# Usage Space Sampling for Fringe Customer Identification

Kunxiong Ling
BMW Group
Catholic University of Eichstätt-Ingolstadt
kunxiong.ling@bmw.de

Jan Thiele
BMW Group
jan.thiele@bmw.de

Thomas Setzer
Catholic University of Eichstätt-Ingolstadt
thomas.setzer@ku.de

## 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.*

## 1. Introduction

To develop customer-centric technical products and services, it is essential to know the customers, where information on customer behavior is increasingly derived by analyzing customer usage data to better understand how and to what extend products and services are used [1, 2]. For instance, Haselgruber *et al.* [3] analyzed the usage space of customers to evaluate engine reliability testing programs. Schoch *et al.* [4] applied customer usage statistics to generate profiles for electric vehicle service analytics.

However, the ongoing deluge of usage and sensor data discourages computational-intensive customer analytics. First, logging vast amounts of time series-based data is expensive and could raise ethical issues. An example in the automotive industry is driver fingerprinting [5]. Second, advanced analytical methods are often non-linear and computationally expensive (e. g. large-scale simulations).

A common solution to preserve customer privacy and lower computational costs is to reduce the number of features (dimensions) by data aggregation. For instance, long time measurement values from customer vehicles are aggregated in histogram values through aftersale diagnostics, which allows for a massive reduction of the number of features per customer [6].

After dimensionality reduction, in case individual analysis per customer is required, the number of customers ought to be further reduced, typically by selecting a smaller set of reference customers from the customer population, i. e. sampling. For reasons of representativeness and unbiasedness, a sound coverage of the usage behavior of all customers is required here.

Sampling is widely applied in various branches, such as uncertainty quantification [7], banking [8], and customer relationship management [9], to name only a few. However, in many applications, not all customers are equally important, in particular when applications are related to anomaly detection, risk assessment, and the determination of critical or extreme behavior. For instance, in the automotive engine development, usage behavior of different customers lead to different engine health conditions.

Typically, many customers have similar usage behavior, which then allows for determining suitable product requirements. These customers are the target for product marketing, called core customers. Considering the representativeness of distributions over customer attributes, potentially stratified random sampling is appropriate to represent core customers [10]. However, unusual usage behavior of customers leads to increased risks of subsequent product damage or even failure. The respective usage behavior is to be identified to guide reliability design and testing. These groups of unusual customers, coined fringe customers, are the references for product quality considerations.

Unfortunately, usual approaches such as confidence intervals of random samples in survey techniques can

URI: https://hdl.handle.net/10125/70824
978-0-9981331-4-0

hardly indicate whether one of the fringe customers is contained or not. Moreover, random sampling is a stochastic process, which is not deterministic and not reproducible. Together with the non-linearity and high complexity of today's technical or service systems, it is particularly challenging to identify fringe customers [11]. Hence, a sampling technique is demanded that aims at minimizing the number of samples required while selecting the latent fringe customers and such that the overall distribution of customers is still represented to an acceptable extend.

Having the usage data of customers available, our hypothesis is that sampling in the usage space (on features related to the service or product usage behavior of customers) can improve the representativeness of fringe customers' impact on latent indicators (e. g. damages) that can then be used to perform further analytical tasks.

In this paper, we present a novel method to perform sampling in the usage space of customers focusing on fringe customers. Given a high-dimensional usage space, first, we reduce its dimensionality using singular value decomposition. Afterwards, we select fringe customers based on geometrical properties in the compressed space, and select core customers using non-discrepancy sequences.

To further approximate the distribution of customer population, we compute the weights of selected customers based on their mutual distances. To evaluate the feasibility of the method, we consider three benchmark functions to simulate latent indicators according to the distribution of fringe customers in their usage space, i. e., on the boundary, near the boundary, and randomly. Finally, we will discuss the impact of each step above on the representativeness of sampling.

The remainder of this paper is structured as follows. In Section 2, we will introduce the notation together with related work on sampling, and highlight the contribution of this work. In Section 3, we will propose the treatments and algorithmic design of our method. Then, we will describe the experimental settings and loss functions considered in Section 4. In Section 5, we will compare the representativeness of different sampling schemes and discuss the mechanism, the feasibility in experimental studies, and sketch sensitivities of the method to different treatment combinations. Finally, we conclude and outline promising research directions in Section 6.

## 2. Preliminaries and Related Work

Suppose a dataset $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ is available, which describes the usage behavior of $p$ customers with e. g. statistics of sensor measurements. For customer $i = 1, \ldots, p$, the usage data is aggregated in a vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_x})^\top$, which consists of $n_x$ features. Hence, the column space of $X$ spans the usage space of a customer.

Applying a model or function $f$ to each customer $i$, customer usage features are regarded as input to yield $n_y$ outputs in $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_y})^\top$, i. e. $f : \boldsymbol{x}_i \mapsto \boldsymbol{y}_i,\ \boldsymbol{x}_i \in X$. These outputs could indicate various performance metrics, providing references to support decision-making. When applying the model to all $p$ customers, the outputs can be represented in an indicator space $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p\}$.

However, if model $f$ is computationally complex and $p$ is large, $\widehat{p}$ customers are selected as samples, where the number of samples can be much lower then the number of customers, i. e. $\widehat{p} \ll p$. As our sampling method is carried out in usage space $X$, we call it usage space sampling.

After usage space sampling, the dataset becomes $\widehat{X} \subset X$, and the outputs become $\widehat{Y} \subset Y$. Typically, these models are carried out independently from the customers. Hence, $f$ should have a linear time complexity of $O(n)$. Regardless of the time of sampling, elapsed time for this analytic task is reduced by a factor of $\widehat{p}/p$.

