Kumulative Dissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
Doktor rerum politicarum (Dr. rer. pol.)
an der Wirtschaftswissenschaftlichen Fakultät
der Katholischen Universität Eichstätt-Ingolstadt

DATA AGGREGATION AND SAMPLING PROCEDURES FOR USAGE PROFILING AND CUSTOMER-CENTRIC AUTOMOTIVE SYSTEMS ENGINEERING

vorgelegt von
Kunxiong Ling, M.Sc.

Referent
Prof. Dr. Thomas Setzer

Korreferent
Prof. Dr. Hansjörg Fromm

Tag der mündlichen Prüfung
12. Dezember 2022

München, 2023

Data Aggregation and Sampling Procedures
for Usage Profiling and Customer-Centric Automotive Systems Engineering

Abstract Specifying or deriving customers' needs into detailed requirements becomes essential for holistic customer-centricity in automotive systems engineering, relying heavily on simulation models. However, with the increase of system complexity, customer diversity, and difficulty of customer data acquisition, it becomes challenging to specify model inputs that represent individual customer usage behavior across the whole product lifecycle. To let these challenges be tackled, this dissertation addresses the problem of customer usage profiling to support decision-making in the context of automotive systems engineering. First, two fundamental data engineering procedures are investigated, including data aggregation for feature reduction and sampling for representing the customer fleets with a few selected reference customers. After data preprocessing, a method for preparing the model inputs from aggregate customer data, i.e., usage profiling, is developed. Usage profiling applies meta-heuristics, synthesizing sufficiently representative from aggregate fleet data. Furthermore, a decision support system is developed to deploy the aggregation, sampling, and usage profiling into model-based automotive system engineering processes. As both data and various models are connected, digital twinning is applied. Using real-world fleet data, an evaluation case study for determining lifetime requirements indicates that the methodology is plausible and capable of consolidating customer-centricity into automotive systems engineering processes.

Keywords Aggregation, Automotive, Customer Centricity, Sampling, Systems Engineering, Usage Profiling.

Datenaggregations- und Samplingverfahren
für Nutzungsprofilierung und kundenorientierte Fahrzeugsystemtechnik

Kurzfassung Die Spezifizierung oder Ableitung von Kundenbedürfnissen in Form von detaillierten Anforderungen ist für eine ganzheitliche Kundenorientierung in der Fahrzeugsystemtechnik, die sich stark auf Simulationsmodelle stützt, unerlässlich. Mit zunehmender Systemkomplexität, Kundenvielfalt und der Schwierigkeit, Kundendaten zu erfassen, wird es jedoch schwieriger, Modelleingaben zu spezifizieren, die das individuelle Nutzungsverhalten der Kunden über den gesamten Produktlebenszyklus hinweg repräsentieren. Um diese Herausforderungen zu bewältigen, befasst sich diese Dissertation mit dem Problem der Erstellung von Kundennutzungsprofilen zur Unterstützung der Entscheidungsfindung im Kontext der Fahrzeugsystemtechnik. Zunächst werden zwei grundlegende Verfahren der Datenverarbeitung untersucht, darunter die Datenaggregation zur Merkmalsreduktion und das Sampling zur Repräsentation der Kundenflotten mit wenigen ausgewählten Referenzkunden. Nach der Datenvorverarbeitung wird eine Methode zur Aufbereitung der Modelleingaben aus den aggregierten Kundendaten, das sogenannte Nutzungsprofilierung, entwickelt. Bei der Nutzungsprofilierung werden Meta-Heuristiken eingesetzt, die es ermöglichen, aus den aggregierten Flottendaten ausreichend repräsentative Daten zu synthetisieren. Darüber hinaus wird durch den Einsatz von Aggregation, Sampling und Nutzungsprofilierung in modellbasierten Fahrzeugentwicklungsprozessen ein Entscheidungsunterstützungssystem entwickelt. Da sowohl die Daten als auch die verschiedenen Modelle miteinander verbunden sind, wird der Rahmen der Erstellung digitaler Zwillinge angewendet. Anhand von realen Flottendaten zeigt eine Evaluierungsfallstudie zur Ermittlung von Lebensdaueranforderungen, dass die Methodik plausibel und in der Lage ist, die Kundenorientierung in die Systementwicklungsprozesse der Automobilindustrie zu integrieren.

Schlüsselwörter Aggregation, Fahrzeugsystemtechnik, Kundenorientierung, Nutzungsprofilierung, Sampling.

用于使用行为综合和以客户为中心的汽车系统工程的数据聚合和抽样过程

摘要 在汽车系统工程中，细化客户的需求，对于以客户为中心的整体工程来说是至关重要的，这在很大程度上依赖于仿真模型。然而，随着系统复杂性、客户多样性和客户数据获取难度的增加，指定能代表整个产品生命周期中的客户使用行为的模型输入变得具有挑战性。为了应对这些挑战，本论文讨论了在汽车系统工程背景下支持决策的客户使用概况问题。首先，研究了两个基本的数据预处理过程，包括用于减少特征的数据聚合和客户车队的抽样方法。在数据预处理之后，我们开发了一种从聚合的客户车队统计数据中合成模型输入的方法，即使用行为综合。使用行为综合应用了元启发式方法，允许从总的车队数据中综合出足够的代表性。此外，在将聚合、抽样和使用行为综合部署到模型驱动的汽车系统工程流程中时，开发了一个决策支持系统。由于数据和各种模型都被连接起来，数字孪生的框架得以应用。使用真实世界的车队数据，确定寿命要求的评估案例研究表明，该方法是可行的，能够将以客户为中心整合到汽车系统工程流程中。

关键词 抽样，汽车，使用行为分析，数据聚合，系统工程，以客户为中心。

Acknowledgments.

