scientific reports

OPEN

Research on a multivariate measurement system for polyethylene gas pipelines utilizing a variable weight UMAP model

Chenjia Zong[1], Juan Zhou[1✉], Qiang Wang[2], Haiting Zhou[1], Yun Song[1] & Zhilong Yu[1]

In producing gas-polyethylene pipelines, the five major quality indicators—wall thickness, inner diameter, outer diameter, concentricity, and ovality—exhibit complex interactions, making it challenging for traditional methods to comprehensively evaluate the measurement system's performance. To address this issue, this study introduces a multivariate measurement system analysis model based on Variable-Weight Uniform Manifold Approximation and Projection (VUMAP) dimensionality reduction. First, a dynamic weighting method is proposed, which adaptively adjusts the weights of quality characteristics using K-Means + + clustering and the Particle Swarm Optimization (PSO) algorithm, thereby overcoming the limitation of traditional uniform weighting in accurately reflecting the contributions of different quality characteristics. Second, the VUMAP dimensionality reduction algorithm is refined to enhance its capability to preserve the nonlinear relationships within high-dimensional data, ensuring the integrity and reliability of the dimensionality reduction results. Finally, an integrated multivariate measurement system analysis model is developed to comprehensively assess the measurement system's performance, effectively overcoming the limitations of traditional methods in data structure exploration and capability evaluation. The model's effectiveness is validated using a gas-polyethylene pipeline production process dataset.

Keywords Gas polyethylene pipeline, Variable-Weight uniform manifold approximation and projection, Measurement system, Dimensionality reduction

Gas-polyethylene (PE) pipes are widely utilized in gas transmission and distribution systems due to their excellent corrosion resistance and flexibility. However, critical quality indicators, including wall thickness, inner diameter, outer diameter, concentricity, and ovality, are highly susceptible to variations in process parameters (e.g., barrel temperature and head pressure). These variations impact installation quality and construction efficiency and reduce the pipe's pressure-bearing capacity, posing potential risks to system safety[1–3]. Hence, maintaining a precise and reliable measurement system in gas-polyethylene (PE) pipe production is paramount. Measurement System Analysis (MSA) is a crucial tool for evaluating the reliability of measurement systems. Traditional univariate MSA assesses a single quality indicator (e.g., height or weight) and employs statistical techniques such as analysis of variance (ANOVA) to evaluate the measurement system's performance from multiple perspectives. While this method is simple and efficient, making it highly effective in single-dimensional quality control[4], its applicability diminishes as manufacturing processes become increasingly complex. The intricate interactions among multidimensional quality indicators hinder univariate MSA from fully capturing the measurement system's performance, rendering it insufficient for addressing the demands of complex manufacturing scenarios[5].

Multivariate Measurement System Analysis has become a prominent research focus[6,7]. Currently, the predominant methodologies for multivariate measurement system analysis are grounded in dimensionality reduction techniques, with principal component analysis (PCA) and projection pursuit (PP) being among the most commonly used approaches. PCA reduces data dimensionality by extracting principal components through a linear transformation, effectively capturing the most significant data features. He et al[8]. proposed an online multivariate measurement system analysis method based on PCA, in which multivariate measurement data are transformed into independent principal components to assess the performance of testing instruments.

[1]College of Quality and Standardization, China Jiliang University, Hangzhou 310018, China. [2]College of Energy Environment and Safety Engineering and College of Carbon Metrology, China Jiliang University, Hangzhou 310018, China. ✉email: zhoujuan@cjlu.edu.cn

This approach significantly enhances the efficiency of multidimensional data analysis. However, PCA assumes a linear relationship among variables, making it less suitable for complex nonlinear datasets. Gewers et al[9]. highlight that PCA may result in substantial information loss when applied to data exhibiting strong nonlinear interactions. The PP method, on the other hand, identifies the optimal projection direction by maximizing the variability in the data, making it well-suited for datasets with non-normal distributions or nonlinear structures. Li Xiaoping et al[10]. introduced a multivariate measurement system capability analysis method based on projection pursuit, which accounts for subjective preferences and correlations among quality characteristics, providing a more accurate assessment framework for complex measurement systems. However, PP's sensitivity to parameter selection and high computational complexity limit its practical application in industrial settings[11]. While both PCA and PP substantially simplify multidimensional data analysis, their inherent limitations—particularly the potential loss of critical information during the dimensionality reduction process—constrain their ability to evaluate multivariate measurement systems comprehensively.

With the continuous advancement of dimensionality reduction techniques, nonlinear dimensionality reduction methods have become essential tools for analyzing high-dimensional complex data. In recent years, nonlinear algorithms such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have demonstrated notable advantages in capturing nonlinear relationships and handling non-normally distributed data[12]. t-SNE preserves the local neighborhood structure of data by minimizing the Kullback-Leibler divergence between high-dimensional and low-dimensional distributions, making it widely applicable in fields such as life prediction[13], fault diagnosis[14], and data visualization[15]. However, t-SNE has several limitations, including weak global structure preservation, susceptibility to blurred class boundaries in multi-clustered data, high computational complexity, and sensitivity to hyperparameters, which may compromise its stability and efficiency when processing large-scale datasets[16]. In contrast, UMAP offers superior dimensionality reduction performance and computational efficiency. Leveraging Riemannian geometry and algebraic topology, UMAP constructs a high-dimensional neighborhood graph and optimizes its low-dimensional mapping, thereby preserving local data structures while maintaining global relationships[17]. UMAP more effectively delineates the boundaries between different quality characteristic clusters in industrial multivariate production data while minimizing local detail loss. Compared to t-SNE, UMAP exhibits lower computational complexity and greater efficiency in handling large-scale datasets[17]. Moreover, UMAP demonstrates reduced sensitivity to hyperparameter selection, often yielding stable results with default settings. Within the context of smart manufacturing, UMAP has been leveraged to analyze multidimensional quality characteristic data, offering superior dimensionality reduction efficiency and visualization performance compared to t-SNE[18]. Despite its advantages in processing nonlinear and non-normally distributed data, UMAP has an inherent limitation: it assigns equal weight to all features by default, overlooking the differential significance of various quality characteristics in measurement system analysis. In multivariate measurement system analysis, individual quality characteristics may exert differing influences on system performance, particularly in the presence of significant interactions. Assigning appropriate weights to features can enhance the accuracy of the analysis. Consequently, this limitation of UMAP constrains its applicability in multivariate measurement system analysis[19].

In light of the limitation of the UMAP algorithm, which assigns uniform weights to all quality characteristics by default, this paper proposes an improved UMAP algorithm with variable weights. The proposed algorithm optimizes and assigns adaptive weights to multiple quality characteristics by introducing a weight optimization mechanism. Moreover, traditional PCA algorithms exhibit limited structural preservation and applicability when dealing with manufacturing data with significant nonlinear relationships and complex interactions. In addition, existing multivariate MSA methods generally lack a unified and integrated analytical framework, hindering the implementation of an end-to-end workflow encompassing data preprocessing, nonlinear dimensionality reduction, and system capability evaluation in real-world manufacturing contexts. The fragmented structure of such methods compromises the analysis's practicality and interpretability. Based on this, an integrated analysis model for multiple measurement systems based on Variable-Weight Uniform Manifold Approximation and Projection (VUMAP)is developed. The key contributions of this study are outlined as follows:

(1) A dynamic assignment method based on the VUMAP algorithm is proposed to address the difference in the importance of quality characteristics. The method combines the K - Means + + clustering technique and the Particle Swarm Optimization (PSO) technique, divides quality characteristics into categories through spatial similarity clustering, and optimizes the weight allocation based on the spatial distribution characteristics. This way, it can characterize the actual contribution of quality characteristics more accurately than the traditional uniform weighting method.

(2) An improved method based on VUMAP is proposed to address the problem of the loss of nonlinear information during data dimensionality reduction. This method achieves the differentiated processing of quality characteristics by optimizing the intrinsic topology of high-dimensional data and integrating it with the dynamic empowerment mechanism. To effectively preserve the nonlinear relationships and complex data features, and mitigate the information distortion and error accumulation in traditional methods (e.g., PCA). Experiments demonstrate that VUMAP outperforms traditional methods in maintaining the structural integrity of data and characterizing the interaction of quality characteristics.

(3) Addressing the limitations of the traditional multivariate measurement system analysis model, an integrated analysis framework is proposed, which comprises three core modules: data preprocessing and correlation analysis, dynamic assignment VUMAP dimensionality reduction, and multivariate capability assessment. By integrating anomaly detection and multi-dimensional correlation analysis, the model can accurately identify the correlation characteristics of quality features, utilize the improved VUMAP algorithm for preserving the non-linear data structure, and establish an assessment system that integrates visualization and

quantitative indicators. The model overcomes the limitations of analyzing complex data structures in traditional methods, enhances the comprehensiveness and accuracy of measurement system analysis through modular synergy, and offers a reliable quality management tool for intelligent manufacturing.

