle usage, for example, it is possible to check whether the vehicle is a potential candidate for a particular damage causality (Reinecke, 2021). In addition, the remaining useful life can be determined more precisely on the basis of the simulated usage data.

In this paper the implementation of a concept for the multivariate simulation of products and product fleets based on the Monte Carlo Method in the context of LEV usage data is presented. In Chapter 2, an overview of stochastic simulation based on the Monte Carlo method and its application to field data simulation is shown. The case study, the simulation of LEV usage data, is presented in Chapter 3. The analysed LEV usage data was recorded over a period of several days and contains multiple signals and is based on recorded trips of a Kumpan electric 54 e-moped. Chapter 4 presents the results of the case study and Chapter 5 concludes with a summary of this work.

2. Baseline

Stochastic simulation using the Monte Carlo Method (MCM) is an important tool for modelling and analysing technically complex products due to its ability to represent a closer adherence to a reality (Zio, 2013). MCM uses random sampling to simulate the behavior of random systems on the computer by randomly generating system-describing variables (Christiane Lemieux, 2009). The fundamental aspect of MCM is the generation of pseudo random numbers from a $U[0,1]$-uniform distribution. To ensure the randomness of the generated numbers, a suitable algorithm or generator is used to generate the random numbers. Algorithms such as the inversion method, the composition method or the rejection method are used (Zio, 2013; Waldmann and Helm, 2016). Because of its capabilities, MCM is used in many areas of science, including physics, finance and engineering (Christiane Lemieux, 2009; Waldmann and Helm, 2016).

In the last decade, several papers have been published in the domain of field data simulation based on MCM. In (Feijóo et al., 2011) a developed method for the simulation of correlated wind speeds is presented, where the MCM is used to generate wind speed series based on specified distribution functions for different locations. In (Feijóo and Villanueva, 2016) an overview of methods for the simulation of wind speed time series is given. In particular, the simulation based on descriptive wind speed distribution functions is discussed.

In (Subbiah and Turrin, 2015) and (Turrin et al., 2015), a data-driven Monte Carlo simulation approach was presented that can be used to predict the health status and remaining useful life (RUL) of a product based on its condition monitoring data.

Additionally, Hienzsch (2016) and Hinz et al. (2016) developed two methods for simulating field data on a univariate basis in the form of time series. The simulated time series include the driving speeds of different vehicles. In the work, the simulation is based on alternating extrema and the polynomials fitted in between using MCM. In the second paper, the journeys are simulated on the basis of data transformed by the discrete Fourier transform and the resulting frequency and amplitude values using the MCM. Both methods were presented using case studies from the automotive industry.

In the simulation concept developed by Reinecke (2021) based on the MCM, multivariate, stochastic relationships between the signals can be taken into account on the basis of correlation analyses and classification models. This allows special usage states (e.g., longer dwell times at the same speed) to be modelled, as well as dependencies between historical and current values. Compared to the work presented above, this enables a prognosis of the usage states at product and fleet level (Reinecke, 2021).

3. Case Study: Multivariate Simulation of Light Electric Vehicle Usage Data

This chapter presents the case study 'Multivariate simulation of LEV usage data', which was carried out in this work. The underlying data is first presented in section 3.1. The method used for the simulation, the simulation concept of Reinecke (2021), is then presented and fundamentally explained (cf. section 3.2). Finally, the underlying LEV usage data is analysed and prepared for the simulation (cf. section 3.3).

3.1. *Base of operation*

The underlying LEV usage profiles are derived from the usage of a Kumpan electric 54 e-moped, whose journeys were recorded for various parameters using external hardware and software. For data recording, the CANedge2 Logger (manufacturer: CSS Electronics) was utilised. The logger was connected to the vehicle's CAN bus system.