Time flies. Originated from the deep mountain region near Shanghai, listening to "Overpass Graffiti" from Ed Sheeran, the "Eismann" is bringing his dissertation to an end, despite the pandemic, in a quiet night near the heart of Munich, one of the most beautiful city on earth, probably.

Things started from that summer in 2018. I worked on my master thesis, simulating the water behavior inside over 100 pipelines in a fuel cell. Each has a diameter of less than a millimeter and is filled with wet hydrogen and air mix flow. Then I saw the doctoral job description from BMW Group, which is about "Customer Centricity" but for engine development. I wanted just to figure out why these two fields could work together, as I studied everything about "Energy", including management and machine learning as well. Then I passed five rounds of job interviews and got supported and accepted by my doctoral father, Prof. Dr. Thomas Setzer, my supervisor Jan Thiele, and my group leader Dr. Falk Hannemann. Thank you all for trusting me and giving me a chance to exactly figure out that tricky topic and bring the worlds of business informatics and mechanical engineering together.

I could not forget that Falk told me to "focus". It let my innovation engine turn without overheating. It keeps me alive until now. And I believe that will allow me to keep going further safely.

Jan, I appreciate you so much. As the nicest supervisor I have seen in my life, you did not just guide me on the proper paths but a real sense of teamwork. Working together with you and our colleagues in the team of AAE (Anforderungsgerechte Auslegung und Erprobung, requirements-driven design and testing) makes up the best working experience I have ever had in my life!

Thomas, with unclear directions and a bunch of trials and errors in my first year, you gave me unlimited patience and confidence. You taught me what research is on uncountable Thursdays in Ingolstadt, where we discussed the works and papers together. You even opened R Studio and tested the ideas at once. You are the coolest professor I have ever seen in my life. From you, I also learned a lot about being a person and the sense of being kind and humble!

Although a doctoral degree means the candidate can conduct independent research activities, teamwork cannot be neglected. Thank you so much, Zhao Song and Nishant Shah! You spent over a half year working on your Master's theses to help investigate my research fields. I learned a lot from you both in system modeling and stochastic optimization. Really appreciate. Also,

thank you so much, Zihan Kong, for applying your extraordinary skills in web development and bringing our decision support system alive! Moreover, thank you, Daniel Barta, for supporting me in implementing the automated model verification program and John Vicente for investigating various decomposition methods for histogram data, including that tricky tensor one. I wish all of you all the best in the future!

A special thank goes to Laurent Chauvigné, Dr. Guang Rao, and Dr. Cornelis Wirth for encouraging me to broaden the network and supporting me to begin my career in the fantastic company.

Also, I would like to thank my colleagues at BMW Group who supported my work and provided valuable discussion, especially Helena Alder, Dr. Christine Spannagl, Christian Vetter, Dr. Benjamin Esterl, Maximilian Spannagl, and Thomas Stadlbauer.

For the smooth administrative process at the university, thank you very much, Waltraud Fischermeier! For the fruitful discussion and the support of the enrollment process in the doctoral program, I would like to thank Felix Schulz, Nathalie Balla, and Prof. Dr. Benjamin Buchwitz.

For the rigorous review and supporting my doctoral defense, I appreciate the committee, including Prof. Dr. Hansjörg Fromm, Prof. Dr. Ulrich Küsters, and Prof. Dr. Jens Hogreve!

For the happy hours in work-life balancing, many thanks to Zhengtian Ai, Dr. Fengmin Du, Giovanni Filomeno, Melissa Gresser, Yuran Liang, Zhenzheng Lu, Xiong Xiao, Yuxin Xiong, Ding Zhang, and all my friends! My psychological health is therefore kept in a great condition.

Last but not least, my family's dedicated support consolidates my determination to pursue a doctoral degree. Furthermore, thank you, Ziqing, for being in my life. Now, your "Eismann" is evolving into "Dr. Eismann"!

Das Leben ist schön. Niemals aufgeben 😄

Contents

Introduction.

This cumulative dissertation was accomplished at the Chair for Business Administration and Business Informatics, Ingolstadt School of Management, Catholic University of Eichstätt-Ingolstadt, in cooperation with the BMW Group. It addresses the problem of customer usage profiling to support decision-making in the context of automotive systems engineering, in particular, through two fundamental data engineering procedures: aggregation and sampling.