Methods

Uniform manifold approximation and projection (UMAP)

UMAP (Uniform Manifold Approximation and Projection)[17,20,21] is an emerging dimensionality reduction algorithm aimed at preserving both global and local structural characteristics of high-dimensional data as much as possible. The algorithm integrates algebraic topology and Riemannian geometry to achieve high-to-low-dimensional mapping by establishing a weighted network framework in high-dimensional space based on an optimization objective. The UMAP algorithm exhibits high result stability and strong adaptability; however, it cannot assign variable weights to different features of the input data, which constrains its application in complex industrial measurements. In this paper, the term low-dimensional space refers to the target projection space resulting from the dimensionality reduction process. In contrast, low-dimensional embeddings denote the coordinate representations of the sample points in this space. The primary procedures of the algorithm include:

Construct a neighborhood graph for high-dimensional data

Given a high-dimensional input dataset $X = \{x_1, x_2, \cdots, x_N\}$, the neighborhood of any sample point x_i is defined as follows:

$$\rho_i = min \{d_E(x_i, x_j) | 1 \leq j \leq k, d_E(x_i, x_j) > 0\} \tag{1}$$

$$\sigma_i = \frac{1}{k-1} \sum_{j=1}^{k} max\left(0, d_E(x_i, x_j) - \rho_i\right) \tag{2}$$

Where k is the number of neighbors, $d_E(x_i, x_j)$ represents the Euclidean distance between x_i and x_j, ρ_i is the distance to the nearest neighbor, and σ_i is the scale factor.

Specifically, a neighbor refers to a sample point close to the target point in high-dimensional space. At the same time, a neighborhood denotes the collection of such neighbors, typically determined by the number of nearest neighbors k. This concept is illustrated in Fig. 1.

Define the similarity probability in high-dimensional space:

$$p_{ij} = exp\left(-\frac{max(0, d_E(x_i, x_j) - \rho_i)}{\sigma_i}\right) \tag{3}$$

Construct an undirected weighted neighborhood graph with edge weights determined by symmetrized similarity probabilities:

$$w_{ij} = p_{ij} + p_{ji} - p_{ij}p_{ji} \tag{4}$$

Define the similarity distribution in the low-dimensional space

In the low-dimensional space, the similarity between embedded sample points is modeled by the following probability distribution:

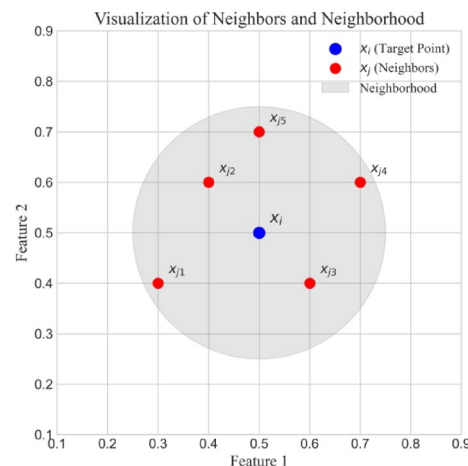$$q_{ij} = \frac{1}{1 + a \,||\, y_i - y_j \,||^{2b}} \tag{5}$$

Fig. 1. Illustration of "neighbor" and "neighborhood" in UMAP.

where $\| y_i - y_j \|$ represents the Euclidean distance between sample points y_i and y_j in the low-dimensional space, $a = 1.929$ and $b = 0.7915$ are empirically determined hyperparameters[17].

The objective function is optimized by minimizing the difference in entropy between the distributions in high- and low-dimensional spaces:

$$C = \sum_{i \neq j} w_{ij} log(\frac{w_{ij}}{q_{ij}}) + (1 - w_{ij}) log(\frac{1 - w_{ij}}{1 - q_{ij}}) \tag{6}$$

Optimize low-dimensional embeddings and result output

The gradient descent method minimizes the objective function C and iteratively updates the positions of the low-dimensional embedding points.

The gradient of the objective function is computed, incorporating both the attractive and repulsive forces, with the attractive force defined as:

$$\nabla C_{attractive} = \frac{-2ab \| y_i - y_j \|_2^{2(b-1)}}{1 + \| y_i - y_j \|_2^2} \omega (x_i, x_j) (y_i - y_j) \tag{7}$$

Here, a and b represent the hyperparameters. The repulsive force is expressed as:

$$\nabla C_{repulsive} = \frac{2b [1 - \omega (x_i, x_j)] (y_i - y_j)}{(\epsilon + \|y_i - y_j\|_2^2)(1 + a \| y_i - y_j \|_2^{2b})} \tag{8}$$

Among these, ϵ is a tiny constant that prevents the denominator from becoming zero.

In summary, the gradient of the objective function C with respect to the embedding point y_i is composed of an attractive force and a repulsive force. The total gradient expression is:

$$\frac{\partial C}{\partial y_i} = \nabla C_{attractive} + \nabla C_{repulsive} \tag{9}$$

Update the position of the point based on the gradient:

$$y_i^{(t+1)} = y_i^{(t)} - \eta \frac{\partial C}{\partial y_i} \tag{10}$$

Here, η represents the learning rate.

The optimization process continues iteratively until the objective function C converges or the predefined iteration limit is attained, resulting in an embedding in a lower-dimensional space.

Variable-weight uniform manifold approximation and projection (VUMAP)

VUMAP integrated strategy and improvement framework

The uniform manifold approximation and projection (UMAP) algorithm employs Euclidean distance as the metric for dimensionality reduction. It assigns uniform weights to all data characteristics, failing to fully leverage the distinctions in data distributions. We propose an enhanced method to address this limitation: variable-weight uniform manifold approximation and projection (VUMAP). This approach introduces the Minkowski distance, dynamic weighting, and an optimization strategy to achieve effective dimensionality reduction of multivariate data, enhancing the preservation of data characteristics and overall reduction performance. The specific improvement framework is shown in Fig. 2.

First, the Minkowski distance is employed as the distance metric to yield greater relative distance variations than Euclidean distance[19]. Then, based on the distribution of these distances, a weighting procedure is applied
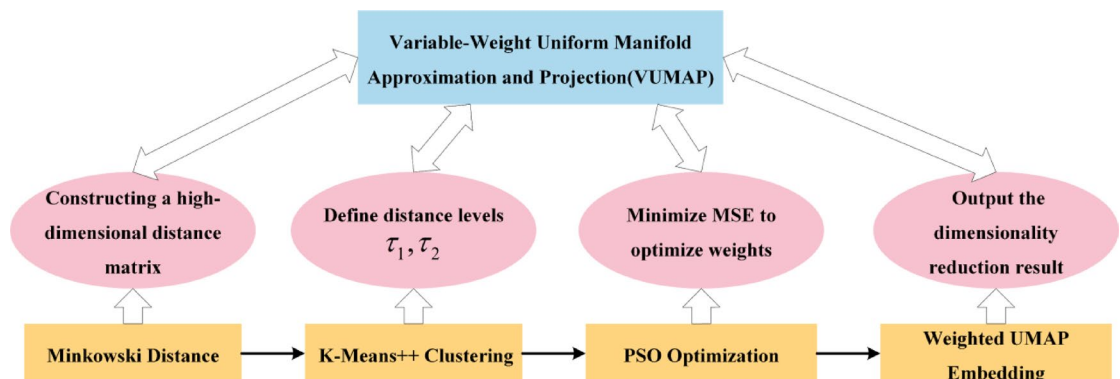

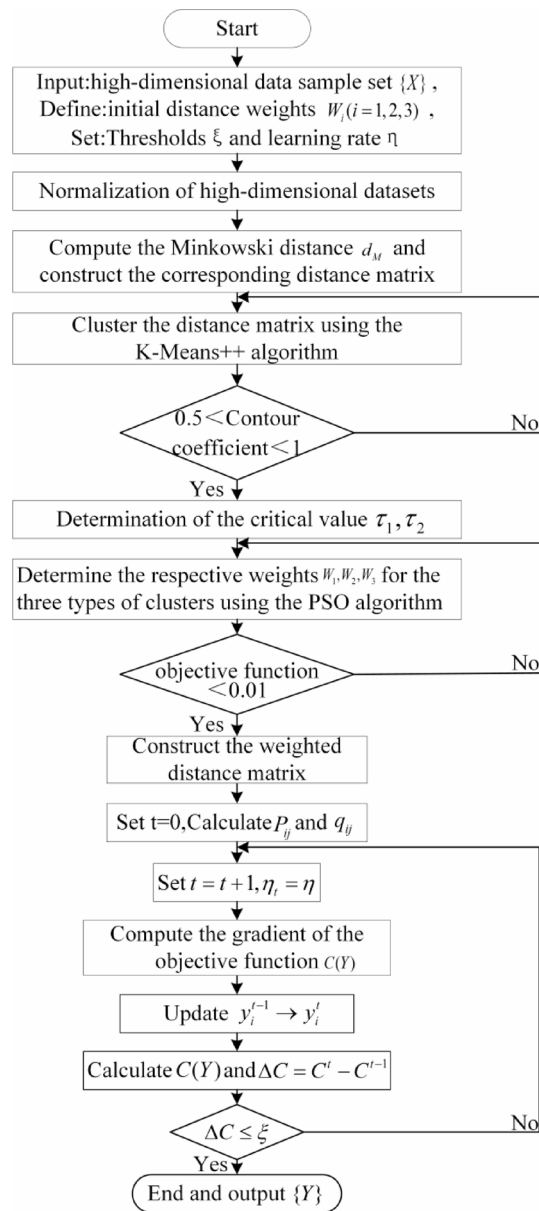**Fig. 2.** VUMAP improvement framework diagram.

