

Generating Synthetic Procurement Data for Small- And Medium-Sized Enterprises Using a Zero-Inflated Conditional Variational Autoencoder

Julian Stang^{1*}, Jannik Stoeber¹, Daniel Schnitzler¹,
Jan Joppien², Marc Wegmann¹, Christina Reuter¹

¹Institute for Machine Tools and Industrial Management (*iwb*), School
of Engineering and Design, Technical University of Munich (TUM),
Boltzmannstr. 15, Garching near Munich, 85748, Bavaria, Germany.

²Research Unit Production Management, FIR (Institute for Industrial
Management) at the RWTH Aachen University, Campus-Boulevard 55,
Aachen, 52074, North Rhine-Westphalia, Germany.

*Corresponding author(s). E-mail(s): julian.stang@iwb.tum.de;

Abstract

Small- and medium-sized enterprises (SMEs) often face significant data scarcity in procurement planning, which limits the effective use of advanced analytics for forecasting and decision support. This study addresses the data scarcity by operationalizing a zero-inflated Hurdle model through a conditional variational autoencoder to generate realistic synthetic procurement data. The model explicitly represents both the structural mass of on-time deliveries and the asymmetric distribution of early and late deviations using a mixture of Gaussians. Following the cross-industry standard process for data mining, the conditional variational autoencoder was trained and validated on industrial procurement records. The evaluation results demonstrate strong fidelity: the generated distribution of delivery deviations achieved a Wasserstein distance of 1.92 days relative to the real data, with higher accuracy for early deliveries (1.84 days) than for late deliveries (1.93 days). The zero-inflated component reached a precision of 74.41 % and an $F_{0.5}$ -score of 62.86 % when identifying on-time deliveries. These findings confirm that the conditional variational autoencoder effectively

captures both structural zeros and multimodal deviations. By enriching limited datasets while preserving domain-specific behavior, the approach enables SMEs to train forecasting models, perform supplier risk simulations, and deploy procurement-focused decision-support tools even in data-scarce environments.

Keywords: Synthetic data generation; Procurement planning; Zero-inflated Hurdle model; Conditional variational autoencoder (cVAE); Machine learning; Small- and medium-sized enterprises (SMEs)

1 Introduction

1.1 Background to the study

In recent years, the importance of robust procurement planning [1] in manufacturing companies has grown tremendously, as manufacturing companies increasingly face volatile supply conditions [2], geopolitical uncertainties [3, 4], and global logistics disruptions [4]. An efficient procurement planning, therefore, significantly influences the competitiveness of manufacturing companies, and especially of small- and medium-sized enterprises (SMEs) [5], as they often lack the redundancy of production systems, bargaining power, and digital infrastructure of larger manufacturing companies [6]. The lack of digital infrastructure is particularly problematic for SMEs, as it leads to a limited data availability and a poor data quality, hindering the application of advanced data analytics methods, such as artificial intelligence (AI), to improve the procurement planning efficiency [7]. Applying advanced analytics has been shown to reduce supply chain errors by 20 to 50 % [7]. The data scarcity becomes especially critical during production ramp-up phases [8, 9], which significantly contribute to a manufacturing companies' success – House and Price [10] stated that "well over 50% of sales [at Hewlett-Packard] come from products introduced during the past three years", highlighting the importance of new and innovative products, which require more frequent and successful production ramp-ups.

1.2 Aim, design, and scope of the study

This study addresses the problem of data scarcity in SMEs during production ramp-up phases by employing a conditional variational autoencoder (cVAE)-based implementation of a Hurdle model to generate realistic synthetic procurement data. The Hurdle model is a type of zero-inflated model that is particularly well-suited for data with an excess of structural zeros, as it separates the process generating on-time deliveries from the continuous distribution of early or late deviations. Unlike Poisson-based count models, which couple zero frequency with overall dispersion, the Hurdle formulation allows independent modeling of punctual and deviating deliveries, thus more accurately reflecting the asymmetric and continuous nature of procurement performance. The central performance variable is the *delivery deviation*, defined as the difference between the *planned* and *actual* delivery date. While deviations are ideally zero in practice (i.e., on-time delivery), non-zero deviations (positive or negative) occur frequently and have a critical impact on production continuity, procurement efficiency, and overall supply chain performance.

The study’s research design builds upon the cross industry standard process for data mining (CRISP-DM) framework [11], which provides a structured and practitioner-oriented process suitable for deployment in SMEs. Employing a structured research design ensures adherence to the key scientific quality criteria of *objectivity*, *reliability*, and *validity* [12].

The scope of this study encompasses the procurement function in SMEs, considered as an integral part of the overall value chain [1]. The procurement becomes particularly critical during production ramp-up phases, which are decisive for the long-term success of an SME [9]. A production ramp-up, defined as the period from the start of production until the full output quantity is achieved [13, 14], is characterized by high vulnerability to disruptions due to uncertain processes and product conditions [8]. The systematic use of data has the potential to increase efficiency and reduce risks

139 in this sensitive phase. However, data scarcity often hinders the practical application
140 of advanced analytics [8]. To address this challenge, Stang and Zaeh [15] proposed a
141 framework for improving production ramp-up efficiency that comprises three research
142 elements aimed at enabling data-driven decision-making. This study contributes to
143 the research element *synthetic data generation* of the framework proposed by Stang
144 and Zaeh [15] by introducing a cVAE model for generating synthetic procurement
145 data in SMEs. The resulting datasets are designed to support downstream applica-
146 tions, including delivery reliability forecasting, supplier risk simulation, and training
147 procurement-focused decision-support models.
148

155 1.3 Summary of the contributions

157 Building on the theoretical foundation of a zero-inflated Hurdle model, this study
158 employs a cVAE to implement the Hurdle model and generate realistic delivery devi-
159 ations in procurement data. By conditioning on supplier and article embeddings, the
160 model captures entity-specific behavior patterns, while its generative capacity enables
161 the creation of realistic synthetic datasets under conditions of data scarcity. This inte-
162 gration of zero-inflated modeling, mixture-based likelihoods, and entity embeddings
163 constitutes a technological advance over prior work, offering a scalable and transfer-
164 able approach that SMEs can apply to improve the procurement planning, scenario
165 analysis, and the decision-support model training when historical data is limited.
166
167
168
169
170
171

173 1.4 Structure of the paper

175 The remainder of this article is organized as follows: Section 2 details the fundamentals,
176 i.e., the materials and methods necessary for this study. Section 3 reviews related
177 work and outlines the state of research on synthetic data generation for procurement
178 planning in SMEs. Section 4 presents and discusses the results in the context of the
179 state of the art. Section 5 highlights the study’s limitations and directions for future
180 research. Finally, section 6 provides a concise summary of the article.
181
182
183
184

2 Materials and methods	185
	186
2.1 Processes and methodologies	187
	188
	189
The research design follows the CRISP-DM framework [11], which provides a practical	190
and iterative standard for data-driven projects. The framework was selected because	191
it ensures methodological transparency, modularity, and alignment with deployment	192
requirements in SMEs. Within the present context, the six CRISP-DM phases –	193
<i>business understanding, data understanding, data preparation, modeling, evaluation,</i>	194
and <i>deployment</i> – are adapted to address the specific characteristics of procurement	195
data and the constraints of manufacturing SMEs. In this study, <i>business understand-</i>	196
<i>ing</i> is operationalized as delivery reliability forecasting; <i>data understanding</i> involves	197
exploration of procurement transaction records; <i>data preparation</i> includes data clean-	198
ing and feature engineering; <i>modeling</i> applies the architecture of zero-inflated cVAE;	199
<i>evaluation</i> compares synthetic and real distributions; and <i>deployment</i> considers the	200
integration into enterprise resource planning (ERP) systems. A detailed explanation	201
of each CRISP-DM phase, along with its role in the synthetic data generation pipeline,	202
is presented in section 4.	203
	204
	205
	206
	207
	208
	209
	210
	211
	212
	213
2.2 Type of the statistical analysis	214
	215
This subsection explains the necessary fundamentals, i.e., the Hurdle model in	216
section 2.2.1 and the approach to model a conditional probability using a cVAE in	217
section 2.2.2.	218
	219
	220
	221
	222
2.2.1 Hurdle model	223
	224
The statistical analysis employs a zero-inflated probabilistic model [16], more specif-	225
ically a Hurdle model [17], to capture the structural tendency of zero deviations in	226
procurement (i.e., on-time deliveries). The conditional distribution of a target variable	227
	228
	229
	230