This dissertation accords § 8 (2) in the specialized doctoral regulations (FachPromO) of the Ingolstadt School of Management, Catholic University of Eichstätt-Ingolstadt, dated 13. Sep. 2019, and accumulates the following four research articles:

- Article 1 (Aggregation): K. Ling, J. Thiele, and T. Setzer, "Loss-Aware Histogram Binning and Principal Component Analysis for Customer Fleet Analytics", (Working Paper).
- Article 2 (Sampling): K. Ling, J. Thiele, and T. Setzer, "Usage Space Sampling for Fringe Customer Identification", in: *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 1748-1757, HICSS, 2021, doi: 10.24251/HICSS.2021.212 (VHB-JQ3: C).
- Article 3 (Profiling): K. Ling, N. Shah, and J. Thiele, "Customer-Centric Vehicle Usage Profiling Considering Driving, Parking, and Charging Behavior", in: *Proceedings of the 23rd IEEE International Conference on Intelligent Transportation Systems*, pp. 146-151, IEEE, 2020, doi: 10.1109/ITSC45102.2020.9294669 (H-Index: 73).
- Article 4 (Integration): K. Ling, "Digital Twinning from Vehicle Usage Statistics for Customer-Centric Automotive Systems Engineering", (Working Paper).

Motivation

Systems engineering coordinates complex components, functions, products, and services, thus playing a central role in the automobile industry. In automotive systems engineering, specifying or deriving customers' needs into detailed system requirements becomes essential towards holistic customer-centricity. According to the International Council on Systems Engineering (INCOSE), systems engineering "focuses on defining customers needs and required functionality early in the development cycle, documenting requirements, and then proceeding with design synthesis and system validation while considering the complete problem" [1]. In the management society, Fader (2020) outlines the importance of building representative customer profiles for pursuing customer-centricity, i.e., to ensure the coverage of products or services on the wide range of customer usage behavior [2].

Currently, two types of methods exist for the extraction of usage behavior from customer fleets, i.e., logging time-series sensor signal traces (logging) and aggregating long-term statistic data (aggregation), schematically shown in Table 1.

Table 1: An exemplary comparison between logging data and aggregate data for one vehicle.

Exemplary data for one vehicle through logging:

Timestamp	Velocity in km/h	Acceleration in m/s^2	Slope in °	External Temperature in °C	···
2021-03-05 13:58:00	33.2	0.72	2.2	26	···
2021-03-05 13:58:01	35.8	0.31	1.5	26	···
···	···	···	···	···	···

Exemplary data for one vehicle through aggregation:

Total mileage in km	Duration in h when velocity $\leqslant 10$ km/h	···	Duration in h when velocity > 200 km/h	Trips with distance $\leqslant 5$ km	···	Trips with distance > 400 km	···
67345	320	···	0	1450	···	6	···

Logging data preserves the sequential information of a vehicle with a predefined temporal resolution (e.g., 1 Hz or one sample per second). Logging from customer fleets helps reproduce customers' usage history and can directly serve as inputs for systems simulation models. However, the following limitations have pushed back the wide-ranging implementation of installing data loggers in customer vehicles worldwide on sale: (i) time-series data with a growing number of sensors continuously accumulate throughout the vehicle lifetime, which significantly increases the operational costs of automotive service providers in data cleansing and time-series analysis; (ii) long-term logging data are generally too long for system simulations, especially for what-if analysis; and (iii) personal information could be reconstructed based on those logging data (i.e., driver fingerprinting [3–5]), which could seriously threaten customers' privacy and violate Article 89 of the European General

Data Protection Regulation (GDPR) [6].

Instead, it is common to aggregate the signal traces into statistical usage data, to save them on control units for service purposes, and to acquire them via vehicle telemetries over the air [7, 8]. These aggregate data range from average, summary (e.g., the total mileage shown in Table 1), deviation, quantiles, up to histogram values for all measurements. Recital 162 of GDPR implies that aggregate data, instead of personal data, could "safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality" [6]. Esser et al. (2018) [9] highlighted that data aggregation protects privacy, as it suppresses precise sequential information and disables reconstructing personal information from customers. Furthermore, data aggregation is particularly suitable for customer fleet analytics, as they significantly reduces storage volumes and operational costs. More importantly, data aggregation allows for combining individual customer histograms into a joint fleet histogram by simply adding up their histogram values, enabling rapid usage profile synthesis investigated in this dissertation.

Let us dive deeper into the customer-centric automotive systems engineering. Automotive manufacturers, suppliers, and engineering service providers are transforming document-based development processes into systems engineering based on data-driven or simulation models [10].

With the rise of machine learning and deep learning, data-driven models are popular in software-related fields such as social networks and internet services. These models are generally interpretative yet, up to now, hardly explainable and rely heavily on data [11]. Also, Articles 13-15 of the GDPR require "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" to be available [6]. Therefore, it is infeasible to train machine learning or deep learning models alone to make explainable predictions and provide guidance for vehicle systems that are complex but understandable, as they are engineered by humans.

Alternatively, system simulation models are well explainable, supporting engineers and project managers to analyze the inner mechanisms with their domain expertise under various scenarios. However, compared to machine learning models, physics-based system simulation models are dynamic (or quasi-stationary). Thus, they strictly require time-series signals for relevant physical quantities as inputs. Typically, the simulation inputs come from legal testing procedures, e.g., Worldwide harmonized Light vehicles Test Procedure (WLTP) [12] and Real Driving Emissions (RDE) Regulation [13]. Those procedures are relatively short and not customer-centric. Hence, it is worthwhile to investigate how to provide customer representative inputs for those system simulation models based on the data acquired from customer vehicles (fleets).