Fig. 3. VUMAP algorithm flowchart.

to enhance the discrimination between data instances, thereby improving the retention of high-dimensional information and enhancing the effectiveness of the dimensionality reduction process. Additionally, the Minkowski distance can accommodate variations in scale and statistical characteristics across features by tuning the order parameter p, thus exhibiting enhanced representational capacity.

To objectively categorize the pairwise distances among samples, all unique distance values were extracted from the high-dimensional Minkowski distance matrix and clustered into three groups using the K-Means++ algorithm[19,23,24]. The resulting clusters were then ordered in ascending order of their centroid values, and the midpoints between adjacent centroids were calculated to determine the distance thresholds for classification(τ_1 and τ_2).

In the VUMAP dimensionality reduction algorithm, weight allocation optimization relies on the particle swarm optimization algorithm[25]. PSO adaptively optimizes the weights assigned to the three distance categories, thereby minimizing structural discrepancy between the weighted and original distance matrices. The mathematical details of PSO, including the update equations and optimization objective, are presented in the next subsection (Eqs. 14–16).

Compared to the original UMAP algorithm, which uses Euclidean distance by default and assigns equal weights to all features, the VUMAP framework proposed in this paper introduces several key improvements in three major aspects. First, the Minkowski distance addresses the limitations of Euclidean distance in handling feature scale differences by amplifying relative distance disparities, thereby improving the discriminative capacity in similarity computation. Second, the K-Means++ clustering algorithm automatically partitions sample
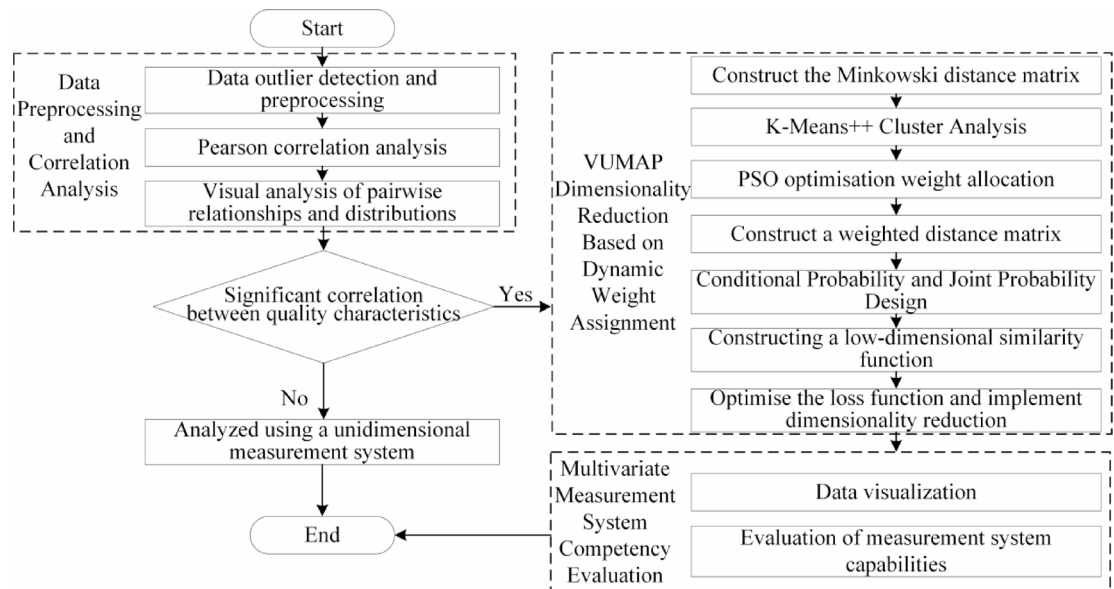
Fig. 4. VUMAP-Based multivariate measurement system analysis flowchart.

distances into three tiers ("near," "medium," and "far"), thereby replacing manually defined thresholds and determining two critical thresholds (τ_1 and τ_2) to guide subsequent weighting. Finally, the PSO algorithm adaptively optimizes the weights for different distance levels by minimizing the structural error (MSE) between the weighted and original distances, ensuring that the resulting embedding faithfully represents the data structure. The synergistic integration of these three components significantly enhances VUMAP's capabilities in global structure preservation, feature weighting, and preservation of variable interaction information, making it suitable for precise analysis of complex and diverse measurement systems.

VUMAP overall process

The flowchart of the VUMAP algorithm is illustrated in Fig. 3. The key steps are outlined as follows:

Step 1: Data preprocessing

The high-dimensional dataset is normalized to mitigate the impact of varying characteristic scales on the dimensionality reduction process.

$$x'_{i,u} = \frac{x_{i,u} - min(X_u)}{max(X_u) - min(X_u)} \tag{11}$$

Here, $x_{i,u}$ denotes the value of the u-th characteristic of the sample x_i; while $min(X_u)$ and $max(X_u)$ represent the minimum and maximum values of the u-th characteristic, respectively.

Step 2: Minkowski distance

To better differentiate features, the Minkowski distance is employed in place of the Euclidean distance, as follows:

$$d_M(x_i, x_j) = \left(\sum\nolimits_{u=1}^{d} |x'_{i,u} - x'_{j,u}|^p \right)^{\frac{1}{p}} \tag{12}$$

Here, p is the order parameter for the Minkowski distance, and the significance of dimensional characteristics is regulated by adjusting p.

Step 3: Dynamic weighting based on distribution characteristics

Weights are assigned to distances based on their distribution characteristics. K-Means++ is applied to cluster the distances into three categories: near, medium, and far. These distances are then weighted, and the PSO algorithm is employed to determine the optimal weights for each category:

$$w_{ij} = \begin{cases} w_1 \times d_M(x_i, x_j), & d_M < \tau_1 \\ w_2 \times d_M(x_i, x_j), & \tau_1 \leqslant d_M < \tau_2 \\ w_3 \times d_M(x_i, x_j), & d_M \geqslant \tau_2 \end{cases} \tag{13}$$

Here, τ_1 and τ_2 are thresholds that differentiate near, medium, and far distances, defining the three categories; $w_1(> 1)$ is the weight for near distances; $w_2(= 1)$ is the weight for medium distances; $w_3(< 1)$ is the weight for far distances.

PSO is an optimization method based on swarm intelligence that simulates the stochastic movements of optimization vectors across the feasible domain to locate Pareto optima. Assuming that the position of the particle is x and its velocity is v, the updated formulas for its position and velocity are:

$$v_i^{t+1} = \omega\, v_i^t + c_1 r_1 \left(p_i^t - x_i^t \right) + c_2 r_2 \left(g^t - x_i^t \right) \tag{14}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{15}$$

Here, ω, c_1, c_2 are weight coefficients, r_1, r_2 are random factors, and p_i^t, g^t represents the individual and global optimal positions, respectively.

The optimization objective of the PSO algorithm is to minimize the structural discrepancy between the weighted Minkowski distance matrix and the original one. Structural consistency is measured by the mean square error (MSE), which is mathematically defined as follows:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i,j} \left(d_M^{(w)} \left(x_i, x_j \right) - d_M(x_i, x_j) \right)^2 \tag{16}$$

Here, $d_M^{(w)}(x_i, x_j)$ represents the Minkowski distance after weight optimization, and N indicates the total number of non-redundant distance pairs.

Step 4 Dimensionality reduction and optimization

The similarity probabilities, cross-entropy objective, and gradient update in VUMAP follow the same formulations as in UMAP (Eqs. 3–10). The only modification is that Euclidean distances d_E are replaced by the optimized weighted Minkowski distances $d_M^{(w)}$. This substitution enables VUMAP to preserve the same optimization pipeline as UMAP while improving feature weighting and structural fidelity.

Model Building

Overall process of the measurement system model

In producing gas-polyethylene pipes, the complex interactions among the five quality characteristics—wall thickness, inner diameter, outer diameter, concentricity, and ovality—make it challenging for traditional measurement system analysis to assess their performance fully. This paper proposes a multivariate measurement system analysis model based on VUMAP to address this issue. The model comprises four modules: data preprocessing and correlation analysis, correlation discrimination of quality characteristics, VUMAP dimensionality reduction based on dynamic weighting, and multivariate measurement system capability evaluation. Data preprocessing reveals the linear and nonlinear relationships among quality characteristics, providing a basis for dimensionality reduction analysis. The VUMAP algorithm, based on dynamic weighting optimization, further extracts nonlinear information from the data and reduces its dimensionality. Finally, the dimensionality reduction data are visualized and quantitatively evaluated to comprehensively assess the performance of the multivariate measurement system under the influence of complex interactions. The general flowchart of the model is demonstrated in Fig. 4.