Despite the dataset cardinality reduction, the quality of sampling procedures is indicated by measures of representativeness of the drawn samples. The distribution of outputs among the samples should be similar to that among the whole customer population. In addition, as discussed above, the so-called fringe customers are crucial for determining the margins of indicators and are hence important to be identified and considered appropriately.

A large body of works has been conducted on improving the representativeness of sampling methods. For instance, Arnst *et al.* [12] found that random sampling tends to perform better in approximating the distribution of population when the dimensionality is reduced. Yet, Park *et al.* [13] indicated the limitation of random sampling for data visualization. Furthermore, Loyola R. *et al.* [14] investigated different combinations of sampling techniques in discrepancy and statistical moments. They found that Halton sequences outperform random sampling and other geometric-based sampling methods, considering the trade-off between computational efficiency, uniformity and high-dimensional capabilities.

Inspired from the related works, we found that performing dimensionality reduction before using Halton sequences can improve the sampling representativeness for high-dimensional data. With

the input of usage statistics, in addition, our sampling method complements the state of the art by enhancing fringe customer identification.

## 3. Usage Space Sampling

This section introduces the usage space-based sampling method we propose in this work. The sampling procedure, illustrated in Figure 1, will now be described.

In step 1, usage space analysis, singular value decomposition is applied to dataset $X$ to linearly project the data into a lower-dimension space, yielding $\widetilde{X}$ with $\widetilde{n}_x$ dimensions.

In step 2, fringe sampling, we separately select $\widehat{p}_{\mathrm{F}}$ samples from the borderline area of the cloud of data points in usage space, in order to target fringe customers. The determination of samples and $\widehat{p}_{\mathrm{F}}$ are pre-selected using a so-called convex hull-based approach.

Given the number of required samples $\widehat{p}$, the remained quantity $(\widehat{p} - \widehat{p}_{\mathrm{F}})$ is arranged for the core customers by an extra step of sampling in step 3.

Note that the samples are equally weighted in $\widehat{X}$. In the original usage space $X$, instead, each of them represents different number of customers. Considering their unequal weights, in step 4, we can group the $p$ customers into $\widehat{p}$ segments using Voronoi tessellation. The carnality of each segment determines the corresponding weight as market volume.

So far, $\widehat{p}$ samples are selected from $p$ customers in their usage space with their weights, and prepared for further analytics. The individual steps will now be described in detail.

### 3.1. Usage Space Analysis

The objective of step 1 is to reduce the dimensionality of $X$ from $n_x$ to $\widetilde{n}_x$. First, we concatenate dataset $X$ directly according to its natural sequence, building a matrix $\boldsymbol{X} = (x_{ij})$ with $p$ rows and $n_x$ columns (dimensions). Then, we approximate matrix $\boldsymbol{X}$ with truncated singular value decomposition (tSVD), a widely-used linear dimensionality reduction technique [15].

To ensure balanced importance of features, before applying tSVD, we perform column-wise $z$-score normalization of $\boldsymbol{X}$ as shown in (1).

$$\boldsymbol{X}_{\mathrm{n}} = \frac{\boldsymbol{X} - \overline{\boldsymbol{X}}}{\sqrt{\mathrm{Var}(\boldsymbol{X})}}, \tag{1}$$

where $\overline{\boldsymbol{X}}$ and $\mathrm{Var}(\boldsymbol{X})$ are column-wise mean values and variances. This allows that the tSVD can capture the maximum variance in the matrix.

We perform tSVD on the normalized matrix according to

$$\boldsymbol{X}_{\mathrm{n}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \tag{2}$$

where $\boldsymbol{U} \in \mathbb{R}^{p \times n_x}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n_x \times n_x}$. The unitary weight matrix $\boldsymbol{V} \in \mathbb{R}^{n_x \times n_x}$ can be represented with $n_x$ vertical weight vectors $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_x})$. Then, we use the first $\widetilde{n}_x$ vectors to build $\widetilde{\boldsymbol{V}} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\widetilde{n}_x})$. The normalized matrix $\boldsymbol{X}_{\mathbf{n}}$ can be compressed into $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{p \times \widetilde{n}_x}$ by

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X}_{\mathbf{n}}\widetilde{\boldsymbol{V}}. \tag{3}$$

Matrix $\widetilde{\boldsymbol{X}}$ represents the usage space in a compact fashion and serves as the input for sampling.

### 3.2. Fringe Sampling

After step 1, core customers with usual usage behavior, are located near the center of the cloud of data points in usage space. However, compared to core customers, customers exhibiting unusual usage behavior, i. e. fringe customers, are typically observed less frequently.

As we are particular interested in fringe customers, we perform an extra sampling step before sampling the

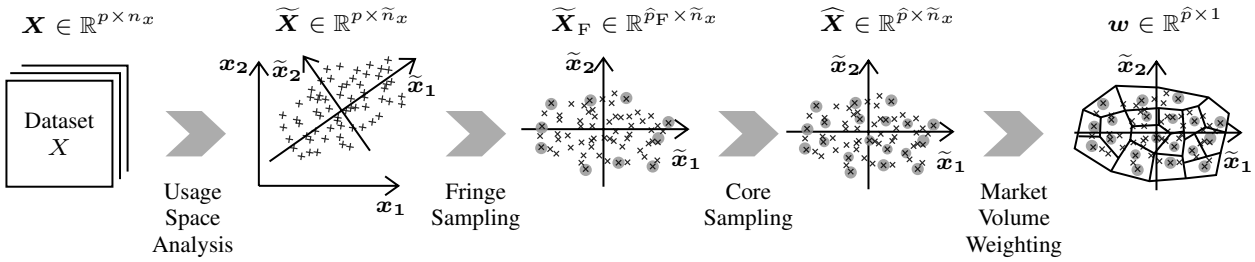



**Figure 1. A schematic overview of the usage space sampling proposed in this work. Grey circles are the sampled fringe and core customers. The grid between those samples represents the segmentation for weighting. Piror to sampling, the $n_x$-dimensional usage space is reduced to $\widetilde{n}_x - d$ dimensions. In the reduced space, $\widehat{p}$ samples are determined, in which $\widehat{p}_{\mathrm{F}}$ fringe customers are selected. Every sample is then weighted by its market volumes $w$.**

core customers, focusing on potential fringe customer candidates which are far from the cloud center. To outline the fringe of a vector space, we use the concept of a convex hull of the cloud of data points, which is a straight-forward approach [16].

