IEEE Open Journal of
Intelligent Transportation
Systems

Received 16 July 2023; revised 20 November 2023; accepted 12 February 2024. Date of publication 15 February 2024; date of current version 29 February 2024.

Digital Object Identifier 10.1109/OJITS.2024.3366279

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

KUNXIONG LING [1,2], JAN THIELE [1], AND THOMAS SETZER [2]

[1] Research and Innovation Center, BMW Group, 80788 Munich, Germany

[2] Ingolstadt School of Management, Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany

CORRESPONDING AUTHOR: K. LING (e-mail: kunxiong.ling@bmw.de)

This work was supported by BMW Group.

ABSTRACT We propose a method to estimate information loss when conducting histogram binning and principal component analysis (PCA) sequentially, as usually done in practice for fleet analytics. Coarser-grained histogram binning results in less data volume, fewer dimensions, but more information loss. Considering fewer principal components (PCs) 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. We use Monte Carlo simulations to regress information loss on the number of bins and PCs, given few parameters of a dataset related to its scale and correlation structure. A sensitivity study shows that information loss can be approximated well given sufficiently large datasets. Using the number of bins, PCs, 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, continuously distributed, and can be represented by Beta-distributions, we recommend not to coarsen the histogram binning before PCA.

INDEX TERMS Fleet analytics, histogram, information loss, Monte Carlo, principal component analysis.

NOMENCLATURE

Functions and scalars

α, β	Parameters of a Beta distribution
λ	Column-wise blending factor
μ	Row-wise blending factor
ρ	Average correlation coefficient
σ	Singular value
$\varkappa$	Number of principal components
A	Clustering accuracy
$F(x)$	Empirical cumulative density function (ECDF)
i	Index of vehicle
j	Index of sensor
K	Performance indicator of binning or PCA
k	Number of bins
l	Information loss of a random variable
L	Information loss of the whole dataset
m	Number of sensors of each vehicle
N	Number of samples for computing information losses
n	Number of customer vehicles
p	Number of samples per sensor per vehicle
$p(x)$	Empirical probability density function
r	Number of vertical vectors
t	Index of bin
X	Random variable of the raw acquired data

Matrices and vectors

$\boldsymbol{u}$	Vertical coordinate vector
$\boldsymbol{v}$	Vertical weight vector
Σ	Singular value matrix
M	Matrix of binned and flattened dataset
R	Pearson correlation matrix
U	Coordinate matrix
V	Unitary weight matrix

The review of this article was arranged by Associate Editor Hyunbum Kim.

Sets

$\boldsymbol{\pi}$	Bin value set of a histogram
$\boldsymbol{A}$, $\boldsymbol{B}$	Generated set of Beta-parameters
$\boldsymbol{b}$	Sample set of the raw acquired data
$\boldsymbol{D}$	Dataset after histogram binning
$\boldsymbol{e}$	Interval set of a histogram
$\boldsymbol{h}$	Histogram
$\boldsymbol{X}$	Target dataset for n vehicles
Ω	Set of real-world customer fleets

Subscripts and superscripts

$(\cdot)^{(ij)}$	Variable for vehicle i and sensor j
$(\cdot)^{(x)}$	The x^{th} element of a set
$(\cdot)_{\text{c}}$	Column-wise variable
$(\cdot)_{\text{n}}$	Normalized variable
$(\cdot)_{\text{r}}$	Row-wise variable
$(\cdot)_{ij}$	Set element for vehicle i and sensor j
$\tilde{(\cdot)}$	Reconstructed variable using the first $\varkappa$ principal components.

I. INTRODUCTION

TO UNDERSTAND the usage of vehicles across the lifetime, companies acquire and analyze measurements from various sensors inside the vehicle, providing data-driven decision support for customer-centric automotive development. For instance, Wilberg et al. [1] highlighted the potential of sensor data analysis in supporting requirement engineering and reliability evaluation. Albers et al. [2] showed the prospective of automotive development processes driven by sensor data. More specifically, Reicherts et al. [3] used naturalistic driving studies to identify vehicle dynamics. Tanshi and Söffker [4] proposed a method for determining the takeover time budget based on the analysis of driver behavior.

However, companies aiming to exploit the customer data face policies related to data protection and privacy preservation. To ensure privacy by design, data thriftiness and obligating procedures are usually required. For example, Viktoriya et al. [5] identified the automobile industry's ethical issue. Enev et al. [6] showed that driver fingerprinting is possible with sensor data, which would strongly violate customer privacy. Furthermore, multivariate sensor data with fine-grained temporal resolution prohibit conducting analytics on the raw data. These data need to be reduced, often by orders of magnitude, in respective preprocessing procedures.

A common approach to manage both, preparing the data for analytics and privacy preservation, is to aggregate the operational data from customers and solely keep the aggregated data as historical sensor data, e.g., by binning the data [7], [8]. Binning temporal data is typically an initial step in preprocessing sensor data, often already conducted directly on vehicle control units, i.e., on the customer side.

It is, however, still challenging to use the heterogeneous, high-dimensional data, although binned, in exploratory or supervised analytical models – a problem coined the "curse of dimensionality" [9]. As each histogram bin accounts for a dimension, in total, the average number of bins times the number of sensors considered easily results in a dimensionality of thousands and prohibiting the application of most visual procedures and analytical models. Hence, the aggregated dataset needs to be shrunk to a manageable number of dimensions. One of the most widespread means of further reducing the dimension is to perform principal component analysis (PCA), an unsupervised low-rank matrix approximation technique, as the second step. Together with binning, after all, we would like to reduce the number of principal components (dimensionality) without losing much information.

With significantly large amount of samples, coarser-grained histogram binning results in fewer dimensions, but loses more information. However, with the same outcome dimensionality, should we perform coarser binning or finer binning at first to optimize the performance of upstream analytical models? Although there are already evaluation metrics for PCA, do they really represent the influence of binning on the pre-processed dataset? If not, which evaluation metric can consider the whole process and provide representative decision support for configuring the binning?

So far, it remains a research gap to understand the mechanism of the decision support problem that exist in the two-step process. On the one hand, there is a lack of large amount of raw data before binning from customer fleets, as they are previously binned on-board (inside of control units) [10]. On the other hand, binning and PCA are combined with the spread of distribution patterns of various customer fleets. This makes it especially complex to investigate the research questions using theoretical analysis and mathematical proofs.

In this paper, we tackle these challenges from the perspective of information losses, measured by the Kullback-Leibler divergence between original, binned, and PCA approximated data. Considering the difficulty of theoretical analysis, we estimate the information losses by simulating the raw data based on Monte Carlo approaches. In the following, we highlight the contributions of our work.

- For the purpose of loss estimation of binning and PCA, we model raw customer fleet data using three scale parameters and the correlation structures between histograms within a row and between rows. Based on simulations with sensor data drawn from Beta-distributions with varying degrees of correlation between and within rows, sensitivity study shows the influence of each parameter on information losses.
- Based on the sensitivity study, we derive an empirical model that guides how to set the number of bins in combination with the order of dimensions to be considered. The model can determine appropriate values for the number of bins, number of principal components, and total loss, given two of the values are set.

- Using a case study of real-world fleet data from 1454 vehicles, we found the benchmark evaluation metrics from binning (Kolmogorov-Smirnov statistic) and PCA (variance unexplained) cannot be considered as decision support metric. Taking k-means clustering as an example, neither of those two metrics is capable of representing the accuracy of fleet analytics right after binning and PCA. Instead, we demonstrate how the estimated total information loss outperform those metrics.
- For fleet analytics with sufficient number of samples, it is recommended not to coarsen the finely-binned histograms before performing PCA. Note that all the findings in this paper are valid only when the raw fleet sensor data are continuously distributed and could be represented by Beta-distributions.

The remainder of this article is organized as follows. In Section II, we review the related work on data binning followed by PCA and loss estimation and highlight our position in the research field. In Section III, we introduce the notation followed by an overview of our methodology, the model assumptions, the algorithms proposed, and loss functions considered. In Section IV, we then present the results of various investigations, including sensitivity studies with the proposed model, empirical modeling of the information loss mechanism throughout the procedures, and a case study on k-means clustering. We will then summarize the key findings obtained, conclude, and outline future research direction in the realm of sensor data preprocessing in Section V.

II. RELATED WORK

Histogram binning and subsequent principal component analysis (PCA) are popular data aggregation techniques for preprocessing customer fleet data. To locate the scope of our work, we review the related work in three parts: histogram binning for fleet analytics, PCA for histogram data, and loss-aware perspective for determining the cardinality of principal component after PCA. Afterward, we highlight the contribution of our work to the related fields.

Histogram binning allows the removal of the temporal dimension of sensor information, reduces the cardinality of sensor values, and can mitigate the effects of noise (observation inaccuracy). Numerous research activities have been conducted with binned customer data for customer-centric decision support, especially in the automobile industry. Schoch et al. [11] optimized charging strategy for longer battery cell lifetimes using binned customer data. Huang and Meng [12] put out a decision support framework for pricing automobile insurance based on binned telematics driving data. Ling et al. [13] used binned data for customer vehicle usage profiling.