Data preprocessing and correlation analysis

Data outlier detection and preprocessing

To maintain data integrity and enhance the reliability of subsequent analytical processes, outliers in the collected production dataset must be detected and processed. This study employs the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm[26] to detect outliers in the dataset. DBSCAN identifies and removes low-density regions (outliers) by defining the density characteristics of sample points.

The specific formula description and the definition of core points are as follows:

$$|N_\epsilon(p)| \geq MinPts \tag{17}$$

where: $N_\epsilon(P)$ is the number of points within the ϵ radius neighborhood of P; and $MinPts$ is the density threshold.

This preprocessing step ensures the dataset has good quality consistency and provides a reliable basis for subsequent analysis.

Pearson correlation analysis

Pearson's correlation analysis[27] measures the degree of linear correlation among various quality characteristics. The correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{18}$$

Where: x_i, y_i is the i-th sample data point; $\bar{x}$, $\bar{y}$ is the variable's mean.

The precise analytical protocol comprised the following implementation stages: first, calculate the Pearson correlation coefficient for each pair of quality characteristics and present the results in a matrix; then, draw a heatmap, where the degree of correlation is reflected by the color intensity, to visually display the strength of the correlation between the characteristics and provide a reference for data dimensionality reduction.
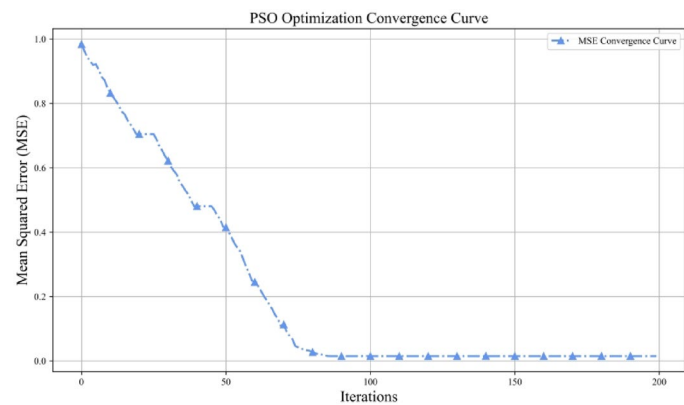
Fig. 5. PSO optimization convergence curve.

Acceptability Evaluation indicators	acceptable	conditional acceptance	unacceptable
%R&R	$\%R\&R < 10\%$	$10\% < \%R\&R \leq 30\%$	$\%R\&R > 30\%$
NDC	$NDC \geq 5$	$3 \leq NDC < 5$	$NDC < 3$

Table 1. Performance evaluation criteria for measurement systems.

Visual analysis of pairwise relationships and distributions

A pair plot is used to visualize the data and further explore the potential relationship between its distribution structure and its quality characteristics. The pair plot displays the two-dimensional scatter distribution between each pair of characteristics, and combined with the univariate distribution histograms of each characteristic on the diagonal, it is possible to determine whether there is a nonlinear correlation between the characteristics and analyze the distribution type of the data.

The data preprocessing and correlation analysis module integrates outlier detection and distribution correlation analysis methods, which are key steps in analyzing multivariate quality characteristic data. This module provides stable and reliable data support for the subsequent VUMAP dimensionality reduction analysis and lays the foundation for multivariate measurement system analysis.

Multivariate quality characteristic dimension reduction judgment and dynamic weighting dimension reduction
Correlation test for qualitative characteristics

The correlation test between multiple quality characteristics, based on the data preprocessing and correlation analysis module, determines the next step: where statistically significant inter-characteristic associations are detected, the VUMAP-based multivariate measurement system analysis model is applied; if the quality characteristics are independent and uncorrelated, the traditional univariate measurement system analysis is used.

Convergence analysis of weight allocation via PSO optimization

The PSO algorithm is employed to optimize the distance weight parameters in VUMAP to achieve an optimal weighted configuration of quality characteristics. To ensure the algorithm's convergence and stability, the following hyperparameters were set: $\omega = 0.7298$, $c_1 = 1.4945$, $c_2 = 1.4945$, $iterations = 200$. Additionally, to verify the weight optimization process's stability and convergence, the convergence trajectory of the objective function (MSE) across PSO iterations is illustrated in Fig. 5.

As shown in Fig. 5, the PSO algorithm significantly reduced the objective function value within the first 50 iterations, with the MSE decreasing to approximately 0.4, demonstrating strong global search capability during the early optimization phase. After the 85th iteration, the error stabilizes and eventually converges to a low value of approximately 0.01, indicating that the algorithm has entered a steady state. This convergence trend highlights the reliability of the weight allocation process and the robustness of the final optimization outcome, thereby providing a solid foundation for the subsequent VUMAP-based dimensionality reduction.

Multivariate measurement system competency evaluation
Data visualization

Based on the data reduced to two dimensions, a scatter plot is used to visually analyze and classify the measurement system, intuitively displaying the fluctuation and distribution characteristics of the multivariate measurement system. The specific steps are as follows: first, the reduced data is classified and labeled based on the number

of measurements, the measurement part number, and the measurement personnel (factory) number; then, a scatter plot is used to visually represent the data points of different classifications in two-dimensional space. The formula for data distribution is defined as follows:

$$S_{ijk} = \{(x_n, y_n) \,|\, n \in \ category\ i, j, k\} \tag{19}$$

Where: x_n, y_n are the coordinates of the nth data point in two-dimensional space; i, j, k denotes the number of measurements, the measurement part number, and the measurement personnel (factory) number, respectively.

Finally, an intuitive analysis is conducted. The scatter plot clearly illustrates the distribution of different data types in two-dimensional space. By observing the clustering, dispersion, and distribution patterns of the points, the repeatability and reproducibility of the measurement system can be qualitatively analyzed.

Following dimensional visualization through bivariate coordinate mapping, the dataset undergoes further condensation into univariate projections for subsequent MSA. These statistical process control (SPC) tools establish quantifiable evaluation criteria that systematically expose the inspection system's variation components and operational metrics via multi-dimensional graphical representations. The synergistic integration of spatial pattern recognition in 2D visualization with probabilistic assessment in 1D MSA enables comprehensive evaluation of measurement reliability and operational validity across methodological dimensions.

Evaluation of measurement system capabilities

For data reduced to one dimension, a univariate measurement system analysis method is employed to quantitatively evaluate the measurement system's performance, focusing on the two key indicators: %R&R and NDC.

First, analysis of variance (ANOVA) is employed to quantify the total variation σ_T, the variation attributable to the measurement system σ_{MS}, and the variation among parts σ_P.

$$\sigma_{ms} = \sqrt{\frac{SS_{MS}}{n_{MS}}}, \sigma_p = \sqrt{\frac{SS_P}{n_P}} \tag{20}$$

$$\sigma_T^2 = \sigma_P^2 + \sigma_{MS}^2 \tag{21}$$

In the formula, SS_{MS} denotes the sum of squares due to the measurement system, and n_{MS} represents its degrees of freedom; SS_P denotes the sum of squares among parts, and n_P denotes the corresponding degrees of freedom.

(1) Percentage of Repeatability and Reproducibility (%R&R)

The %R&R is a critical metric for assessing the variation in a measurement system[28]. It measures the proportion of the measurement system's total error relative to the total variation. The formula is defined as:

$$\%R\&R = \frac{\sigma_{MS}}{\sigma_T} \times\ 100\% \tag{22}$$

(2) Number of Distinct Categories (NDC)

NDC quantifies the maximum discernible categories achievable by the measurement system, where "discernible" specifically precludes dimensional misclassification between larger and smaller specimens. This parameter characterizes the metrological sensitivity threshold, representing the minimal detectable dimensional variation. The formal mathematical representation is expressed as:

$$NDC = \lfloor 1.414 \times\ \frac{\sigma_P}{\sigma_{MS}} \rfloor \tag{23}$$

In the analysis of multivariate measurement systems, repeatability and reproducibility (%R&R) and the number of distinct categories (NDC) are the core metrics for evaluating the measurement system's performance. This paper establishes a grading standard, as shown in Table 1, to classify the measurement system's performance accurately.

Experimental verification
Experimental design and parameter setting

In producing gas-fired polyethylene (PE) pipes, the measurement system's accuracy and stability directly determine the pipes' quality pass rate and application performance. This experiment used polyethylene gas pipes produced in three factories (A, B, and C) as the research objects. The measurement data were analyzed and evaluated based on the VUMAP multivariate measurement system capability analysis model. The experiment employed a high-precision polyethylene gas pipe production extruder (see Fig. 6), which can stably produce gas pipes with a diameter of 161 ± 0.1 mm and exhibits excellent dimensional control and extrusion rate control capabilities.