A widely-applied method to construct convex hulls, used in our procedure, is the Quickhull algorithm [17]. In a data space, points on the convex hull are found by using a divide and conquer approach. Similar to several convex hull algorithms, it has a time complexity of $O(n \log n)$, in the worst case $O(n^2)$. Given a large number of customers as population (e. g. $10^5$), it is inappropriate to compute the convex hull directly. This is because, (i) numerous points lead to computational inefficiency, and (ii) local maldistribution of these points results in large amount of fringe samples to be chosen than needed.

**Algorithm 1:** Fringe sampling.

**Require** $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{\widetilde{n}_x})$
**for** $i = 1, \ldots, (\widetilde{n}_x - 1)$ **do**
  **for** $j = (i+1), \ldots, \widetilde{n}_x$ **do**
    $\boldsymbol{\varphi} \in \mathbb{R}^{p \times 1}, \boldsymbol{r} \in \mathbb{R}^{p \times 1} \leftarrow$ map Cartesian coordinate vectors $(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j)$ to polar coordinate vectors;
    Split the data evenly into $k$ segments according to $\boldsymbol{\varphi}$;
    For each segment, find the maximum element of $\boldsymbol{r}$;
  **end**
**end**
$\widetilde{\boldsymbol{X}}_{\mathrm{F}} \leftarrow$ Concatenate all the samples found in the previous step;
Compute the $\widetilde{n}_x$-dimensional convex hull of the samples in $\widetilde{\boldsymbol{X}}_{\mathrm{F}}$;
$\widetilde{\boldsymbol{X}}_{\mathrm{F}} \leftarrow$ Update the samples by removing those who are not on the convex hull;
**Return** $\widetilde{\boldsymbol{X}}_{\mathrm{F}} \in \mathbb{R}^{\widehat{p}_{\mathrm{F}} \times \widetilde{n}_x}$

Algorithm 1 shows how these problems are addressed in our work. We transfer the Cartesian coordinates of customers in usage space $\widetilde{\boldsymbol{X}}$ to polar coordinates. For each combination of two dimensions, we split the customers in $k$ segments according to their polar angles $\boldsymbol{\varphi}$. Per segment, we determine the customer with the largest radius. By computing the convex hull using the identified customers, $\widehat{p}_{\mathrm{F}}$ discrete fringe customers $\widetilde{\boldsymbol{X}}_{\mathrm{F}}$ are found in the usage space on the hull. Other customers inside the hull are omitted. This method could approach a time complexity of $O(n)$ and outline the fringe customers potentially missed by traditional sampling approaches.

### 3.3. Core Sampling

Beyond the $\widehat{p}_{\mathrm{F}}$ fringe customers, $\widehat{p} - \widehat{p}_{\mathrm{F}}$ core customers are sampled from inside the hull in the usage space.

Therefore, we apply quasirandom sequence as a replacement of uniformly distributed random samples, as (i) the method is deterministic, (ii) the sequence is generated in a given space, and (iii) samples are chosen according to a sequence with low discrepancy, seemingly random even with a few samples [18].

One of the most popular quasirandom sequence that is implemented in our procedure is Halton sequence. The sequence of numbers are generated based on coprime numbers. Loyola R. *et al.* [14] show that Halton sequence (i) can generate samples that fill the space in a highly uniform manner, and (ii) is computational efficient also in high-dimensional spaces.

First, we generate a Halton quasirandom point set with $\widehat{p} - \widehat{p}_{\mathrm{F}}$ samples in $\widetilde{n}_x$ dimensions, denoted $H$. As the point set ranges between $[0, 1]$, we then rescale the populations $\widetilde{\boldsymbol{X}}$ in the usage space with that range, yielding $\widetilde{\boldsymbol{X}}_{\mathrm{n}}$. Afterwards, for each generated point in set $H$, find the nearest neighbor from the rows of $\widetilde{\boldsymbol{X}}_{\mathrm{n}}$, measured by Euclidean distances. These neighbors found are regarded as our core samples, denoted by $\widehat{\boldsymbol{X}} \in \mathbb{R}^{\widehat{p} \times \widetilde{n}_x}$. The matrix builds a vector space $\widehat{X}$ with a cardinality of $\widehat{p}$.

### 3.4. Market Volume Weighting

The selected $\widehat{p}$ individual samples from $p$ customers are prepared for further analytics. Assuming that $p$ customers represent the population, however, the customers represented by each sample, coined market volumes, are not equal. For instance, a fringe sample could represent much less customers compared to a core sample. Without weighting, the effect of fringe segments could be overestimated and is not representative for the population.