Currently, the investigation of binning strategy is mainly conducted using available raw samples, e.g., Boulle [14] considered the influence of samples on the bins and the frequency-based binning. However, if the raw data samples are not available, information loss cannot be computed. Hence, in this paper, we enable the binning loss estimation by simulating the samples and reconstructing the correlation structure to estimate.

PCA based on histogram data has been systematically investigated, whereas the histograms are treated as symbolic data [15], driven by the methodological approach for symbolic data analytics [16], [17]. Billard and Le-Rademacher [18] put out the PCA method for interval data. Makosso-Kallyth [19] extended the scope from interval data to symbolic histogram variables.

Yet, this research field has not been widely applied in industrial contexts such as customer fleet analytics, where interpretability and robustness of methods are essential. Hence, we focus on a relative more conventional, but more popular context, i.e., concatenate the histogram bins and then perform matrix factorization using PCA. Haselgruber et al. [20] used PCA to aggregate engine load data for evaluating reliability testing. Bartłomiejczyk [21] analyzed the driving behavior of bus drivers by aggregating the measurement signals followed by PCA. Schoch et al. [22] implemented PCA for binned sensor data to reduce the features for electric vehicle service analytics. Ling et al. [23] applied PCA to improve the representativeness of customer sampling based on aggregated usage data and thus identified fringe customers.

From the perspective of variances, the approximation error with a PCA-reconstructed matrix, for instance, can be measured by the Frobenius norm of the approximation error matrix. It can be computed as the sum of all squared singular values minus the sum of the first few selected singular values squared, as a singular value squared captures the variance explained by the associated dimension. According to predefined fraction of variance explained, it is common practice to determine the number of principal components [24].

From the perspective of information theory, there is a principle for guiding model selection, namely minimum description length (MDL) based on Kullback-Leibler divergences [25]. Tavory integrated the MDL principle into PCA [26]. Bruni et al. [27] reviewed the methodology using three test cases and pointed out that MDL outperforms most of the model selection methods such as variance explained. On the contrary, specifying the model parameters has unclear influences on MDL performance. The influences behave sometimes explicit, but sometimes also implicit.

However, spanning binning and PCA as a sequence, the binning resolution and number of PCs influence on each other. Adding to the difficulty of changing the data acquisition strategy of customer fleets with various control units, it becomes increasingly time-consuming for big data analytics of customer fleets. Still, the parameter choosing for histogram binning followed by PCA is merely investigated, which remains interesting even for fleet analytics.

So far, Vaiciukynas et al. [28] investigated histogram binning followed by dimensionality reduction using fleet data with more than 20,000 vehicles. Although they thoroughly

investigated various dimensionality reduction techniques on the performance of feature representation, the influence of binning resolution remains unknown, especially in the context of customer fleet analytics.

Therefore, we address the problem from the information theory perspective, but with simpler metrics, i.e., Kullback-Leibler divergences. We focus on the estimation and comparison of the information losses between the raw data, binned data and PCA approximated data.

III. METHODOLOGY

This section introduces our simulation-based research design to estimate information loss when preprocessing with changing characteristics using Monte Carlo sampling. First, we describe binning and principal component analysis (PCA) with notations. Then, we present our simulation procedure and algorithms. Afterward, we define loss functions based on Kullback-Leibler divergences for these binned data to evaluate the information losses.

A. PRELIMINARIES

1) HISTOGRAM BINNING

Consider a target dataset $\boldsymbol{X}$ for n vehicles. Each vehicle consists of m sensors. For vehicle $i = 1, \ldots, n$ and sensor $j = 1, \ldots, m$, we represent the acquired data using a random variable X_{ij}, so that the samples of sensor values can be allocated to discrete intervals (binning).

Assuming (i) the sensor data per vehicle to be binned with the same number of intervals k, and (ii) the number of samples p remains to be consistent for each sensor and each vehicle, we represent random variable X_{ij} by p samples, and the data by set $\boldsymbol{b_{ij}} = \{b_{ij}^{(1)}, \ldots, b_{ij}^{(p)}\}$.

Considering the feasibility of saving the whole sample set across the whole life of a vehicle, we aggregate $\boldsymbol{b_{ij}}$ and represent them as histogram $\boldsymbol{h_{ij}}$. Each histogram consists of k rescaled intervals $\boldsymbol{e_{ij}} = \{[0, \frac{1}{k}), \ldots, [\frac{k-1}{k}, 1]\}$ and k bins $\boldsymbol{\pi_{ij}} = (\pi_{ij}^{(1)}, \ldots, \pi_{ij}^{(k)})$, where $\sum_{t=1}^{k} \pi_{ij}^{(t)} = p$ and $\pi_{ij}^{(t)} \geq 0$. In this way, $\boldsymbol{\pi_{ij}}$ need to be stored on board, e.g., in the control units. In total, m histograms from n vehicles can be acquired and collected together in dataset

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\pi_{11}} & \cdots & \boldsymbol{\pi_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\pi_{n1}} & \cdots & \boldsymbol{\pi_{nm}} \end{bmatrix}, \tag{1}$$

with the histogram values alone. We refer to this procedure as k-fold binning.

After k-fold binning, the sequential information is completely lost. However, as the distributions are discretized by intervals, the distribution information is lost as well. Histogram $\boldsymbol{h_{ij}}$ can be described by an empirical cumulative density function (ECDF) $F_{ij}(x)$, where the range of x is identical to the range of $\boldsymbol{e_{ij}}$. The relative difference of two ECDFs can be quantified by the Kolmogorov-Smirnov (K-S) statistic $D_{\text{K-S}}$, i.e., supremum absolute ECDF difference [29]. The supremum represents the maximum across all x values. For $\boldsymbol{D}$, the overall K-S statistic is the mean of all mn histograms, i.e.,

$$K_{\text{K-S}} = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \sup_{x} |F_{ij}(x) - F_{ij}^{\text{ref}}(x)|, \tag{2}$$

where the reference ECDF $F_{ij}^{\text{ref}}(x)$ could be the original ECDF of raw samples $\boldsymbol{b_{ij}}$ or another finely binned histogram.

2) PRINCIPAL COMPONENT ANALYSIS OF BINNED DATA

Once the sensor data is aggregated through k-fold binning, the dataset can be represented as a matrix. One row represents a vehicle. One column represents the average number of bins per sensor times the number of sensors as the number of columns (also coined dimension of the data). However, the high dimensions might prohibit the direct usage of the matrix for data mining purposes. By applying principal component analysis (PCA), we can, however, exploit correlations within the matrix to represent the principal structure of the matrix more concisely.

Consider a k-fold binned dataset M consisting of $k \cdot m$ column vectors by concatenating each vector element from $\boldsymbol{D}$ in the row direction, i.e.,

$$M = \left[\begin{array}{ccc:c:ccc} \pi_{11}^{(1)} & \cdots & \pi_{11}^{(k)} & & \pi_{1m}^{(1)} & \cdots & \pi_{1m}^{(k)} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \pi_{n1}^{(1)} & \cdots & \pi_{n1}^{(k)} & & \pi_{nm}^{(1)} & \cdots & \pi_{nm}^{(k)} \end{array}\right]. \tag{3}$$

To simplify the denotation, we define r as the number of columns of M, i.e., $r = k \cdot m$. We now represent $M \in \mathbb{R}^{n \times r}$ in a $\varkappa$-dimensional latent space using PCA. To balance the weights between the columns, we normalize M to $M_{\text{n}} \in \mathbb{R}^{n \times r}$ by centering and scaling the histogram values into probability densities, where element $\pi_{ij}^{(t)}$, $t = 1, \ldots, k$, is transformed to

$$\pi_{ij,\text{n}}^{(t)} = \frac{k}{p} \left(\pi_{ij}^{(t)} - \frac{1}{n} \sum_{i=1}^{n} \pi_{ij}^{(t)} \right), \quad t = 1, \ldots, k, \tag{4}$$

as centering elements by subtracting the mean of the elements' column entries moves the novel basis vectors towards maximum variance directions.

After normalizing M to M_{n}, we approximate the matrix with lower rank to determine latent basis vectors for M_{n}'s column and row space, as done, with truncated singular value decomposition [30]. These latent basis vectors are located in directions where the data is most widely spread to capture the maximum amount of information in the matrix (the variance of its elements) with as few latent dimensions as possible. The intuition is to then only consider the primary latent dimensions derived and discard higher-order dimensions considered as noise. Hence, M_{n} is decomposed into three parts, i.e.,

$$M_{\text{n}} = U \Sigma V^{\top}. \tag{5}$$

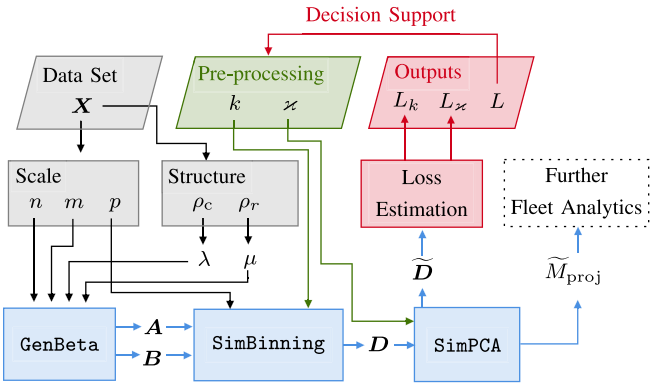


FIGURE 1. Methodology overview.