The measurement objects are the same batch of gas-polyethylene pipes from three factories. Ten pipes were randomly selected as experimental samples to ensure the representativeness and comprehensiveness of the data. Three cross-sections were evenly selected along the axis of each pipe as measurement points, and five key quality characteristics at each cross-section were measured. The five quality characteristics measured include wall thickness (e), outer diameter (D), inner diameter (d), concentricity (δ), and ovality (ϵ). Among them, the wall thickness is used to evaluate the strength and durability of the pipe; the outer diameter and inner diameter affect

Fig. 6. Physical diagram of the extruder.

the connection compatibility and fluid transportation capacity of the pipe, respectively; and the concentricity and ovality are closely related to the symmetry and structural integrity of the pipe.

This experiment employed an ultrasonic testing device for measurement, with an accuracy of ± 0.001 mm. The device automatically performs sampling and measurement to ensure the accuracy and stability of the data. In the experimental design, a total of 10 pipes ($p = 10$) were selected, three factories were chosen for production ($o = 3$), and each pipe was measured three times ($r = 3$) to obtain representative measurement data. The entire measurement process maintained consistent environmental conditions to minimize the influence of external factors on the data.

Dimensionality reduction algorithm comparison and evaluation

Inter- and Intra-Class distances

Intra and inter-class distances are important metrics for evaluating the effectiveness of dimensionality reduction. Intra-class distance (S_w) measures the similarity between samples within the same class, while inter-class distance (S_b) measures the difference between samples from different classes. By optimizing the ratio of intra-class to inter-class distances, the separability of the reduced-dimensional data can be significantly enhanced.

The formulae for calculating the intra-class distance S_w and inter-class distance S_b are:

$$S_w = tr\left(S_w\right) = \sum_{i=1}^{L}\sum_{j=1}^{N_i} ||x_{ij} - \mu_i||^2 \tag{24}$$

$$S_b = tr\left(S_b\right) = \sum_{i=1}^{L} N_i ||\mu_i - \mu||^2 \tag{25}$$

Where: x_{ij} represents the j-th sample vector of the i-th class in the feature space; μ_i represents the mean vector of the ith class samples; and μ is the overall mean vector of all samples.

To compare and evaluate the effects of different dimensionality reduction algorithms, this study selects four methods: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and the variable-weight UMAP (VUMAP) proposed in this paper. Each method's intra-class distance S_w, inter-class distance S_b, and inter-class-to-intra-class ratio S_b/S_w are compared.

Appropriate parameters were configured for each dimensionality reduction algorithm to ensure a fair comparison, and the reduction was uniformly conducted to two dimensions across all experiments. The parameter configurations for each algorithm are summarized as follows: PCA is a linear dimensionality reduction method, using the default Euclidean metric; t-SNE was configured with $perplexity = 30$, $n\_iter = 1000$, metric = "Euclidean" (default); UMAP was set with $n\_neighbors = 15$, $min\_dist = 0.1$, metric = "Euclidean" (default); VUMAP was optimized using $n\_neighbors = 15$, $min\_dist = 0.1$, metric = "Minkowski".

As shown in Table 2, the intra-class distance S_w of the UMAP and VUMAP dimensionality reduction algorithms is small, indicating that similar samples are closely clustered in the reduced-dimensional space. In contrast, the inter-class distance S_b is considerable, indicating significant differences between samples of different categories. Thus, UMAP and VUMAP exhibit superior performance in category distinction. In particular, the inter-class-to-intra-class ratio S_b/S_w of VUMAP is significantly higher than that of other algorithms, highlighting its advantage in category distinction.

Visualization of the dimensionality reduction of the four dimensionality reduction algorithms

To further qualitatively evaluate the effects of dimensionality reduction, this study employs data visualization analysis, comparing the results of different algorithms using scatter plots of the reduced-dimensional data in two-dimensional space. The four algorithms are analyzed across three dimensions: the number of measurements,

algorithm name	S_w	S_b	S_b/S_w
PCA	0.1258	11.9447	94.9835
t-SNE	6.6535	208.5685	31.7772
UMAP	1.3687	136.3697	93.6345
VUMAP	0.1901	139.3284	732.9216

Table 2. Parameter values for four standard dimensionality reduction algorithms.

the measurement plant, and the measurement part number. Figure 7 illustrates the visual embedding results of the four dimensionality reduction algorithms across different dimensions. The horizontal and vertical axes indicate the projected positions of high-dimensional data points within a two-dimensional space following dimensionality reduction.

As shown in Fig. 7, VUMAP demonstrates the best performance in visualizing the data distribution after dimensionality reduction. The embedding results cluster samples within each class more tightly and separate samples between classes more distinctly. This further validates the superiority of the proposed VUMAP algorithm in data dimensionality reduction tasks.

In summary, through quantitative analysis of intra-class and inter-class distances and qualitative evaluation of visual analysis, the proposed VUMAP algorithm demonstrates significantly greater effectiveness than other algorithms in reducing the dimensionality of high-dimensional data, particularly in feature separability, providing more valuable results for subsequent measurement system analysis.

Comparison of nonlinear structure retention capabilities

To further validate the superiority of the proposed VUMAP method over traditional PCA in preserving nonlinear structures, the classic Swiss Roll dataset is employed as a benchmark for nonlinear structure validation. A quantitative evaluation is conducted by calculating the trustworthiness and continuity[29,30] metrics for both methods, which measure their ability to preserve high-dimensional adjacency relationships. Trustworthiness quantifies the extent to which high-dimensional neighbors remain close in the low-dimensional space. At the same time, Continuity assesses the degree to which low-dimensional neighbors correspond to original high-dimensional neighbors. Both metrics have been widely adopted in prior studies to assess topological preservation performance.

Table 3 presents the structural preservation metrics of PCA and VUMAP on the Swiss Roll dataset, averaged over multiple experimental runs. Considering the inherent randomness in dimensionality reduction methods due to initialization and stochastic sampling, each method was executed multiple times, and the results were reported in terms of mean ± standard deviation. The results indicate that VUMAP significantly outperforms PCA across both metrics, with P less than 0.01, thereby confirming the superior capability of VUMAP in retaining nonlinear information.

Analysis of the VUMAP-based multivariate measurement system
Module 1 data preprocessing and correlation analysis

Step1 Data outlier detection and preprocessing

During the production of gas polyethylene pipes, an ultrasonic thickness gauge is employed to accurately measure key production parameters, including wall thickness, outer diameter, inner diameter, concentricity, and ovality. To ensure the accuracy of data analysis, outlier detection was first conducted. Potential outliers were identified using the DBSCAN clustering algorithm, and its parameters were optimized through grid search (epsilon = 0.30004, min_samples = 1) to ensure data integrity and consistency. After processing, it was confirmed that the original dataset contained no significant outliers. As shown in Table 4, the preprocessed dataset provides reliable data for subsequent analysis.

Step2 Pearson correlation analysis

A Pearson correlation analysis was conducted on the preprocessed dataset, and a Pearson correlation coefficient matrix was constructed. A heatmap was then generated to explore the linear relationships between key production parameters.

Figure 8 shows the strongest positive correlation between wall thickness and ovality, with a correlation coefficient of 0.83. This suggests that an increase in wall thickness may lead to uneven deformation of the pipe cross-section, thereby increasing ovality. Additionally, wall thickness exhibits slight (−0.27) and moderate (−0.35) negative correlations with outer and inner diameter, respectively. This implies that, under certain production conditions, an increase in wall thickness may correspond to a reduction in outer and inner diameters. A strong positive correlation (0.80) between outer and inner diameter reflects their close physical relationship. Similarly, a moderate positive correlation (0.53) between outer diameter and concentricity suggests that changes in outer diameter may influence pipe concentricity.