Therefore, weighting based on market volumes is conducted, where we estimate market volumes using Voronoi tessellation in the compressed usage space. For each customer, the nearest sample is found with smallest Euclidan distance and assigned to the customer. Afterwards, we consider the number of customers, assigned in sample $i = 1, \ldots, \widehat{p}$, as the corresponding market volume, or weight $w_i$. The weighting can be applied for estimating the distributions of outputs. For output $j = 1, \ldots, n_y$ in the indicator space $Y$, the empirical cumulated distribution function (ECDF) of population is expressed in (4).

$$F_j(\theta) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}_{y_{ij} \leqslant \theta}, \tag{4}$$

where $i$ represents the customer, $\theta$ represents the function value of output $j$, and $\mathbf{1}$ is the indicator function to be activated under the condition written in the subscript. After the previous sampling steps, the ECDF of $\widehat{p}$ sample is

$$\widehat{F}_{j,0}(\theta) = \frac{1}{\widehat{p}} \sum_{i=1}^{\widehat{p}} \mathbf{1}_{y_{ij} \leqslant \theta}. \tag{5}$$

With weighted samples, $\widehat{F}_{j,0}(\theta)$ becomes

$$\widehat{F}_{j}(\theta) = \frac{1}{\sum_i w_i} \sum_{i=1}^{\widehat{p}} w_i \mathbf{1}_{y_{ij} \leqslant \theta}. \tag{6}$$

To evaluate the performance of the proposed method, we now describe the setup and the results of an experiment using several non-linear benchmark functions.

## 4. Experimental Study

To study the influence of the treatments (usage space analysis, fringe sampling, core sampling, and market volume weighting) on the representativeness of the samples, we regard relevant combinations of the treatments as sampling schemes, outlined in Section 4.1.

Considering reproducibility, we simulate the population ($p$ customers with $n_x$ features) in a stochastic fashion by generating the values randomly from Gaussian distributions. In addition, we use three benchmark functions to represent corresponding complex analytical models as will be presented in Section 4.2. After computing the function outputs (indicators) for $p$ customers, we perform different sampling schemes on those customers.

With the indicators available, we evaluate their similarities according to two loss functions the will be forumlated in Section 4.3. Both focus on fringe customer identification and their distributions.

### 4.1. Sampling Schemes

To reduce the degrees of freedom, we will not consider the impact of different parameterization of the steps on the representativeness, but use default parameters for each treatment that turned out to be suitable in preliminary tests. In this case, the sampling scheme solely considers whether each step is activated.

For usage space sampling, the dimension of $\widetilde{n}_x$ is set to three, the maximum possible dimension for intuitive visualization.

For fringe sampling, the number of segments $k$ is set to 16. This is done such that for each combination of two column vectors, their polar plane is split into 16 folds, each of which accounts for 22.5 degrees. After fringe sampling, up to 16 samples are selected for each two dimensions, i. e., $\widehat{p}_{\mathrm{F}} \leqslant 16 \binom{\widetilde{n}_x}{2}$ in total.

As $\widehat{p}_{\mathrm{F}}$ could increase with $\widetilde{n}_x$ squared, we perform fringe sampling only if usage space sampling is activated.

For core sampling, the default configuration is to position the $\widetilde{n}_x$-dimensional Halton sequence in the usage space and select the nearest neighbor as the sample. In contrast, random permutation is used when core sampling is deactivated.

The market volume weighting is non-parametric and conducted after having the samples available. Hence, the influence of weighting will be investigated for every sampling scheme based on the treatments above.

We represent the sampling scheme with four digits, where each digit represents whether a certain treatment is activated (hot) or not. For instance, scheme 0001 means that all the treatments are deactivated except the weighting is activated. In this paper, ten sampling schemes are defined in Table 1, where (i) fringe sampling is only conducted jointly with usage space analysis, and (ii) there is no influence from usage space analysis on random sampling (core sampling deactivated). Hence, cases with 01** and 100* are omitted, where * is a wildcard for $\{0, 1\}$.

**Table 1. Overview of the sampling schemes.**

| Case | Usage Space Analysis | Fringe Sampling | Core Sampling | Market Volume Weighting |
|---|---|---|---|---|
| 0000 | | | | |
| 0001 | | | | ✓ |
| 0010 | | | ✓ | |
| 0011 | | | ✓ | ✓ |
| 1010 | ✓ | | ✓ | |
| 1011 | ✓ | | ✓ | ✓ |
| 1100 | ✓ | ✓ | | |
| 1101 | ✓ | ✓ | | ✓ |
| 1110 | ✓ | ✓ | ✓ | |
| 1111 | ✓ | ✓ | ✓ | ✓ |

### 4.2. Benchmark Functions

To test the performance of the sampling procedure, we choose three different benchmark functions. These functions serve as the outputs of analytical model $f$, hence $n_y = 3$, in which the fringe customers are

differently spanned. In addition, these functions are non-convex and high-dimensional configurable, which enables simpler experimental procedure for different $n_x$. For customer $i$, the $n_x$-dimensional versions of the functions are shown below.

**Ackley Function.** We apply Ackley function with the parameters recommended in [19], i. e.,

$$\begin{aligned} y_{i1} = &-20\exp\left(-0.2\sqrt{\frac{1}{n_x}\sum_{j=1}^{n_x} x_{ij}^2}\right) \\ &-\exp\left(\frac{1}{n_x}\sum_{i=1}^{n_x}\cos(2\pi x_{ij})\right) + 20 + \mathrm{e}. \end{aligned} \tag{7}$$

This function is symmetric, where its largest values are found on the boundary of the cloud of data points– presumably the convex hull.