231 y given input features x is defined as

$$232$$

$$233$$

$$234 \quad P(y | x) = \begin{cases} p_{d,0}(x) & y_d = 0, \\ (1 - p_{d,0}(x)) \cdot f(y | x), & y_d \neq 0, \end{cases} \quad (1)$$

$$235$$

$$236$$

$$237$$

$$238$$

239 where $p_{d,0}(x)$ denotes the probability of a structural zero, and $f(y | x)$ represents the
 240 conditional probability density function (PDF) of non-zero delivery deviations.
 241

242 2.2.2 Probabilistic modeling using a zero-inflated conditional 243 variational autoencoder

244 To model the conditional distribution $P(y | x)$ in the zero-inflated Hurdle model
 245 defined in equation (1), a cVAE [18] is employed. This generative model extends
 246 the standard variational autoencoder (VAE) [19] by conditioning both encoder and
 247 decoder on input features x , thereby enabling the approximation of complex, multi-
 248 modal distributions typical of procurement delivery deviations. A cVAE introduces
 249 a latent variable z , with an encoder $q_\phi(z | y, x)$ and a decoder $p_\theta(y | x, z)$. In the
 250 proposed architecture, the decoder represents non-zero delivery deviations through a
 251 Gaussian mixture:
 252

$$253 \quad f(y | x) := p_\theta(y | x, z) = \sum_{k=1}^K w_k(x, z) \mathcal{N}(y | \mu_k(x, z), \sigma_k^2(x, z)), \quad (2)$$

$$254$$

$$255$$

$$256$$

$$257$$

258 where the mixture weights satisfy $w_k(x, z) \geq 0$ and $\sum_{k=1}^K w_k(x, z) = 1$. The Gaussian
 259 mixture captures the empirical clustering of deviations into early (negative) and late
 260 (positive) deliveries, which cannot be represented adequately by a single Gaussian
 261 distribution. To incorporate the structural zeros of on-time deliveries, the decoder is
 262 extended with a second output head that estimates the logit of the zero probability
 263 $p_{d,0}(x)$. Thus, the decoder jointly produces (i) the parameters of the Gaussian mixture
 264

distribution for non-zero deviations, and (ii) the zero-inflation probability, consistent with the likelihood structure in equation (1).

3 Related work and necessity of the study

A systematic literature review (SLR) was conducted following the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines [20] to identify existing research on synthetic data generation for procurement planning in SMEs. The search, carried out on July 13, 2025, initially retrieved 10,794 articles. After filtering by publication year (post-2015), language (English or German), disciplinary categories (Engineering, Computer Science, Decision Sciences, Business, Management and Accounting, or Multidisciplinary), and open access availability, 1,087 articles remained. The AI-based ASReview tool [21] was then applied for efficient screening of the identified publications. Following the stopping criterion of 10 % consecutively irrelevant articles and an additional manual check of 20 papers, nine articles were identified as relevant. One was excluded due to access restrictions, leaving eight articles for full-text assessment. These were evaluated using the Joanna Briggs Institute (JBI) qualitative analysis and review instrument (QARI) methodology [22], which formed the evidence base for this review.

Recent research has explored various strategies to address imbalanced and sparse datasets in industrial settings. Several studies applied deep generative models to fault classification: for instance, Shon et al. [23] proposed a convolutional VAE for wafer defect classification in semiconductor manufacturing, while Karamti et al. [24] combined sparse coding with VAEs for machinery fault detection. These approaches demonstrate the potential of generative models to mitigate data imbalance, but they remain limited to sensor-rich production domains and do not extend to transactional procurement data. Another body of work has applied zero-inflated or Hurdle models to count data with excess zeros. Zhang et al. [25] used a Hurdle regression model

for defect data in steel manufacturing, and Workie and Azene [26] and Aragaw et al. [27] applied similar models in public health contexts. Akkas et al. [28] employed a zero-inflated negative binomial model to study product expiration in consumer goods retail. While these methods effectively capture sparsity, they focus on statistical analysis rather than synthetic data generation, and thus cannot directly support the training of AI-based procurement models. Hybrid approaches have also emerged that integrate statistical and neural network methods for monitoring non-linear processes. Tang et al. [29] developed a Gaussian mixture VAE for fault detection across multiple process modes, and Ibrahim et al. [30] combined deep learning with control chart techniques in high-yield manufacturing. These works highlight the value of combining interpretability with generative flexibility. However, their scope remains confined to process monitoring rather than generating procurement data.

In summary, while existing research provides strong foundations for handling imbalance, zero-inflation, and hybrid modeling, no study has yet addressed the generation of synthetic procurement data for SMEs. This study is the first to propose a zero-inflated Hurdle model with a cVAE, a modeling approach that can generate realistic synthetic data to both increase the data availability and improve the data quality in procurement planning, thereby enabling SMEs to develop robust decision-support models under realistic constraints.

4 Results and discussion

This section presents the results obtained by applying the research design within the structured framework of the CRISP-DM. The execution of the six CRISP-DM phases is summarized in figure 1, which illustrates how each phase was instantiated in the context of generating synthetic procurement data. The outputs of individual phases either provide inputs for subsequent steps or directly support the development of the

Phase	Step(s)	Output(s)
Business understanding	Define target variable	Delivery deviation y_d selected as target
Data understanding	Data analysis Correlation analysis	Distribution of y_d Correlation table
Data preparation	NaN removal and threshold filtering Outlier detection (MAD) Feature scaling (Z-Score for y_d , integer encoding for numbers)	Cleaned dataset Scaled inputs and outputs
Modeling	Embedding layers for supplier and article number Encoder network Decoder with zero-inflation and mixture-of-Gaussians	Trained cVAE model
Evaluation	Optimizer setup (AdamW) Beta scheduler Loss monitoring	Loss curves Histogram comparisons Wasserstein metric
Deployment	Sample latent vector z Generate synthetic delivery deviations $\tilde{y}_i$	Synthetic procurement data

Fig. 1: The overall CRISP-DM proposed in this article with phases, steps, and a definition of the outputs of each phase.

generative pipeline. All models were implemented and trained in the *Python* programming language [31] using the *JAX* [32] and the *Flax* [33] libraries, ensuring an efficient numerical computation and hardware acceleration on modern accelerators. The following subsections present the outcomes of each phase, including data characteristics, preprocessing steps, model training, and evaluation of synthetic data quality.

4.1 Business understanding

In the business understanding phase, the focus is on defining the objectives of the procurement planning tasks and translating business requirements into analytical targets. In this context, the primary objective is to forecast the delivery reliability in

order to support the operational planning and the supplier management under conditions of limited data availability. The dataset analyzed in this study originates from one German SME. Table A1 categorizes the company according to its key characteristics defined by Schuh and Stich [34]. The company provided structured procurement records covering multiple years of operations, with the earliest procurement record starting in 2018 until the latest procurement records from 2024. The dataset comprises several thousands of procurement records. The following attributes were available: *order number*, *supplier number*, *article number*, *planned delivery date*, *actual delivery date*, *ordered quantity*, and *delivered quantity*. From these attributes, the *delivery deviation* y_d was derived as the primary dependent variable:

$$y_d = y_{\text{actual}} - y_{\text{planned}} \in \mathbb{Z}, \quad (3)$$

where $y_d > 0$ indicates late deliveries, $y_d < 0$ indicates early deliveries, and $y_d = 0$ corresponds to on-time deliveries. Although multiple dependent variables could be constructed from the dataset (e.g., quantity deviation), the present analysis focuses on delivery deviation as a representative case.

4.2 Data understanding

4.2.1 Data analysis

Before the modeling and the correlation assessment, the distribution of delivery deviations was examined to understand the structural characteristics of the target variable y_d . The dataset consists of a total of 18,078 procurement records. Some procurement records have a *planned delivery date* but no *actual delivery date* since these dates were in the future when the dataset was created. The earliest *planned delivery date* is on May 18, 2020. The *actual delivery date* corresponding to this procurement record is May 24, 2023, indicating either a poor data quality or extreme delivery delays due

to, e.g., Covid-19. The latest procurement record that was actually delivered has a *planned delivery date* on March 24, 2025 and has an *actual delivery date* on April 4, 2024.