Step3 Visual analysis of pairwise relationships and distributions

To better understand the interrelationships and structural distributions among key production parameters, pairwise relationships between each parameter were further analyzed and visualized. As shown in Fig. 9, the results reveal linear correlations and more complex nonlinear interactions between the parameters. For instance, the weak correlation (0.08) between outer diameter and ovality suggests no direct relationship. In contrast, the moderate negative correlation (−0.47) between inner diameter and ovality implies that an increase in inner
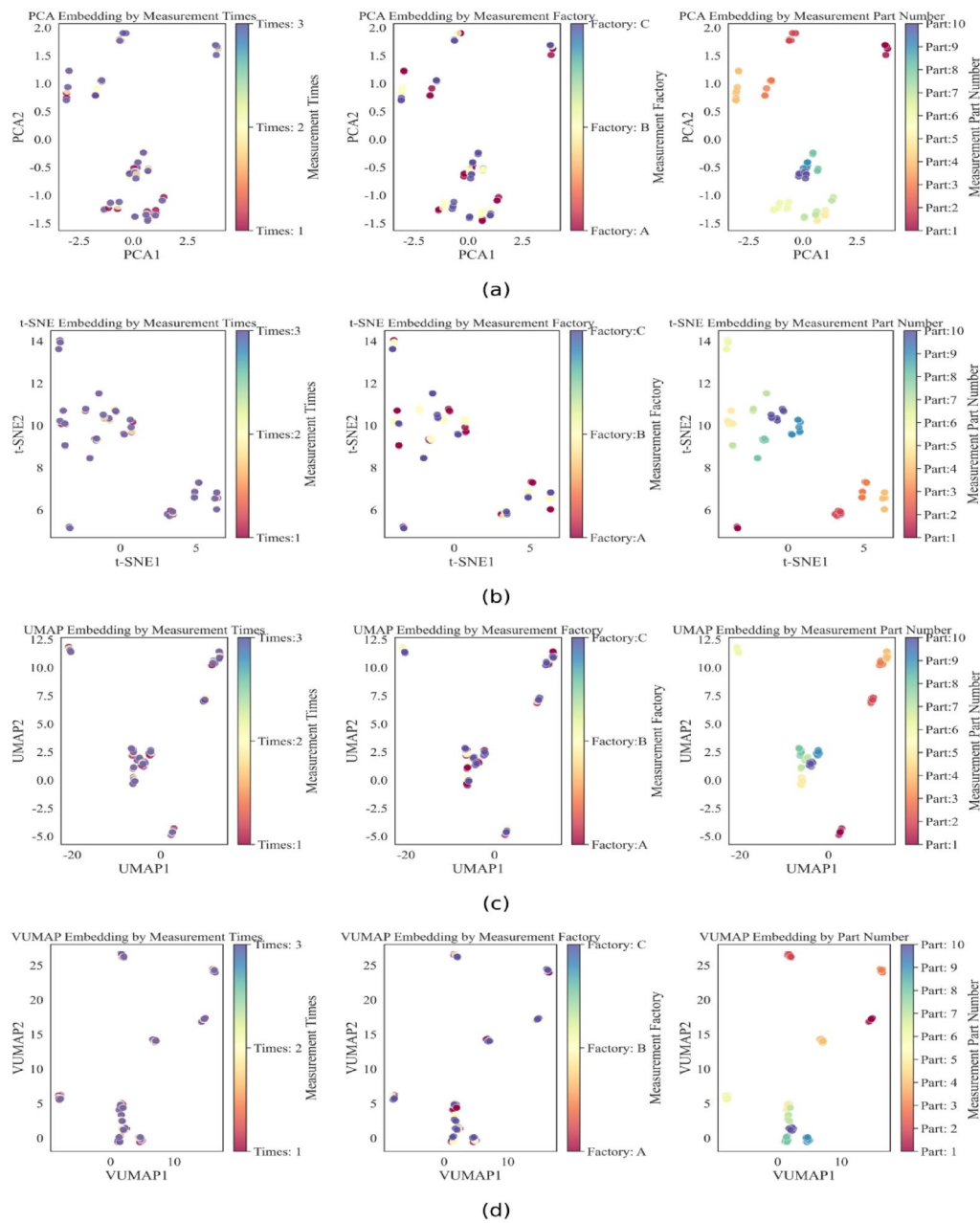
Fig. 7. Two-Dimensional (2D) embedding visualization results of the four dimensionality reduction algorithms.(**a**)PCA, (**b**)t-SNE, (**c**)UMAP, (**d**)VUMAP.

indicator	PCA	VUMAP	P
Trustworthiness	$0.9730\pm<0.0001$	0.9984 ± 0.0001	$P < 0.01$
Continuity	$0.3939\pm<0.0001$	0.7058 ± 0.0047	$P < 0.01$

Table 3. Comparison of topological preservation between VUMAP and PCA.

diameter may reduce ovality. This analysis highlights the complexity of interactions in the production process, indicating that analyzing parameters individually may fail to capture their intrinsic relationships fully.

The above analysis demonstrates that the key parameters in the production process of gas polyethylene pipes exhibit complex interactions. Traditional univariate measurement system analysis methods are insufficient to fully reveal the measurement system's true capability. Therefore, the VUMAP-based multivariate measurement system analysis model proposed in this study is employed to evaluate the system's performance more accurately.

serial number	Number of measurements	e	D	d	δ	ϵ
1	1	9.804	160.932	141.323	97.650	0.287
2	2	9.804	160.932	141.324	97.651	0.287
3	3	9.804	160.931	141.324	97.649	0.288
4	1	9.805	160.925	141.316	97.645	0.287
5	2	9.804	160.925	141.317	97.645	0.287
6	3	9.804	160.926	141.318	97.645	0.287
7	1	9.806	160.923	141.311	97.639	0.287
8	2	9.806	160.923	141.311	97.637	0.287
9	3	9.806	160.923	141.310	97.638	0.287

Table 4. Pretreatment dataset for the production process of gas polyethylene pipes (partial data).

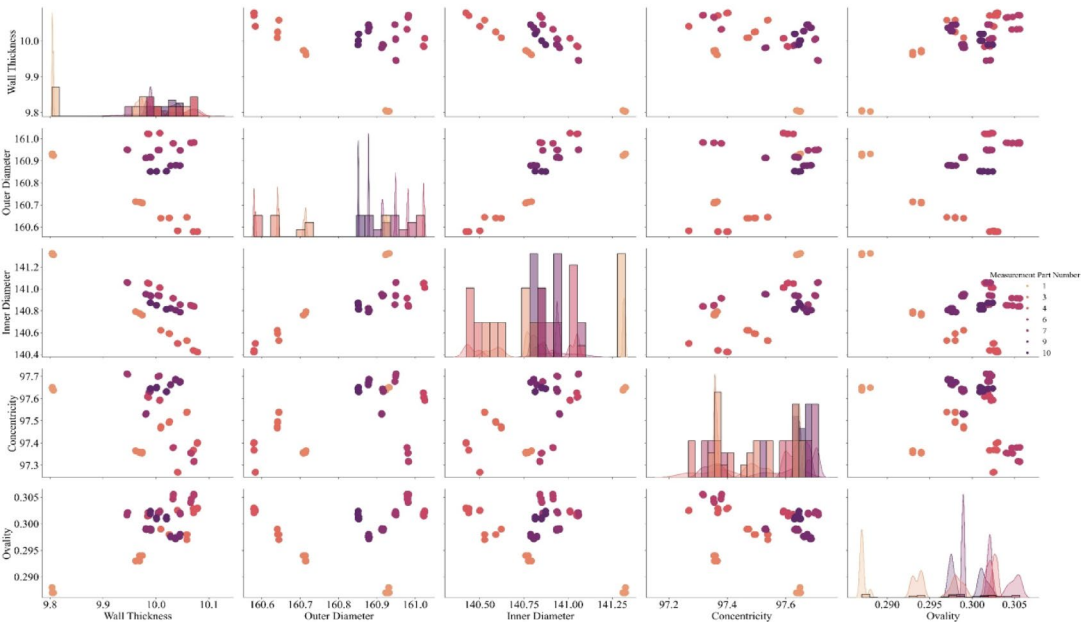**Fig. 8**. Heatmap of pearson correlation coefficients for key production parameters.

Fig. 9. Pairwise relationship diagram of key production parameters.

serial number	Number of measurements	$VUMAP_1$
1	1	24.1424427
2	2	24.26433754
3	3	23.77777863
4	1	24.01725006
5	2	23.97200775
6	3	24.2247467
7	1	24.00163269
8	2	24.07838249
9	3	23.86462593

Table 5. One-Dimensional data (partial) after VUMAP-based dimensionality reduction.

serial number	Number of measurements	$VUMAP_1$	$VUMAP_2$
1	1	14.555439	17.01202774
2	2	14.66190338	17.11843872
3	3	14.87879467	17.3352108
4	1	14.37575245	16.83219528
5	2	14.69468117	17.15104675
6	3	14.48497677	16.94153595
7	1	14.88752747	17.34385109
8	2	14.65681362	17.11327171
9	3	14.76489067	17.22128868

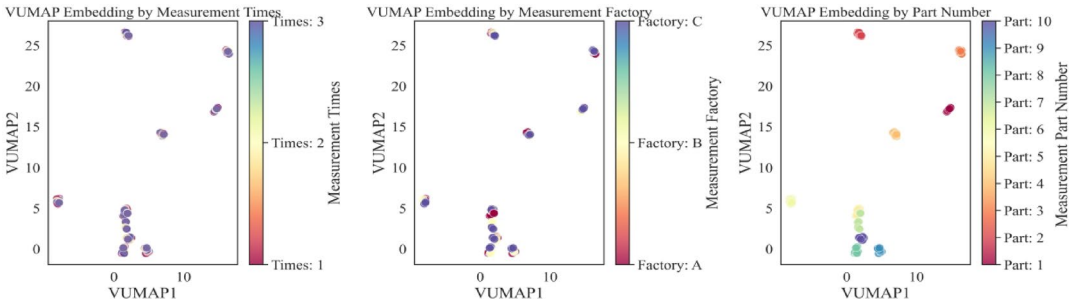**Table 6**. Two-Dimensional data (partial) after VUMAP-based dimensionality reduction.