**Rescaled Schwefel Function.** Schwefel function [20] is asymmetric and exhibits several local maxima and minima. As it is typically evaluated in a hypercube $x_{ij} \in [-500, 500]$, we rescale $x_{ij}$ with a factor of 250 such that its hypercube range lies in $[-2, 2]$, which is comparable to the range of the Ackley function. The modified Schwefel function is written as

$$y_{i2} = 419\, n_x - \sum_{j=1}^{n_x} 250\, x_{ij} \sin\left(\sqrt{|250\, x_{ij}|}\right). \tag{8}$$

The largest values of rescaled Schwefel function are between the boundary and the center.

**Random Function.** The most non-linear function we use is random, which shows no dependence on $x_{ij}$. We model the random function using a Uniform distribution in $[0, 1]$, i. e.,

$$y_{i3} \xleftarrow{\text{random}} \mathcal{U}(0, 1). \tag{9}$$

This function is asymmetric. Its largest values are dispersed in the usage space.

Figure 2 visualizes the benchmark functions in a two-dimensional usage space, where the distribution properties of the points with the largest values can be observed intuitively. The 1000 points are randomly generated from the Gaussian $\mathcal{N}(0, 1)$. Their function values $y_{i1}$, $y_{i2}$, and $y_{i3}$ are then separately calculated as in to equations (7-9).

Under these conditions, we are able to evaluate the influence of our individual treatments on the sampling representativeness.

### 4.3. Evaluation

To build the population of customers, we individually and randomly generate $n_x$-dimensional usage space values for $p$ customers from $\mathcal{N}(0, 1)$. This is the worst case for sampling in the usage space, as the usage space values are uncorrelated. It is expected that, with a stronger correlation, the sampling representativeness could be improved. Furthermore, considering the variable region of $\mathcal{N}(0, 1)$ and the properties of our benchmark functions, the effects of sample imbalance can be estimated. This represents, to some extent, the imbalance between core and fringe customers.

In this experiment, we set $p = 50000$ and analyse a low-dimensional ($n_x = 5$) and a high-dimensional ($n_x = 500$) case. In both cases, we apply each sampling scheme from the lowest possible number of samples up to 500 samples, i. e., $\widehat{p} \leqslant 500$. To allow for a reliable comparison of deterministic (**1*) to stochastic random sampling schemes (**0*), we repeat each experiment configuration 100 times.

The representativeness of the sampling is measured by the two loss functions introduced below.

**Fringe Loss.** A successful identification of the fringe customers with our method requires that – at least to a large extend – the data points with the largest indicator values are those of the fringe customers. We quantify this requirement by positioning the sample with largest quantile from the sample ECDF $\widehat{F}_j$ into the ECDF of all customers $F_j$.

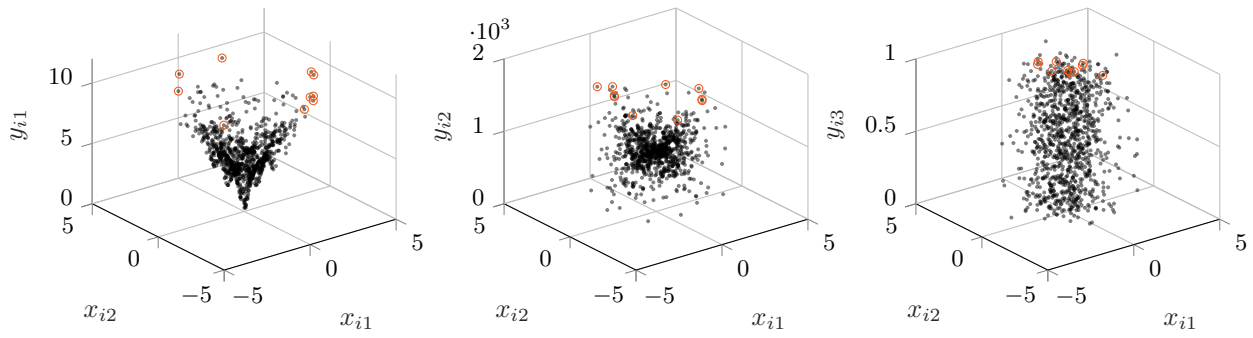
For indicator $j \in \{1, 2, 3\}$, we therefore define the fringe loss as

$$L_{\mathrm{f}} = \max\left\{1 - F_j\left(\widehat{F}_j^{-1}(1)\right), 10^{-6}\right\}, \tag{10}$$

where $\widehat{F}_j^{-1}$ is the inverse ECDF of indicator $j$. The closer to the largest quantile of all customers this data point, the smaller the fringe loss. To enable the visualization of losses in a logarithmic fashion, we limit the minimum to $10^{-6}$, such that the sample with largest indicator value exceeds 99.9999% of the customers. Similarly, a $L_{\mathrm{F}}$ of $10^{-2}$ indicates a tolerance of 99%.

**Distribution Loss.** Despite the fringe loss, for indicator $j$, the distribution loss measures the difference between the ECDF of the samples and that of the population. We choose the root of the relative sum of squared residuals to represent the difference, i. e.,

$$L_{\mathrm{d}} = \sqrt{\frac{\sum_{i=1}^{p}\left(\widehat{F}_j^{-1}(y_{ij}) - F_j^{-1}(y_{ij})\right)^2}{\sum_{i=1}^{p}\left(F_j^{-1}(y_{ij})\right)^2}}. \tag{11}$$




(a) Ackley function (b) Rescaled Schwefel function (c) Random function

**Figure 2. Visualization of benchmark functions in a two-dimensional usage space. For each subplot, a black point represents a customer, in total 1000 customers. Customers trimmed with red circles are those with function values larger than 99% of the customers.**

However, as $\widehat{p} \neq p$, prior to calculating the loss, we interpolate the quantile vectors of them from $\{1, \ldots, \widehat{p}\}$ to $\{1, \ldots, p\}$ towards their previous neighbors. The smaller the loss is, the closer the ECDFs are.

