# A novel virtual sample generation method to improve the quality of data and the accuracy of data-driven models

Zhiwen Chen, Zhigang Lv, Ruohai Di, Peng Wang, Xiaoyan Li, Xiaojing Sun, Yuntao Xu

# A novel virtual sample generation method to improve the quality of data and the accuracy of data-driven models

Zhiwen Chen[a], Zhigang Lv[a,b], Ruohai Di[a]*, Peng Wang[a], Xiaoyan Li[a], Xiaojing Sun[c], Yuntao Xu[a]

[a] School of Electronics Information Engineering, Xi'an Technological University, Xi'an 710021, China
[b] School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China.
[c] General Office, Northwest Institute of Mechanical and Electrical Engineering, Xianyang 712099 China
*Corresponding author: Ruohai Di (E-mail address: xfwtdrh@163.com)

**0. Abstract**

Small data volume and data imbalance often lead to statistical failure and seriously restrict the accuracy of data-driven models, which has become a bottleneck problem, needing to be solved, in small sample modeling. The data expansion method has become the main way to solve small sample modeling. However, the randomness in the process of virtual sample generation and combination leads to many invalid data, resulting in poor consistency between the expanded data and the original data. For this reason, this paper proposes a virtual sample generation method based on acceptable area and joint probability distribution sampling (APS-VSG) to limit the randomness in the data expansion method, reduce the proportion of invalid data, improve data consistency after expansion, and improve the accuracy of the data-driven model under the condition of small samples. Firstly, the concept of "compact range of interaction (CRI)" was proposed, which further limits the domain estimation range of data to approximate the valid area of the data. Secondly, the prior knowledge was used to improve mega-trend-diffusion (MTD), and the CRI is delineated according to the trend dispersion to obtain the acceptable area of the virtual data. Finally, a joint probability distribution was constructed based on the true values of small samples in the acceptable area, and data sampling was conducted based on the probability distribution to generate virtual data. The experimental results of standard function datasets show that the virtual samples generated by the proposed method can ensure validity of more than 85%. The experimental results of the NASA li-ion battery dataset show that, compared with Interpolation, Noise, MD-MTD, GAN, and GMM-VSG methods, the error of the data-driven model trained with virtual data generated by the proposed method is significantly reduced. Compared with GAN and GMM-VSG, MSE, RMSE, MAE, and MAPE are reduced by at least 19.3%, 10.6%, 15.4%, and 16.7%, respectively.

**Keywords:** small sample; data-driven model; virtual sample generation (VSG); acceptable area; joint probability distribution sampling

**1. Introduction**

Machine learning is the focus in the field of artificial intelligence and pattern recognition. Various novel machine-learning algorithms emerge in endlessly. In the age of big data, data-driven machine learning algorithms have gradually become research focus, which are widely used to solve complex problems in scientific fields and engineering applications. However, there are still many fields, which cannot obtain a large amount of data due to factors, such as high experiment cost and long test cycle. Therefore, it is difficult to apply advanced deep learning algorithm to solve problems [1], such as voice print recognition [2] in multimedia field, disease diagnosis [3][4]

and water analysis [5] in biological and medical field, product sales prediction [6] in the economic field and life prediction of fuel cells [7] in the industrial and military fields. In the process of using machine learning to deal with the above problems, there are problems such as small data volume and data imbalance, etc. All the above belong to small sample problems, because the sample size of the target object is too small to train a model that meets the accuracy requirements [8, 9].

For the small sample problem, scholars have carried out data-centric research. Data expansion is the main method used in data-centric research. It is the most direct method to deal with small sample problems by constructing virtual samples to increase sample size and balance data sets. The main ideas for constructing virtual samples are knowledge-based, disturbance-based, and distribution-based [10].

The knowledge-based idea is mainly to artificially generate virtual samples based on expert knowledge in the research field [11]. Xu [12] found, in the engineering problems of the double-cylindrical shell structure, that the signals of the double-cylindrical shell structure collected by sensors are often mixed with the noise from the exciter and the seawater pump. According to the frequency characteristics of these noises, he proposed a method, based on the frequency response function and Fourier transform, to highly simulate the sampling signals of the double-cylindrical shell structure with the noise signals of the exciter and the seawater pump. These simulated virtual signal used to train the model can effectively improve the noise recognition rate of the model for the double cylindrical shell structure. The threshold for creating virtual samples based on knowledge is very high, which requires rich expert knowledge and experience, and is difficult for nonprofessional personnel to use. In various methods of constructing virtual samples based on the idea of disturbance, Chris M. Bishop [13] found that a model with higher generalization performance was obtained by adding a certain amount of noise to the input data and inputting it into the neural network for training. Guozhong An [14] further confirmed that adding noise to input samples can effectively improve the generalization performance of classification and regression problems. Wang [15] added disturbance based on training samples to obtain new virtual samples and used these virtual samples to make the model have a better recognition rate. The idea based on distribution is mainly to determine the range and probability of virtual sample generation according to the domain distribution of small sample data [16-26]. Der Chiang Li [16] proposed mega-trend-diffusion (MTD), which uses a common diffusion function to spread a group of data and determine the possible coverage of the data set based on group consideration to generate reasonable virtual data. Chongfu Huang [17] and Der Chiang Li [18] [19] used information diffusion, MTD, and other methods to turn a clear value sample point into a fuzzy set, turning a small number of single value sample points into a large number of set-valued sample points. This method can make the generated virtual samples have more information. Zhu Bao [20] [21] proposed multi-distribution mega-trend-diffusion (MD-MTD) to generate virtual samples, in which uniform distribution and triangular distribution are added to describe the characteristics of small sample data. Compared with MTD, the virtual samples

generated by MD-MTD are more realistic. Ling Li [23] proposed a Gaussian mixture model-based virtual sample generation (GMM-VSG) method to generate virtual samples under multiple working conditions. Der-Chiang Li [24] proposed a genetic algorithm-based virtual sample generation (GABVSG), which generated more valid virtual samples by considering the overall integrated effects of the attributes. Qun Xiong Zhu [25] proposed a VSG method based on singular value decomposition (SVD) feature decomposition and gradient boosting decision tree (GBDT) prediction model. Yan Lin He [26] generated virtual samples by using the proposed t-SNE-based virtual sample generation (t-SNE-VSG). SVD-VSG and t-SNE-VSG both use the distribution of the original data as a reference to generate virtual samples, which is difficult to use for small sample data with unknown distribution. In addition, Der Chiang Li [27] proposed a nonlinear virtual sample generation technology based on a hypersphere parameter equation in combination with group discovery technology, which can effectively improve the learning accuracy of the model. He also used the interval kernel density estimator to generate more similar virtual samples, overcoming the problem of learning difficulties when data is insufficient [28]. Chen Zhongsheng [29] proposed a new virtual sample generation method QRCGAN, which embedded quantile regression into a conditional generation countermeasure network. Yan LinHe [30] proposed a novel virtual sample generation method embedding a deep neural network as a regression into conditional Wasserstein generative adversarial networks with gradient penalty (rCWGAN). Embedding networks can improve the validity of virtual samples to some extent, but it takes a long time and is expensive to calculate.

Comparing virtual data generated by different methods, knowledge-based virtual samples are limited by the accuracy and diversity of expert knowledge, obtained in the research field. Manual sample production is expensive and time-consuming, which is difficult to apply to most scenarios. Disturbance-based virtual samples mainly expand the boundary area of small samples by adding disturbances to improve the generalization of the model. However, there is no unified answer to how much disturbance is appropriate for different problems, which needs to be determined through a large number of experiments. For this reason, we proposed an improved acceptable area estimation method in this paper. This method delimited a more compact range of interaction (CRI) of multi variable within the theoretical upper limit and lower limit of each variable which is called the wide range of permissible (WRP) to ensure that the generated multidimensional data is acceptable in this area. This method can avoid the problem that the generated virtual samples deviate from the real samples too much. The basis for creating virtual samples based on distribution is statistics, but there is still doubt whether the conclusions obtained by using statistics for small samples that do not meet the large sample theorem can express the properties of large samples. For this reason, this paper reduced statistical methods for estimating data distribution through small samples and artificially constructed a probability distribution function with universality to describe the distribution of data according to small samples in the acceptable area, from which the problem that the sample size is too small to accurately estimate the sample

distribution was solved.

A large number of scholars have found that in the process of generating multidimensional virtual data, the method of first generating single-dimensional virtual data and then reconstructing high-dimensional data often makes combination errors, resulting in a large number of invalid virtual samples [31]. This part of invalid virtual samples will greatly reduce the accuracy of the model and it is difficult to build an appropriate screening mechanism to screen out invalid data. For this reason, this paper analyzes the probability distribution between multiple variables and target variables and then determines the joint distribution between multiple variables by target variables to solve the problem of data combination. The long-standing problem with virtual samples is that it is difficult to consider both validity and robustness. In this paper, edge virtual samples were added during sampling to explore the edge information inconsistent with small samples to improve the robustness of virtual samples in the data-driven model.