Fig. 10. Scatter plot based on VUMAP dimensionality reduction embedding.

Module 2 VUMAP dimensionality reduction based on dynamic weight assignment

Building on the previous overview and process analysis of the VUMAP dimensionality reduction algorithm, this section employs Python to reduce the dimensionality of the preprocessed dataset after outlier detection. The VUMAP algorithm was applied to generate data in two dimensions, and one dimension was used to support both qualitative visualization and quantitative system evaluation. The results of the dimensionality reduction are presented in Tables 5 and 6, respectively.

Module 3 multivariate measurement system competency evaluation

Step1 Data visualization

To evaluate the capabilities of the multivariate measurement system visually, the reduced-dimensional data is first classified and analyzed using a scatter plot. Specifically, the data is classified based on three key labels: the number of measurements, the measurement part number, and the measurement plant. The results are visualized in the scatter plot shown in Fig. 10. The horizontal and vertical axes indicate the projected positions of high-dimensional data points within a two-dimensional space following dimensionality reduction.

Combined with the data distribution after VUMAP-based dimensionality reduction shown in Fig. 10, the following observations can be made: In the graph labeled by the number of measurements, the closely distributed point clusters indicate high consistency in repeated measurements over time. In the graph labeled by the measurement factory, the lack of significant differences in the distribution of point clusters across factories suggests good reproducibility of the measurement system. In the graph labeled by the measured part,

14

the clear separation of point clusters for different parts demonstrates the measurement system's strong ability to distinguish between parts, with variation primarily stemming from part differences rather than random measurement errors.

Based on the scatter plot analysis in Fig. 10, the high-dimensional data is further reduced to one dimension, and the reduced data is analyzed using the R&R method, as illustrated in Fig. 11.

The following conclusions can be derived from Fig. 11:

(1) Variance Component Plot: The variance between parts is significantly higher than in repeatability and reproducibility, indicating that the primary source of variation in the measurement system stems from part differences. This further demonstrates the system's strong discrimination ability.
(2) R Control Chart: Most measurement points fall within the control limits, indicating that error fluctuations in the measurement process are well-controlled across factories. This demonstrates the system's high consistency.
(3) Xbar Control Chart: Most points lie outside the control limits, indicating that the sample selection effectively captures the variability between parts, with significant differences in average values.
(4) Part Measurement Value Graph: The minor differences between multiple measurements of the same part confirm the system's stability and consistency in repeated measurements.
(5) Operating Plant Measurement Value Diagram: The near-parallel average values across measurement plants indicate the system's reproducibility and consistent measurement results across different plants.
(6) Interaction Diagram: The near-overlapping measurement lines of the operating factories, with no significant crossings, indicate no significant interaction between factories and parts. The measurement error primarily arises from part differences.

Step2 Evaluation of measurement system capabilities

Based on the data visualization and analysis, the second step involves evaluating the measurement system capability of the reduced-dimensional dataset. For the high-dimensional quality characteristic data of the gas-polyethylene pipeline production process, this study compares the measurement system performance before and after dimensionality reduction, validating the effectiveness of the multivariate measurement system analysis model based on VUMAP. In this study, the original high-dimensional dataset and its reduced-dimensional counterpart are consistently analyzed using MINITAB, where the measurement system variance is computed alongside other key statistical parameters. The evaluation indices %R&R and NDC for each measurement system were calculated using Eqs. (22)-(23), with the specific results presented in Table 7.

As shown in Table 7, the %R&R and NDC values in the original high-dimensional dataset exhibit significant variation across different quality characteristics. For instance, the %R&R reaches 29.49%, nearing the upper limit, while the NDC is only 4, indicating significant measurement system variation and weak discrimination ability for specific characteristics. After dimensionality reduction to one dimension, the %R&R drops significantly to 1.94%, and the NDC increases to 72, demonstrating the measurement system's enhanced efficiency. Moreover,
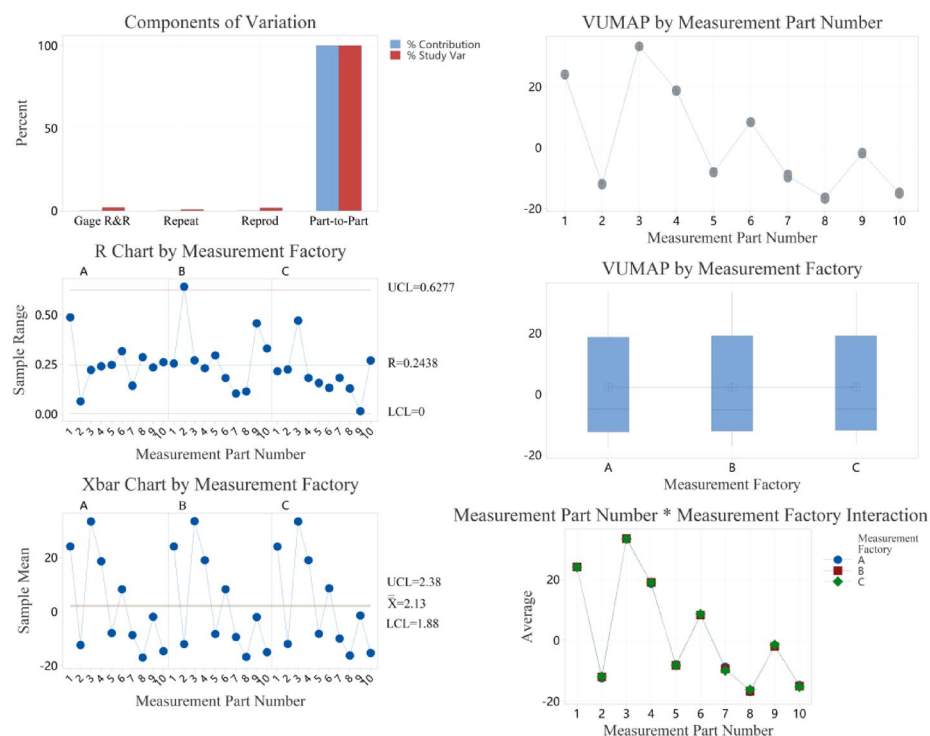

Fig. 11. Results of the R&R analysis for the measurement system's gauges.

	ANOVA					
	e	D	d	δ	ϵ	VUMAP
$\widehat{\sigma}_P$	0.0722370	0.150123	0.236381	0.139804	0.0052294	17.8711
$\widehat{\sigma}_{MS}$	0.0223377	0.002386	0.044696	0.035438	0.0004797	0.3463
$\widehat{\sigma}_T$	0.0756119	0.150142	0.240570	0.139804	0.0052514	17.8744
$\%R\&R$	29.49%	1.59%	18.58%	25.35%	9.13%	1.9374%
NDC	4	88	7	5	15	72

Table 7. Evaluation of measurement system capability for original and reduced data.

indicator	e	D	d	δ	ϵ	VUMAP
$\%R\&R$	30.01 ± 5.68**	1.35 ± 0.31**	18.42 ± 2.52**	25.40 ± 3**	8.72 ± 1.14**	1.85 ± 0.26
NDC	4.63 ± 0.88**	111.32 ± 30.75*	7.68 ± 1.16**	5.44 ± 0.67**	16.38 ± 2.11**	77.72 ± 10.25

Table 8. Bootstrap-based robustness evaluation of %R&R and NDC. Table note: Significance markers: $P < 0.01$ (**), $P < 0.05$ (*), based on t-tests versus VUMAP.

the correlation variation of the measurement system post-reduction ($\widehat{\sigma}_{MS}$ = 0.3463) is significantly lower than the variation between parts ($\widehat{\sigma}_P$ = 17.8711), indicating effective control of measurement errors and retention of inter-part variation information. The VUMAP-based dimensionality reduction method enhances the measurement system's repeatability, reproducibility, and discrimination ability while reducing the source of system variation.

Robustness analysis
Since a novel multivariate measurement system evaluation method based on VUMAP was proposed, a hierarchical bootstrap analysis[31] was conducted to assess the robustness and consistency of its evaluation results. A nested structure of "parts–measurers–repeated measurements" was adopted, and 30 bootstrap iterations were performed. The %R&R and NDC indicators were calculated separately, and the results were reported as mean ± standard deviation (SD). Statistical significance was assessed via independent-samples t-tests.

As shown in Table 8, the %R&R and NDC indicators obtained by the VUMAP method exhibit minimal variation across multiple resampling iterations, with consistently low standard deviations, demonstrating high stability and consistency. Furthermore, the VUMAP method exhibits statistically significant differences from traditional dimensionality reduction methods, indicating that the measurement capability assessment results are statistically robust and demonstrate strong repeatability and generalizability.