Here, $U \in \mathbb{R}^{n\times r}$ represents the coordinate matrix, consisting of r vertical coordinate vectors $U = [\boldsymbol{u_1}, \ldots, \boldsymbol{u_r}]$. Singular value matrix $\Sigma \in \mathbb{R}^{r\times r}$ is positive, semi-definite and diagonalized with $\sigma_1 \geq \cdots \geq \sigma_r$. $V \in \mathbb{R}^{r\times r}$ is the unitary weight matrix with r vertical weight vectors $V = [\boldsymbol{v_1}, \ldots, \boldsymbol{v_r}]$. Using the first $\varkappa$ principal components, we reconstruct M_n to $\widetilde{M}_\mathrm{n} \in \mathbb{R}^{n\times r}$ with

$$\widetilde{M}_\mathrm{n} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top, \tag{6}$$

where $\widetilde{U} = [\boldsymbol{u_1}, \ldots, \boldsymbol{u_\varkappa}]$, $\widetilde{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_\varkappa)$ and $\widetilde{V} = [\boldsymbol{v_1}, \ldots, \boldsymbol{v_\varkappa}]$. Moreover, M_n can be projected in the $\varkappa$-dimensional latent space $\widetilde{M}_\mathrm{proj} \in \mathbb{R}^{n\times\varkappa}$ by

$$\widetilde{M}_\mathrm{proj} = M_\mathrm{n}\widetilde{V}, \tag{7}$$

to work with fewer dimensions in further analytical tasks. We refer to this procedure as $\varkappa$-rank PCA.

As mentioned in the related work, the most popular selection metric of $\varkappa$ is the variance explained [24], which is the proportion of variance remained from all variance, i.e., the ratio from the sum of first $\varkappa$ singular value squared to the sum of all r singular value squared. For the consistency of comparison to the losses after binning, we regard the variance unexplained K_var, the opposite indicator, as the standard metric to quantify PCA performance, i.e.,

$$K_\mathrm{var} = 1 - \frac{\sum_{t=1}^{\varkappa}\sigma_t^2}{\sum_{t=1}^{r}\sigma_t^2}. \tag{8}$$

3) METHODOLOGY OVERVIEW

Based on the formulation of histogram binning and PCA for customer fleet analytics, we formulate our objective and illustrate the overview of the methodology shown in Fig. 1.

We provide decision support in determining preprocessing parameters k and $\varkappa$ from the perspective of information losses, in particular the total information loss after preprocessing L. Without having the raw samples of dataset $\boldsymbol{X}$, our hypothesis is that there exists general behavior for customer fleet data. First, we characterize $\boldsymbol{X}$ by three scale parameters and two structure parameters. Then, we simulate the data distribution via `GenBeta` and get the Beta-parameter set $\boldsymbol{A}$ and $\boldsymbol{B}$. After generating binned histogram data $\boldsymbol{D}$ in `SimBinning`, we perform PCA in `SimPCA`. With the first $\varkappa$ principal components, we project $\boldsymbol{D}$ to $\widetilde{M}_\mathrm{proj}$ for further analytics and approximate the generated binned dataset $\boldsymbol{D}$ by $\widetilde{\boldsymbol{D}}$. The core lies in estimating the information losses between generated Beta distributions characterized by $\boldsymbol{A}$ and $\boldsymbol{B}$, binned dataset $\boldsymbol{D}$ and PCA-approximated dataset $\widetilde{\boldsymbol{D}}$.

B. DATASET CHARACTERIZATION

We represent the time-series sensor data measurements by Beta distributed random variables, i.e., $X_{ij} \sim \mathrm{Beta}(\alpha_{ij}, \beta_{ij})$, as (i) according to the study from Greene [31] and Lin et al. [32], sensor information in vehicles can be well synthesized with unimodal Gamma distributions, which can also be represented using Beta-distributions; (ii) compared to normal distributions, the distribution can describe different shapes; (iii) their probability density can be zero as apparent in many practical settings; (iv) realizations are between 0 and 1, facilitating the data preprocessing in our simulation study and result interpretation.

We restrict Beta-parameters $\alpha_{ij} \in \boldsymbol{A}$, $\beta_{ij} \in \boldsymbol{B}$ to the interval $[1, 10]$. That is because distributions with $\alpha_{ij} < 1$ or $\beta_{ij} < 1$ are U-shaped, a distribution hardly occurring for sensor data as the intervals are larger than real conditions.

We model k as a global parameter such that the binning strategy, in terms of the number of bins, is identical for all vehicles and sensors.

The samples from a single vehicle i and a single sensor j describe a random variable X_{ij} that we aggregate into a histogram $\boldsymbol{h_{ij}}$ to receive matrix M.

We model sensor correlation by letting sensor value follow beta distributions with a varying similarity of their probability density functions (PDFs), and then calculate the (resulting) correlation matrices based on the mean value of each random variable X_{ij} in $\overline{b}_{ij} = \frac{1}{p}\sum \boldsymbol{b_{ij}}$. The mean values of the whole dataset $\boldsymbol{D}$ is represented as matrix $\overline{D} \in \mathbb{R}^{n\times m}$. $\overline{D}$ can be decomposed in m column vectors $\{\boldsymbol{d_{c1}}, \ldots, \boldsymbol{d_{cm}}\}$ or n row vectors $\{\boldsymbol{d_{r1}}^\top, \ldots, \boldsymbol{d_{rn}}^\top\}$. The column-wise Pearson correlation matrix of $\overline{D}$ is represented with $R \in \mathbb{R}^{m\times m}$, in which the element ij of the correlation matrix is

$$R^{(ij)} = \frac{\mathrm{cov}(\boldsymbol{d_{ci}}, \boldsymbol{d_{cj}})}{\sqrt{\mathrm{var}(\boldsymbol{d_{ci}})}\sqrt{\mathrm{var}(\boldsymbol{d_{cj}})}}. \tag{9}$$

Here, cov stands for covariance of two vectors and var represents the variance of a vector. Hence, we get coefficient $\rho_\mathrm{c} \in [0, 1]$ as the total average of all the correlation coefficients without the diagonal elements, i.e.,

$$\rho_\mathrm{c} = \frac{\mathbb{1}^\top(R - \mathrm{diag}(R))\mathbb{1}}{m(m-1)}, \tag{10}$$

where $\mathbb{1} \in \mathbb{R}^{m\times 1}$ represents the unity vector.

Correspondingly, the row-wise Pearson correlation matrix of $\overline{D}$ is represented with $R_\mathrm{r} \in \mathbb{R}^{n\times n}$, in which the element ij of the correlation matrix is

$$R_\mathrm{r}^{(ij)} = \frac{\mathrm{cov}(\boldsymbol{d_{ri}}, \boldsymbol{d_{rj}})}{\sqrt{\mathrm{var}(\boldsymbol{d_{ri}})}\sqrt{\mathrm{var}(\boldsymbol{d_{rj}})}}. \tag{11}$$

Algorithm 1 Generating Beta-Parameters

function GenBeta(n, m, λ, μ)
▷ Get beta-parameters for each X_{ij}.
  $\boldsymbol{A} \leftarrow n \times m$ random numbers in $[1, 10]$;
  $\boldsymbol{B} \leftarrow n \times m$ random numbers in $[1, 10]$;
▷ Construct the correlation of the beta-parameters.
  for $i = 2, \ldots, n$ **do**
    $\boldsymbol{A}_{\text{row}\,i} \leftarrow \mu \boldsymbol{A}_{\text{row}\,1} + (1-\mu)\boldsymbol{A}_{\text{row}\,i}$;
    $\boldsymbol{B}_{\text{row}\,i} \leftarrow \mu \boldsymbol{B}_{\text{row}\,1} + (1-\mu)\boldsymbol{B}_{\text{row}\,i}$;
  end for
  for $j = 2, \ldots, m$ **do**
    $\boldsymbol{A}_{\text{column}\,j} \leftarrow \lambda \boldsymbol{A}_{\text{column}\,1} + (1-\lambda)\boldsymbol{A}_{\text{column}\,j}$;
    $\boldsymbol{B}_{\text{column}\,j} \leftarrow \lambda \boldsymbol{B}_{\text{column}\,1} + (1-\lambda)\boldsymbol{B}_{\text{column}\,j}$;
  end for
  Return $\boldsymbol{A}, \boldsymbol{B}$
end function

Row-wise correlation coefficient $\rho_{\text{r}} \in [0, 1]$ can be calculated with

$$\rho_{\text{r}} = \frac{\mathbb{1}^{\top}(R_{\text{r}} - \text{diag}(R_{\text{r}}))\mathbb{1}}{n(n-1)}, \tag{12}$$

where $\mathbb{1} \in \mathbb{R}^{n\times 1}$, $R_{\text{r}} \in \mathbb{R}^{n\times n}$.