The main contributions of this paper are shown as the following:

(1) A virtual sample generation method, based on CRI and joint probability distribution sampling, was proposed in this paper. This method can expand the data for small samples with unknown data characteristics, limit the randomness in the process of virtual sample generation, and greatly improve the validity of virtual samples;

(2) An estimation method of acceptable area was proposed, which uses prior knowledge to improve MTD and estimates CRI according to trend dispersion to comprehensively estimate the generation space of virtual samples, to provide a guarantee for generating high-quality virtual samples;

(3) The conditional distribution of each variable under specific conditions was constructed to ensure a high sampling probability near the real sample and increase the overall sampling probability in the acceptable area to include part of the edge virtual samples to improve the robustness;

(4) Joint probability sampling was used to generate high-dimensional data as a whole to avoid the problem of data combination.

The rest organization of this paper is shown: In section 2, the basic theory of this paper is introduced. In section 3, the proposed method is described in detail. In section 4, the experimental verification is finished. In section 5, the conclusions are drawn.

## 2. Basic theory

In this section, the idea source and basic theory of the method in this paper is briefly introduced, including MTD, confidence interval, and sampling methods.

### 2.1 Mega-trend-diffusion

MTD [16] was proposed by Der Chiang Li to determine the data coverage, which uses a common diffusion function to spread a group of data and determine the possible coverage of the data set based on group consideration. Zhu added uniform distribution and triangular distribution to describe the characteristics of small samples. The new method was called MD-MTD. The core

idea of MTD and MD-MTD is to determine the floating range of data according to the number and distribution density of data. Fig. 1 shows that the extrapolation boundary of virtual samples was estimated based on small samples under the triangular diffusion function.

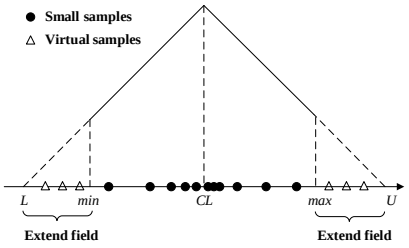


Fig. 1. Diagram of MTD

For a given small sample set $X=\{x_1,x_2,...,x_n\}$, the extrapolation formulas of MTD are as follows:

$$L=CL-\frac{n_L}{n_L+n_U}\times\sqrt{-2\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n_L(n-1)}\times\ln(10^{-20})} \tag{1}$$

$$U=CL+\frac{n_U}{n_L+n_U}\times\sqrt{-2\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n_U(n-1)}\times\ln(10^{-20})} \tag{2}$$

$$CL=\frac{\min+\max}{2} \tag{3}$$

where, $\min$ and $\max$ are the minimum and maximum values in a small sample set, $n_L$ and $n_U$ are the number of samples in $[\min,CL]$ and $[CL,\max]$ in a small sample set.

It is meaningful to use the MTD to estimate the acceptable area of data, and also effective to generate one-dimensional virtual data according to the extension field. However, it will cause a combination error if aligning multi-dimensional data after generating each one-dimensional virtual data. Especially, when there is a large distribution difference between each one-dimensional data, most of the virtual data generated is even invalid. For this reason, this paper optimizes MTD based on prior knowledge, proposes CRI to limit the excessive acceptable area, and uses joint probability sampling to avoid combination error.

**2.2 Confidence interval**

Confidence interval is a common interval estimation method. It is an estimation interval of population parameters constructed by sample statistics, with upper and lower confidence limits of statistics as upper and lower bounds respectively. It refers to the range where the true value appears with the measured value as the center under a certain confidence, which is the probability that the true value will occur within a certain range. The real data are often not known, and we can only estimate. For example, in $[a,b]$, the probability that the real value appears between $a$ and $b$ is 95% (there is also a 5% probability that it appears outside this interval). The smaller the interval is, the lower the confidence is, and the higher the accuracy is. The 95% confidence level is general. [32] [33]

In regression analysis, confidence intervals are usually used to judge the reliability of data and increase the robustness of the analysis. This paper uses this idea for reference and sets a certain confidence level to expand the range of small samples according to the data distribution statistical chart of small samples, which is regarded as one of the criteria for defining the acceptable area.

### 2.3 Sampling

#### 2.3.1 Stratified sampling

Hierarchical sampling is a method of randomly selecting samples from different layers according to the specified proportion of a sample population that can be divided into multiple layers. The advantage of this method is that the sample is representative and the sampling error is small. For data $X$, if $n$ samples were obtained by stratified sampling, it is divided into $k$ layers according to certain rules, then $n/k$ samples are sampled from each layer, and finally, all samples are integrated.

#### 2.3.2 Acceptance-Rejection Sampling

Acceptance-Rejection Sampling (ARS) is a basic technique used to generate observations from distributions, and it is a Monte Carlo method. The specific operations of ARS are as follows: Set a constant $k$ and a function $q(x)$ for convenient sampling which makes the known distribution $p(x)$ always below $kq(x)$. The symbol of $z_0$ is sampled from $q(x)$ and $u_0$ is sampled from a uniform distribution $(0, kq(z_0))$. If $u_0$ falls above $p(x)$, reject this sampling, otherwise accept this sampling. Repeat the above process to obtain $\{x_1, x_2, ..., x_n\}$ close to distribution $p(x)$.

In the case of high-dimensional, there are two problems in ARS. First, it is difficult to obtain the general form of multidimensional distribution when there is only conditional distribution $p(x_1 \mid y)$, $p(x_2 \mid y)$, … , $p(x_n \mid y)$. The second is that it is difficult to find a suitable distribution $q(x)$.

## 3. Proposed method

In this section, the methods proposed in this paper will be described in detail. The object of this algorithm is $X_{(m\times n)}$ and $Y_{(1\times n)}$. When the target object is $Z_{((m+1)\times n)}$, try to divide a vector with a strong correlation with other vectors by correlation analysis, and take this vector as $Y_{(1\times n)}$, the remaining m-dimensional vector is taken as $X_{(m\times n)}$. In this paper, we first use prior knowledge and trend dispersion to delineate a more reasonable acceptable area where generating virtual samples can greatly improve the validity of virtual samples. Then, two-stage sampling is conducted on the joint distribution constructed in this paper to obtain virtual samples, to ensure that a large amount of valid virtual samples is generated, and a certain number of edge virtual samples can also be generated to improve the robustness.

$$Z=\begin{bmatrix} Z1 \\ Z2 \\ \vdots \\ Z(m+1) \end{bmatrix}=\begin{bmatrix} Z1_1, Z1_2, \cdots, Z1_n \\ Z2_1, Z2_2, \cdots, Z2_n \\ \vdots \\ Z(m+1)_1, Z(m+1)_2, \cdots, Z(m+1)_n \end{bmatrix}$$

$$partition \Rightarrow \begin{cases} X=\begin{bmatrix} X1 \\ X2 \\ \vdots \\ Xm \end{bmatrix}=\begin{bmatrix} X1_1, X1_2, \cdots, X1_n \\ X2_1, X2_2, \cdots, X2_n \\ \vdots \\ Xm_1, Xm_2, \cdots, Xm_n \end{bmatrix} \\ Y=[Y_1, Y_2, \cdots, Y_n] \end{cases} \tag{4}$$

### 3.1 Algorithm framework

Different from MTD, the proposed method estimates not only the domain range of small samples but also the compact range of interaction (CRI) between multiple variables to limit the acceptable area of data. To avoid the problem of data combination and generate high-dimensional data as a whole, the proposed method constructs the conditional distribution of each variable and multidimensional probability distribution according to the true value of small sample, which provides technical support for the generation of virtual data. The algorithm framework is shown in Fig. 2.