CAN bus data was collected via soldered wires on the vehicle's wiring harness near the battery connections. The installation was based on the schematics and technical specifications provided by the manufacturer of the Kumpan electric 54 e-moped. A total of three loggers were used simultaneously. The data is stored on the logger's internal 64 GB memory card. These are unprocessed raw CAN bus signals that require further interpretation before analysis. The logger includes a WLAN interface, which connects automatically to known networks. An S3 file server, compatible with the logger, was used for data transmission. Data recording was performed at sampling rates of 10 Hz and 1 Hz.

The recorded raw data was translated into physical values using a CAN Bus Database File (DBC). This conversion was automated using the ASAM MDF software application. The physical measurement data was subsequently stored in a time-series database (InfluxDB). This database enables effective visualisation of data using Grafana. Based on the data stored in InfluxDB, driving profiles were automatically generated for each ride. Calculated driving profiles were stored together with a pseudonymised user ID in a relational database for further analysis.

The database comprises 16 rides that were recorded continuously over a period of four days and approx. 98 hours. Signals from 13 different parameters were recorded for the rides. Table 1 shows the various recorded parameters, which include both parameters at the battery system level and parameters at the overall vehicle level. The signals were recorded at a frequency of 2 Hz (500 ms), therefore almost 319,000 data points are available for each parameter. The data set was manually extended by the parameter *Timestep*, which takes into account the time in whole steps and is needed for the simulation.

Table 1.: The recorded parameters and their units.

Parameter	Unit
Battey_Current_1	[mA]
Battey_Current_2	[mA]
Battey_Current_3	[mA]
Battery_Temperature_1	[°C]
Battery_Voltage_1	[V]
Battery_Voltage_2	[V]
Battery_Voltage_3	[V]
Brake_Light	Binary
Motorcontroller_Temperature	[°C]
Odometer	[m]
Remaining_Distance	[km]
Speed	[km/h]
Torque	[%]
Timestep	-

3.2. *Methodology*

The multivariate simulation of LEV field data presented in this paper is based on the simulation concept of Reinecke (2021). This chapter gives a short overview of the concept. The simulation concept is separated into the four sub-aspects, preparation of the data and simulation plan, the simulation of values used in the concept, the univariate simulation of stochastic independent variables and the multivariate simulation of stochastic related variables. A more detailed description can be found in the corresponding paper.

3.2.1. *Preparation*

The first step involves analysing and preparing the underlying signal and sensor data in the form of time series in order to prepare them for the simulation. This involves identifying implausible characteristics in the underlying data and imputing missing data points in order to generate consistent time series without impurities.

In order to be able to map and simulate different usage states in the concept, these are determined from the time series using the k-means clustering algorithm (Hartigan and Wong, 1979). For this purpose, the time series are segmented into sections of defined length and their stochastic characteristic values (e.g., mean, median, standard deviation) are grouped into the various usage states by the cluster algorithm. The evaluation is carried out using the elbow criterion (Thorndike, 1953) to find the optimal number of usage states.

The simulation is carried out using a simulation plan, which defines the sequence of the parameters to be simulated, and the input parameter required for each simulation of a parameter. The starting point for the simulation plan is a hierarchical cluster analysis (here: agglomerative clustering with the single linkage method) of the underlying parameters in order to determine the correlations between the variables on the basis of the spearman correlation factor. Based on the correlation distance values, the variables are then categorised into groups to be simulated independently. Within each group, the simulation sequence is selected according to the correlation between the variables.

3.2.2. *Simulation of target parameters in both concepts*

The simulation of target parameters, which is used in both simulations concepts, is based on the inversion method for discrete random values and the empirical distribution function of the original values in the respective usage state. To account for distances between the value to be simulated and its k-nearest neighbours (cf. section 3.2.1) in the usage state, the distribution function is adapted according to the spatial distance between the different values.