In a nutshell, the previous formulation shows that dataset $\boldsymbol{D}$ can be characterized with scale parameters n, m, p as well as structure parameters $\rho_{\text{c}}, \rho_{\text{r}}$.

C. SIMULATION PROCEDURE

We simulate dataset $\boldsymbol{D}$ with combinations of parameters described above using Monte Carlo methods. For each simulation run, the procedure consists of two parts: Beta-parameter generation to derive $\boldsymbol{D}$ as `GenBeta` in Algorithm 1, and preprocessing as `SimBinning` and `SimPCA` in Algorithm 2.

In Algorithm 1, we generate $n \times m$ random numbers uniformly distributed in the range $[1, 10]$, assigning them to matrix $\boldsymbol{A}$. Subsequently, we repeat this process, generating a new set of random numbers and assigning them to matrix $\boldsymbol{B}$. At this point, Beta-parameter matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are initialized for further processing. Due to the shape property of Beta-parameters [33], the closer the parameters are, the stronger their correlation is. Hence, we add a treatment to the correlation structure using blending about the first column or row. Regarding the first row and the first column as references, we subsequently blend the other rows and columns towards the references, controlled by the column-wise and row-wise blending factors $\lambda, \mu \in [0, 1]$. The larger the blending factors are, the nearer the rows (columns) to the reference row (column) are. However, λ, μ are not identical to the correlation coefficients. By performing a parameter study, the blending factors, as expected, exhibit strong positive associations with correlation coefficients, i.e., $\lambda \propto \rho_{\text{c}}$ and $\mu \propto \rho_{\text{r}}$. In the sensitivity study, we use λ and μ to represent ρ_{c} and ρ_{r}. When estimating information loss with an observed matrix of sensor values, we can empirically determine estimates of the corresponding λ and μ values.

Algorithm 2 Simulation of Histogram Binning and PCA

function SimBinning($\boldsymbol{A}, \boldsymbol{B}, n, m, p, k$)
▷ Generate the samples for each X_{ij} and perform k-fold binning.
  for $i = 1, \ldots, n$ **do**
    for $j = 1, \ldots, m$ **do**
      $\boldsymbol{b_{ij}} \leftarrow$ randomly generate p samples according to $X_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij})$;
      $\boldsymbol{\pi_{ij}} \leftarrow$ aggregate the samples in $\boldsymbol{b_{ij}}$ with k equally distributed intervals in $[0, 1]$;
    end for
  end for
  Construct $\boldsymbol{D}$ with $\boldsymbol{\pi_{ij}}$ according to (1);
  Return $\boldsymbol{D}$
end function

function SimPCA($\boldsymbol{D}, n, p, k, \varkappa$)
▷ Perform $\varkappa$-rank approximation using PCA.
  Flatten $\boldsymbol{D}$ to M according to (3);
  $M_{\text{n}} \leftarrow$ Normalize M into probability densities by (4);
  Perform truncated singular value decomposition for M according to (5);
  $\widetilde{M}_{\text{n}} \leftarrow$ Approximate M with $\varkappa$ rank according to (6);
  $\widetilde{\boldsymbol{D}} \leftarrow$ Reconstruct $\widetilde{M}_{\text{n}}$ into the nested dataset structure similar to $\boldsymbol{D}$;
  $\widetilde{M}_{\text{proj}} \leftarrow$ Compress M using the first $\varkappa$ principal components according to (7);
  Return $\widetilde{\boldsymbol{D}}, \widetilde{M}_{\text{proj}}$
end function

As a result, we get two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ and perform the simulation in the next step based on these simulated beta parameters.

As shown in Algorithm 2, we randomly generate the samples that obey the Beta-distributions from Algorithm 1. By binning the samples, we simulate the histogram values and approximate them with a given lower rank. Here, k and $\varkappa$ serve as the variables that control the preprocessing procedure. Typically we project the data into the latent space to reduce their dimension according to (7). In this case, the evaluation of information losses requires the reconstructed dataset. Hence, we take the reconstructed normalized matrix $\widetilde{M}_{\text{n}}$ from (6), de-normalize it into $\widetilde{M}$ by solving (4). Hence, M is approximated by $\widetilde{M}$ with $\varkappa$ principal components. As a result, we obtain the binned dataset $\boldsymbol{D}$ and the approximated dataset $\widetilde{\boldsymbol{D}}$ which has the same nested structure as $\boldsymbol{D}$, according to (1).

The simulation parameters are summarized in Table 1. Column-wise parameters are the parameters among the sensors. Row-wise parameters stand for the parameters among the vehicles.

After each simulation case, we keep the beta-parameters $\boldsymbol{A}$, $\boldsymbol{B}$, and the binned dataset $\boldsymbol{D}$ as well as the approximated dataset $\widetilde{\boldsymbol{D}}$ for the evaluation of the information losses.

D. LOSS ESTIMATION

To compare the distribution of X_{ij} before and after each preprocessing step, their probability density functions (PDFs) are used. For the Beta-distributions, we represent the PDF

TABLE 1. Preprocessing variables and simulation parameters of sensor data aggregation.

Type	Expression	Reference Value
Preprocessing variable	$k \in \mathbb{N}$	8
	$\varkappa \in \mathbb{N},\ \varkappa \leqslant mk$	3
Scale parameter	$n \in \mathbb{N}$	100
	$m \in \mathbb{N}$	10
	$p \in \mathbb{N}$	1000
Structure parameter (Correlation)	$\rho_{\mathrm{c}} \in [0,1]$	0.77
	$\rho_{\mathrm{r}} \in [0,1]$	0.42
Structure parameter (Treatment)	$\lambda \in [0,1]$	0.5
	$\mu \in [0,1]$	0.5

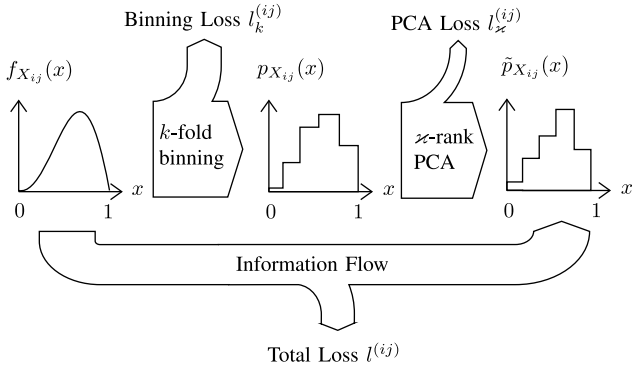


FIGURE 2. Information flow of random variable X_{ij} before and after the preprocessing steps.

of X_{ij} with $f_{X_{ij}}(x)$, $0 \leq x \leq 1$. After k-fold binning, we represent the empirical PDF of the binned X_{ij} as

$$p_{X_{ij}}(x) = \frac{k}{p}\,\pi_{ij}^{(\lceil xk \rceil)}, \tag{13}$$

where $\lceil xk \rceil$ represents the ceiling function, and $\pi_{ij}^{(\lceil xk \rceil)}$ is an element of de-normalized M_{n} towards M. Similarly, we get the empirical PDF for the reconstructed X_{ij} after its $\varkappa$-rank approximation as

$$\widetilde{p}_{X_{ij}}(x) = \frac{k}{p}\,\widetilde{\pi}_{ij}^{(\lceil xk \rceil)}, \tag{14}$$

where $\widetilde{\pi}_{ij}^{(\lceil xk \rceil)}$ represents the reconstructed histogram value as an element of $\widetilde{M}$, de-normalized from $\widetilde{M}_{\mathrm{n}}$.

As shown in Fig. 2, we can determine the loss of the binning and the PCA steps.

Hence, we have

- binning loss $l_k^{(ij)}$ from $f_{X_{ij}}(x)$ to $p_{X_{ij}}(x)$,
- PCA loss $l_\varkappa^{(ij)}$ from $p_{X_{ij}}(x)$ to $\widetilde{p}_{X_{ij}}(x)$, and
- total loss $l^{(ij)}$ from $f_{X_{ij}}(x)$ to $\widetilde{p}_{X_{ij}}(x)$.

The information loss function used between two PDFs is the Kullback-Leiber divergence (relative entropy), i.e., the expectation of the logarithmic difference [34]. In this paper, the logarithmic function base is 2, and the unit is Shannon (Sh).