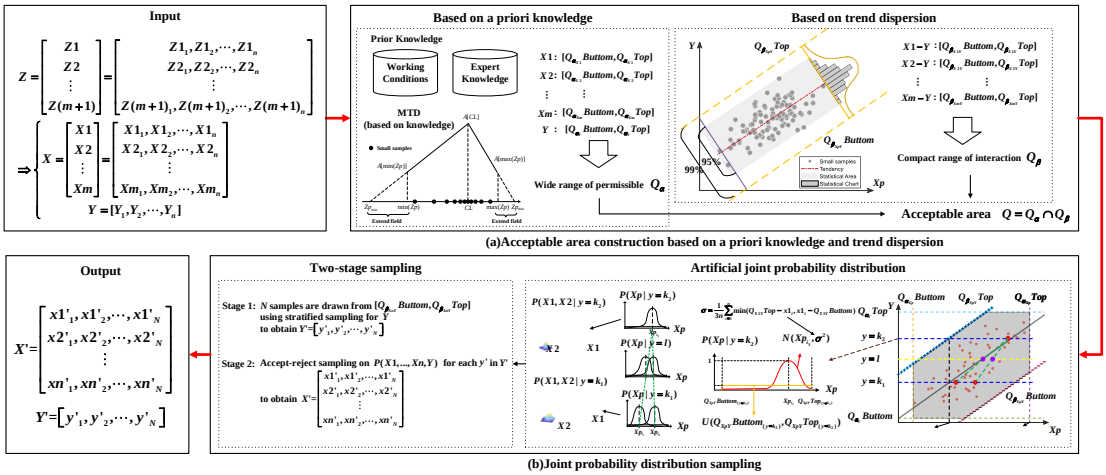


Fig. 2. The algorithm framework

### 3.2 Acceptable area construction based on a priori knowledge and trend dispersion

All virtual sample generation methods need to estimate the distribution range of data before generating virtual data. In this paper, this range is referred to as the acceptable area. This paper proposed an estimation method of acceptable area. Firstly, the upper and lower limits of each variable are determined according to prior knowledge, and a large acceptable area of data is determined according to the upper and lower limits of each variable. This large acceptable area of data is referred to as the wide range of permissible (WRP). Then, determine the trend of each $X$ and $Y$, and fit the distribution curve according to the distribution histogram of small samples in the data trend direction. Take the 99% or 100% confidence interval of the distribution curve as CRI. Finally, final the acceptable area will be obtained by using CRI to limit this large acceptable

area. Therefore, an acceptable area $Q$ is the intersection of WRP $Q_\alpha$ and CRI $Q_\beta$.

$$Q = Q_\alpha \cap Q_\beta \tag{5}$$

### 3.2.1 Estimation method of WRP

In practical application, the measured values collected have practical meanings. Generally, these measured values have critical values, which are objective limitations and are usually determined by physical factors such as material and shape. Especially for some products produced by assembly lines, each parameter has a rated working range and a maximum fluctuation range. In this paper, the interval bounded by the theoretical upper limit and lower limit is called the wide range of permissible (WRP), which is the large acceptable area mentioned above. When the theoretical upper limit $Zp_{\max}$ and lower limit $Zp_{\min}$ of data is obtained through prior knowledge, WRP $Q_{\alpha_{Zp}}$ can be directly obtained.

$$Q_{\alpha_{Zp}} = [Zp_{\min}, Zp_{\max}] \tag{6}$$

where $Zp_{\min} \le \min(Zp) < \max(Zp) \le Zp_{\max}$.

When one of the theoretical upper and lower limits of data can be obtained through prior knowledge, this paper improves the MD-MTD algorithm to estimate WRP by incorporating prior knowledge.

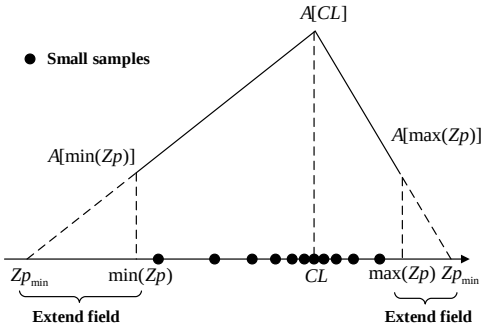


Fig. 3. The estimation method for WRP

As shown in Fig. 3, is a diagram of the estimation method for WRP, where,

$$CL = \begin{cases} Z_{(n+1/2)} & if \quad n \quad is \quad odd \\ \dfrac{1}{2}(Z_{n/2} + Z_{n/2+1}) & if \quad n \quad is \quad even \end{cases} \tag{7}$$

$$A[CL] = 1 \tag{8}$$

For the case that only the theoretical lower limit $Zp_{\min}$ of $Z_p$ is known, but the theoretical upper limit $Zp_{\max}$ is not known, calculate $A[\min(Zp)]$ according to $Zp_{\min}$.

$$A[\min(Zp)] = A[CL] \cdot \frac{\min(Zp) - Zp_{\min}}{CL - Zp_{\min}} \tag{9}$$

$$A[\max(Zp)] = 1 - A[\min(Zp)] \tag{10}$$

According to $A[\max(Zp)] = 1 - A[\min(Zp)]$ and the principle of an equal proportion distribution:

$$\frac{Zp_{\max} - \max(Zp)}{Zp_{\max} - CL} = \frac{A[\max(Zp)]}{A[CL]} \tag{11}$$

Calculate $Zp_{\max}$ by the following ormula.

$$
\begin{aligned}
Zp_{\max} &= \frac{A[CL]\cdot\max(Zp)-A[\max(Zp)]\cdot CL}{A[CL]-A[\max(Zp)]} \\
&= \frac{A[CL]\cdot\max(Zp)-(1-A[\min(Zp)])\cdot CL}{A[\min(Zp)]} \\
&= \frac{A[CL]\cdot\max(Zp)-(1-A[CL]\cdot\frac{\min(Zp)-Zp_{\min}}{CL-Zp_{\min}})\cdot CL}{A[CL]\cdot\frac{\min(Zp)-Zp_{\min}}{CL-Zp_{\min}}} \\
&= \frac{-CL^2+[\max(Zp)+\min(Zp)]\cdot CL-\max(Zp)\cdot Zp_{\min}}{\min(Zp)-Zp_{\min}}
\end{aligned}
\tag{12}
$$

Similarly, when only the theoretical upper limit $Zp_{\max}$ of $Z_p$ is known but the theoretical lower limit $Zp_{\min}$ is not known.

$$
\begin{aligned}
Zp_{\min} &= \frac{A[CL]\cdot\min(Zp)-A[\min(Zp)]\cdot CL}{A[CL]-A[\min(Zp)]} \\
&= \frac{CL^2-[\max(Zp)+\min(Zp)]\cdot CL+\min(Zp)\cdot Zp_{\max}}{Zp_{\max}-\max(Zp)}
\end{aligned}
\tag{13}
$$

**3.2.2 Estimation method of CRI**

It is useful that MTD generates virtual samples $X'_{(1\times N)}$ on the estimated acceptable area of $X_{(1\times n)}$ and uses the hyperplane to give the corresponding $Y'_{(1\times N)}$. When the dimension of $X$ increases, many of $X_{(m\times n)}$ obtained by data combinations after the expansion of $X1_{(1\times n)}, X2_{(1\times n)}, \cdots, Xm_{(1\times n)}$ will deviate from the real samples. The reason is that there may be some correlation factor between $X1_{(1\times n)}, X2_{(1\times n)}, \cdots, Xm_{(1\times n)}$, which makes $X1_{(1\times n)}$ change with the change of $X2_{(1\times n)}$. This change in coupling can greatly narrow the acceptable area of multivariate data. Therefore, this paper further reduces the acceptable area by estimating the area of such influence. This paper will refer to this area as a compact range of interaction (CRI).

As is known to all, in the feature selection of the neural network model, input features with a strong correlation with output features will be selected. Based on this, this paper analyzes the relationship between each input feature $X_{(1\times n)}$ and output feature $Y_{(1\times n)}$, and uses this relationship to construct the CRI of $X_{(1\times n)}-Y_{(1\times n)}$, to clarify the $X1_{(1\times n)}, X2_{(1\times n)}, \cdots, Xm_{(1\times n)}$ range corresponding to $Y_{(1\times n)}$, to limit the acceptable area of the input feature and prevent the phenomenon of combination error.

Firstly, $X1_{(1\times n)}, X2_{(1\times n)}, \cdots, Xm_{(1\times n)}$ is extracted from $X$ in order, and takes vector $Xp_{(1\times n)}(p\in[1,m]\text{且}p\in Z^*)$ and vector $Y_{(1\times n)}$ to form $XpY_{(n\times 2)}$. You take the rows of $XpY_{(n\times 2)}$ minus $\overline{XpY}_{(1\times 2)}$ which is the mean of the columns of $XpY_{(n\times 2)}$ to get $\varepsilon_{XpY(n\times 2)}$.

$$
\varepsilon_{XpY(n\times 2)} = XpY_{(n\times 2)} - \overline{XpY}_{(1\times 2)} \tag{14}
$$

Then, calculate the positive operator $C_{XpY(2\times 2)}$ and one of its largest eigenvectors $V_{XpY(2\times 1)}$.

$$
C_{XpY(2\times 2)} = \varepsilon^T_{XpY(2\times n)} \times \varepsilon_{XpY(n\times 2)} / n \tag{15}
$$

Finally, calculate the maximum $\max(\varepsilon V_{XpY(2\times 1)})$ and minimum $\min(\varepsilon V_{XpY(2\times 1)})$ of $\varepsilon_{XpY(XpY)}\times V_{XpY(2\times 1)}$, then $O_{XpY(2\times 2)}$ is:

$$O_{XpY(2\times2)} = \overline{XpY}^{T}_{(2\times1)} + [\max(\varepsilon V_{XpY(2\times1)}), \min(\varepsilon V_{XpY(2\times1)})]_{(1\times2)} * V_{XpY(2\times1)} \times \varsigma \quad (16)$$

where $\varsigma$ is 1.05.