3.2.3. *Univariate Simulation*

The univariate simulation is used when no parameter (within the corresponding group) has been simulated before or when there is no stochastic dependence on other parameters. The univariate simulation consists of different sub-simulations. Firstly, the previously determined usage states are simulated. This is done on the basis of the sequence of state changes between the different usage states, using it to simulate the next state sequence. For each simulated state, an extreme point simulation is performed, alternately simulating minima and maxima. Finally, the values between two extreme points are simulated with interpolated cubic splines. The result is a simulated time series of the underlying variable with its specific characteristics.

3.2.4. *Multivariate Simulation*

The multivariate simulation is used if influencing variables have already been simulated for the target parameter, which allows the inclusion of stochastic correlations between the target and the influencing parameter. In the first step, a classification model is trained for the identified usage states (cf. section 3.2.1) of the target parameter on the basis of the most highly correlated influencing parameters. For this purpose, a decision tree according to the C4.5-Algorithm (Salzberg, 1994) is used. Based on the classification model, the usage states of the target parameter to be simulated are then classified on the basis of the segmented time series of the influencing parameters. The simulation of the target parameter begins with the random selection of a starting value from the original values of the parameter, for which the original and the classified usage states are identical. Starting from the initial value, the following values and corresponding dwell times are simulated. The result is a simulated time series of the target variable, taking into account the dependencies on the influencing variables.

3.3. *Data Analysis and Simulation Preparation*

Prior to the multivariate simulation, several steps are taken to prepare the data for simulation. First, the parameter signals are cleaned of erroneous values. This applies in particular to sections between the rides, as implausible, missing and incorrect values were recorded by the recording system. The two parameters *Motorcontroller_Temperature* and *Remaining_Distance* are heavily affected by this type of errors, so they are completely removed from further consideration. Additionally to these two parameters, the parameter *Odometer* was also removed from the simulation in order to determine the distance travelled based on the speed over time.

After data preparation, the simulation plan is generated according to the methodology described in section 3.3. The spearman correlation analysis performed for this purpose shows that the three voltage parameters are very strongly correlated ($r = 0.96$ and $r = 0.99$). In order to reduce the simulation time without losing much information, only the parameter *Battery_Voltage_1* is used for the simulation in the following. Compared to this, The parameter *Brake_Light* has a low correlation with the other parameters, which results in a high dividing line within the hierarchical agglomerative clustering algorithm (cf. figure 1).

If a spearman correlation of $|r| \geq 0.25$ is chosen for the classification of the simulation groups according to Reinecke (2021), the parameter *Brake_Light* is divided into a single cluster and simulated univariate. To further reduce the simulation time, the parameter *Brake_Light* is also not simulated due to lack of multivariate correlations with the other parameters. This means that only the right cluster is used for the simulation, which reduced the number of parameters to be simulated to seven.

Starting from one parameter, which is simulated univariate and forms the basis of the simulation, all other parameters in the cluster are then simulated according to their correlation with the influencing parameter using the multivariate simulation concept. The overall simulation is carried out in two steps: Firstly, only 70% of the parameter signals are used for the simulation. Some simulations are carried out for validation and compared with the remaining 30% of the signals. In the second step, after evaluating the simulation results, the entire simulation was carried out, which can then be used for further investigations.

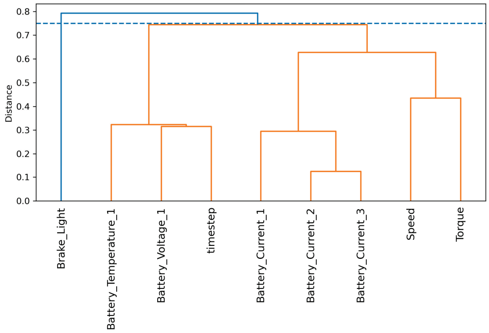


Fig. 1.: Hierarchical agglomerative clustering based on the spearman distance of the recorded parameters. The parameters are splitted at a distance of 0.75 (blue line).