In the following, we will present both losses with different sample cardinalities $\widehat{p}$ and the sampling schemes introduced in 4.1.

## 5. Results and Discussion

The fringe and distribution losses with different sampling schemes are compared under two scenarios: low-dimensional ($n_x = 5$, henceforth: 5D ) and high-dimensional ($n_x = 500$, henceforth: 500D) usage spaces. Further parameters are introduced in Section 4.3. Afterwards, the robustness of our methods for real-world data is discussed, according to noises, missing data and scale errors.

### 5.1. Sampling Representativeness in Low-Dimensional Usage Space

The properties of sampling losses can be observed in the random scheme without any treatment (0000). As shown in Figure 3, with the increase of $\widehat{p}$, both losses are reduced quasi-exponentially for all three functions, which are almost straight along with the log-log axes. However, after 100 repetitions, the losses fluctuate due to the uncertainty of random sampling.

Adding the treatment of fringe sampling together with usage space analysis (1100), the fringe customers are successfully identified in Ackley and rescaled Schwefel functions. However, their distribution losses are larger, which is biased due to the inference of fringe samples for the whole group, resulting in maldistribution.

In random function, fringe sampling increases the robustness of random sampler by reducing the upper bound of error bands for $\widehat{p} \leqslant 100$, as the majority of samples are selected according to the geometry property of usage space. For $\widehat{p} > 100$, the losses yield scheme without usage space treatments (0000).

Sampling schemes with the treatment of core sampling but without fringe sampling (*010) generally outperform random sampling. In concrete, ten samples are sufficient to reach 99.9% in Ackley function, as well as 99% in rescaled Schwefel function. This implies that the deterministic core sampling is capable of identifying the customers which are not exactly on the boundary (rescaled Schwefel function), as the low-discrepancy sequences manage to take samples evenly in the low-dimensional space.

For random function with no dependency on the fringe customers in the usage space, the results with treatments remain similar to the median of random scheme (0000). However, such small number of samples could be due to luck, as the global maximum is a neighbor of the pre-defined sequence. Furthermore, with the increase of number of samples, the loss further reduces until the customer with global maximum is included, where around 200 customers are sampled. Moreover, as the dimensionality of usage space is relatively low, no significant influence of usage space analysis on both losses is observed.

Combining the core sampling with fringe sampling, the sampling scheme (1110) identifies the fringe customers as that with core sampling deactivated (1100). However, the distribution loss remains at a high level due to the deterministic property of Halton sequence in the core sampling.

The sampling schemes which are compared above are without market volume weighting. On the one hand, a step of weighting after sampling has no influence on fringe loss. The reason is that, the quantile from the

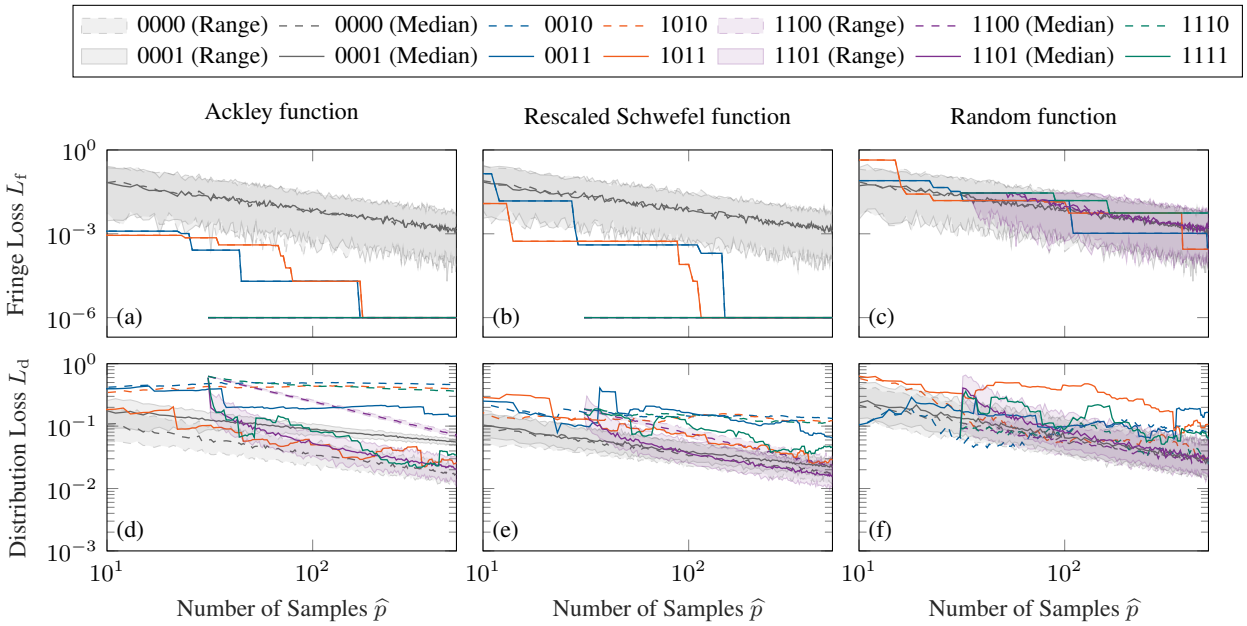



**Figure 3. Sampling representativeness with different number of samples in five-dimensional usage spaces. The sampling schemes are coded using four digits, each representing the activation of a treatment, i. e., usage space analysis, fringe sampling, core sampling, and market volume weighting. Results of schemes with uncertainties (**0*) are represented with medians and their error bands from their fifth to their 95$^{\text{th}}$ percentiles (ranges).**

sample ECDF, utilized in equation (10), is the maximum value of the samples and compared to the ECDF of the customer population. On the other hand, combining weighting can significantly improve the distribution loss for Ackley and rescaled Schwefel functions. Most customers are close to the center of the usage space, while the fringe customers for those two functions are away from center. This also explains why weighting has no significant influence on distribution losses in case of random functions.