In the direction of $O_{XpY(2\times2)}$, the distribution curve is fitted according to the distribution histogram of small samples, and the 99% or 100% confidence interval of the distribution curve is taken as the CRI $Q_{\beta_{XpY}} = [Q_{\beta_{XpY}} Buttom, Q_{\beta_{XpY}} Top]$.

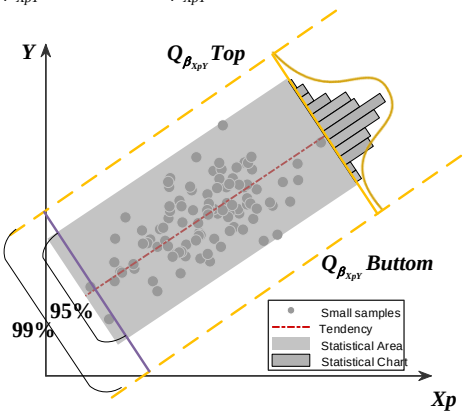


Fig. 4. The estimation method for CRI

### 3.3 Joint probability distribution sampling

Tested by the author, it will cause a combination error if aligning multi-dimensional data after generating each one-dimensional virtual data. Especially when there is a large distribution difference between each one-dimensional data, most of the virtual data generated is even invalid. To solve this problem, this paper constructed the artificial multi-variable joint distribution based on the acceptable area, to simultaneously generate $X1, X2, \cdots, Xm$ as a whole to avoid combination error.

#### 3.3.1 Artificial joint probability distribution

In small-sample modeling, virtual samples combined with small sample sets are used for model training. The basic reason why the accuracy is not improved or decreased is the low quality of the data. To improve the data quality, it is necessary to start from the data validity and generate virtual samples conforming to the characteristics of small samples. Secondly, to improve the fault tolerance ability of the model, it is necessary to generate a small amount of differentiated virtual samples to improve the robustness.

To make the generated virtual data conform to the characteristics of small samples, the conditional probability distribution $P(Xp \mid y = k)$ of $Xp(p \in [1, m] 且 p \in Z^{*})$ is constructed for each value $k$ in $Y$ in the acceptable area $Q_{XpY}$, and then the conditional probability distribution $P(Xp \mid l)$ corresponding to the value $l$ in the middle of adjacent $k$ values is calculated. Finally, $P(X1X2\cdots Xm \mid y)$ is calculated according to the conditional probability distribution $P(Xp \mid y)$ of $Y$ and $X1, X2, \cdots, Xm$, and the joint distribution $P(X \mid Y)$ is obtained by integration. Fig. 5 shows the diagram of the process of constructing a joint probability distribution.

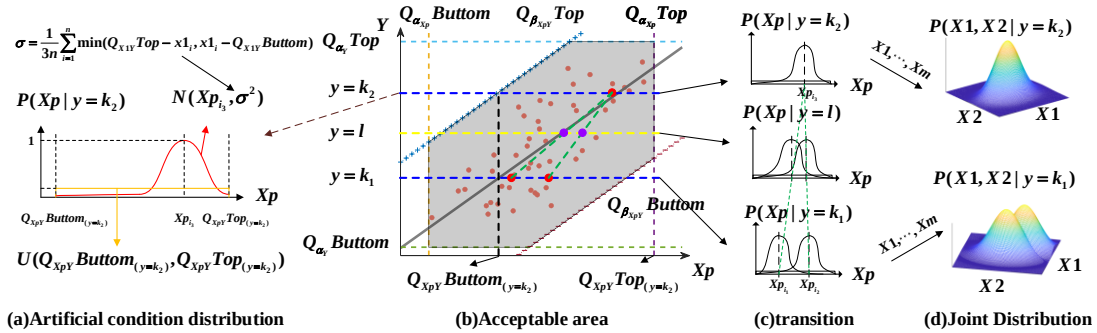


Fig. 5. The process of constructing a joint probability distribution

where $Xp(p\in[1,m]\text{且}p\in Z^*)$ corresponds to the conditional probability distribution of each value of $k$, $P(Xp\,|\,k)$ is constructed as follows:

Firstly, determine whether there is the same $y$ in $Y$, denoting the number of different values $k$ in $Y$ as $n_k$, $n_k\le n$, then

$$k_j\in[y_1,y_2,\cdots,y_n]\quad and\quad 1\le j\le n_k \tag{17}$$

Assume that sample $i$ is $X_i=[x1_i,x2_i,\cdots,xm_i]$, $Y=[y_i]$, $y_i=k_j$, and construct the probability distribution of $x1_i,x2_i,\cdots,xm_i$ when $y_i=k_j$ respectively. Take $(x1_i,y_i)$ as an example, first determine the acceptable range $[Q_{X1Y}Buttom,Q_{X1Y}Top]$ of $X1$ when $y_i=k_j$, then calculate the standard deviation $\sigma_i$ of the normal distribution to be constructed.

$$\sigma=\frac{1}{3n}\sum_{i=1}^{n}\min(|\,Q_{X1Y}Top-x1_i\,|,|\,x1_i-Q_{X1Y}Buttom\,|) \tag{18}$$

The normal distribution constructed according to sample truth value $(x1_i,y_i)$ and acceptable area boundary $[Q_{X1Y}Buttom,Q_{X1Y}Top]$ is $N(x1_i,\sigma^2)$, When the same $k$ value corresponds to multiple sample truth values, a normal distribution $N(x1_{i2(y=k_j)},\sigma^2)$ is constructed for each sample truth value. To include part of the edge virtual samples to improve the robustness, a uniform distribution is superimposed based on normal distribution. So when $y=k_j$,

$$X1\sim N(x1_i,\sigma^2)\cup N(x1_{i2},\sigma^2)\cup\cdots\cup U(Q_{X1Y}Buttom,Q_{X1Y}Top) \tag{19}$$

From this, the conditional probability distribution $P(X1\,|\,y=k_j)$ is obtained, and by analogy, $P(X2\,|\,y=k_j)$, $P(Xm\,|\,y=k_j)$ is obtained. Similarly, the conditional probability distribution $P(Xp\,|\,k)$ of $Xp(p\in[1,m]\text{且}p\in Z^*)$, corresponding to different $k$ values, can be obtained.

The discrete probability distribution is difficult to describe the probability distribution of continuous variables. Therefore, the conditional probability distribution $P(Xp\,|\,l)$ corresponding to the value $l$ in the middle vacancy of adjacent $k$ values should be calculated, where $l\notin Y$ and $l\in[Q_{\alpha_Y}Buttom,Q_{\alpha_Y}Top]$. The more number $l$ is calculated, the denser the distribution, and the smoother the transition.

Then, the joint probability distribution $P(X\,|\,y)$ is constructed.

$$\begin{aligned}&P(X\,|\,y=k(l))=P(X1X2\cdots Xm\,|\,y=k(l))\\&=P(X1\,|\,y=k(l))\cdot P(X2\,|\,y=k(l))\cdot\cdots\cdot P(Xm\,|\,y=k(l))\end{aligned} \tag{20}$$

Finally, the joint probability distribution $P(X\,|\,y)$ of all values in $[Q_YButtom,Q_YTop]$ is integrated to get the final joint probability distribution $P(X\,|\,Y)$.

### 3.3.2 Two-stage sampling

Since all the methods used above construct the joint probability distribution of $X1, X2, \cdots, Xm$ on the premise of determining the value of $y$. Therefore, it is necessary to take $N$ samples from $[Q_Y Buttom, Q_Y Top]$ to get $Y'_{(1\times N)}$, and then sample $X1', X2', \cdots, Xm'$ from the constructed artificial joint probability distribution according to each value in $Y'_{(1\times N)}$, and finally get the virtual sample set $D_{(m\times N)}$.

Stage 1: Latin hypercube sampling from $[Q_Y Buttom, Q_Y Top]$

Divide $[Q_Y Buttom, Q_Y Top]$ into $N/10$ intervals, and sampling was conducted based on the uniform distribution in each interval. The sampling results are integrated into $Y'_{(1\times N)}$.

$$Y' = [y'_1, y'_2, \cdots, y'_N] \tag{21}$$

Stage 2: ARS according to each $y$ in $Y'_{(1\times N)}$ based on the joint probability distribution constructed in the previous section

Step 1: Pick $y'$ from $Y'_{(1\times N)}$ and find the corresponding $P(X \mid y = y')$;

Step 2: $Xp'_{(m\times 1)}$ is randomly generated based on uniform distribution on $(Q_{X(y=y')}, P)$;

Step 3: Judge whether the point is in $P(X \mid y = y')$. If so, accept it; if not, reject it, and return to Step 2.