4. Results

The results of the case study are presented in this chapter, focusing on representative data from the first simulation step and a comparison between the final 30% of the original signal and the simulated signals. A total of 17 simulations were carried out to validate the simulation results. Figure 2 shows a segment of the original usage profile (highlighted in blue) based on the distance and the 17 usage profiles simulated from 70% of the signals. The simulated usage profiles show plausible behaviour compared to the original profile, whereby the larger proportion achieves a greater distance with a similar number of rides. The downtimes between rides were also mapped to a suitable amount.

A similar behaviour can be identified by comparing the simulated signals and the respective original signals. Figure 3 shows the orig-

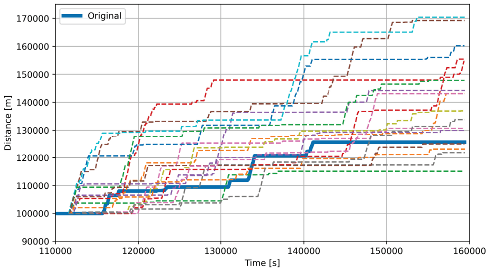
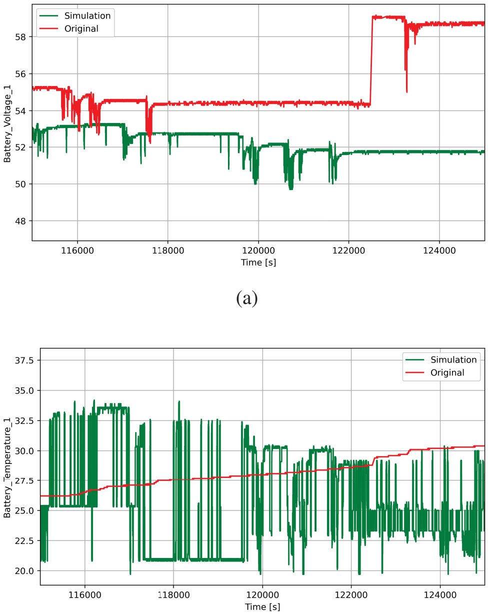


Fig. 2.: Segment of the travelled distance of the original and the simulated usage profiles.

inal and simulated signals for the parameters *Battery_Voltage_1* (a) and *Battery_Temperature_1* (b) in red and green respectively. In (a), the characteristics are adequately simulated. The individual voltage drops and the continuous slight noise can be seen in both curves. A sudden rise in the original voltage curve at around 125,000 seconds is not simulated in the short segment, but can be found in the other parts of the simulations. In addition to the voltage signal, the characteristics of the signals and the specific operating and standstill times have also been modelled accurately for the other parameters. The only exception is the parameter *Battery_Temperature_1*, which shows greater fluctuations in the simulation (cf. figure 3 (b)). One possible cause can be found in the upward trend of the temperature signal due to the many rides in the short recording time, which prevented the battery system from cooling down completely. According to Reinecke (2021), the simulation concept cannot be used for trended parameters, which is the case here and could lead to poor results.

To analyse the correlations between the original and simulated signals, a correlation matrix based on the spearman correlation is used to plot the correlations between all parameters. The correlation matrices are shown in figure 4. The improvable simulations of the parameter *Battery_Temperature_1*, are also reflected in the corresponding correlations with the other parameters. The correlations with the *Timestep* and the current parameters cannot be correctly reproduced by the simulations. The correlations between the parameter *Battery_Voltage_1* and *Timestep* are also in need of improvement. A possible explanation for the weak correlations is the simulation plan division using hierarchical agglomerative clustering (cf. figure 1). The three parameters mentioned above show weak correlations with the other parameter group, which is close to the chosen limit of $|r| \geq 0.25$. However, as their correlations distances are below the selected limit, all parameters are joined in a simulation group. The division into an independent simulation group could possibly improve the correlations of the parameters *Battery_temperature_1*, *Battery_Voltage_1* and *Timestep*. The correlations of the other parameters, in particular for *Speed* and *Torque*, were simulated with high accuracy.