For low-dimensional usage spaces and indicator functions where the fringe values are on or close to the outer boundaries of customer population, the conclusions are summarized as follows. Usage space analysis has less influence on sampling performances. Fringe sampling significantly reduces the fringe loss. Core sampling improves the robustness of selected samples due to its deterministic property. Market volume weighting compensates its drawback on distribution losses, matching the representativeness of simple random sampling without any treatment.

## 5.2. Sampling Representativeness in High-Dimensional Usage Spaces

In 500-dimensional usage spaces, most conclusions from the 5D case are found, which can be observed in Figure 4. However, fringe sampling does not exactly identify the fringe customer with maximal indicator values, but limits the upper range of losses compared to simple random sampling (scheme 0000).

With the core sampling activated, sampling in compressed three-dimensional space (scheme 101*) has lower fringe loss than directly sampling in the high-dimensional space (scheme 001*). This can be observed with all functions. As the sampling is performed in a compact usage space with maximized variances in each dimension compared to 5D usage spaces, fewer samples are required to cover the usage space with similar density.

With the help of fringe sampling, the fringe loss further decreases when considering the Ackley function. Yet, no significant improvement of fringe loss is observed in rescaled Schwefel and random functions, where the fringe customers are not exactly on the boundary. As the usage space analysis performed in this paper is an linear approach, the geometric structure inside the usage space cannot be clearly represented, with a low-level linear approximation of 500 dimensions using three axes. For indicator functions without local maxima on the boundary, therefore, fringe identification hardly reduces the fringe loss in 500D usage spaces. Yet, no negative impact is observed.

The geometry of 500D data is less interpretative

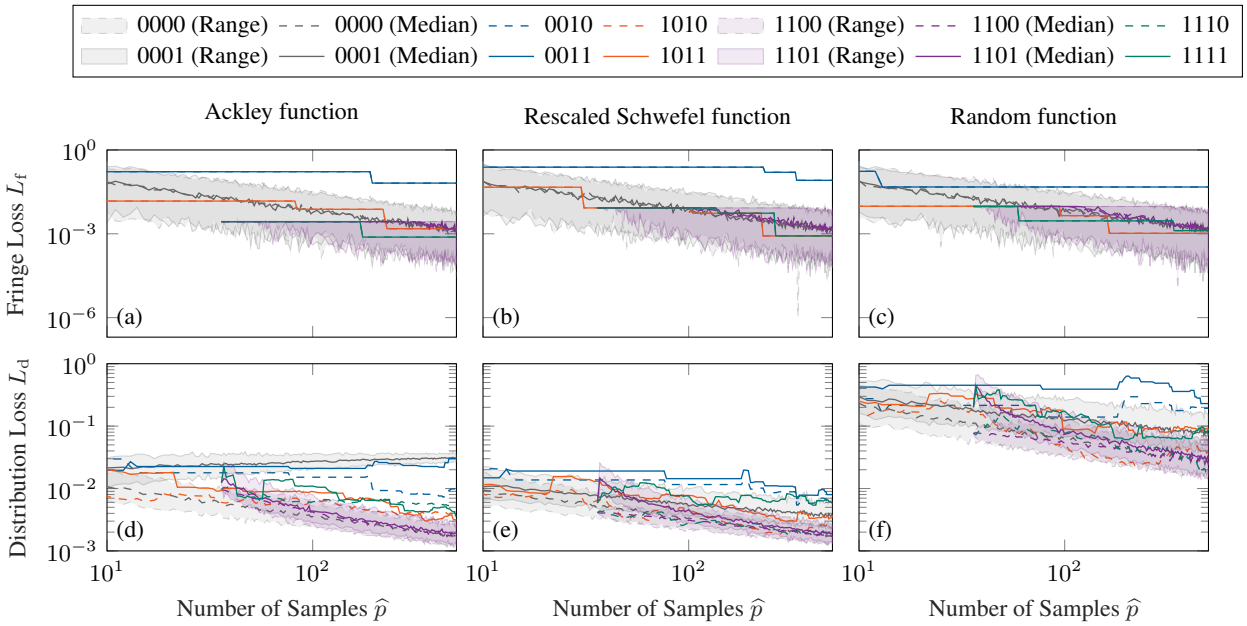



**Figure 4. Sampling representativeness with different number of samples in 500-dimensional usage spaces. The sampling schemes are abbreviated using four digits, each of which represents the activation of each treatment, i. e., usage space analysis, fringe sampling, core sampling, and market volume weighting. Results of schemes with uncertainties (**0*) are represented with medians and their error bands from fifth to 95$^{th}$ percentiles (ranges).**

and more randomized if we represent them in a low-dimensional space. Compared to the former case with 5D usage spaces, distribution losses in this case are generally lower in Ackley and rescaled Schwefel functions. Correspondingly, the distribution losses of random cases are generally similar to that in 5D case, as the data are originally randomized.

Different from the 5D case with core sampling activated, an interesting aspect is observed about the influence of usage space analysis on distribution losses. Schemes 1*1* represent the distribution of outputs from Ackley and rescaled Schwefel functions better than schemes 0*1*, which is opposite to the findings in 5D usage spaces. This indicates that the customers sampled according to Halton sequence (core sampling) is less representative in high-dimensional space.