When every $Xp'_{(m\times 1)}$, corresponding to $y$, is obtained, $X'$ can be integrated.

$$X' = \begin{bmatrix} x1'_1, x1'_2, \cdots, x1'_N \\ x2'_1, x2'_2, \cdots, x2'_N \\ \vdots \\ xm'_1, xm'_2, \cdots, xm'_N \end{bmatrix}_{(m\times N)} \tag{22}$$

Finally, a virtual sample set $D$ is obtained.

$$D = \left\{ \begin{array}{l} X' = \begin{bmatrix} x1'_1, x1'_2, \cdots, x1'_N \\ x2'_1, x2'_2, \cdots, x2'_N \\ \vdots \\ xm'_1, xm'_2, \cdots, xm'_N \end{bmatrix}_{(m\times N)} \\ Y' = [y'_1, y'_2, \cdots, y'_N]_{(1\times N)} \end{array} \right\} \tag{23}$$

The process of ARS is equivalent to using the CRI and distribution of real samples as a filter to remove invalid samples from the virtual samples obtained by the WRP.

### 3.4 Algorithm description

In summary, the structure and details of the algorithm in this paper are expanded, and then the pseudo-code of the algorithm is given.

**Algorithm**: APS-VSG

**Input**: Small samples $X_{(m\times n)}, Y_{(1\times n)}$

**Output**: Virtual samples: $X'_{(m\times N)}, Y'_{(1\times N)}$

1: **for** each $Xp_{(1\times n)}$ of $X_{(m\times n)}$ and $Y_{(1\times n)}$ **do**

2: Set $Q_{\alpha_{Zp}}$ according to a priori knowledge

3: **end for**

4: **for** each $Xp_{(1\times n)}$ of $X_{(m\times n)}$ **do**

5: Calculate $O_{XpY(2\times 2)}$ for $XpY_{(2\times n)}=[Xp_{(1\times n)};Y_{(1\times n)}]$

6: Set $Q_{\beta_{XpY}}$ according to $O_{XpY(2\times 2)}$ and confidence interval

7: **end for**

8: $Q=Q_{\alpha}\cap Q_{\beta}$

9: **for** each $XpY_{(2\times n)}$ **do**

10: Set $P(Xp|Y)$ according to $XpY_{i(2\times 1)}$ and $Q_{XpY}$

11: **end for**

12: $P(X|Y)=P(X1|Y)\cdot P(X2|Y)\cdots\cdots P(Xm|Y)$

13: Get $Y'=[y'_1,y'_2,\cdots,y'_N]_{(1\times N)}$ by Latin Hypercube Sampling on $Q_Y$

14: **for** each $y'_j$ of $Y'_{(1\times N)}$ **do**

15: Get $[x1'_j;x2'_j;\cdots;xm'_j]_{(m\times 1)}$ by Sampling on $Q_X$ according to $P(X|Y)$

16: **end for**

17: $$X'=\begin{bmatrix} x1'_1,x1'_2,\cdots,x1'_N \\ x2'_1,x2'_2,\cdots,x2'_N \\ \vdots \\ xm'_1,xm'_2,\cdots,xm'_N \end{bmatrix}_{(m\times N)}$$

18: Get a virtual sample set $D=\{X'_{(m\times N)},Y'_{(m\times N)}\}$

## 4. Experiments

The method presented in this paper was tested on three standard functions and NASA lithium battery data sets to analyze the validity of the acceptable area and the virtual samples. Compared with five other methods, including Interpolation, Noise, MTD, GAN and GMM-VSG, the advantages and disadvantages of the proposed method are discussed in the experimental analysis.

### 4.1 Datasets

#### 4.1.1 Standard function datasets

The three artificial standard function datasets used in the experiment have different data characteristics. The first standard function is a classical linear model, called Linear. The second standard function is a classical nonlinear model called Nonlinear. The third standard function is the model with sinusoidal oscillation, called Oscillation. The three models have three inputs and one output and contain a certain amount of noise to simulate the data collected in the real environment. The applicability of the proposed method to various types of data can be explored and the factors that affect the performance of the method can be analyzed by testing on different standard function datasets. Evenly take 2000 samples from the standard function as the real sample set, and take 50 samples from the real sample set as the small sample set.

Table 1 Definition of standard function

| Dataset | Standard function | Define interval | | | |
|---|---|---|---|---|---|
| | | $x_1$ | $x_2$ | $x_3$ | $y$ |

| | | | | | |
|---|---|---|---|---|---|
| Linear | $y=\frac{x_1+x_2+x_3}{3}+\varepsilon$ | [1,5] | [1,10] | [1,20] | [0.5,12.5] |
| Nonlinear | $y=\frac{1}{3}x_1^2+\frac{1}{6}x_2^3-\frac{2}{3}x_3+\varepsilon$ | [1,5] | [1,10] | [1,20] | [-2,160] |
| Oscillation | $y=\frac{1}{3}\sin(\pi x_1)+\frac{2}{3}\sin(\pi x_2)+\frac{1}{3}x_3+\varepsilon$ | [1.7,4.3] | [1.7,9.3] | [1.7,19.3] | [0,7] |

**4.1.2 NASA li-ion battery dataset**

The ultimate purpose of generating virtual samples is to use the right amount of data to train a data-driven model that meets the target performance. Therefore, to verify the validity of the method proposed in this paper, from the performance degradation data of NASA li-ion battery [34], data of two batteries of the same model were selected to build a real sample set with five inputs and one output, and take 34 samples from the real sample set as the small sample set.

Table 2 NASA Li-ion Battery Dataset (Real sample set)

| Data source | Feature | Meaning | Length | Remarks |
|---|---|---|---|---|
| B0005<br>B0007 | F1 | Time for the voltage to rise to 4.2V during the constant current charging process | 326 | the theoretical upper limit of the capacity of the battery is 2Ahr |
| | F2 | Time for the current to drop to 20mA during the constant voltage charging process | | |
| | F3 | Time for the voltage to drop to 2.7V during the constant current discharge process | | |
| | F4 | Total time spent in the discharge process | | |
| | F5 | The average temperature during the discharge process | | |
| | Capacity | Capacity of battery | | |

**4.2 Evaluation**

**4.2.1 Evaluation of Acceptable area**

Compare the degree of overlap between the acceptable area and the valid area. If the acceptable area overlaps highly with the valid area, the valid coverage level is high. If the acceptable area does not completely cover the valid area and contains lots of invalid areas, the valid coverage level is low. $\eta_{ca}$ , $\eta_{cv}$ , and $\eta_v$ are used to evaluate the acceptable area. $\eta_{ca}$ is the ratio of the valid area covered by the acceptable area to the acceptable area $\eta_{cv}$ is the ratio of the valid area covered by the acceptable area to the total valid area $\eta_v$ is the ratio of the valid area covered by the acceptable area to the total space of the valid area and acceptable area. The expressions are as follows:

$$\eta_{ca}=\frac{V_{va}}{V_a}\times 100\% \tag{24}$$

$$\eta_{cv}=\frac{V_{va}}{V_v}\times 100\% \tag{25}$$

$$\eta_v = \frac{V_{va}}{V_a + V_v - V_{va}} \times 100\% \tag{26}$$

Where, $V_{va}$ is the size of the valid area within the acceptable area, $V_a$ is the size of the acceptable area, $V_v$ is the size of the valid area.

**4.2.2 Direct evaluation of virtual sample**

It is judged whether the generated virtual samples are in the valid area, if yes, they are considered as valid virtual samples, if not, they are considered as invalid virtual samples. Count the number of valid virtual samples and the number of invalid virtual samples. The ratio of the number of valid virtual samples to the total number of virtual samples, $\gamma_v$, is used to evaluate the validity of the virtual samples. The expression of $\gamma_v$ is as follows:

$$\gamma_v = \frac{n_v}{n_v + n_i} \tag{27}$$

Where, $n_v$ is the number of valid virtual samples, $n_i$ is the number of invalid virtual samples.

**4.2.3 Indirect evaluation of virtual sample**

In this paper, we use different virtual sample generation methods to obtain the same number of virtual samples, integrate the virtual samples with small samples as the training set of the neural network model, and evaluate the virtual samples by the prediction results of the model. The evaluation criteria used are MAE, MSE, RMSE, and MAPE. MAE is the average of absolute errors, which can better reflect the actual situation of prediction value errors. MSE is the summation average of the square of the difference between the true and predicted values, and RMSE is the root of the MSE. MSE and RMSE are used to detect the deviation between the predicted and true values of the model. MAPE is the mean of the absolute percentage error between the predicted and actual values. The four evaluation criteria are described as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \bar{y}_i - y_i | \tag{28}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\bar{y}_i - y_i)^2 \tag{29}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\bar{y}_i - y_i)^2} \tag{30}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{\bar{y}_i - y_i}{y_i} \right| \times 100\% \tag{31}$$

**4.3 Results and Analysis**

**4.3.1 Experiment on the standard function datasets**

The method proposed in this paper is used to estimate the acceptable area and generate 200 virtual samples through the small sample set of standard functions to evaluate the validity of the acceptable area and the validity of the virtual samples. The experimental results are shown in Fig. 6, 7 and 8.