Let us first describe the binning loss. The loss function is defined as

$$l_k^{(ij)} = \int_0^1 f_{X_{ij}}(x) \log_2\left(\frac{f_{X_{ij}}(x)}{p_{X_{ij}}(x)}\right) \mathrm{d}x. \tag{15}$$

Here only if $p_{X_{ij}}(x) = 0$, for all x, the divergence is defined as $f_{X_{ij}}(x) = 0$. To compute the integral, we regard it as a sum of the function values of N samples of variable x with equal distance, i.e., we apply Quasi-Monte Carlo approach [35]. Hence, we can approximate the loss by giving a finite value of N, which acts as a hyper-parameter. Over our n times m random variables, we evaluate the whole dataset using the average of their absolute values, yielding the average binning loss

$$L_k \approx \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{N}\sum_{t=1}^{N} f_{X_{ij}}(t/N)\log_2\left(\frac{f_{X_{ij}}(t/N)}{p_{X_{ij}}(t/N)}\right)\right|. \tag{16}$$

Correspondingly, we evaluate the PCA loss and total loss by computing

$$L_\varkappa \approx \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{N}\sum_{t=1}^{N} p_{X_{ij}}(t/N)\log_2\left(\frac{p_{X_{ij}}(t/N)}{\widetilde{p}_{X_{ij}}(t/N)}\right)\right|, \tag{17}$$

and

$$L \approx \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{N}\sum_{t=1}^{N} f_{X_{ij}}(t/N)\log_2\left(\frac{f_{X_{ij}}(t/N)}{\widetilde{p}_{X_{ij}}(t/N)}\right)\right|. \tag{18}$$

IV. RESULTS AND DISCUSSION

Before putting our approach into applications, the sensitivities of information losses (binning, PCA, and total losses) on the parameters in Table 1 are analyzed. Based on the sensitivity study, we will derive an empirical model from estimating these losses without simulation. Furthermore, we will demonstrate how the loss model benefits the decision-making of choosing parameters of aggregation.

A. SENSITIVITY STUDY

Table 1 shows the parameters in the simulation procedure. However, with the data generation and simulation procedure described, both correlation structure parameters ρ_{c}, ρ_{r} can only be calculated when the dataset already exists. Instead, to generate the dataset, we use the treatment parameter λ, μ to control the correlation structure.

To start sensitivity analysis, we define the reference case configurations, shown in the outer-right column in Table 1. Then, we change parameters individually and derive its impact on information loss.

As shown in the preprocessing reference configuration, we choose eight-fold binning and three-rank dimensionality approximation. It is relatively easy to identify the distribution shape with lower noise with eight bin histograms, and a rank of three allows us to visualize the data intuitively. The scale and structure parameters represent a typical dataset, exhibiting a weak correlation structure.

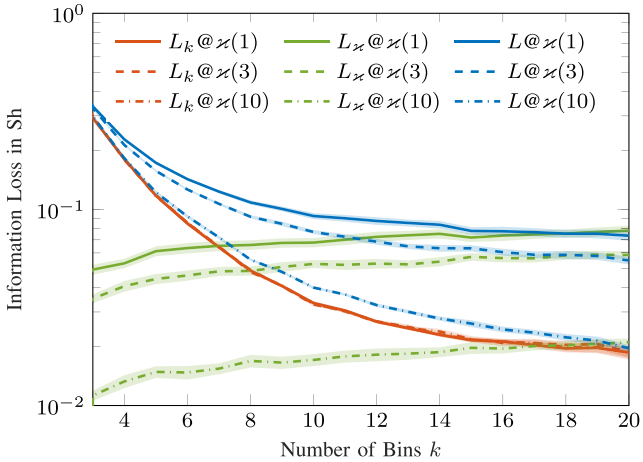


FIGURE 3. Information losses with different resolutions of the preprocessing. The areas around the curves are 95% confidence bands. L_k, $L_\varkappa$ and L represent binning, PCA and total losses. $\varkappa$ is the order of the approximated rank, set 1, 3, or 10.

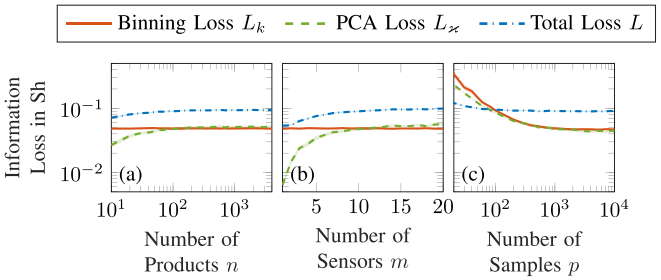


FIGURE 4. Information losses with different scales of the dataset. The areas around the curves are 95% confidence bands.

The solver settings play a role in the sampling resolution. The number N affects the accuracy to compute the loss functions in (16)–(18), due to the quasi Monte Carlo approximation of the integrals. Based on results with extensive preliminary studies, we found a N of 1000 to be sufficient to approximate the Kullback-Leibler divergence with an accuracy of at least 99%. Furthermore, with a basis of stochastic processing, we repeat each simulation case 50 times, i.e., $n_{\text{sim}} = 50$.

In the following, we present the results from the simulations with respect to the type of parameters: resolution of the preprocessing, scale of the dataset, and the correlation structure.

1) PREPROCESSING PARAMETERS

Based on the reference case, we modify $\varkappa$ to values between one and ten. The number of bins, k, is varied from three to 20, representing an increase in granularity. As k rises, more detailed information about the data distribution is captured. Fig. 3 shows the resulting losses (binning, PCA, and total losses) over various levels.

With increasing k, a reduction of binning loss and an increase of PCA loss can be observed. The aggregated histograms contain more information, resulting in a closer difference to the original distributions. However, with an identical objective (e.g., $\varkappa = 3$), a higher order of dimension brings more information and more linearly uncorrelated dimensions. Furthermore, no influence of $\varkappa$ on the binning loss is shown, as the binning is the step before PCA. Due to the growing variance explained with a higher order of $\varkappa$, the PCA loss reduces.

Another interesting aspect is that $L_\varkappa$ will surpass L_k with a higher k. It indicates that with a higher k, the information loss in the PCA is higher than that induced from the aggregation process. The turning point that $L_\varkappa$ goes over the L_k is positively correlated with $\varkappa$, implying that if we accept a higher order of dimension for the preprocessing, then it makes sense to aggregate the data with higher order.

Additionally, the trend of the total loss L is dominated by binning loss with up to 20-fold aggregation. It seems that we can minimize the whole information loss by merely increasing k. However, L yields to $L_\varkappa$, although L reduces with an increasing k. At high values of k, the dimension decrease brings more information loss to the dataset. In this case, the PCA dominates the divergence induced from the whole preprocessing.

2) SCALE PARAMETERS

In our study, the scale parameters n, m, and p are systematically varied within the range of 1 to 10000. This variation allows for an exploration of the algorithm's sensitivity to different scales of input features. Their dependencies to the information losses are presented in Fig. 4.

As shown in Fig. 4(a), no correlation between n and L_k is found. Similarly, $L_\varkappa$ and L become less dependent to n when n exceeds 100 approximately. After reaching this threshold, $L_\varkappa$ yields L_k. With small n, the smaller losses are due to the correlation structure. With identical correlation coefficients or treatment parameters, the linear dependency of the whole matrix after aggregation becomes stronger if the matrix is small. According to Fig. 4(b), a similar phenomenon is observed between m and the losses, but the yield threshold is found at roundly over ten. This threshold is much less than that for n, as the dimension order is k times higher than m.

The trends become different when it comes to p. Fig. 4(c) shows a reduction of all the information losses with a larger number of samples. It saturates with more than 1000 samples. With lower p, the sampling affects the noise and the histogram binning accuracy, which further influences L_k. Another effect is that we have 80 columns of the matrix after aggregation (according to the reference case). With a p fewer than the number of columns, aggregated histogram values are noisy. At the same time, large amounts of missing values exist in the dataset. Based on this matrix, another level of noise is included after PCA. The PCA performed here indirectly approximated the missing values, resulting in a lower total loss. With higher p, this missing-value effect disappears, and the losses go stable. In summary, before

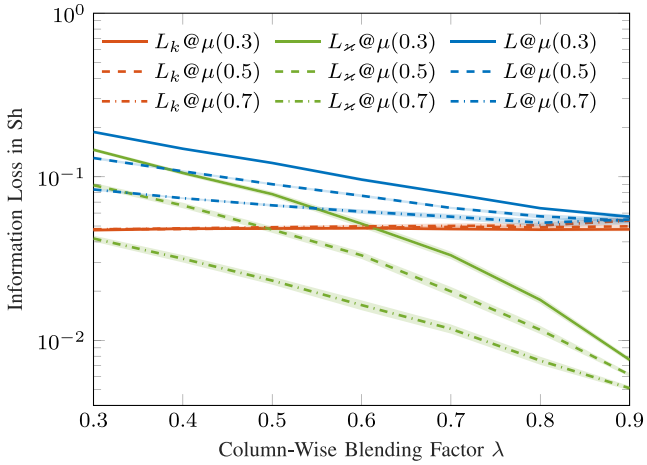


FIGURE 5. Information losses with different correlation structures. The areas around the curves are 95% confidence bands. L_k, $L_\varkappa$ and L represent binning, PCA and total losses. μ is the row-wise blending factor, set 0.3, 0.5, or 0.7.

decision-making for the data acquisition, it is necessary to keep scale parameters above the thresholds and disable the noise resulted from missing values.