Currently, various research activities about vehicle usage profiling focus solely on synthesizing a small number of driving cycles as system simulation inputs [14]. However, little guidance exists on deriving comprehensive usage profiles from aggregate customer fleet data and integrating those customer usage profiles into novel cyber-physical model-based automotive development processes. More essentially, systematic know-hows remain to be investigated in preprocessing the customer

data before usage profiling, in particular in (i) aggregating operational data for the whole fleets with the awareness of information losses and (ii) selecting representative customer samples for large-scale customer simulations with a focus on vehicles with potentially critical loads induced by fringe customer usage behaviors.

This dissertation addresses the highlighted research gaps above by four individual research articles. Their relationship is described in the following from the perspective of the lifecycle of customer fleet data.

Overview

Towards customer-centric automotive systems engineering, the data acquired from customer fleets shall go through the product development process and benefit themselves via customer aftersales services. In addition to data acquision from aftersales services, if the research gaps identified in the previous section could be filled, the lifecycle of customer fleet data, shown in Figure 1, would be enclosed.

The lifecycle begins with the acquisition of customer fleet data. Suppose there is a fleet of n customer vehicles, each of which has m sensors to describe the vehicle usage behavior. For vehicle i and sensor j, until the moment of data acquisition, p_{ij} measurements are acquired to describe its operational history, represented by X_{ij}. All of the customer vehicles and their sensors together can be regarded as a field operational dataset $\boldsymbol{X}$. Due to the issues of data logging outlined in the previous section, in practice, this dataset is not fully acquired nor ingested to backend servers. The following paragraphs briefly describe the four essential parts in the lifecycle, which are in the scope of this dissertation. The parts also correspond to the four research articles to be summarized later on.

Aggregation As p_{ij} becomes huge throughout the lifetime of vehicle i, the high order of dataset $\boldsymbol{X}$ ought to be reduced for usage analytics of the whole fleet, which is done by data aggregation. After data aggregation such as histogram binning, tensor $\boldsymbol{X}$ is aggregated in a matrix with n customers (rows) and n_x attributes (columns). Assuming that each sensor is binned with k intervals, n_x equals mk, which typically reaches or exceeds a magnitude of 10^2 and could lead to "the curse of dimensionality" [15]. Hence, n_x attributes are to be represented by $\widetilde{n}_x$ attributes, where $\widetilde{n}_x \ll n_x$.

Sampling The aggregated dataset has $\widetilde{n}_x$ features and n observations. For complex data processing such as system simulations, the temporal and spatial complexity is $O(n)$, as the processing takes place for those customer vehicles one by one individually. Despite having the dimensionality reduced, a huge n could make the simulation tasks over budget, especially when using them to determine requirements for different market regions. Therefore, it is essential to perform sampling, i.e., n customers are represented by $\widetilde{n}$ samples before preparing profiles for simulation, where $\widetilde{n} \ll n$.