(a)

(b)

Fig. 3.: (a) Segment of the comparison between the original (red) and simulated (green) signals for the *Battery_Voltage_1* parameter; (b) analogous comparison for the *Battery_Temperature_1* parameter.

After validating the simulations with the last 30% of the original signals, finally a total of 84 simulations are carried out on the basis of the original usage profile, presenting different pos-

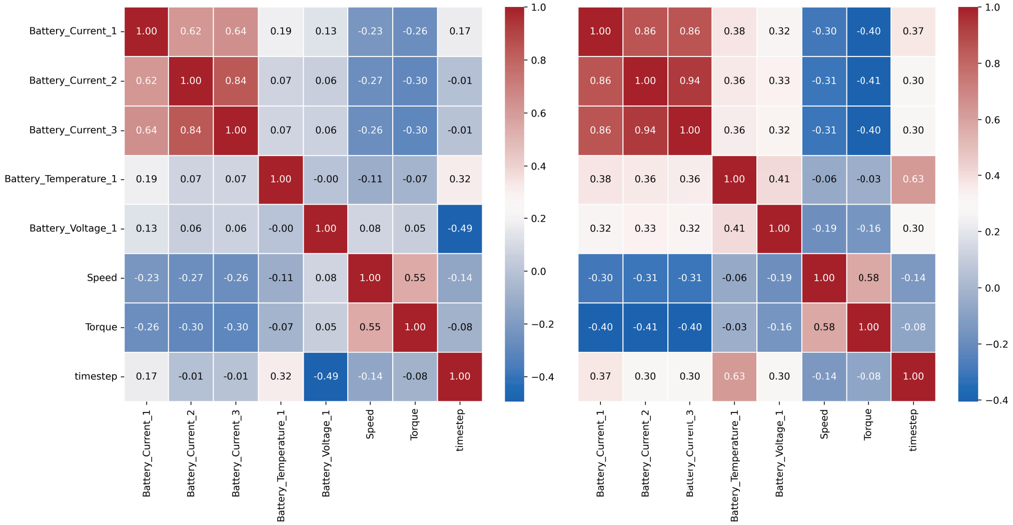


Fig. 4.: Comparison of the spearman correlation matrices between the original signals (left) and simulated signals (right).

sible prospective usage profiles for the user in the given dataset. Each of the 84 simulation contains the individual simulations for each parameter. Figure 5 shows the final result of the case study, exemplified by the travelled distance of the original dataset in blue, as well as the subsequent simulated distances on the basis of the original dataset. Since the travelled distance is not simulated directly (cf. section 3.3), it is calculated as the simulated speed over time.

The multivariate simulation shows overall positive results for a first simulation of LEV usage data within the case study, which can be further optimised by more investigations.

5. Summary and Outlook

In this study, a multivariate simulation of LEV usage data from several signals and the prediction of a usage profile was carried out, which can be used for field and usage failure analysis. For the presented case study, the simulation concept from the work of Reinecke (2021) was used, where stochastic relationships are taken into account in the simulation using different classification and clustering methods. Overall, the usage profile was simulated well and the specific characteristics (e.g., usage and downtime) of the different signals could be accurately mapped in the simulations. A correlation analysis also showed that the mapping of the correlations between the parameters was taken into account and simulated well. The correlation for the time and the parameter *Battery_Temperature_1* show potential for improvement. Further in-depth investigations, especially in the analysis of the underlying data and the generation of the simulation plan, would further optimise the simulation results.

The simulation is based on 16 rides recorded over four days, which may represent a small amount of data for the simulation, for example, the correlations may fluctuate more due to the small database. Further analysis of other case studies and larger amounts of data would also improve the simulation results.

Acknowledgement

This study was conducted as part of the research project ‘Pilot Factory for End of Life Strategies of Light Electric Vehicles‘ (Pilot4CircuLEV) funded by the Federal Ministry of Education and Research (BMBF) under the funding code 033RK110E.