Another interesting aspect is that schemes with weighting have slightly increased distribution losses, increasing with the number of samples except for the scheme with all treatments activated (1111). As the weights reflect Euclidean distances (see Section 3.4), the weighted ECDFs are less representative in high dimensions due to the well-known "curse of dimensionality" [21].

In summary, in 500D usage spaces, schemes with usage space analysis but without sampling (1**0) show less disadvantage on distribution losses compared to simple random sampling (0000).

### 5.3. Data Quality Management for Practical Applications

The experiments conducted in this work are based on simulated data. As in practical applications the sampling representativeness might be affected by data quality issues, we will now discuss the robustness of the usage space sampling against different data quality issues.

The proposed approach requires acceptable levels of data quality that must be ensured by appropriate data inspection and preprocessing. Specifically, large measurement errors, extreme values as well as large volumes of missing values must be managed prior to applying the method.

However, the method is robust against several types of quality concerns, namely scale errors, noise and smaller amounts of missing values as long as the missing data do not follow different structures than the available data. Different scale magnitudes as well as mean- or stretch-biased measures are handled intrinsically by the initial $z$-score normalization shown in (1). Also, the methods corrects White noise errors, given acceptable signal to noise ratios, due to the application of tSVD-based dimensionality reduction. Finally, in

case of small amounts of missing data, the tSVD-based dimensionality reduction can be approximated using techniques defined on incomplete matrices such as Alternating Least Squares. Several approaches for low-rank matrix approximation with missing data are for instance proposed in (3).

## 6. Conclusion and Future Research

We introduced a sampling method based on customers usage data, which is deterministic and particularly suitable for customer service analytics in case fringe customers are to be identified and considered appropriately.

To evaluate the feasibility of the approach, we conducted experiments with three benchmark functions generating samples in low as well as in high dimensions.

Results show that (i) tSVD-based usage space analysis and convex hull-based fringe sampling can well identify fringe customers when they are near the boundary of their usage space, in which – as expected – it clearly outperforms random sampling; (ii) Halton sequence-based core sampling can enhance the representativeness of the samples in case of high randomness; (iii) at low dimension, Voronoi tessellation-based market volume weighting further reduces the distribution losses.

Promising further research on this topic could be on alternative usage space analysis. For instance, besides tSVD, it would be worthwhile to investigate the influence of histogram binning on fringe identification, or combination of both histogram binning and subsequent tSVD. In addition, appropriate weighting methods suitable in high dimensional spaces must account for the curse of dimensionality in case those are based on distance relationships. Exploring weighting schemes that work in higher dimensions, such as angle-based approaches, is another promising direction of future research to increase the robustness of the sampling technique proposed.

## References


[1] J. Cardoso, H. Fromm, S. Nickel, G. Satzger, R. Studer, and C. Weinhardt, *Fundamentals of Service Systems*. Cham: Springer International Publishing, 2015.

[2] B. Spottke, J. Wulf, and W. Brenner, *Consumer-Centric Information Systems: A Literature Review and Avenues for Further Research*, vol. 36. Fort Worth: 36th International Conference on Information Systems, 2015.

[3] N. Haselgruber, K. Mautner, and J. Thiele, "Usage space analysis for reliability testing," *Quality and Reliability Engineering International*, vol. 26, no. 8, pp. 877–885, 2010.

[4] J. Schoch, P. Staudt, and T. Setzer, "Smart Data Selection and Reduction for Electric Vehicle Service Analytics," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[5] 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.

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

[7] C. Soize and R. Ghanem, "Data-driven probability concentration and sampling on manifold," *Journal of Computational Physics*, vol. 321, pp. 242–258, 2016.

[8] S. Ren, Q. Sun, and Y. Shi, *Customer segmentation of bank based on data warehouse and data mining*. 2nd IEEE International Conference on Information Management and Engineering, 2010.

[9] Y. Kim, "Toward a successful CRM: variable selection, sampling, and ensemble," *Decision Support Systems*, vol. 41, no. 2, pp. 542–553, 2006.

[10] A. Barbu and S.-C. Zhu, *Monte Carlo Methods*. Singapore: Springer Singapore, 1st ed. 2020 ed., 2020.

[11] Kolarova Viktoriya, T. Kuhnimhof, and S. Trommer, eds., *Assessment of real-world vehicle data from electric vehicles – potentials and challenges*, 2017.

[12] M. Arnst, C. Soize, and R. Ghanem, "Hybrid Sampling/Spectral Method for Solving Stochastic Coupled Problems," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 1, no. 1, pp. 218–243, 2013.

[13] Y. Park, M. Cafarella, and B. Mozafari, *Visualization-aware sampling for very large databases: IEEE International Conference on Data Engineering*. Piscataway, NJ: IEEE, 2016.

[14] D. G. Loyola R, M. Pedergnana, and S. Gimeno García, "Smart sampling and incremental function learning for very large high dimensional data," *Neural networks*, vol. 78, pp. 75–87, 2016.

[15] K. P. Murphy, *Machine learning: A probabilistic perspective*. Adaptive computation and machine learning series, Cambridge MA: MIT Press, 2012.

[16] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[17] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software (TOMS)*, vol. 22, no. 4, pp. 469–483, 1996.

[18] L. Kocis and W. J. Whiten, "Computational investigations of low-discrepancy sequences," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 2, pp. 266–294, 1997.

[19] T. Bäck, *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. New York: Oxford Univ. Press, 1996.

[20] G. F. Raggett, "Numerical optimization of computer models," *Optimal Control Applications and Methods*, vol. 3, no. 1, p. 97, 1982.

[21] R. Bellman, *Dynamic Programming*. 1972.