4.2.2 Correlation analysis

From the available features described in section 4.1, the *supplier number* and *article number* were considered further. The *order number* is not a relevant feature since this is a value set internally by the considered SME. A one-way analysis of variance (ANOVA) was conducted to assess the statistical association of the supplier and article number with the delivery deviation, y_d , and to, therefore, assess their predictive capability. Both predictors show highly significant effects on delivery deviation, with p-values well below the threshold of $\alpha = 0.05$. The F-statistic for article number ($F = 7,631.6231$) exceeds that for supplier number ($F = 4,535.1749$), indicating that article identity accounts for a greater share of the variance in delivery deviations. This difference suggests that the characteristics of individual articles (e.g., complexity, sourcing criticality) play a more systematic role in explaining the variation in delivery reliability than supplier identity alone. Nevertheless, the supplier number remains a critical predictor from a business perspective. Suppliers are not interchangeable: each one is responsible for a specific subset of articles and exhibits distinct reliability profiles, operational constraints, and logistical behaviors. These findings support the inclusion of both categorical predictors in subsequent modeling steps. Together, supplier and article identifiers enable the generative model to represent procurement-specific interactions realistically and provide a basis for transferable synthetic data generation in other SME settings.

507 4.3 Data preparation

508

509 4.3.1 Data cleaning

510

511 The data cleaning follows a structured process to ensure the quality and consistency
512 of the inputs prior to modeling:
513

514

515

516 1. Removal of missing values from key fields like *planned delivery date* and *actual*
517 *delivery date*.
518

519

520 2. Threshold-based filtering to exclude implausible values, such as delivery devia-
521 tions outside ± 180 days.
522

523 3. Robust outlier detection within the plausible range using the median absolute
524 deviation (MAD) method with a threshold multiplier of 2 [35].
525

526

527

528 The combined approach ensures that both gross anomalies and subtle statistical out-
529 liers were identified and removed. After the data cleaning, 13,464 procurement records
530 remained. Approximately 18 % of the records represent on-time deliveries ($y_d = 0$),
531 while late deliveries constitute the majority with 51 % ($y_d > 0$), and early deliver-
532 ies account for 30 % ($y_d < 0$). This imbalance highlights the asymmetric nature of
533 procurement performance, where late deliveries occur more frequently than early or on-
534 time deliveries. The presence of structural zeros, combined with distinct positive and
535 negative tails, motivates the use of a zero-inflated, mixture-based modeling approach.
536

537

538

539 4.3.2 Data partitioning

540

541 The dataset was partitioned into training (s_{train}), validation (s_{val}), and test (s_{test})
542 subsets using a stratified shuffle split (SSS) to enable a representative evaluation. The
543 size of each partition is given in table 1. The stratification variable was the categorical
544 delivery deviation label, encoding whether a delivery was on time (0), late (1), or
545 early (2). The stratification preserves the relative frequency of each delivery class
546
547
548
549
550
551
552

across all subsets, preventing minority categories (e.g., early deliveries) from being underrepresented. Table A2 shows the equally split data.

4.3.3 Feature scaling

After the data cleaning and partitioning, the feature scaling was applied to prepare the inputs for the model training and output interpretation. Two scaling procedures were applied depending on the variable type.

Target variable (delivery deviation).

The dependent variable y_d was standardized using Z-score normalization using the mean μ and standard deviation σ estimated from the *training dataset* to obtain the input variable $\tilde{y}_d$ for the cVAE. The same parameters were reused for validation and test data to prevent information leakage. Note that the raw on-time value $y_d = 0$ maps to a non-zero location in model space, i.e., $\tilde{y}^* = -\mu/\sigma$. To restore interpretability after sampling, the inverse Z-score transformation was applied.

Categorical identifiers (supplier and article number).

The supplier and article numbers are nominal categorical variables and were transformed using column-wise integer encoding. Each unique identifier is mapped to a contiguous index starting from zero (e.g., 1,012 $\rightarrow$ 0, 2,003 $\rightarrow$ 1, 3,055 $\rightarrow$ 2). This mapping is invertible and enables numerical embedding in the cVAE architecture. Since learned-only index mappings are created, the categorical scaler was fit on the full dataset (training, validation, and test). Using the full dataset ensures that all supplier and article identifiers encountered at inference can be reliably decoded. Both scaling procedures are implemented as reusable modules and are applied consistently during training, evaluation, and synthetic data generation.

599 4.4 Modeling

600
601 The modeling phase implements the probabilistic approach described in section 2
602 through a cVAE. The cVAE approximates the conditional distribution of delivery devi-
603 ations by combining a latent-variable generative process with a zero-inflated Gaussian
604 mixture decoder. This design enables the model to capture both structural zeros (on-
605 time deliveries) and multimodal non-zero deviations (early or late). The architecture
606 therefore integrates statistical rigor from zero-inflated modeling with the flexibility
607 of deep generative learning, representing the core methodological contribution of this
608 study. The model encodes categorical input features together with the observed target
609 into a latent representation, from which it decodes (i) the parameters of a Gaussian
610 mixture distribution for non-zero deviations, and (ii) the zero-inflation probability (see
611 equation (1)). Figure 2 illustrates the overall architecture.

620 4.4.1 Embedding layer

621
622 Each procurement record was indexed by i , with supplier number $x_{s,i} \in \{1, \dots, N_s - 1\}$
623 and article number $x_{a,i} \in \{1, \dots, N_a - 1\}$. Both categorical variables were mapped
624 into continuous embedding representations of shared dimension E according to
625 equation (4):

$$631 \quad e_{s,i} = \text{Embed}_s(x_{s,i}) \in \mathbb{R}^E, \quad e_{a,i} = \text{Embed}_a(x_{a,i}) \in \mathbb{R}^E. \quad (4)$$

632
633
634 Figure 3 illustrates the distribution of supplier-article pair frequencies in the training
635 dataset using a Pareto-style chart. The green bars show the number of occurrences for
636 the top 60 most frequent supplier-article pairs, sorted in descending order. The orange
637 line indicates the cumulative share of all procurement records across the entire dataset.
638
639 The plot highlights that the data are strongly imbalanced: while most supplier-article
640 pairs occur only rarely, a small subset of pairs dominates the transaction volume, with
641
642
643
644

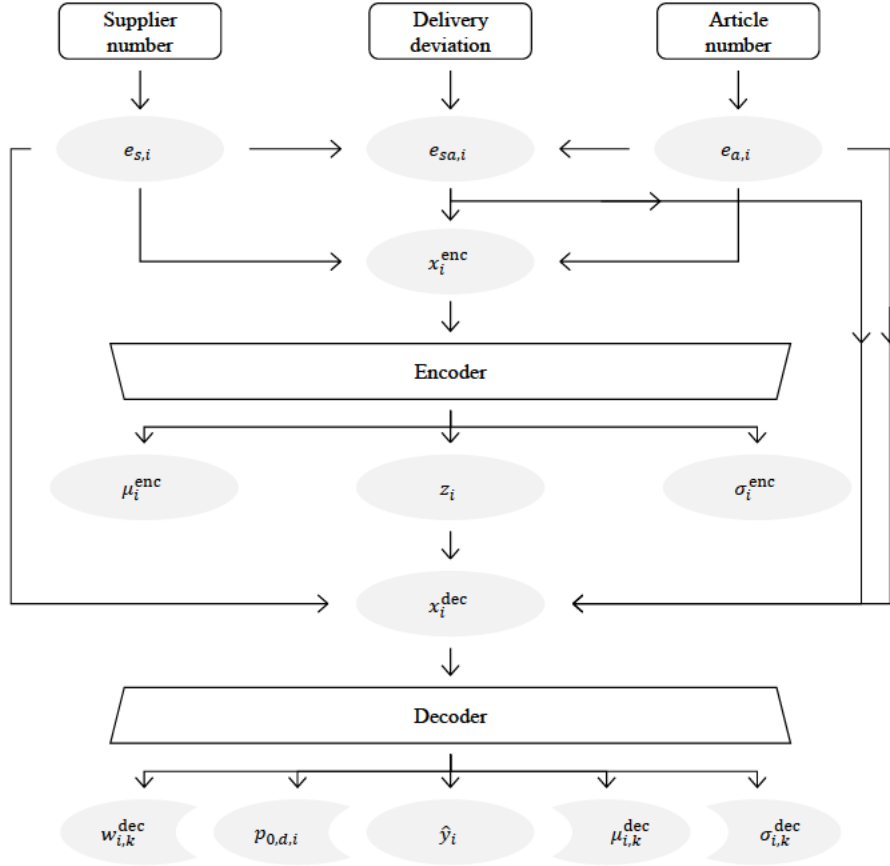


Fig. 2: Architecture of the cVAE with supplier and article embeddings, encoder, and decoder. The outputs of the encoder and decoder are explained in detail in the sections 4.4.1 to 4.4.3.