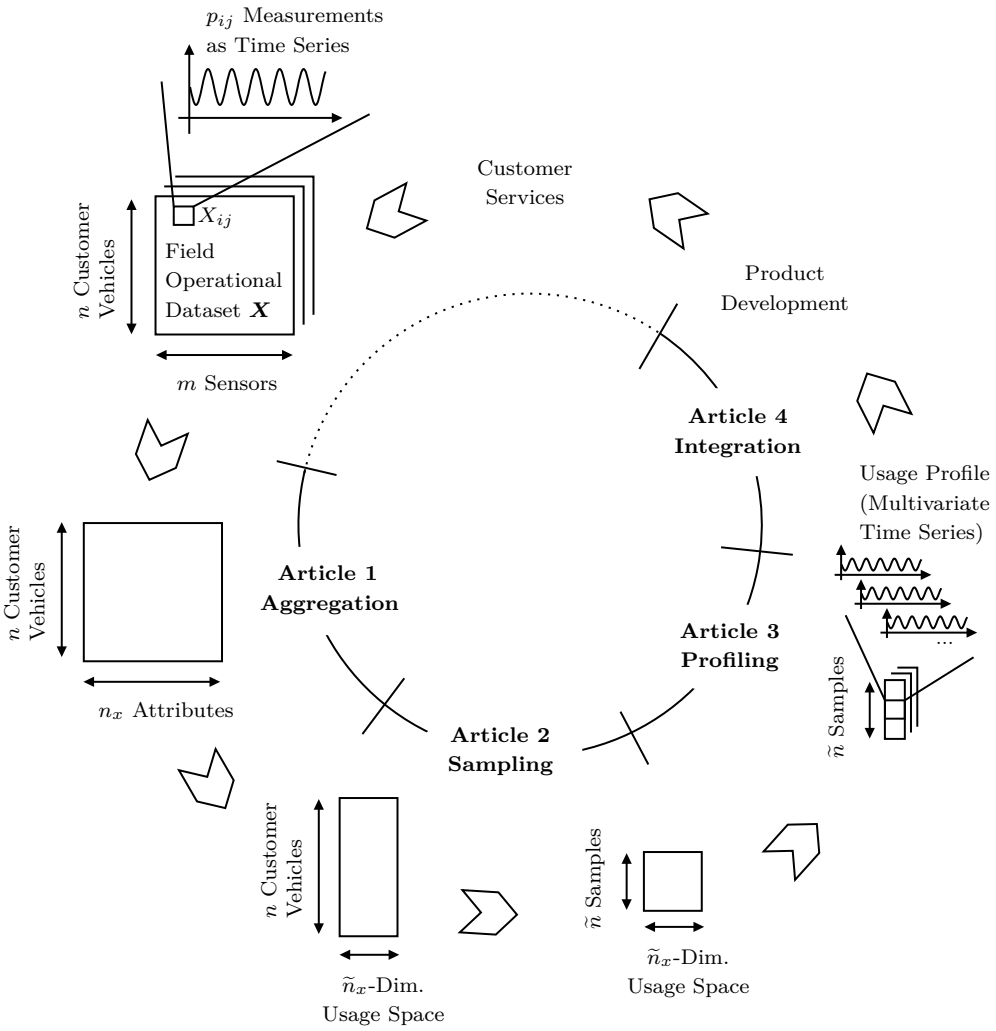

Figure 1: The lifecycle of customer fleet data in customer-centric automotive systems engineering.

Profiling Given $\widetilde{n}$ customer samples, their usage statistics are transformed or interpreted into usage profiles which can be regarded as inputs for system simulation models, referred to as usage profiling. These inputs are usually time-series signals, including necessary driving conditions and environmental information such as velocity, road slope, and external temperature.

Integration The synthesized customer usage profiles provide quantitative information and allow customer-centric decision-making in automotive systems engineering. If the scale, complexity, and versatility of automotive products are considered, usage profiling would be widely applied to a large scale. Correspondingly, an information system is necessary to integrate data, models, and to automate decision support pipelines including aggregation, sampling, profiling and further evaluation.

Together with customer services from which $\boldsymbol{X}$ is acquired, the lifecycle of customer fleet data is accomplished. This dissertation contributes to the fields of fleet data analytics and decision support systems in the form of four research articles, summarized in the following.

Article 1 Aggregation

K. Ling, J. Thiele, and T. Setzer, "Loss-aware histogram binning and principal component analysis for customer fleet analytics", (Working Paper).

We consider a typical aggregation process, i.e. histogram binning followed by principal component analysis (PCA), one of the most widely used dimension reduction techniques. Both of them reduce the dataset's dimensionality, whereas a part of information is lost. With sufficient measurements in general, finer binning (larger n_x) reduces the information loss in binning. More principal components (larger $\widetilde{n}_x$) also reduce the information loss from binned data to PCA aggregated data. However, the granularity of binning influences the total information loss across the two-step aggregation process in a complex fashion. In practice, binning takes place on-board (in the control units of vehicle sensors). It remains unclear if we should make binning coarser to improve the variance explained without increasing $\widetilde{n}_x$. Therefore, guidance should be provided on configuring the aggregation process (binning and PCA) with the awareness of information losses without logging the sensor data, which is addressed in Article 1.

Article 1 enables information loss estimation for the two-step process using Monte-Carlo simulation. First, the measurements of sensor data according to the vehicle usage behavior is modeled by Beta distributions. Subsequently, the scale and correlation structure of the customer fleet sample dataset is modeled by five parameters. Then, the aggregation process is simulated considering the number of bins and principal components remained. To systematically investigate the information loss mechanism, a sensitivity study is carried out. As a result, we found that for fleet usage data, the loss behavior of binning followed by PCA could be quantified empirically. Under the assumption of Beta distributions etc., the result implies that a higher resolution of histogram bins per sensor can generally improve the performance of PCA. The finding is also validated in a case study using a real-world aggregated dataset with 1454 vehicles from two different car segments. In addition, the limitation of this method is discussed using two counterexamples. Furthermore, the approach guides the quantity and relative trends of information loss for engineers and data scientists to configure the preprocessing steps for customer fleet analytics beneficially.

K. Ling contributed to the experimental investigation, methodology development, and most writing tasks. J. Thiele supervised the investigation and contributed to the design of the Monte-Carlo simulation. T. Setzer supervised the result interpretation and contributed to the introduction, hypothesis making, and conclusion sections.

Article 2 Sampling

K. Ling, J. Thiele, and T. Setzer, "Usage Space Sampling for Fringe Customer Identification", in: *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 1748-1757, HICSS, 2021, doi: 10.24251/HICSS.2021.212.

Key reliability indicators (e.g., bearing wear) for automotive engineering usually depend on extreme vehicle loads (e.g., engine start gradients). Those customers who potentially reach those loads are called fringe customers. However, due to the complex relationships between various physical components, those extreme loads are mostly not linearly dependent on the vehicle usage behavior. However, research activities of sampling algorithms are primarily targeted at representing population distribution, whereas the coverage of latent fringe customers remains a challenge. As addressed in Article 2, a novel sampling method has to be developed, which should be deterministic, representable for distribution, and should be able to cover the fringe customers.