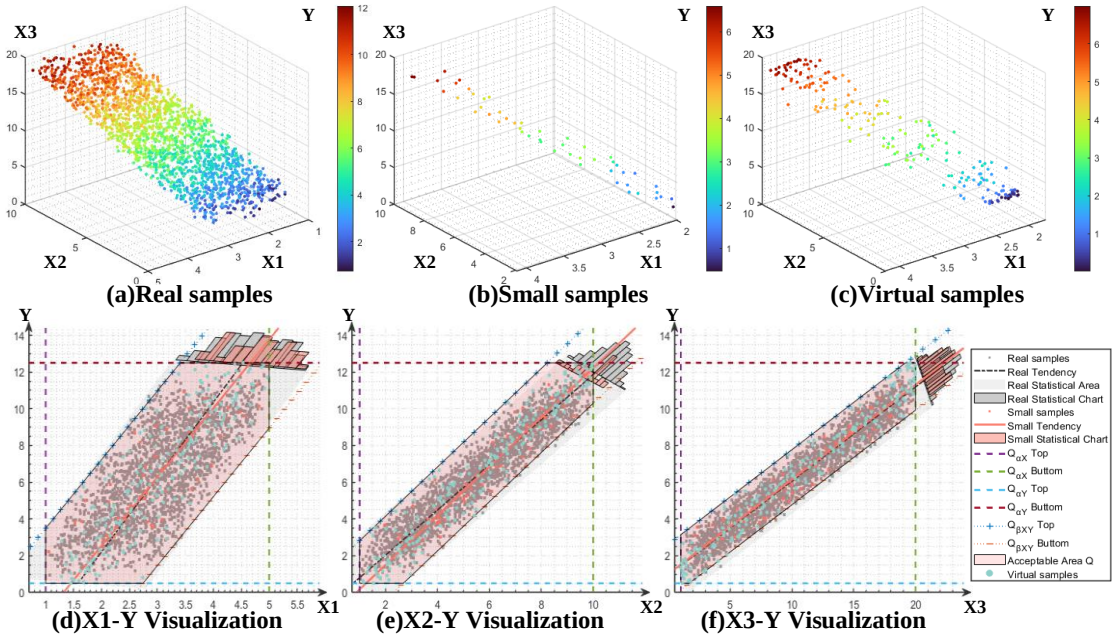


Fig. 6. Experimental results of Linear

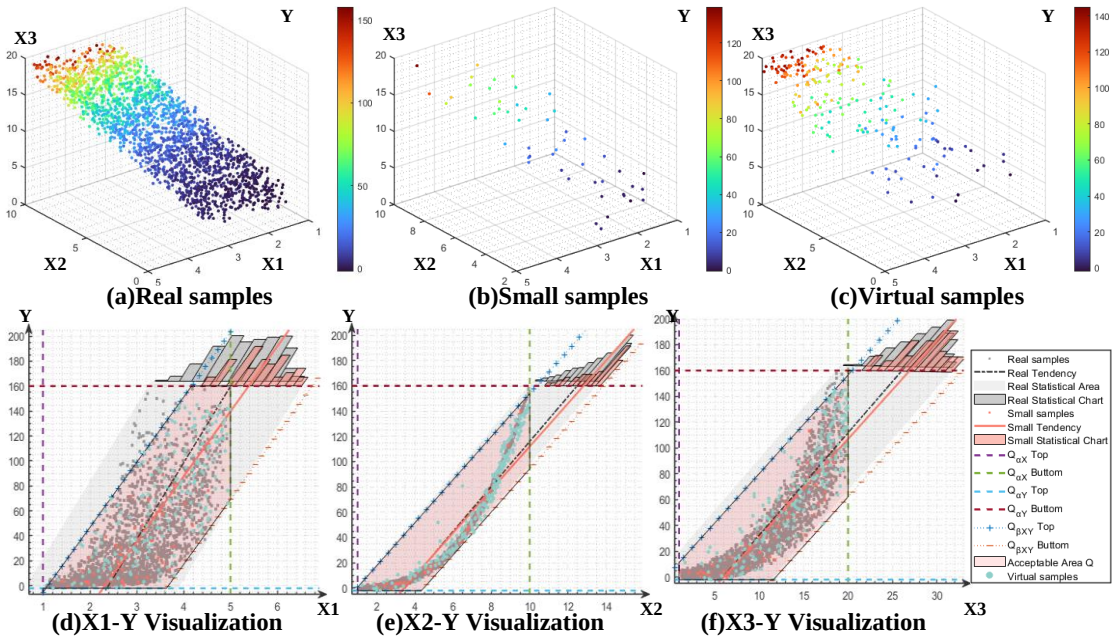


Fig. 7. Experimental results of Nonlinear

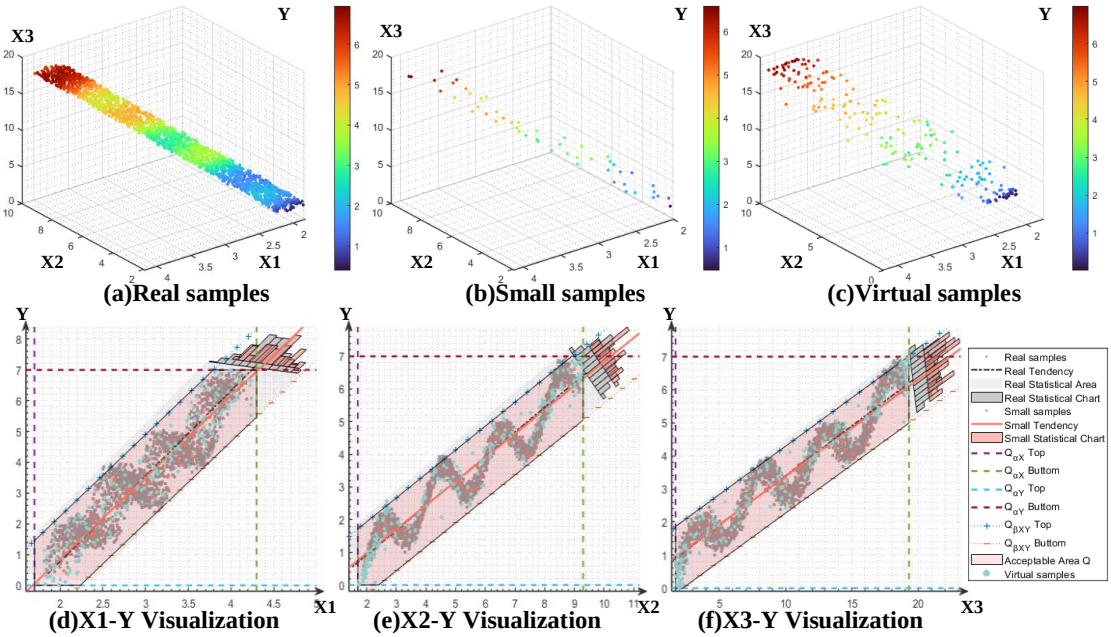


Fig. 8. Experimental results of Oscillation

Table 3 Evaluation of the acceptable area

| Dataset | Acceptable area | Evaluation criteria | | |
|---|---|---|---|---|
| | | $\eta_{ca}$ | $\eta_{ncv}$ | $\eta_v$ |
| Linear | $Q_{x_1-y}$ | 0.9962 | 0.9623 | 0.9588 |
| | $Q_{x_2-y}$ | 0.9666 | 0.9062 | 0.8787 |
| | $Q_{x_3-y}$ | 0.9980 | 0.8653 | 0.8638 |
| Nonlinear | $Q_{x_1-y}$ | 0.9427 | 0.8586 | 0.8160 |
| | $Q_{x_2-y}$ | 0.1548 | 0.9797 | 0.1543 |
| | $Q_{x_3-y}$ | 0.6742 | 0.9609 | 0.6562 |
| Oscillation | $Q_{x_1-y}$ | 0.7502 | 0.9945 | 0.7471 |
| | $Q_{x_2-y}$ | 0.4104 | 0.9731 | 0.4058 |
| | $Q_{x_3-y}$ | 0.5392 | 0.9760 | 0.5321 |

Table 4 Direct evaluation of virtual samples

| Dataset | Evaluation criteria | | |
|---|---|---|---|
| | $n_v$ | $n_i$ | $\gamma_v$ |
| Linear | 286 | 14 | 95.33% |
| Nonlinear | 258 | 42 | 86.00% |
| Oscillation | 281 | 19 | 93.67% |