a relatively low number of pairs already accounting for the majority of occurrences. The first bar on the left, for example, corresponds to the supplier number 197 and article number 1,440, which were ordered 332 times. The purple dashed lines mark the point at which 80 % of the transaction volume is reached, both on the x-axis (number of distinct pairs needed) and the y-axis (cumulative share). This concentration pattern emphasizes that supplier-article relationships are sparse and specific, such that marginal embeddings of suppliers and articles alone cannot capture their joint behavior. To address this, the cVAE incorporates an *interaction term* in the conditioning

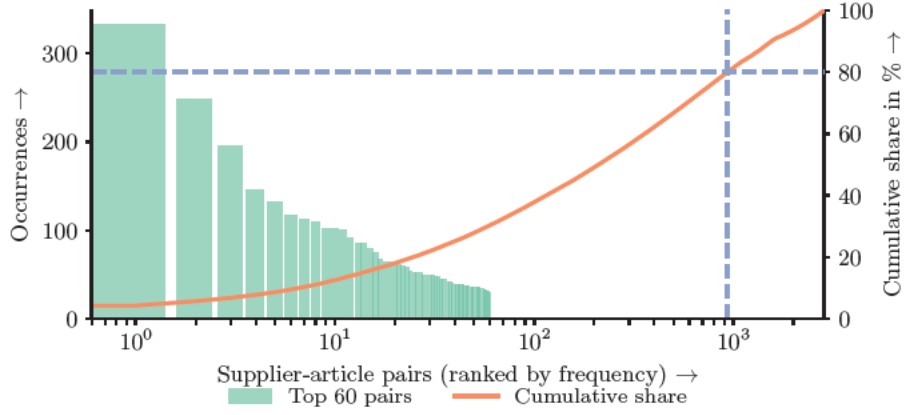


Fig. 3: Pareto plot of supplier-article pair frequencies. The bars show the number of occurrences for the top 60 most frequent pairs, while the orange line indicates the cumulative share of all transactions across the full dataset. A small fraction of pairs accounts for the majority of occurrences, reflecting the highly skewed and long-tailed nature of procurement interactions.

vector, defined in equation (5) as the element-wise product of supplier and article embeddings:

$$e_{sa,i} = e_{s,i} \odot e_{a,i} \in \mathbb{R}^E. \quad (5)$$

This interaction explicitly models multiplicative effects between suppliers and articles, allowing the latent representation to encode joint characteristics that are not recoverable from embeddings in isolation. Such an approach is particularly beneficial in the presence of the observed long-tailed co-occurrence patterns, as shown in figure 3.

To evaluate whether the learned embeddings capture a meaningful behavioral structure, a two-dimensional principal component analysis (PCA) [36] was applied to the final-layer embeddings of both supplier ($e_{s,i}$) and article ($e_{a,i}$) embeddings of the whole dataset. The resulting projections are shown in figure 4, with each point representing either a supplier or an article number. The x- and y-axis represent the two principal components (PCs). The supplier and article numbers were color-coded using a rule-based classification into three categories: *Reliable* (green), *frequently late* (orange), and *frequently early* (blue). The classification was based on the entire dataset

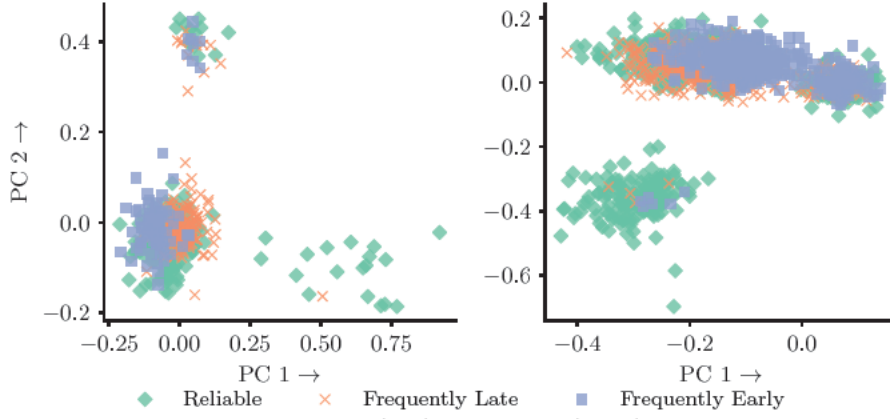


Fig. 4: Visualization of the supplier (left) and article (right) number embeddings using a PCA.

of procurement records (13,464) using two pragmatic thresholds chosen for interpretability: the average delivery deviation and the proportion of frequently late or early deliveries (defined using a ± 1 -day cutoff). Specifically, a supplier or article was classified as *frequently late* if more than 50 % of deliveries were late by at least one day or if the mean deviation exceeded 2 days. Conversely, a supplier or article was classified as *frequently early* if more than 50 % of deliveries were early by more than one day or if the mean deviation was below -2 days. Suppliers or articles not meeting either condition were labeled as *reliable*. The visualization reveals distinct clusters for suppliers and articles, particularly in the *frequently late* and *frequently early* categories, suggesting that the cVAE has internalized behavioral differences in delivery patterns into the embedding space. Overall, the separation of suppliers by delivery behavior supports the interpretability and validity of the learned embeddings, which are trained jointly with the cVAE and used as conditioning inputs in the decoder.

783 4.4.2 Encoder network

784

785 The encoder conditions on both categorical embeddings $e_{s,i}$; $e_{a,i}$, the interaction term
786 $e_{sa,i}$, and the scaled target value $\tilde{y}_i$:

788

789

$$790 \quad x_i^{\text{enc}} = [e_{s,i}; e_{a,i}; e_{sa,i}; \tilde{y}_i] \in \mathbb{R}^{3E+1}. \quad (6)$$

791

792

793 Equation (6) defines the covariate $x = x_i^{\text{enc}}$ that is used to model the encoder
794 $q_\phi(z_i \mid x_i^{\text{enc}})$ (see section 2.2.2). The latent variable z is obtained by applying the
795 reparameterization trick [19]:

796

797

$$798 \quad z_i = \mu_i^{\text{enc}} + \sigma_i^{\text{enc}} \odot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, I). \quad (7)$$

799

800

801 In equation (7), μ_i^{enc} and σ_i^{enc} are obtained by two distinct dense layers. In practice,
802 standard deviations may be lower-bounded for numerical stability.

803

804

805 4.4.3 Decoder network

806

807 As outlined in section 2.2.2, the decoder outputs the (i) parameters of the Gaussian
808 mixture distribution (see equation (2)) for non-zero deviations, and (ii) the zero-
809 inflation probability $p_{d,0,i}$ (see equation (1)). The decoder conditions on the latent
810 variable z , the categorical embeddings $e_{s,i}$; $e_{a,i}$, and their interaction $e_{sa,i}$:

811

812

$$813 \quad x_i^{\text{dec}} = [z_i; e_{s,i}; e_{a,i}; e_{sa,i}] \in \mathbb{R}^{D+3E}. \quad (8)$$

814

815

816 In the following, the decoder's outputs are explained in detail.

817

818

819

820

821

822

823

824

825

826

827

828

Parameters of the Gaussian mixture (i).

The non-structural part of the conditional density is the Gaussian mixture

$$p_{\theta}(\tilde{y}_i | z_i, x_i^{\text{dec}}) = \sum_{k=1}^K w_{i,k}^{\text{dec}} \mathcal{N}(\tilde{y}_i; \mu_{i,k}^{\text{dec}}, (\sigma_{i,k}^{\text{dec}})^2), \quad (9)$$

consistent with the general form in equation (2). In equation (9), $w_{i,k}^{\text{dec}}$, $\mu_{i,k}^{\text{dec}}$, and $\sigma_{i,k}^{\text{dec}}$ are obtained by three distinct dense layers.