3) STRUCTURE PARAMETERS

From the reference case described in Table 1, we modify the column-wise (between sensors) structure treatment λ from 0.3 to 0.9 at three levels of row-wise (between vehicles) structure treatment μ, namely 0.3, 0.5, and 0.7. Their information losses are presented in Fig. 5.

It is observed that a stronger correlation between the sensors reduces the PCA losses, resulting in a decrease in total losses. Since the aggregation is performed for each sensor and each vehicle individually, the matrix structure does not affect the aggregated histogram values. Hence, no clear dependence is observed in this study. These findings are also valid for the correlation between the vehicles.

B. LOSS MODEL DERIVATION

According to the findings in Section IV-A, when $p \gg mk$, m and n are sufficiently large, we can outline the dependence between the parameters and the information losses empirically in (19).

$$L = L_k + 0.777\, L_\varkappa. \tag{19}$$

The binning loss term is shown in (20).

$$L_k = 2.364 \cdot \exp\left(-\frac{k}{1.450}\right) + 0.028. \tag{20}$$

The PCA loss term is a function of preprocessing parameters and structure parameters. According to Fig. 5, the effect from structural parameters is considered as a multiplier in the loss function, based on a function of the preprocessing parameters. Hence,

$$L_\varkappa = C \cdot \left[-\exp\left(-\frac{k}{3.235}\right) + 1.494 \cdot \exp\left(-\frac{\varkappa}{7.647}\right)\right]. \tag{21}$$

The left term C represents the structural multiplier, which combines the row-wise and column-wise correlation effects as

$$C = 1.290 \cdot \exp\left(-\frac{\lambda}{0.323} - \frac{\mu}{0.271}\right). \tag{22}$$

Based on the mean values from the sensitivity study cases, we perform curve-fitting on the equations by minimizing R^2. The empirical functions from (19)–(22) predict the information losses with an accuracy over 97.8%. This further supports the findings about the parameter dependencies.

C. CASE STUDY

In this section, we demonstrate how to use the derived information loss model and evaluate if the estimated losses imply the accuracy of upstream analytics. First, we present the data basis of the real-world fleet data and the evaluation case configuration. Subsequently, we project the data onto the first two principal components and visualize it with five binning levels. This step is taken to investigate the influence of binning levels on data interpretation, aiming to understand how varying levels of granularity impact the visual representation. Afterward, we evaluate the two-step preprocessing using our loss model. Furthermore, we compare our loss-aware method to two benchmark reference metrics, i.e., the variance unexplained for PCA and the K-S statistic for histogram binning.

1) CASE DESCRIPTION

Let us take an example from exploratory data analysis based on usage statistics from customer fleets. We take 1454 customer vehicles as the dataset with two segments, in which 727 BMW 740i limousine vehicles from the United Arab Emirates (UAE), denoted as set Ω_1, and 727 BMW 540i limousine vehicles from Japan, denoted as Ω_2 are given. In this case, these two segments are with identical cardinalities, i.e., $|\Omega_1| = |\Omega_2| = 727$. The long-term statistical data is acquired from dealers via on-board diagnostics [36], [37] or vehicle telemetries [10]. As shown in Table 2, the statistical usage data includes binned histogram values from ten measurements, with 24 bins for each measurement (sensor) histogram in average, 240 dimensions in total.

According to the prior knowledge, the usage behavior of the given two customer segments are quite different. Solely based on those binned data, we try separating the whole dataset into two segments after the preprocessing steps introduced in Section III-A. As the number of segments (two) is known, we group the 1454 vehicles by minimizing the squared Euclidean distances between them and their cluster centroids or means, i.e., k-means clustering [38]. According to Tselentis and Papadimitriou [39], k-means is one of the most commonly used methodologies for driver profile identification and driving pattern detection.

Denote the resulted cluster sets $\widetilde{\Omega}_1$ and $\widetilde{\Omega}_2$ from clustering based on the binned data after $\varkappa$-rank PCA. After clustering, we count the proportion of correctly clustered

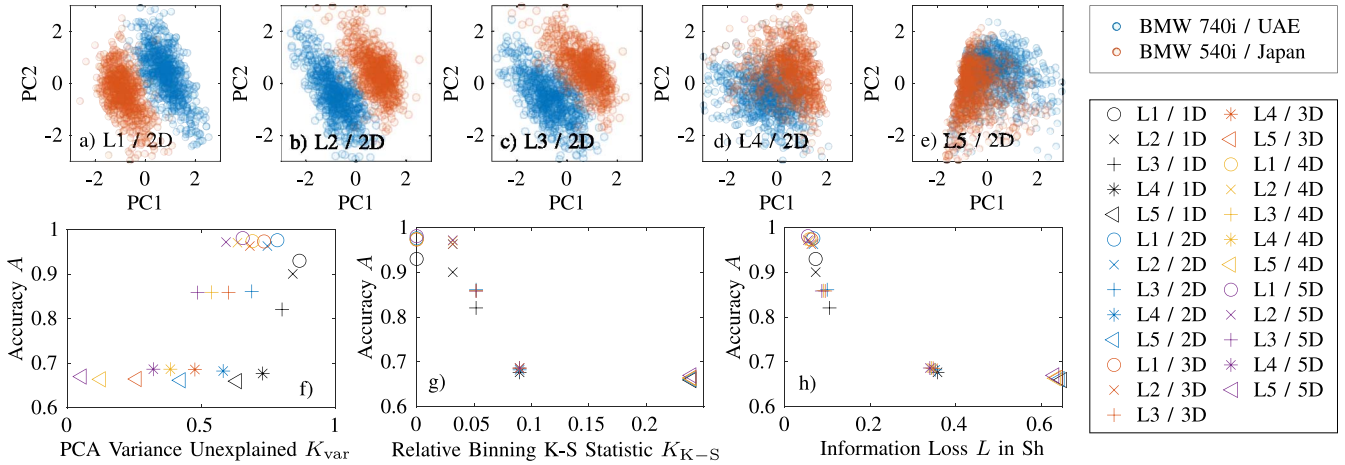


FIGURE 6. Experimental results of case study. (a)-(e) show the scatter plots of the 1424 customer fleets projected by the first two principal components (PC1 and PC2, $\varkappa = 2$) with five different binning levels (L1 to L5). For example, Fig. 6b is the scatter plot of PC1 and PC2 with binning level L2, denoted as L1 / 2D. (f)-(h) show three metrics (PCA variance unexplained K_{var}, relative binning K-S statistic $K_{\mathrm{K-S}}$, and information loss L in Sh) and their correlation to the accuracy A for clustering the two fleet segments over 25 experiment configurations (L1 to L5 combined with 1D to 5D).

TABLE 2. Binning configurations for case study. # is the abbreviation of "number of". L1–L5 are binning levels, where L1 is the original dataset without coarsening bins in the case study.

Measurement	Count	# bins				
		L1	L2	L3	L4	L5
Driving duration	# Sessions	24	12	6	3	2
Daily driving dur.	# Days	30	15	7	3	2
Daily mileage	# Days	34	17	8	4	2
Driving mileage	# Sessions	25	12	6	3	2
Parking duration	# Sessions	25	12	6	3	2
Stopping duration	# Sessions	16	8	4	2	1
Driving clock time	Duration	24	12	6	3	2
Stopping duration	Duration	16	8	4	2	1
Velocity	Duration	25	12	6	3	2
Acceleration	Duration	21	10	5	2	1
Total dimensions		240	118	58	28	17
Average # bins (k)		24	12	6	3	2

customers to all customers. As the clustered segment sets are unsupervised, their mapping to our reference segments can be represented by either the intersection of corresponding segment and cluster numbers or vice versa. We take the larger proportion as the clustering accuracy A, i.e.,

$$A = \max\left\{ \frac{|\Omega_1 \cap \widetilde{\Omega}_1| + |\Omega_2 \cap \widetilde{\Omega}_2|}{|\Omega_1| + |\Omega_2|}, \frac{|\Omega_1 \cap \widetilde{\Omega}_2| + |\Omega_2 \cap \widetilde{\Omega}_1|}{|\Omega_1| + |\Omega_2|} \right\}. \tag{23}$$

To evaluate the relationship between estimated information loss and the accuracy of k-means clustering, we compare them under different number of bins k and principal components $\varkappa$. As time-series raw data before binning are not available, we coarsen the fine-binned data by adding every two bins up. If the k for a sensor is odd, in the end, we add the last three bins up to make the coarsening more conservative. If there is only one bin, we keep it as it is without further coarsening. Regarding the original data as binning level one (L1), we repeat the coarsening steps until all bins to be concatenated, yielding five binning levels (L1,. . . , L5), whose number of bins for each random variable are listed in Table 2. Furthermore, we test the clustering with $\varkappa$ from one up to five, where the "curse of dimensionality" hardly occurs.