Article 2 introduces a sampling method that is deterministic and particularly suitable for appropriately identifying non-linear latent fringe customers based on their usage data. This method integrates four treatments, including (i) PCA-based usage space analysis (addressed in Article 1), (ii) convex hull-based fringe sampling to capture samples on the outer boundary of reduced usage space, (iii) Halton sequence-based core sampling to enable the representativeness of sample distribution for high randomness, and (iv) Voronoi tessellation-based market volume weighting to reduce the distribution losses. Using three benchmark functions (Ackley, Schwefel, and random) under low dimension (5D) and high dimension (500D), various combinations of the four treatments are tested. For fringe customers which are geometrically near the boundary of their usage space, results indicate that usage space analysis combined with fringe sampling clearly outperforms random sampling. The representativeness of sample distribution is enhanced by core sampling. Market volume weighting further improved the distribution representativeness, whereas only at low dimension.

K. Ling contributed to the integration of sampling methodology, design and implementation of experimental validation, and all writing tasks. J. Thiele developed the fringe sampling algorithm and supervised the investigation. T. Setzer supervised the hypothesis making, evaluation, literature review, and writing.

Article 3 Profiling

K. Ling, N. Shah, and J. Thiele, "Customer-Centric Vehicle Usage Profiling Considering Driving, Parking, and Charging Behavior", in: *Proceedings of the 23rd IEEE International Conference on Intelligent Transportation Systems*, pp. 146-151, IEEE, 2020, doi: 10.1109/ITSC45102.2020.9294669.

As a representative customer usage profile, a driving cycle alone can hardly represent how the

customer usage of vehicles affects product loads and system reliability. With the rise of electromobility and the demand for privacy preservation, novel approaches are necessary to derive realistic usage behavior solely based on long-term sensor measurement aggregates, e.g., histogram values. Surprisingly, Esser et al. (2018) [14] found that given massive testing data from vehicle manufacturers, which are time-series signals, driving profiles can be generated by selecting trips or micro-trips from those signals and recombining them to represent the usage statistics optimally. However, considering lifetime reliability and electromobility, it remains unclear how to integrate parking and charge behavior into the profiles.

Targeting to this research gap, Article 3 introduces an integrated profiling method. For a selected sample as a result from Article 1-2, first, a time table (e.g., one week) is allocated. The time table consists of driving and parking sections, targeting to various duration histograms as a part of aggregate data. Then, meta-heuristics is applied to select driving sections from a trip library established from in-house testing fleets. Afterward, the selected sections as multivariate time-series data are concatenated according to the time table. The concatenated profile can be regarded as inputs for system simulation. At the end of each driving section during simulation, it is stochastically inferred if the vehicle is charged during the parking section. Regarding two testing vehicles as customer vehicles, their logging data over six months is compared to the corresponding synthesized profiles. Evaluation shows that also complex indicators such as engine starts for plug-in hybrid electric vehicles reach a relative error within 5%.

K. Ling developed the profile integration method, contributed to the evaluation design, and formulated the majority of paper. N. Shah developed the optimal trip selection algorithm and implemented the evaluation. J. Thiele developed the trip allocation and supervised the work. In addition, the simulation model in Article 3 supports another research article: M. Dietrich, K. Ling, R. Schmid, Z. Song, and C. Beidl, "Design and evaluation of an engine-in-the-loop environment for developing plug-in hybrid electric vehicle operating strategies at conventional test benches," *Automotive and Engine Technology* 6, pp. 247–259, Springer, 2021, doi: 10.1007/s41104-021-00090-5.

Article 4 Integration

> K. Ling, "Digital Twinning from Vehicle Usage Statistics for Customer-Centric Automotive Systems Engineering", (Working Paper).

Article 1-3 contribute to the methodology development and enable the derivation of quantitative requirements and verification programs tailored to individual usage behavior of customer vehicles. Given the high flux of requirements from development engineers and managers of vehicle projects, the decision support processes, i.e., aggregation, sampling, profiling and further quantitative evaluation, should be integrated into a decision support system considering data management, model management, and process automation. To integrate operational data and models, the concept of digital twins, i.e., simulated virtual representations for real physical products with connectivity, has

been pioneered in product development. However, the majority of current research activities focus on real-time data acquisition and real-time feedback from simulation results, which are feasible for business-to-business (B2B) use cases, but infeasible for large-scale customer fleets where no lifetime data logging but data aggregation is possible. Thereby, it remains unclear if it is feasible to industrialize the fleet analytics for aggregate usage statistics by building digital twins (digital twinning) and to integrate the techniques into an information system for potential use cases in automotive systems engineering.

Article 4 addresses the research gap by presenting a decision support system framework, aiming at building and applying digital twins for various customer markets. The digital twinning is enabled by usage profiling and system simulation using aggregate data from customer fleets and logging data from testing fleets. The feasibility of proposed system framework is evaluated by a proof of concept. Given a customer fleet of 57110 vehicles from three market regions and 657909 trip recordings from an in-house testing fleet of 823 vehicles, four metrics which are available form their control units are selected, i.e., average velocity, engine starts, fuel consumption, and time fraction of recuperation. Result shows that the prediction of digital twin reference profiles covers 99% of vehicles from the real-world customer fleet. Furthermore, two potential use cases are presented which are complex but possible to be implemented based on the framework: (i) requirement localization for supplier selection, and (ii) recall prioritization for predictive maintenance.