Zero-inflation probability (ii).

As introduced in equation (1), the probability of a structural zero is modeled as $p_{d,0,i}$ and obtained by a fourth distinct dense layer in the decoder.

Complete Hurdle model likelihood.

Because the target is standardized, the raw on-time value $y_i = 0$ corresponds to $\tilde{y}_i^* = -\mu/\sigma$ (see section 4.3.3). Therefore, according to equation (1), the scaled Hurdle model can be formalized as

$$P(\tilde{y}_i | z_i, x_i^{\text{dec}}) = \begin{cases} p_{0,d,i}(x_i^{\text{dec}}) & \tilde{y}_i = \tilde{y}_i^*, \\ (1 - p_{0,d,i}(x_i^{\text{dec}})) \cdot p_{\theta}(\tilde{y}_i | z_i, x_i^{\text{dec}}), & \tilde{y}_i \neq \tilde{y}_i^*. \end{cases} \quad (10)$$

Practically, the condition in equation (10) is enforced by creating a boolean mask based on the unscaled target values y_d , i.e., $\mathbb{I}[y_i = 0]$ for the structural zero case and $1 - \mathbb{I}[y_i = 0]$ for the non-zero case. In summary, the decoder of the cVAE denoted by $p_{\theta}(\tilde{y}_i | z_i, x_i^{\text{dec}})$ represents the conditional PDF of the non-zero delivery deviations denoted by $f(y | x)$ in equation (1).

875 *Auxiliary point estimate.*

876

877 For evaluation purposes, a conditional mean estimate can be computed as

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

$$\hat{y}_i = (1 - p_{0,d,i}) \sum_{k=1}^K w_{i,k}^{\text{dec}} \mu_{i,k}^{\text{dec}}. \quad (11)$$

This estimate is not part of the generative process but is useful for forecasting metrics.

4.4.4 Loss terms

The overall training objective combines multiple components, each targeting a different aspect of the learning problem. Specifically, the model is trained using (i) a Kullback–Leibler (KL) divergence to regularize the latent space ($\mathcal{L}_{\text{KL}}$), (ii) a weighted binary cross entropy (BCE) loss ($\mathcal{L}_{\text{BCE}}$) to capture structural zeros $p_{d,0,i}$, (iii) a negative log likelihood (NLL) over a mixture-of-Gaussians for non-zero deviations ($\mathcal{L}_{\text{NLL}}$), (iv) a mean squared error (MSE) for auxiliary point estimation ($\mathcal{L}_{\text{MSE}}$), and (v) additional penalties on skewness ($\mathcal{L}_{\text{skew}}$) and kurtosis ($\mathcal{L}_{\text{kurt}}$) to align higher-order moments with empirical data. Together, these components enable the cVAE to learn both the structural and stochastic properties of the delivery deviations. The composite objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{BCE}} + \lambda_{\text{point}} \mathcal{L}_{\text{MSE}} + \beta_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{skew}} \mathcal{L}_{\text{skew}} + \lambda_{\text{kurt}} \mathcal{L}_{\text{kurt}}, \quad (12)$$

where λ_{point} , β_{KL} , λ_{skew} , and λ_{kurt} are hyperparameters. The model parameters ϕ and θ are optimized by minimizing the loss function using gradient descent. The formulation in equation (12) extends standard variational inference by integrating additional supervision and distributional regularization, ensuring that the model captures structural zeros, represents multimodal deviations, and generates synthetic data aligned with real-world delivery patterns.

4.5 Evaluation

4.5.1 Hyperparameter optimization

The hyperparameters of the cVAE model were optimized using the *Optuna* [37] framework, designed for automated hyperparameter search based on a Bayesian optimization algorithm. The search space included architectural parameters (e.g., embedding size, latent dimensionality, number of layers), training-related parameters (e.g., learning rate, dropout rate, batch size), and weighting coefficients for the different loss terms. Table A3 outlines the whole search space of the optimization procedure. The model training employed the *AdamW* optimizer [38] using a weight decay coefficient of 10^{-5} . To stabilize training and improve latent space utilization, a linear KL-annealing schedule was applied: the weight β_{KL} was gradually increased from zero to a maximum value of $\beta_{\text{max}} = 10^{-2}$ over the first $w_f \cdot T$ training epochs, where w_f denotes the warm-up fraction and T the total number of epochs. After the warm-up phase, β_{KL} remained constant for the remainder of the training. This scheduling strategy prevented the posterior collapse and ensured the effective use of the latent space during the optimization. The model was trained for a maximum of 100 epochs with an early stopping patience of 25 epochs. The multi-objective optimization included minimizing the MSE of the non-zero model (p_θ) on the test data, maximizing the precision of the zero-inflated model ($p_{0,d}$) on the test data, and minimizing the overall validation loss ($\mathcal{L}_{\text{sval}}$) during training. After conducting 1,000 optimization trials, the best-performing configuration was selected according to the Pareto front of the multi-objective optimization. A high precision for the zero model ensures that predicted on-time deliveries correspond to actual punctual deliveries, which is critical for maintaining operational trust in procurement planning. At the same time, minimizing the MSE of the non-zero predictions reduces the forecast error for early and late deliveries, directly supporting the production scheduling and supplier evaluation. Incorporating the overall validation loss into

Table 1: Overview of the optimal set of hyperparameters obtained following the optimization procedure using *Optuna* [37]

Hyperparameter	Symbol	Value
Embedding dimension	E	1024
Latent dimension	D	32
Number of epochs	T	40
Warmup fraction	w_f	0.20
Batch size	B	64
KL divergence weight	β_{KL}	0.0001
Encoder dropout rate	p^{enc}	0.00
Decoder dropout rate	p^{dec}	0.00
Encoder hidden dimensions	H^{enc}	128, 64, 32
Decoder hidden dimensions	H^{dec}	128, 64, 32
Learning rate	η	0.0001
Decoder Gaussian heads	K	5
MSE point loss weight	λ_{point}	1.20
Training split	s_{train}	0.66
Validation split	s_{val}	0.10
Test split	s_{test}	0.24
Skewness loss weight	λ_{skew}	0.10
Kurtosis loss weight	λ_{kurt}	0.01

the optimization further balances these objectives by favoring configurations that generalize well across both structural zeros and non-zero deviations. Table 1 summarizes the final set of hyperparameters obtained through this procedure. These values were then used for all subsequent training and evaluation steps. Figure A1 shows a heatmap of the relative importance of each hyperparameter for each of the three optimization objectives.

4.5.2 Training

The hyperparameters described in section 4.5.1 were used to train and evaluate the cVAE. Figure 5 reports the evolution of the composite loss' components ($\mathcal{L}_{\text{NLL}}, \mathcal{L}_{\text{BCE}}, \mathcal{L}_{\text{MSE}}, \mathcal{L}_{\text{KL}}, \mathcal{L}_{\text{skew}}, \mathcal{L}_{\text{kurt}}$) over epochs. The mixture NLL ($\mathcal{L}_{\text{NLL}}$) decreases steadily, indicating an effective latent-space utilization without a posterior collapse. The BCE term stabilizes as the zero head calibrates around the empirical on-time frequency, whereas the mixture NLL continues to decrease at a slower rate, reflecting the difficulty of fitting the heavier positive tail of the delay distribution.