For this dataset, the number of observations (customer fleets) n equals $|\Omega_1| + |\Omega_2| = 1454$, the number of sensors (measurement variables) m is ten, the number of samples $p \to \infty$ as they are long-term sensor measurements which could span over several years. As the influence of structural parameters on losses are a multiplier, there is no impact on the trends with different k and $\varkappa$. Without the raw data before binning, we approximate the row-wise and column-wise blending factors λ, μ with the correlation coefficients of the L1 dataset. Hence, $\lambda = 0.117$, $\mu = 0.822$, which implies that the bin values are weakly correlated and the customers are strongly correlated.

Given the scale and structure parameters, we estimate the average information loss per measurement variables per vehicle (shortly "information loss" in the following) using our empirical loss model according to (19)–(22) and their parameters.

2) DATA VISUALIZATION

On each aggregation level (L1 to L5), we plot the scatter snapshots of reduced fleet data in Fig. 6 (a-e), using the first two principal components (PC1 and PC2) with given segment information. Corresponding to clustering accuracies shown in Fig. 6h, we can perform exploratory data analysis by observing those snapshots and trying identifying the two segments visually, assuming that the segments are not given.

The information losses and accuracies of L1 and L2 close to each other. Their snapshots also show nearly identical pattern despite their opposite directions of PC1. Compared the snapshot of L2 to that of L3, we observe that the two segments can be decently identified, whereas their centroids are closer. This implies the tiny loss gain and the small accuracy drop. Moving to higher aggregation levels (fewer bins and lower dimensions before PCA), the sensitivity of loss estimation to clustering performance slightly decreases. However, with a tripled information loss from L3 to L4, the accuracy decreases from a general acceptable level (over 80%) to less than 70%, i.e., nearly two thirds of customers are correctly clustered, which shows limited ability of clustering. The snapshot of L4 shows a larger intersection area between two customer segments. Without given segment information, it is already difficult to identify two segments with our eyes. In L5 where the loss is doubled to L4, the points are so close that we can hardly identify the distance between the centroids of both segments.

3) LOSS-AWARE EVALUATION

Figure 6h shows the relationship between the accuracies for k-means clustering and our estimated information losses.

When the average number of bins k decreases, the accuracy for clustering two customer segments decreases significantly. At the same time, with the increase of principal components (PCs) remained $\varkappa$ for clustering, the clustering accuracy A increases from a single PC to two PCs, and then converges with minor fluctuation due to the algorithmic uncertainty of k-means clustering and higher dimensionality. Comparing the converged A for each binning level to their next level (after pairwise coarsening), the increase accuracy by explaining more variance with more PCs did not compensate the influence of binning. In other words, the clustering accuracy with lower binning resolution generally decreases to an extent that more principal components could not explain the original pattern before k-fold binning. Hence, other than $\varkappa$-rank PCA, k-fold binning dominates the clustering performance.

From the information loss perspective, the empirical estimated total information loss L follows the trends of clustering accuracy. With higher binning levels and lower k, higher L is estimated from 0.07 Sh up to over 0.5 Sh when most of the effective information is lost (two bins). On the other side, for lower binning levels with higher k (mainly from L3 to L1), the loss converges without further increase, as the variance explained could cover the most information after binning. With higher $\varkappa$, L also decreases but it decreases comparatively milder than that with higher k. In addition to clustering performance, binning dominates the information loss as well.

In this case in practice, twelve bins per sensor in average (L2) keeps the most of information and ensures the clustering performance. Hence, L2 shows a trade-off between clustering accuracy (or visualization) and the dimensionality.

4) COMPARISON OF EVALUATION METRICS

As mentioned in Section II, there are already well-known metrics of histogram binning and PCA for evaluation. For histogram binning, we usually regard such histograms as empirical distributions and compare them to original distribution via K-S statistics, expressed in (2). For PCA, the variance unexplained quantifies the fraction of variance lost by projecting the high-dimensional dataset into a low-dimensional latent space described by those principal components (PCs), formulated in (8). Spanning binning and PCA as a sequence where binning does not take place on the analytics site but in the vehicle control units, data scientists usually focus on PCA guided by variance explained or unexplained. In this case study, therefore, we first compare our loss-aware approach to the variance unexplained, then to the K-S statistics. As we do not have the original time-series data from customer fleets over the years, we choose the finest binning level from our raw binned data (L1) as the reference for computing K-S statistics. For each comparison, we focus on two trends, more number of PCs, or finer binning.

As shown in Fig. 6f, with more PCs, we observe lower variance explained and higher accuracy. However, with less than 60% variance explained, we can hardly identify the correlation between the accuracy and variance unexplained. This implies that the variance unexplained is representable for PCA induced losses, but insensible for lower variance unexplained. With lower binning levels (finer binning), higher variance unexplained are estimated for the PCA. However, clustering accuracy are higher, indicating the variance unexplained cannot represent the influence of binning granularity on clustering accuracy. To conclude, the variance unexplained is unsuitable for evaluating PCA for histogram data.

For the other benchmark metric (binning K-S statistic), the experiment results are plotted in Fig. 6g. With more PCs, no influence of the relative binning K-S statistic has been observed on the accuracy. With finer binning, the metrics are lower, indicating the higher clustering accuracy, which means that the K-S statistic can properly identify the impact of binning on clustering accuracy. Although binning dominates the overall impact, the minor impact of PCA on the accuracy cannot be indicated by binning statistics, as it happens after binning. In summary, it is partially suitable as a criterion for preprocessing histogram data, but it provides no guidance on how to perform PCA afterward.

Compared to the both reference metrics, shown in Fig. 6h, our loss-aware approach shows good correlation between the estimated total information loss and the clustering accuracy. In summary, our empirical loss model explains the experiment results from k-means clustering well.

D. APPLICABILITY AND LIMITATIONS

From the perspective of overall preprocessing, aggregation with fewer bins is not the proper choice to improve the performance of PCA. If more bins are available without losing more information, fewer dimensions are required,

TABLE 3. An exemplary illustration of limitations of the loss-aware methods. In addition to k, the variance unexplained and total information loss are shown with the first one or two principal components, expressed in $(\cdot)@\varkappa(1 \text{ or } 2)$.

	A: No advantage for higher k			B: No advantage for higher $\varkappa$		
						
k	12	6	3	12	6	3
$K_{\text{var}}@\varkappa(1)$	0.305	0.305	0.305	0.248	0.048	0.033
$K_{\text{var}}@\varkappa(2)$	0	0	0	0	0	0
$K_{\text{var}}@\varkappa(3)$	0	0	0	0	0	0
$L@\varkappa(1)$	0.108	0.156	0.422	0.054	0.070	0.336
$L@\varkappa(2)$	0.098	0.144	0.406	0.051	0.069	0.334
$L@\varkappa(3)$	0.089	0.132	0.391	0.048	0.069	0.333

due to the enhanced correlation structure between the bins and between the sensors. In practice, if more bins per sensor are aggregated, the long-term statistical data can be preprocessed more compactly, improving the performance of further analysis such as clustering.

However, these findings are typically valid for customer fleet analysis or similar use cases where the raw fleet sensor data are continuously distributed, and they could be described by Beta distributions. To better indicate the limitations, we illustrated two counterexamples. Their snapshots, characteristics, and relevant indicators are presented in TABLE 3. The sparkline bar plots in the header show the histogram pattern of exemplary counterexamples A and B. These data for counterexample A and B are with $n = 4$, $m = 1$, and $p \to \infty$. Similar to the case study, both cases have been coarsened twice, yielding the number of bins k from twelve to six and then three.

Counterexample A shows no advantage with more than three bins per histogram. The histograms all have only three different levels, which are cut equidistantly. By coarser binning resolution where $k = 3$, we do not lose any information. Hence, the estimated losses should not be applied to counterexample A.

Counterexample B shows no advantage with more than two principal components (PCs). All four observations of histograms can be described by linear combinations of the first two histograms. Hence, more than two PCs do not bring any additional information gain here. All variance are explained with the first two PCs, whereas our estimated losses show slightly drop with more PCs. Hence, the method proposed in this paper does not imply the real behavior when preprocessing counterexample B.

V. SUMMARY AND OUTLOOK

This paper showed a comprehensive, information loss-based perspective to support decision-making in configuring preprocessing for customer fleet analytics. The preprocessing includes (i) data binning on the customer side, and (ii) principal component analysis (PCA) based on the binned data without given raw sensor measurements. First, we characterized the data using three scale parameters (number of vehicles, number of sensors, and number of samples) and two structural parameters (average correlation coefficients between the vehicles and the sensors each other). Based on these parameters, the scale and correlation structure of the dataset can be modeled. To estimate the information losses across binning and PCA, we generated the sample datasets stochastically. We then simulated the preprocessing parameterized with the number of bins, and the order of dimension remained.

A sensitivity study identified the impact of preprocessing, scale, and structure parameters on the binning, PCA, and total losses. If the scale parameters are sufficiently large, their effects on information losses are negligible. By performing empirical regression, we identified the mechanisms of the loss formulation. The total loss consists of both binning loss and weighted PCA loss. The binning loss depends primarily on the number of bins in a negative exponential fashion. The PCA loss was modeled using two parallel loss terms of the preprocessing parameters, weighted by a structural multiplier, in which the structural parameters are combined serially. A case study based on the customer fleet data, which is acquired in real-world and binned, manually configured various binning levels by coarsen the bins, and applied k-means clustering and exploratory data analysis. The case study demonstrated that the estimation of information loss could support decision-making in properly configure histogram binning and PCA without having the raw time-series measurement logging.