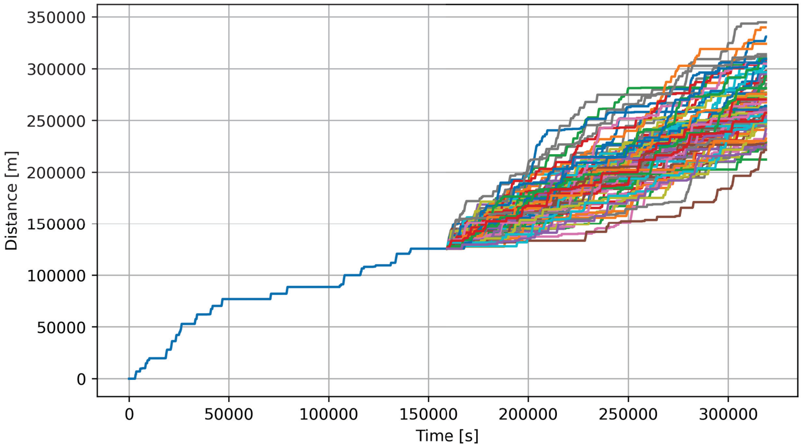


Fig. 5.: The travelled distance of the original usage profile and the simulated usage profiles.

References

Christiane Lemieux (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer New York.

Feijóo, A. and D. Villanueva (2016). Assessing wind speed simulation methods. *Renewable and Sustainable Energy Reviews 56*, 473–483.

Feijóo, A., D. Villanueva, J. L. Pazos, and R. Sobolewski (2011). Simulation of correlated wind speeds: A review. *Renewable and Sustainable Energy Reviews 15*(6), 2826–2832.

Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics 28*(1), 100.

Hienzsch, F. (2016). *Development of methods for simulating individual car rides and determining the behaviour of a car fleet in the field (In german: Entwicklung von Methoden zur Simulation von PKW-Einzelfahrten und Bestimmung des Verhaltens einer Automobilflotte im Feld)*. Masterthesis, University of Wuppertal.

Hinz, M., F. Hienzsch, and S. Bracke (2016). Development of two methods for the characterisation of an automotive fleet behaviour based on the simulation of single car rides. In L. Walls, M. Revie, and T. Bedford (Eds.), *Risk, Reliability and Safety: Innovating Theory and Practice*, pp. 1593–1598. Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742: CRC Press.

LEVA-EU (2023). Rapex warnings 2023. `https://leva-eu.com/rapex-warnings-2023/` (accessed on 29.12.2024).

Precedence Research (2024). Light electric vehicles market size, share, and trends 2024 to 2034. `https://www.precedenceresearch.com/light-electric-vehicles-market` (accessed on 28.12.2024).

Reinecke, F. (2021). *Contribution to the development of a concept for multivariate simulation of the use of technically complex products on the basis of analysed field data (in german: Beitrag zur Entwicklung eines Konzepts zur multivariaten Simulation der Nutzung technisch komplexer Produkte auf Basis analysierter Felddaten)*. Dissertation, Shaker Verlag and University of Wuppertal.

Salzberg, S. L. (1994). C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning 16*(3), 235–240.

Subbiah, S. and S. Turrin (2015). Extraction and exploitation of r&m knowledge from a fleet perspective. In *2015 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–6. IEEE.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika 18*(4), 267–276.

Towards Automotive (2024). Light electric vehicle market size, shares and trends insight. `https://www.towardsautomotive.com/insights/light-electric-vehicle-market-sizing` (accessed on 28.12.2024).

Turrin, S., S. Subbiah, G. Leone, and L. Cristaldi (2015). An algorithm for data-driven prognostics based on statistical analysis of condition monitoring data on a fleet level. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pp. 629–634. IEEE.

Waldmann, K.-H. and W. E. Helm (2016). *Simulation stochastischer Systeme*. Springer Berlin Heidelberg.

Zio, E. (2013). *The Monte Carlo Simulation Method for System Reliability and Risk Analysis*. Springer London.