As the sole author, K. Ling contributed to the design, demonstration, and evaluation of the decision support system and the whole paper.

Conclusion

This dissertation of four research articles enables seamless integration of customer-centricity in model-based automotive systems engineering driven by usage profiling from aggregate customer fleet data. To ensure quality and reduce the cost of decision support, the preprocessing of data aggregates is deeply investigated, particularly histogram binning, principal component analysis, and sampling method. The aggregated and sampled usage statistical data are then coupled with simulation models by means of usage profiling. For industrial application, a framework of decision support system driven by digital twins has been put out. The feasibility of the integrated pipeline and system framework is verified by case studies. Promising use cases for customer-centric automotive systems engineering are demonstrated. Based on the methods and the system framework proposed in this dissertation, various new research perspectives are worthwhile of further investigation. For example, sequential aggregate data in histogram time-series may significantly enhance usage profiling quality to a large extent. The profiling method could also evaluate the representativeness of aggregate customer fleet data for long-term reliability or quality issues. Furthermore, the customer profiles generated for decision support could be used for predictive maintenance by coupling the simulated damage indicators with diagnostic events by classification models.

References

[1] C. Haskins, K. Forsberg, M. Krueger, D. Walden, and D. Hamelin, "Systems engineering handbook," in *INCOSE*, vol. 9, pp. 13–16, 2006.

[2] P. Fader, *Customer centricity: Focus on the right customers for strategic advantage*. Wharton digital press, 2020.

[3] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2016.

[4] S. Ezzini, I. Berrada, and M. Ghogho, "Who is behind the wheel? driver identification and fingerprinting," *Journal of Big Data*, vol. 5, no. 1, pp. 1–15, 2018.

[5] A. Bouhoute, R. Oucheikh, K. Boubouh, and I. Berrada, "Advanced driving behavior analytics for an improved safety assessment and driver fingerprinting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2171–2184, 2018.

[6] European Parliament and Council of European Union, "EU general data protection regulation 2016/679 (GDPR)." `http://data.europa.eu/eli/reg/2016/679/oj`, 2016.

[7] B. Schlegel, *Off-Board Car Diagnostics Based on Heterogeneous, Highly Imbalanced and High-Dimensional Data Using Machine Learning Techniques*, vol. 14. Kassel University Press, 2019.

[8] B. Martens and F. Mueller-Langer, "Access to Digital Car Data and Competition in Aftermarket Maintenance Services," *Journal of Competition Law & Economics*, vol. 16, no. 1, pp. 116–141, 2020.

[9] A. Esser, F. Kohnhäuser, N. Ostern, K. Engleson, and S. Rinderknecht, "Enabling a privacy-preserving synthesis of representative driving cycles from fleet data using data aggregation," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1384–1389, IEEE, 2018.

[10] N. Smith, "How to integrate model-based systems engineering across automotive EE domains," in *SAE Technical Paper Series*, SAE International, Apr. 2016.

[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020.

[12] I. Riemersma, "Technical Report on the development of a Worldwide Worldwide harmonised Light duty driving Test Procedure (WLTP)," in *72nd GRPE*, no. GRPE-72-02, UNECE, 2016.

[13] A. Zardini and P. Bonnel, *Real Driving Emissions Regulation: European methodology to fine tune the EU real driving emissions data evaluation method.* Publications Office of the European Union, 2020.

[14] A. Esser, M. Zeller, S. Foulard, and S. Rinderknecht, "Stochastic synthesis of representative and multidimensional driving cycles," *SAE International Journal of Alternative Powertrains*, vol. 7, no. 3, pp. 263–272, 2018.

[15] R. Bellman, *Dynamic Programming*. Princeton University Press, 1972.

Article 1

Aggregation.

Loss-Aware Histogram Binning and Principal Component Analysis for Customer Fleet Analytics

Kunxiong Ling*†, Jan Thiele†, and Thomas Setzer*‡

*Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany

†BMW Group, 80809 Munich, Germany

‡FZI Research Center for Information Technology, 76131 Karlsruhe, Germany.

Working Paper

Abstract We propose a method to estimate information loss when conducting histogram binning and principal component analysis sequentially, as usually done in practice for customer fleet analytics. Coarser-grained histogram binning results in less data volume, fewer dimensions, but more information loss. Considering fewer principal components results in fewer data dimensions but increased information loss. Although information loss with each step is well understood, little guidance exists on the overall information loss when conducting both steps sequentially. That is because binning-granularity impacts information loss with subsequent PCA in a complex fashion. We use Monte Carlo simulations to regress information loss on the number of bins and principal components, given few parameters of a sensor data matrix related to its size (scale) as well as row- and column-wise correlations. A sensitivity study shows that information loss can be approximated well given sufficiently large data matrices. Using the number of bins, principal components, and two correlation measures, we derive an empirical loss model with high accuracy. Furthermore, we demonstrate the benefits of estimating information losses and the representativeness of total loss in evaluating the accuracy of k-means clustering for a real-world customer fleet dataset. For preprocessing sensor data which are aggregated from sufficient number of samples, are continuously distributed, and can be represented by Beta-distributions, we recommend not to coarsen the binning before performing PCA. Using two counterexamples, we illustrate the limitation of the loss-aware method.