When working with histogram binning and PCA, it is valuable to assess information loss (for example, using our derived loss model) and then determine the optimal number of bins and principal components accordingly. Traditional methods like variance explained for PCA may not fully capture the influence of histogram binning on analytical performance. Our approach provides a more nuanced perspective for a thorough understanding of how histogram binning affects PCA and analytical outcomes. With sufficient number of samples, customers, and sensors, a higher resolution of histogram bins per sensor can generally improve the performance of dimensionality reduction without losing more information in a comprehensive view. Furthermore, we discussed the limitation of our methodology and the findings by illustrating two polar cases.

In the closing section of our paper, we highlight avenues for future research that stem from the methodology proposed in this study. A promising area for further investigation involves the exploration of multi-objective optimization techniques for binning followed by PCA. This optimization could aim to simultaneously minimize noise while maximizing retained information, all while avoiding an increase

in dimensionality that might compromise interpretability. Another interesting aspect worthy of future research is that, how the information losses of the original part of the dataset could be affected, when adding new sensors to the vehicle and keeping the number of bins unchanged. Additionally, when dealing with sensors that share physical relationships, there arises a need for aggregating their joint distributions, such as engine maps. Furthermore, understanding the impact of the mixed-variate aggregated dataset on information losses presents a valuable avenue for exploration in subsequent research efforts.

ACKNOWLEDGMENT

The authors would like to thank Helena Alder, Falk Hannemann, Christine Spannagl, and John Vicente for their continuous support and valuable discussion throughout the research.

REFERENCES

[1] J. Wilberg, F. Schafer, P. Kandlbinder, C. Hollauer, M. Omer, and U. Lindemann, "Data analytics in product development: Implications from expert interviews," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, 2017, pp. 818–822.

[2] A. Albers, F. Haug, J. Fahl, T. Hirschter, J. Reinemann, and S. Rapp, "Customer-oriented product development: Supporting the development of the complete vehicle through the systematic use of engineering generations," in *Proc. IEEE Int. Syst. Eng. Symp. (ISSE)*, 2018, pp. 1–8.

[3] S. Reicherts, B. S. Hesse, and D. Schramm, "Use of naturalistic driving studies for identification of vehicle dynamics," *IEEE Open J. Intell. Transp. Syst.*, vol. 2, pp. 195–206, 2021.

[4] F. Tanshi and D. Söffker, "Determination of takeover time budget based on analysis of driver behavior," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 813–824, 2022.

[5] K. Viktoriya, T. Kuhnimhof, and S. Trommer, "Assessment of real-world vehicle data from electric vehicles—Potentials and challenges," in *Proc. 11th Int. Conf. Transp. Surv. Methods ISCTSC*, Quebec, Canada, 2017, pp. 1–11.

[6] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," in *Proc. Privacy Enhanc. Technol.*, 2016, pp. 34–50.

[7] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, and Y. Yang, "Big data meet cyber-physical systems: A panoramic survey," *IEEE Access*, vol. 6, pp. 73603–73636, 2018.

[8] A. Nicolet, R. R. Negenborn, and B. Atasoy, "A Logit mixture model estimating the heterogeneous mode choice preferences of shippers based on aggregate data," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 650–661, 2022.

[9] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1972.

[10] B. Martens and F. Mueller-Langer, "Access to digital car data and competition in aftermarket maintenance services," *J. Competition Law Econ.*, vol. 16, no. 1, pp. 116–141, 2020.

[11] J. Schoch, J. Gaerttner, A. Schuller, and T. Setzer, "Enhancing electric vehicle sustainability through battery life optimal charging," *Transp. Res. B, Methodol.*, vol. 112, pp. 1–18, Jun. 2018.

[12] Y. Huang and S. Meng, "Automobile insurance classification ratemaking based on telematics driving data," *Decis. Support Syst.*, vol. 127, Dec. 2019, Art. no. 113156.

[13] K. Ling, N. Shah, and J. Thiele, "Customer-centric vehicle usage profiling considering driving, parking, and charging behavior," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, 2020, pp. 1–6.

[14] M. Boulle, "Optimal bin number for equal frequency discretizations in supervized learning," *Intell. Data Anal.*, vol. 9, no. 2, pp. 175–188, 2005.

[15] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosoph. Trans. Royal Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, Art. no. 20150202.

[16] E. Diday, "Thinking by classes in data science: The symbolic data analysis paradigm," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 8, no. 5, pp. 172–205, 2016.

[17] P. Brito, "Symbolic data analysis: Another look at the interaction of data mining and statistics," *Wiley Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 4, no. 4, pp. 281–295, 2014.

[18] L. Billard and J. Le-Rademacher, "Principal component analysis for interval data," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 4, no. 6, pp. 535–540, 2012.

[19] S. Makosso-Kallyth, "Principal axes analysis of symbolic histogram variables," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 9, no. 3, pp. 188–200, 2016.

[20] N. Haselgruber, K. Mautner, and J. Thiele, "Usage space analysis for reliability testing," *Qual. Rel. Eng. Int.*, vol. 26, no. 8, pp. 877–885, 2010.

[21] M. Bartłomiejczyk, "Driving performance indicators of electric bus driving technique: Naturalistic driving data multicriterial analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1442–1451, Apr. 2019.

[22] J. Schoch, P. Staudt, and T. Setzer, "Smart data selection and reduction for electric vehicle service analytics," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1592–1601.

[23] K. Ling, J. Thiele, and T. Setzer, "Usage space sampling for fringe customer identification," in *Proc. 54th Hawaii Int. Conf. Syst. Sci.*, 2021, p. 1748.

[24] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100378.

[25] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[26] A. Tavory, "Determining principal component cardinality through the principle of minimum description length," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, 2019, pp. 655–666.

[27] V. Bruni, M. L. Cardinali, and D. Vitulano, "A short review on minimum description length: An application to dimension reduction in PCA," *Entropy*, vol. 24, no. 2, p. 269, 2022.

[28] E. Vaiciukynas, M. Ulicny, S. Pashami, and S. Nowaczyk, "Learning low-dimensional representation of bivariate histogram data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3723–3735, Nov. 2018.

[29] F. J. Massey Jr., "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.

[30] N. Q. V. Hung, H. Jeung, and K. Aberer, "An evaluation of model-based approaches to sensor data compression," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2434–2447, Nov. 2013.

[31] D. L. Greene, "Estimating daily vehicle usage distributions and the implications for limited-range vehicles," *Transp. Res. B, Methodol.*, vol. 19, no. 4, pp. 347–358, 1985.

[32] Z. Lin, J. Dong, C. Liu, and D. Greene, "Estimation of energy use by plug-in hybrid electric vehicles," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2287, no. 1, pp. 37–43, 2012.

[33] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2, 2nd ed., N. L. Johnson, S. Kotz, N. Balakrishnan, Eds. New York, NY, USA: Wiley, 1994.

[34] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[35] W. J. Morokoff and R. E. Caflisch, "Quasi-Monte Carlo integration," *J. Comput. Phys.*, vol. 122, no. 2, pp. 218–230, 1995.

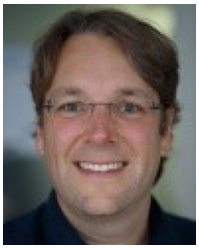
[36] A. Deicke, "The electrical/electronic diagnostic concept of the new 7 series," SAE, Warrendale PA, USA, Rep. 2002-21-0022, 2002.

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

[38] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[39] D. I. Tselentis and E. Papadimitriou, "Driver profile and driving pattern recognition for road safety assessment: Main challenges and future directions," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 83–100, 2023.

KUNXIONG LING received the M.Sc. degree in energy science and engineering from the Darmstadt University of Technology, Germany, in 2018, and the Ph.D. degree in business informatics from the Ingolstadt School of Management, Catholic University of Eichstätt-Ingolstadt, Germany, in 2022. He is currently working with BMW Group, Munich, Germany, focusing on data aggregation and sampling procedures for customer-centric automotive systems engineering.



JAN THIELE received the degree in applied mathematics from the Technical University of Munich (TUM) and the Technical University of Vienna, and the Diploma degree in technomathematics from TUM in 2003. Since then, he is working with the Research and Development Department, BMW Group, Munich, Germany. Over the last years, he succeeded in establishing data science as an important factor in the reliability design and validation of engine components.

THOMAS SETZER received the Diploma degree in business engineering from the Karlsruhe Institute of Technology, Germany, and the Dr.rer.nat. degree in information systems from the Technical University of Munich, Germany. He is currently a Professor of Business Informatics with the Ingolstadt School of Management, Catholic University of Eichstätt-Ingolstadt.