Linear is the classical linear model, which can be perfectly described by linearity. However, the data trends determined based on the small sample set differ slightly from the real sample trends. There is an overall tilt in the acceptable area, which leads to the loss of part of the valid area and the inclusion of part of the invalid area. This leads to the subsequent generation of partially invalid data. The virtual samples generated by APS-VSG mainly surround small samples, which avoids invalid virtual data from being generated in large quantities in sparse areas. Nonlinear is the classical nonlinear model, which is difficult to describe by linearity. Therefore,

the acceptable area is perceptible as too large. Although the acceptable area includes nearly all of the valid areas, most of them are invalid. In particular, the invalid area is as much as 6 times more than the valid area within the acceptable area of $X_2$ and $Y$, which greatly increases the difficulty of generating valid virtual samples. However, most of the virtual data generated by APS-VSG surrounds small samples, which greatly improves the validity of virtual samples, because of its sampling based on the distribution of small samples. Oscillation is a model with sinusoidal oscillations and is one of the most complex models. And the general VSG method tends to have huge problems with this type of model. Due to the presence of sinusoidal oscillations, the acceptable area inevitably includes invalid area of peaks and valleys, which results in the invalid area that accounts for almost half of the acceptable area. However, the acceptable area obtained by APS-VSG includes almost all the valid area. APS-VSG generates a large number of valid virtual samples around sinusoidal oscillations by its property of sampling based on small samples. Although the superimposed uniform distribution-based sampling leads to a small amount of virtual samples out of the valid area.

From the final evaluation results of the experiment, metric $\eta_{ca}$ is not satisfactory. Using $\eta_{ca}$ as a criterion to evaluate the acceptable area, the best is Linear, followed by Oscillation, and finally Nonlinear. However, it cannot be ignored that $\eta_{cv}$ of the acceptable area are high enough to include as many valid areas as possible, which provides strong support for the subsequent sampling. Therefore, it is not feasible to blindly pursue $\eta_{ca}$ while ignoring $\eta_{cv}$. Using $\eta_{cv}$ as a criterion to evaluate the acceptable area, the best is Oscillation, followed by Linear, and finally Nonlinear. The results of the follow-up the direct evaluation of virtual samples also illustrate exactly this.

In summary, the results of virtual data evaluation on three standard function datasets show that APS-VSG can guarantee the high validity of the virtual samples. This proves that the method proposed in this paper can improve the quality of data.

**4.3.2 Experiment on NASA Li-ion battery dataset**

Generate 50 virtual samples based on the constructed NASA li-ion battery small sample set, and train BP neural networks by small sample sets, real sample sets and virtual samples generated by the six methods to verify whether the virtual samples generated by various methods can effectively improve the accuracy of the model. A 3-layer BP neural network model with the structure of 5-10-1, 1000 iterations, a learning rate of $10^{-3}$, a momentum factor of 0.9, and a target error of $4 \times 10^{-3}$ is constructed. The training data is the set of small samples and virtual samples, and the test data is Real samples. The average prediction results and errors of multiple runs are shown in Fig. 9. The performance criteria of the model are shown in Table 5.

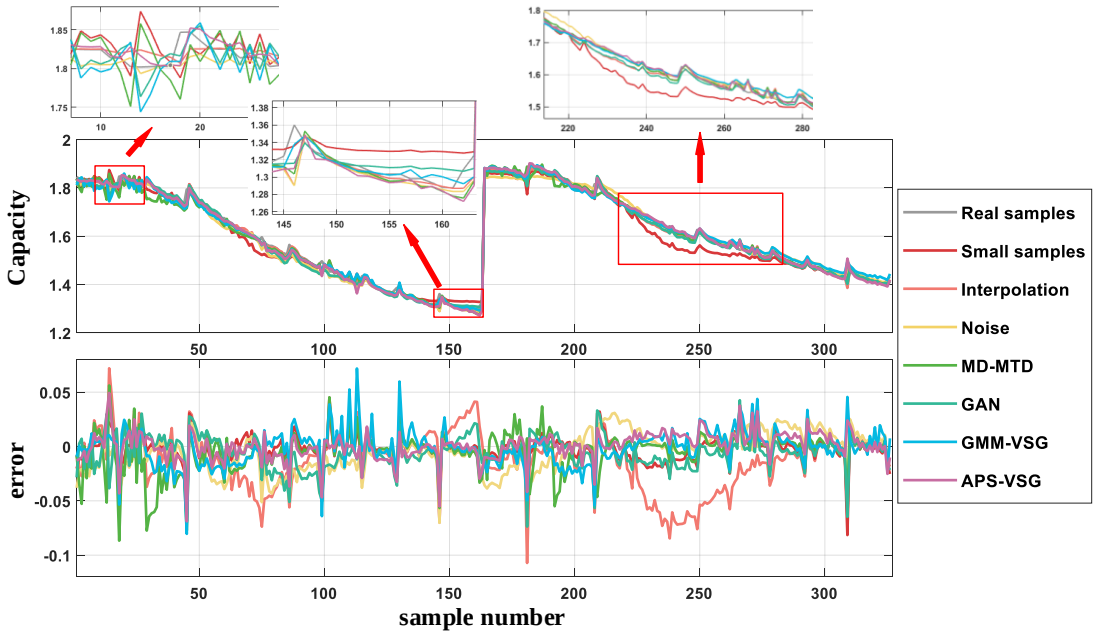


Fig. 9. Prediction results and errors of the BP model trained with different training data

Table 5 Performance criteria of the BP model trained with different training data

| Virtual Data Sources | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| **Small samples** | **$8.29\times10^{-4}$** | **$2.88\times10^{-2}$** | **$2.09\times10^{-2}$** | **$1.30\times10^{-2}$** |
| Interpolation | $1.96\times10^{-4}$ | $1.40\times10^{-2}$ | $0.91\times10^{-2}$ | $0.56\times10^{-2}$ |
| Noise | $3.60\times10^{-4}$ | $1.89\times10^{-2}$ | $1.49\times10^{-2}$ | $0.93\times10^{-2}$ |
| MD-MTD | $3.72\times10^{-4}$ | $1.93\times10^{-2}$ | $1.25\times10^{-2}$ | $0.75\times10^{-2}$ |
| GAN | $2.40\times10^{-4}$ | $1.54\times10^{-2}$ | $1.05\times10^{-2}$ | $0.66\times10^{-2}$ |
| GMM-VSG | $2.02\times10^{-4}$ | $1.42\times10^{-2}$ | $1.04\times10^{-2}$ | $0.66\times10^{-2}$ |
| **APS-VSG** | **$1.63\times10^{-4}$** | **$1.27\times10^{-2}$** | **$0.88\times10^{-2}$** | **$0.55\times10^{-2}$** |
| **Real samples** | **$1.02\times10^{-4}$** | **$1.01\times10^{-2}$** | **$0.69\times10^{-2}$** | **$0.43\times10^{-2}$** |

Data expansion algorithms have their own advantages and disadvantages, but they can improve the accuracy of the model under small sample conditions to a certain extent by analyzing the result of the BP model trained with virtual samples generated by five methods. Through quantitative analysis of experimental results, it can be found that the accuracy of the model trained by the proposed method in this paper is improved more compared with the data-driven model under small sample condition. MSE, RMSE, MAE, and MAPE are decreased by 80.3%, 55.9%, 57.9%, and 57.7%, respectively. Compared with Noise and MD-MTD, MSE, RMSE, MAE, and MAPE are decreased by at least 54.7%, 32.8%, 29.6%, and 26.7%, respectively. Compared with GAN and GMM-VSG, MSE, RMSE, MAE, and MAPE are decreased by at least 19.3%, 10.6%, 15.4%, and 16.7%, respectively.

The error criteria of Interpolation and APS-VSG are very close. However, Interpolation has a fatal flaw in that the BP model trained with virtual samples generated by Interpolation has difficulty in accurately predicting the interval during which some of the battery capacity picks up. The underlying reason is that the virtual samples generated by Interpolation are basically linear, and when the small sample set is not able to characterize the capacity rebound process, then the virtual data generated by Interpolation also does not have the battery capacity rebound property. Both Noise and MD-MTD have relatively high error criteria. The higher deviation of the result of

the BP model trained with virtual samples generated by MD-MTD proves that there are more singular values with large or small values in the prediction results of MD-MTD. If the invalid virtual samples generated by MD-MTD can be further sieved, the accuracy of the model can be greatly improved. In theory, GAN can generate virtual samples that are infinitely close to small sample sets through continuous training. However, due to the small number of samples, the generated virtual samples cannot completely reproduce the real samples in a large number of tests, which leads to an unbreakable bottleneck in the model trained with virtual samples generated by GAN. Due to the stable performance of the Gaussian model, GMM-VSG performs much better than Noise and MD-MTD. However, in localized areas, excessive deviations lead to a slight decrease in the overall performance of the result of the BP model trained with virtual samples generated by GMM-VSG. In many areas where small samples are not well characterized, APS-VSG, like other methods, suffers from large biases. Relatively speaking, APS-VSG delivers excellent performance. The reason is that most of the virtual samples generated by APS-VSG are valid and retain the distribution characteristics of the small sample set well, and the BP model trained with these virtual samples can make accurate predictions for both batteries. Benefiting from both the limits of the acceptable area and the probability distribution based on small samples, the resulting virtual samples are both highly valid and robust.