Article 2

Sampling.

Usage Space Sampling for Fringe Customer Identification

Kunxiong Ling*†, Jan Thiele*, and Thomas Setzer†

*BMW Group, 80809 Munich, Germany

†Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany

Published in Proceedings of the 54th Hawaii International Conference on System Sciences
https://doi.org/10.24251/HICSS.2021.212

Abstract With large numbers of available customers, it is often essential to select representative samples for reasons of computational cost reduction and upstream advanced data analytics. However, for many analytical procedures, the usage behavior observed from a smaller sample of customers must indicate well the fringe of usage and its relation to extreme product loads. Due to the high complexity of technical or service systems, it remains challenging to minimize the number of samples while sufficiently capturing the fringe customers. With the availability of data related to usage behavior, we consider a sampling method to address this problem by analyzing the customer usage space before sampling, then separately sampling fringe and core customers, and weighting the samples afterwards. Experimental results show that the method can identify fringe customers with significantly fewer, yet reproducible samples, while maintaining the distribution representativeness of customer population to a large extend.

Article 3

Profiling.

Customer-Centric Vehicle Usage Profiling Considering Driving, Parking, and Charging Behavior

Kunxiong Ling*†, Nishant Shah*‡, and Jan Thiele*

*BMW Group, 80809 Munich, Germany

†Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany

‡Technical University of Munich, 80333 Munich, Germany

Published in Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)
https://doi.org/10.1109/ITSC45102.2020.9294669

Abstract We introduce a novel method to generate customer vehicle usage profiles, representing driving, parking, and charging behavior. Synthesizing usage profiles is the key to support decision-making for customer-centric vehicle development. So far, current methods for vehicle development focus exclusively on driving cycles, whereas the representability of parking and charging behavior, which is essential for electromobility, remains neglected. In this paper, we perform vehicle usage profiling by (i) allocating time spans for driving and parking sections, (ii) optimally selecting driving sections from a trip library established from testing fleets, (iii) rearranging driving sections with parking sections into e.g. week profiles, and (iv) integrating the inferred charging behavior of the profiles together with simulation. Using a model of an exemplary plug-in hybrid electric vehicle and given raw data from our testing fleets, we demonstrate that our method is capable of estimating the influence of driving, parking and charging behavior on vehicle loads.

Article 4

Integration.

Digital Twinning from Vehicle Usage Statistics for Customer-Centric Automotive Systems Engineering

Kunxiong Ling*†

*Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany

†BMW Group, 80809 Munich, Germany

Working Paper

Abstract Towards customer-centric automotive systems engineering, physical models and vehicle usage behavior need to be fed into decision support systems (DSSs). Such DSSs tend to apply digital twin concepts, in which simulations are parameterized with fine-grained time-series data acquired from customer fleets. However, logging vast amounts of data from customer fleets is costly and raises privacy concerns. Alternatively, aggregating the time-series data into vehicle usage statistics could mitigate the concerns. So far, the feasibility of digital twinning from vehicle usage statistics and corresponding DSSs for systems engineering remains unknown. Hence, we propose a DSS framework that builds digital twins based on aggregate usage statistics from customer fleets and logging data from testing fleets by profiling and simulation. Using a real-world fleet of 57110 vehicles and four evaluation metrics, the proposed DSS framework covers 99% of the critical load of customers and reaches an average fleet twinning accuracy of 91.09%.

Ehrenwörtliche Erklärungen

Hiermit erkläre ich gem. RaPromO § 7 Abs. 3 Satz 3 der KU Eichstätt-Ingolstadt, dass ich die schriftliche Dissertationsleitung selbstständig und ohne unerlaubte fremde Hilfe angefertigt, keine anderen als die in der Arbeit angegebenen Schriften und Hilfsmittel benutzt und die den benutzten werken wörtlich oder inhaltlich entnommenen Stellen kenntlich gemacht habe.

Ferner erkläre ich ebenfalls gem. RaPromO § 7 Abs. 3 Satz 3 der KU Eichstätt-Ingolstadt, dass ich insbesondere nicht die Hilfe von Vermittlungs- oder Beratungsdiensten (Promotionsberater oder Promotionsberaterinnen oder andere Personen) in Anspruch genommen habe.

Hiermit erkläre ich, dass ich gemäß RaPromO § 7 Abs. 3 Satz 4 der KU Eichstätt-Ingolstadt
a) nicht bereits frühere Promotionsversuche unternommen oder Promotionen abgeschlossen habe oder
b) nicht die Dissertation in gleicher oder anderer Form in einem anderen Versuch oder in einem anderen Prüfungsverfahren vorgelegt habe.

München 27.01.23
Ort, Datum

Kunxiong Ling