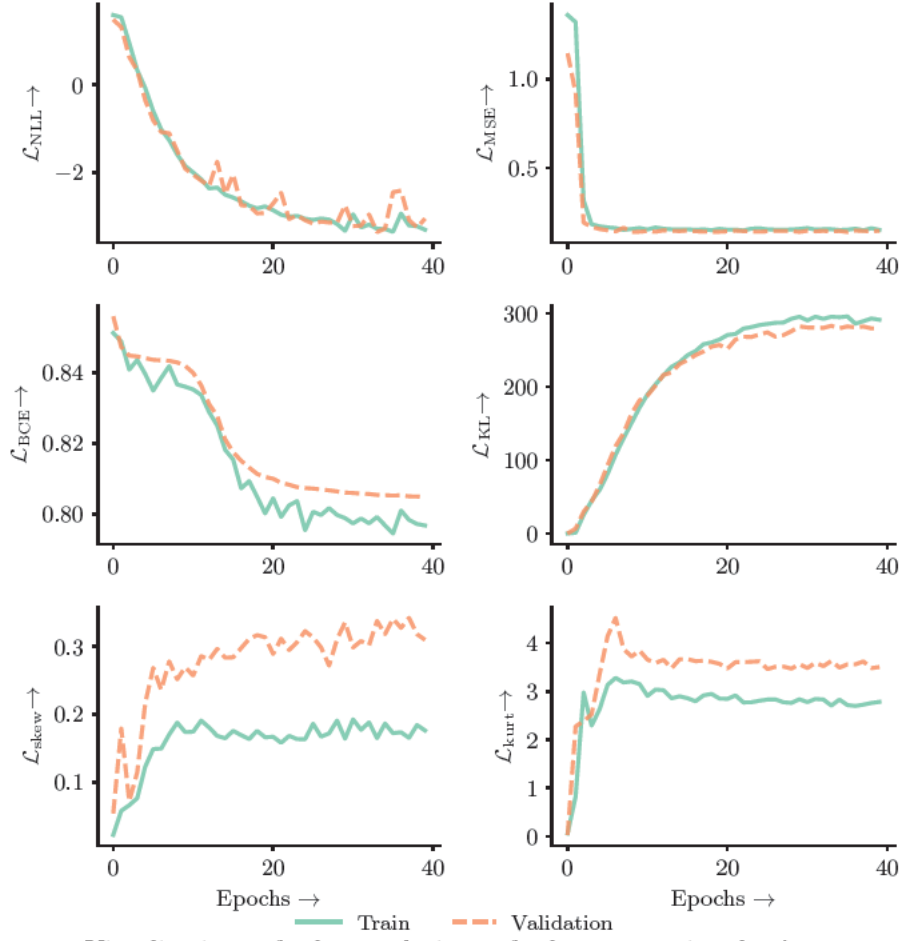


Fig. 5: Visualization of the evolution of the composite loss' components ($\mathcal{L}_{\text{NLL}}$, $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{skew}}$, $\mathcal{L}_{\text{kurt}}$) over epochs.

4.5.3 Zero-inflated component

The performance of the zero-inflated component was evaluated using the test set, focusing on its ability to correctly identify structural zeros in the delivery deviation variable y_d . Following the probabilistic formulation in equation (10), the decoder estimates a structural-zero probability $p_{0,d,i}$, which governs whether a delivery is predicted to be exactly on time, and allocates the remaining mass to the conditional distribution of non-zero deviations. For evaluation, the deliveries with deviations $|y_d| \leq \varepsilon$ were

labeled as structural zeros, while all the others were treated as non-zeros. The predicted probabilities $p_{0,d,i}$ were thresholded at 0.5 to obtain binary decisions. The classification performance was quantified using established metrics from *scikit-learn* [39]: balanced accuracy, recall, precision, F_β -score (with $\beta = 0.5$), and the receiver operating characteristic area under the curve (ROC AUC). The $F_{\beta=0.5}$ score emphasizes precision over recall. In addition, the corresponding confusion matrix was computed to provide a detailed account of true positives, false positives, true negatives, and false negatives. The confusion matrix (left image in figure 6) reveals a clear trade-off in the zero-inflated component’s classification performance. Focusing on non-structural zeros (deliveries with deviation $|y_d| > \varepsilon$), the model achieves high precision (74.41 %), ensuring that most predicted deviations are correct, but suffers from limited recall (47.43 %). Specifically, the 178 false negatives indicate a substantial share of actual deviations is incorrectly classified as structural zeros. While the $F_{0.5}$ -score (62.86 %) reflects this high precision, the overall balanced accuracy (56.31 %) and ROC AUC (57.52 %) suggest only marginal performance, a typical outcome in highly imbalanced settings. In summary, the component provides reliable detection for its predictions of structural zeros but requires future refinements, such as resampling strategies or cost-sensitive training, to improve sensitivity and better balance the precision-recall trade-off.

4.5.4 Non-zero component

The encoder’s latent representations were analyzed to assess whether the model captures the structure of non-zero delivery deviations. A PCA of the posterior means $\{\mu_i^{\text{enc}}\}_{i=1}^N \subset \mathbb{R}^D$ revealed a clear one-dimensional ordering of the target variable $\tilde{y}_{d,i}$. As shown in the right image in figure 6, early deliveries (large negative $\tilde{y}_{d,i}$, shown in dark purple) occupy one end of the first principal component (PC 1), whereas late deliveries (large positive $\tilde{y}_{d,i}$, bright yellow) are concentrated at the opposite end. The familiar *horseshoe* pattern arises because PC 1 captures the dominant variance in

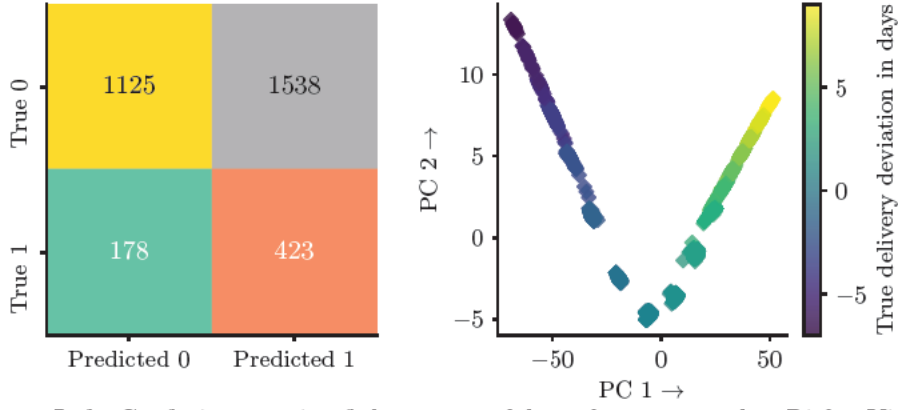


Fig. 6: Left: Confusion matrix of the zero model on the test samples; Right: Visualization of the latent space using a PCA.

delivery deviations, while PC 2 encodes a secondary curvature. The near-monotonic color gradient along PC 1 demonstrates that the encoder has learned a smooth latent manifold in which small movements correspond to gradual changes in predicted delivery delays. This structured latent space facilitates the interpolation, the clustering of delivery behaviors, and the anomaly detection. Furthermore, the absence of the latent collapse confirms that the cVAE model maintains a sufficient representational capacity.

The empirical fidelity of this component was examined by comparing predicted and actual non-zero deviations on the test set. Figure 7 illustrates the predicted histogram, while the quantitative similarity was assessed using the first Wasserstein distance W_1 [40]. The histograms in figure 7 illustrate the extent to which the cVAE model reproduces the empirical distribution of non-zero delivery deviations. For early deliveries (left panel), the generated samples $\tilde{y}_d < 0$ closely match the empirical distribution $y_d < 0$, capturing both the central tendency and the decay toward larger negative deviations. The alignment suggests that the model effectively represents the relatively compact distribution of early arrivals. For late deliveries (right panel), the generated distribution also follows the empirical trend but underestimates the frequency of extreme delays, resulting in lighter tails compared to the observed data. This difference reflects the greater variability and heavier-tailed structure of late deliveries, which

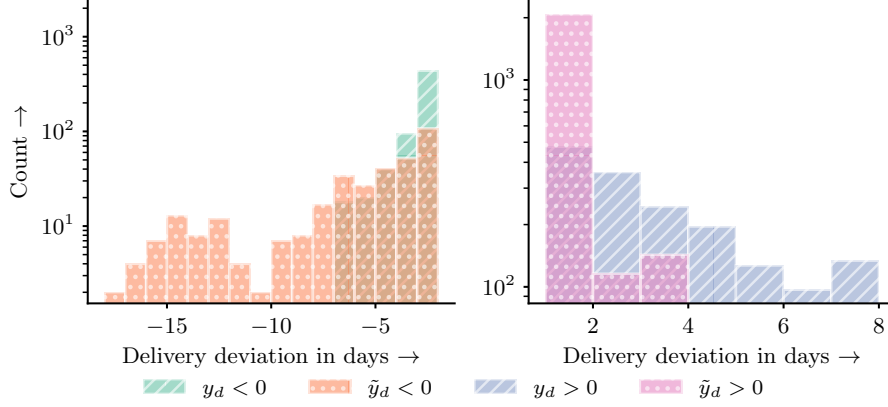


Fig. 7: Comparison of the non-zero delivery deviations, shown separately for early deliveries (negative deviations, left panel) and late deliveries (positive deviations, right panel). The count is plotted on a logarithmic scale to emphasize the tail behavior.

are more difficult to approximate with a finite Gaussian mixture. Overall, the figure confirms that the proposed model reproduces the main asymmetries of procurement deviations – more concentrated early arrivals and more dispersed late delays – while highlighting a scope for further improvement in capturing long positive tails.