In summary, the small sample modeling test results on NASA li-ion battery dataset show that APS-VSG can improve the accuracy of the data-driven model, and the method proposed in this paper is progressiveness compared with other methods.

## 5. Conclusion

This paper proposes a virtual sample generation method based on acceptable area and joint probability distribution sampling (APS-VSG) for the expansion of small samples. Break the bottleneck of low validity of MTD in generating high-dimensional virtual samples. The proposed CRI estimation method and the improved acceptable area estimation method are used to limit the acceptable area of data generation and improve the validity of virtual samples. This paper constructs the reasonable joint probability distribution to ensure a high sampling probability near real samples and increase the overall sampling probability in the acceptable area to include part of the edge virtual samples to improve the robustness. This paper uses two-stage sampling to generate high-dimensional data as a whole to avoid the problem of data combination. Experiments on standard function datasets and NASA li-ion battery datasets show that APS-VSG can improve the quality of data and the accuracy of data-driven models and has better performance than MD-MTD, GAN and GMM-VSG.

## Acknowledgment

This work was supported by National Natural Science Foundation of China (62171360), Shaanxi Science and Technology Department (2022GY-110), Xi'an Key Laboratory of Intelligence (2019220514SYS020CG042), National key research and development program (2022YFF0604900), 2022 Shaanxi University Youth Innovation Team Project, Shandong Key

Laboratory of Smart Transportation (Preparation), 2023 Shaanxi Provincial University Engineering Research Center.

**References**

[1] LU Yiyong; CAI Jianyong; ZHENG Hua; ZENG. Researches on Few-shot Learning Based on Deep Learning:an Overview[J]. Telecommunication Engineering. 2021,61(01):125-130. (in Chinese)

[2] Nasef M M, Sauber A M, Nabil M M. Voice gender recognition under unconstrained environments using self-attention[J]. Applied Acoustics, 2021, 175(1):107823.

[3] Qezelbash-Chamak J, Badamchizadeh S, Eshghi K, et al. A survey of machine learning in kidney disease diagnosis[J]. Machine Learning with Applications, 2022, 10: 100418.

[4] Mirzaei G, Adeli H. Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia[J]. Biomedical Signal Processing and Control, 2022, 72:103293.

[5] Wang X, Ma TM, Yang T, Song P, Xie QJ, Chen ZG. Moisture quantitative analysis with a small sample set of maize grain in filling stage based on near infrared spectroscopy. Journal of Agricultural Engineering, 2018, 34(13):203-210.

[6] Li X, Yin Y, Manrique D V, et al. Lifecycle Forecast for Consumer Technology Products with Limited Sales Data[J]. International Journal of Production Economics, 2021(2):108206.

[7] C. Wang, M. Dou, Z. Li, R. Outbib, D. Zhao, J. Zuo, Y. Wang, B. Liang, P. Wang, Data-driven prognostics based on time-frequency analysis and symbolic recurrent neural network for fuel cells under dynamic load, Reliab. Eng. Syst. Saf. 233 (2023) 109123, https://doi.org/10.1016/j.ress.2023.109123.

[8] ZHAO Kai-Lin; JIN Xiao-Long; WANG Yuan-Zhuo. Survey on Few-shot Learning[J]. Journal of Software. 2021,32(02):349-369. DOI:10.13328/j.cnki.jos.006138. (in Chinese)

[9] He Xulong, Zhang Lei, Zhou Han, Wang Xinlei, Miao Zhun.Virtual sample generation method and its application in reforming data modeling [J]. Petroleum Refining and Chemical Industry,2021,52(06):92-95.

[10] Yu Xu, Yang Jing, Xie Zhiqiang.Research on Virtual Sample Generation Technology [J]. Computer Science,2011,38(03):16-19. (in Chinese)

[11] Jing Y, Xu Y, Xie Z Q, et al. A novel virtual sample generation method based on Gaussian distribution[J]. Knowledge-Based Systems, 2011, 24(6):740-748.

[12] XU Rong-wu, HE Lin, ZHANG Lin-ke, TANG Zhi-yin, TU Song. RESEARCH ON VIRUTAL SAMPLE BASED IDENTIFICATION OF NOISE SOURCES IN RIBBED CYLINDRICAL DOUBLE-SHELLS[J]. Journal of Vibration and Shock. 2008(05):32-35+171-172. (in Chinese)

[13] Bishop C M. Training with Noise is Equivalent to Tikhonov Regularization[J]. Neural Computation, 1995, 7(1).

[14] An G. The Effects of Adding Noise During Backpropagation Training on a Generalization

Performance[J]. Neural Computation, 2014, 8(3):643-674.
[15] Wang Weidong, Yang Jingyu. Quadratic Discriminant Analysis Using Virtual Training Samples [J]. Acta Automatica Sinica,2008(04):400-407. (in Chinese)
[16] Li D C, Wu C S, Tsai T I, et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge[J]. Computers & Operations Research, 2007, 34(4):966-982.
[17] Huang C, Moraga C. A diffusion-neural-network for learning from small samples[J]. International Journal of Approximate Reasoning, 2004, 35(2): p.137-161.
[18] Li D C, Wu C, Chang F M. Using data-fuzzification technology in small data set learning to improve FMS scheduling accuracy[J]. International Journal of Advanced Manufacturing Technology, 2005, 27(3-4):321-328.
[19] Li D C, Wu C S, Tsai T I, et al. Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge[J]. Computers & Operations Research, 2006, 33(6):1857-1869.
[20] ZHU Bao, CHEN Zhongsheng, YU Le'an. A novel mega-trend-diffusion for small sample[J]. CIESC Journal. 2016,67(03):820-826. (in Chinese)
[21] ZHU Bao. Research on Virtual Sample Generation Technologies and Their Modeling Application[D]. Beijing University of Chemical Technology,2017. (in Chinese)
[22] Qiao Junfei, Guo Zihao, Tang Jian.Virtual Sample Generation Method Based on improved General Trend Diffusion and Hidden layer interpolation and its application [J]. Journal of Chemical Engineering,20,71(12):5681-5695.
[23] Li L, Damarla S K, Wang Y, et al. A Gaussian mixture model based virtual sample generation approach for small datasets in industrial processes[J]. Information Sciences, 2021, 581:262-277.
[24] Li D C, Wen I H. A genetic algorithm-based virtual sample generation technique to improve small data set learning[J]. Neurocomputing, 2014, 143: 222-230.
[25] Qun-Xiong Zhu, Xiao-Lu Song, Ning Zhang, Ye Tian, Yuan Xu, Yan-Lin He, Novel SVD integrated with GBDT based Virtual Sample Generation and Its Application in Soft Sensor[J], IFAC-PapersOnLine, 2022, 7: 952-956.
[26] Yan-Lin He, Qiang Hua, Qun-Xiong Zhu, Shan Lu, Enhanced virtual sample generation based on manifold features: Applications to developing soft sensor using small data[J], ISA Transactions, 2022, 126:398-406
[27] Li D C, Fang Y. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere[J]. Expert Systems With Applications, 2009, 36(1): 844-851.
[28] Li D C, Lin Y S. Using virtual sample generation to build up management knowledge in the early manufacturing stages[J]. European Journal of Operational Research 175 (2006) 413–434
[29] CHEN Zhongsheng, ZHU Meiyu, HE Yanlin, XU Yuan, ZHU Qunxiong. Quantile regression CGAN based virtual samples generation and its applications to process modeling[J]. CIESC

Journal. 021,72(03):1529-1538. (in Chinese)
[30] He Y L, Li X Y, Ma J H, et al. A novel virtual sample generation method based on a modified conditional Wasserstein GAN to address the small sample size problem in soft sensing[J]. Journal of Process Control, 2022, 113: 18-28.
[31] Yao-San Lin, Yen-Chuan Chang, Che-Jung Chang, Chien-Chih Chen, Der-Chiang Li. A Virtual Sample Screening Mechanism[C]. Proceedings of the 25th National Grey System Conference. 2014:376-383. (in Chinese)
[32] Wang Fubao Probability Theory and Mathematical Statistics [M] TongJi University Press, 1984
[33] Boslaugh S. Statistics in a Nutshell[M]. O'Reilly & Associates, Inc. 2008.
[34] James C. Chen, Tzu-Li Chen, Wei-Jun Liu, C.C. Cheng, Meng-Gung Li. Combining empirical mode decomposition and deep recurrent neural networks for predictive maintenance of lithium-ion battery[j]. Advanced Engineering Informatics, 2021, 50.