Conclusions and discussion
Conclusions
Given the challenges in comprehensively evaluating the measurement system performance for the five key quality characteristics of gas polyethylene pipe production—wall thickness, inner diameter, outer diameter, concentricity, and ovality—due to their complex interactions, this study proposes a multivariate measurement system analysis model based on VUMAP dimensionality reduction. This model innovatively integrates dynamic weighting, nonlinear information retention, and multi-module system analysis methods. The model comprises three modules: data preprocessing and correlation analysis, VUMAP dimensionality reduction with dynamic weighting, and multi-measurement system capability evaluation, offering a novel approach to quality management in complex production scenarios.

First, the original data undergoes preprocessing, and correlation analysis is conducted to identify significant interactions among quality characteristics. Second, based on the correlation analysis results, the decision is to employ either a univariate or multivariate measurement system analysis method. Then, the VUMAP algorithm with dynamic weighting is applied for dimensionality reduction, preserving key data characteristics to the greatest extent and generating a low-dimensional dataset. Finally, the low-dimensional dataset is comprehensively analyzed by integrating data visualization with quantitative evaluation. The model's superiority is validated using a dataset from gas-polyethylene pipe production: after dimensionality reduction, the %R&R of the measurement system decreased significantly (from 29.49% to 1.94%), while the NDC increased markedly (from 4 to 72), demonstrating substantial improvement in measurement system performance. Additionally, visual analysis reveals a clear separation of point clouds across different parts and minimal variation among point clouds from different factories, further confirming the measurement system's distinguishing ability and consistency. The results demonstrate that the VUMAP-based multivariate measurement system analysis model effectively controls measurement errors, retains part difference information, and excels in handling complex nonlinear data, fully addressing the measurement needs of production applications.

Limitations and application prospects

Despite its effectiveness, the proposed method still has several limitations. First, the efficiency of the dynamic weighting mechanism may degrade when applied to large-scale datasets. Second, quantitatively modeling local quality characteristics' contributions to overall measurement system errors remains challenging. Third, since the weight assignment is derived from unsupervised clustering of pairwise sample distances followed by optimization, it is difficult to establish an explicit correspondence between the assigned weights and the physical significance of individual quality characteristics, which limits the interpretability of the results. Future research should therefore focus on enhancing the algorithm's scalability for industrial-scale data, incorporating causal inference and deep learning techniques to identify error sources better, and introducing domain expertise or physical constraints to improve the model's engineering interpretability and practical applicability.

We plan to develop a National Quality Infrastructure (NQI) demonstration platform for polyethylene gas pipelines in future work. We will embed the proposed model into the quality control panel as the core algorithm for multivariate quality monitoring during production. The model requires no additional hardware and can seamlessly integrate into existing manufacturing systems. It is expected to reduce manual inspection workload, improve the efficiency of multivariate anomaly detection, and enhance product yield through a timely feedback mechanism.

Data availability

The datasets used and analyzed during the current study can be available from the corresponding author on reasonable request.

References

1. Nasiri, S. & Khosravani, M. R. Failure and fracture in polyethylene pipes: Overview, prediction methods, and challenges. *Eng. Fail. Anal.* **152**, 107496 (2023).
2. Wu, S., Zhang, W. & Zhu, Y. Effects of process parameters on the microstructure and mechanical properties of large PE pipe via polymer melt jetting stacking. *Processes* **11**, 2384 (2023).
3. Luo, L. Tensile Strain Hardening Behavior and Strength Failure of Polyethylene Pipes. Master's Thesis, *Xiangtan University* (in Chinese). (2019).
4. Burdick, R. K., Borror, C. M. & Montgomery, D. C. A review of methods for measurement systems capability analysis. *J. Qual. Technol.* **35**, 342–354 (2003).
5. Zhou, Q. Research on the Evaluation Approach and Application of the Ability of Dimensional Reduction Multivariate Measurement System. Master's Thesis, *Jiangsu University of Science and Technology* (in Chinese). (2020).
6. Peruchi, R. S., Junior, M., Fernandes, H., Balestrassi, N. J., Paiva, A. P. & P. P. & Comparisons of multivariate GR&R methods using bootstrap confidence interval. *Acta Sci. Technol.* **38**, 489 (2016).
7. Shi, L., He, Q., Liu, J. & He, Z. A modified region approach for multivariate measurement system capability analysis: L. SHI *ET AL. Qual. Reliab. Eng. Int.* **32**, 37–50 (2016).
8. He, S. G., Wang, G. A. & Cook, D. F. Multivariate measurement system analysis in multisite testing: an online technique using principal component analysis. *Expert Syst. Appl.* **38**, 14602–14608 (2011).
9. Gewers, F. L. et al. Principal component analysis: A natural approach to data exploration. *ACM Comput. Surv.* **54**, 70 (2021).
10. Li, X. P. & Zhou, Q. Capability analysis of projection pursuit multivariate measurement systems considering subjective preferences of quality characteristics. *Industrial Eng. Manage.* **25**, 183–190 (2020). (in Chinese).
11. Fu, Q. & Zhao, X. Y. Principles and applications of projection pursuit models. *Science Press* (2006). (in Chinese).
12. Ayesha, S., Hanif, M. K. & Talib, R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Inform. Fusion.* **59**, 44–58 (2020).
13. Zhong, J. H., Huang, C., Zhong, S. C. & Xiao, S. Remaining useful life prediction of rolling bearings based on t-SNE dimensionality reduction method. *J. Mech. Strength.* **46**, 969–976 (2024). (in Chinese).
14. Chen, X. M., Wang, X. J., Cai, Y. & Zhou, B. Analog fault diagnosis method based on AVMD and t-SNE using HHO-SVM. *J. Electron. Meas. Instrument.* **38**, 233–240 (2024). (in Chinese).
15. Jiménez-Preciado, A. L., Cruz-Aké, S. & Venegas-Martínez, F. Identification of patterns in CO2 emissions among 208 countries: K-Means clustering combined with PCA and Non-Linear t-SNE visualization. *Mathematics* **12**, 2591 (2024).
16. Xu, Z. Y. Application of t-SNE algorithm in dimensionality reduction of high-dimensional data. Master's Thesis, *Xiamen University* (in Chinese). (2021).
17. Mcinnes, L. & Healy, J. U. M. A. P. Uniform manifold approximation and projection for dimension reduction. *J. Open. Source Softw.* **3**, 861–924 (2018).
18. Guo, Z. et al. A multivariate algorithm for identifying contaminated peanut using visible and near-infrared hyperspectral imaging. *Talanta* **267**, 125187 (2024).
19. Li, Y. K., Yang, H. D., Xu, K. K., Lan, Z., Zhang & Y. & R.,N. Temperature prediction of lithium batteries based on weighted UMAP and improved BLS. *Energy Storage Sci. Technol.* **13**, 3006–3015 (2024). (in Chinese).
20. Xue, C. et al. Dimensionality reduction of hydraulic gate safety evaluation index system based on UMAP and CNN. *South-to-North Water Transfers Water Sci. Technol.* **22**, 997–1006 (2024).
21. Marx, V. Seeing data as t-SNE and UMAP do. *Nat. Methods.* **21**, 930–933 (2024).
22. General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. GB/T 8806 – 2008 Determination of Dimensions of Plastic Components in Plastic Piping Systems. (2008).
23. Arthur, D. & Vassilvitskii, S. k-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* 1027–1035 (2007).) 1027–1035 (2007). (2007).
24. Chen, J. F., Bu, F. L. & Wang, Y. F. Text topic recognition of offensive comments based on cosent and improved K-means. *Sci. Technol. Eng.* **24**, 13442–13449 (2024). (in Chinese).
25. Kennedy, J. & Eberhart, R. Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks (ICNN'95)* Vol. 4, 1942–1948 (1995).
26. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)* 226–231 (1996).
27. Pearson, K. On a form of spurious correlation which May arise when indices are used in the measurement of organs. *Proc. R Soc. Lond.* **60**, 489–498 (1897).

28. Silva Soares, W., de Peruchi, O., Silva, R. S. V., Rotella Junior, P., Gage & R. A. & R&R studies in measurement system analysis: A systematic literature review. *Qual. Eng.* **34**, 382–403 (2022).
29. Venna, J. & Kaski, S. Neighborhood preservation in nonlinear projection methods: an experimental study. In *Artificial Neural Networks — ICANN 2001* (eds Dorffner, G. et al.) 485–491 (Springer, 2001).
30. Onclinx, V., Wertz, V. & Verleysen, M. Nonlinear data projection on non-Euclidean manifolds with controlled trade-off between trustworthiness and continuity. *Neurocomputing* **72**, 1444–1454 (2009).
31. Efron, B. Bootstrap methods: another look at the jackknife. *Annals Stat.* **7**, 1–26 (1979).

Author contributions

C.J.Z. was responsible for the overall study design and implementation, including experimental operations, data collection and analysis, and manuscript writing; J.Z. supervised the research, secured funding, reviewed the manuscript, and finalized the version; Q.W., H.T.Z., and Y.S. provided expert advice on experimental framework optimization and data interpretation; Z.L.Y. participated in experimental assistance and technical proofreading of the manuscript. All authors have reviewed and approved the final manuscript.

Funding

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.