Across all non-zero deviations, the model achieved $W_1 = 1.92$ days, indicating that the predicted distribution would need to be shifted by approximately 2 days on average to align with the empirical distribution. A disaggregated analysis showed that early deliveries (negative deviations) were reproduced more accurately ($W_1 = 1.84$ days) than late deliveries ($W_1 = 1.93$ days). This asymmetry reflects the empirical distribution, where late deliveries exhibit greater variability and heavier tails, which are inherently more difficult to approximate. For completeness, the unconditional Wasserstein distance, which combines the structural zero mass with the non-zero deviations, was $W_1 = 1.62$ days. Table 2 summarizes the Wasserstein metrics.

These results confirm that the non-zero component of the model captures both the general structure and the asymmetry of delivery deviations. Nonetheless, some deviations remain challenging to reproduce, particularly in the long-tailed distribution of late deliveries. Overall, the evaluation supports the model’s ability to generate realistic

Table 2: Summary of the Wasserstein metrics.

Component	Unit	Value
Proportion of on-time deliveries (true data)		0.18
Proportion of on-time deliveries (model prediction)		0.09
Difference in on-time delivery proportion (pred - true)		-0.08
Wasserstein distance (all deliveries)	Days	1.62
Wasserstein distance (non-zero deviations only)	Days	1.92
Wasserstein distance (late deliveries only)	Days	1.93
Wasserstein distance (early deliveries only)	Days	1.84
Count of non-zero deliveries in true data		2663.00
Count of non-zero deliveries in model predictions		2950.00

non-zero deviations, which are essential for the generation of synthetic procurement datasets that preserve behavioral heterogeneity.

Following the training and evaluation, the decoder of the cVAE can be deployed to generate realistic synthetic delivery deviations for arbitrary supplier-article combinations. In this setting, the generative process samples from the learned conditional distribution $p_\theta(\tilde{y}_i | z_i, x_i^{\text{dec}})$, where the conditioning vector x_i^{dec} contains the supplier and article numbers, the interaction term, and the output of the encoder’s latent vector (see equation (8)). The simulation of new delivery observations proceeds by first sampling a latent vector $z_i \sim \mathcal{N}(0, I)$, which captures stochastic variability encoded in the latent space. The decoder outputs four relevant quantities: the structural zero probability $p_{d,0,i}$, the mixture weights $w_{i,k}^{\text{dec}}$, the component means $\mu_{i,k}^{\text{dec}}$, and the corresponding standard deviations $\sigma_{i,k}^{\text{dec}}$ (see section 4.4.3). A Bernoulli trial is conducted on $p_{d,0,i}$ to determine whether an on-time delivery (structural zero) occurs. For non-zero outcomes, the process involves a two-step stochastic draw from the Gaussian mixture. First, a single mixture component $k \in \{1, \dots, K\}$ is drawn with probability $w_{i,k}^{\text{dec}}$. Second, a delivery deviation value is sampled from the Gaussian distribution

1243 corresponding to that chosen component k :

1244

1245

1246

1247

1248

1249

1250

$$\tilde{y}_d \sim \begin{cases} 0, & \text{with probability } p_{d,0}, \\ \mathcal{N}\left(\mu_{i,k}^{\text{dec}}, \left(\sigma_{i,k}^{\text{dec}}\right)^2\right), & \text{with probability } 1 - p_{d,0}. \end{cases} \quad (13)$$

1251

1252 The sampling procedure ensures that the output preserves the complexity of the

1253 full mixture distribution, as the component k is stochastically selected based on its

1254

1255 learned weight $w_{i,k}^{\text{dec}}$. Finally, the sampled values are inverse-transformed according

1256

1257 to section 4.3.3 to restore the original measurement scale. The deployment step thus

1258

1259 enables the generation of synthetic procurement datasets that can be directly inte-

1260

1261 grated into downstream applications such as scenario-based planning, supplier risk

1262

1263 assessment, and the augmentation of training data for predictive decision-support

1264

1265 models. The continuous improvement of the model through retraining on new data

1266

1267 ensures that the generated datasets remain relevant and reflective of evolving procure-

1268

1269 ment dynamics. Table 3 shows an example of generated procurement data with the

1270

1271 trained cVAE model.

1272

1273 The resulting synthetic datasets can thus be used in several concrete procurement

1274

1275 use cases. First, they enable the training of forecasting models that predict the delivery

1276

1277 reliability at the level of specific supplier-article combinations, allowing SMEs to proac-

1278

1279 tively adapt production schedules in case of expected delays. Second, they provide

1280

1281 a basis for scenario-based supplier risk simulations, where managers can explore the

1282

1283 impact of repeated late deliveries or supply disruptions without relying solely on scarce

1284

1285 historical data. Third, the datasets enrich training material for procurement-focused

1286

1287 decision-support models, which can be integrated into ERP systems or dashboards to

1288

1289 generate automatic alerts, rank suppliers by reliability, or recommend order reschedul-

1290

1291 ing. Together, these use cases illustrate how the deployment of the proposed cVAE

1292

1293 model supports an actionable decision-making in procurement contexts.

1294

Table 3: Example of synthetically generated procurement data using the trained cVAE model.

Order number	Supplier number	Article number	Planned delivery date	Actual delivery date
1	91	1471	March 5, 2026	March 6, 2026
2	68	1399	May 15, 2026	May 11, 2026
3	29	1483	May 22, 2026	May 21, 2026
4	154	1510	May 27, 2026	May 28, 2026
5	339	1475	June 1, 2026	June 2, 2026
6	348	1246	June 8, 2026	June 11, 2026
7	341	1485	September 8, 2026	August 24, 2026
8	429	1667	November 18, 2026	November 15, 2026
9	197	1440	December 2, 2026	December 3, 2026
10	237	1471	December 7, 2026	December 8, 2026

5 Conclusion

5.1 Limitations

The proposed cVAE model offers a structured and replicable approach for generating synthetic procurement data in SMEs. However, several limitations must be acknowledged. First, the characterization of late deliveries is company-specific. Different organizations apply varying thresholds for defining a delivery as delayed (for example, only after a deviation of more than five days), which affects both model training and interpretation. Second, the current implementation is restricted to suppliers that are already present in the training data. While this limits the applicability to entirely new suppliers, it is important to note that, in practice, the suppliers represented in the model typically cover more than 80 % of the procurement volume in SMEs, meaning that occasional one-off suppliers have a negligible operational impact (see figure 3). Nonetheless, the lack of generalizability to new suppliers may become more relevant in highly dynamic sourcing contexts. Third, the zero-inflation component, although effective in modeling the structural mass of on-time deliveries, may

not be fully calibrated for rare non-zero deviations. In highly imbalanced datasets, this can lead to underestimation of infrequent but operationally significant deviations. Furthermore, the non-zero component assumes a mixture of Gaussian distributions, which may not fully capture the discrete and heavy-tailed nature of delivery deviations, especially in the extremes.

5.2 Future research

Future work should address these limitations along several directions. The feature scope could be expanded to incorporate richer contextual information beyond supplier and article numbers, such as planned lead times, order quantities, transportation modes, contractual terms, or production calendar effects. Including such operational information would likely enhance the realism and causal validity of the generated data, and could also improve generalization to unseen suppliers when paired with attribute-based embeddings. Another scope is the methodological improvement of the zero-inflation handling, for example by exploring cost-sensitive training strategies, calibrated probability outputs, or alternative discrete and tail-aware likelihood functions that better capture the asymmetries observed in practice. The approach is transferable and can be applied to alternative procurement-related targets in future studies. From an application perspective, research should also focus on integrating the synthetic data generation pipeline into ERP systems or decision-support dashboards, enabling real-time scenario simulation and risk assessment in procurement planning. Finally, future studies should systematically quantify the added value of synthetic procurement data in practical applications, for example by measuring its impact on forecasting accuracy, risk mitigation, or decision-support quality in SME environments.

6 Summary	1381
	1382
This article presents a model to generate synthetic procurement data for SMEs, aiming to mitigate data scarcity challenges that hinder the use of AI in procurement planning. Focusing on the delivery deviations, the study presents a Hurdle model operationalized using a cVAE model, combining the probability of a structural zero (on-time delivery) $p_{0,d}(x_i^{\text{dec}})$ with a conditional PDF for non-zero delivery deviations $p_{\theta}(\tilde{y}_i z_i, x_i^{\text{dec}})$. This cVAE model captures both the high frequency of on-time deliveries and the bimodal distribution of early or late deliveries. The research design follows the CRISP-DM framework, covering data cleaning, feature encoding, model training, and synthetic data generation. Results from a German SME demonstrate that the cVAE can effectively replicate real-world patterns and improve data availability and quality for decision-support tools. The study also validates its latent representations through PCA and discusses practical deployment for SMEs.	1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404
Declarations	1405
	1406
Ethics approval and consent to participate	1407
	1408
Not applicable.	1409 1410 1411 1412
Consent for publication	1413
	1414
Not applicable.	1415 1416 1417
Availability of data and materials	1418
	1419
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request. The developed code and the associated guidelines can be accessed via the following repository: (Author's comment: The link will be inserted when the final submission is made)	1420 1421 1422 1423 1424 1425 1426

1427 **Competing interests**

1428

1429 The authors declare that they have no competing interests.

1430

1431

1432 **Funding**

1433

1434 The IGF project "KIbaroP", project number 01IF23054N, is supported by the German
1435

1436 Federal Ministry for Economic Affairs and Climate Action on the basis of a decision

1437

1438 by the German Bundestag. The authors appreciate the support.

1439

1440

1441 **Authors' contributions**

1442

1443 **Julian Stang:** Conceptualization, Data Curation, Formal Analysis, Funding acqui-
1444 tion, Investigation, Methodology, Project administration, Resources, Software, Super-
1445

1446 vision, Validation, Visualization, Writing – original draft, Writing – review & editing.

1447

1448 **Jannik Stoever:** Resources, Software, Visualization, Writing – review & editing.

1449

1450 **Daniel Schnitzler:** Resources, Software, Validation, Visualization, Writing – review

1451 & editing. **Jan Joppien:** Funding acquisition, Project administration, Validation,

1452

1453 Writing – review & editing. **Marc Wegmann:** Writing – review & editing. **Christina**

1454

1455 **Reuter:** Project administration, Resources, Supervision, Writing – review & editing.

1456

1457

1458 **Acknowledgements**

1459

1460 We would like to thank the participating company for providing the necessary research

1461 data. We would also like to thank the anonymous reviewers who provided a number

1462

1463 of important comments that helped improve the overall quality of this manuscript.

1464

1465

1466 **Declaration of generative AI and AI-assisted technologies in**

1467

1468 **the writing process**

1469

1470 During the preparation of this work, the author(s) used ChatGPT (version 4o) to

1471

1472 assist in improving the clarity, style, and readability of the manuscript. The use of the

tool was limited to language editing. The author(s) reviewed and edited all AI-assisted suggestions and take full responsibility for the content and conclusions presented in this publication.

Appendix A Supplementary material

Table A1: Characterization of the participating German SME.

Characteristic	Categorization
Production concept	Make-to-order, assemble-to-order
Duration of the cooperation	One-time, temporary, seasonal, constant
Replaceability	Flexible with low switching costs
Number of suppliers per product	10 to 100
Number of production sites	1 to 5
Number of production lines	1 to 5
Order trigger type	Ordering with individual orders, ordering with blanket orders, customer-anonymous pre-production, customer-related final production
Product range	According to customer specification
Product structure	Multi-part products with a complex structure
Procurement type	Large-scale external procurement
Stockpiling	Order items at lower and higher structural levels
Production type	Individual and small series

Table A2: The SSS ensures that, despite the different proportions of training, validation, and test data, the proportion of zeros, positive values, and negative values in the split data remains consistent.

Data split	Share of total data	Share of zeros	Share of positive values	Share of negative values
Train	65.99 %	18.24 %	51.40 %	30.35 %
Validation	10.00 %	18.26 %	51.37 %	30.36 %
Test	23.99 %	18.26 %	51.40 %	30.33 %

List of Abbreviations

AI artificial intelligence

Table A3: Overview of the search space of the optimization.

Parameter	Symbol	Type	Search Space
Embedding dimension	E	Categorical	64, 128, 256, 512, 1024
Latent dimension	D	Categorical	32, 64, 128, 256, 512
Batch size	B	Categorical	64, 128, 256, 512
KL divergence weight	β_{KL}	Categorical (Log-Spaced)	1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3, 1e-2
Encoder dropout rate	p^{enc}	Categorical	0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5
Decoder dropout rate	p^{dec}	Categorical	0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5
Encoder hidden dimensions	H^{enc}	Categorical	(16), (64, 32), (128, 64, 32), (256, 128, 64), (512, 256, 128)
Decoder hidden dimensions	H^{dec}	Categorical	(16), (64, 32), (128, 64, 32), (256, 128, 64), (512, 256, 128)
Learning rate	η	Categorical (Log-Spaced)	1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3, 1e-2
Decoder Gaussian heads	K	Categorical	2, 3, 4, 5
MSE point loss weight	λ_{point}	Categorical (Log Grid)	0.1, 0.2, 0.3, 0.5, 0.8, 1.0, 1.2, 1.5
Skewness loss weight	λ_{skew}	Categorical (Log Grid)	1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2, 1e-1
Kurtosis loss weight	λ_{kurt}	Categorical (Log Grid)	1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2, 1e-1

ANOVA analysis of variance

BCE binary cross entropy

CRISP-DM cross industry standard process for data mining

cVAE conditional variational autoencoder

ERP enterprise resource planning

JB Joanna Briggs Institute

KL Kullback–Leibler

MAD median absolute deviation

MSE mean squared error

NLL negative log likelihood

PC principal component

PCA principal component analysis

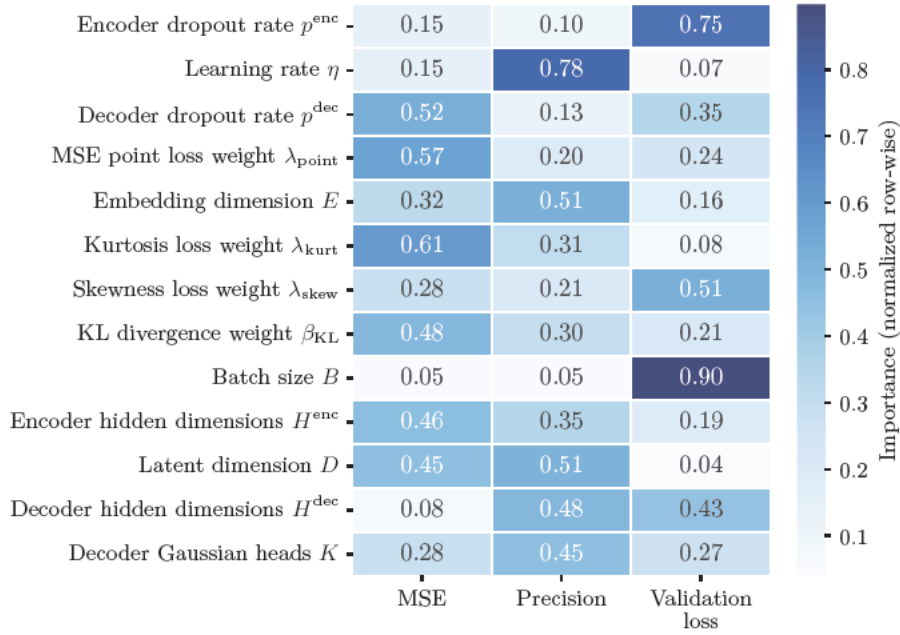


Fig. A1: Visualization of the importance of the hyperparameters outlined in table 1.

PDF probability density function

PRISMA preferred reporting items for systematic reviews and meta-analyses

QARI qualitative analysis and review instrument

ROC AUC receiver operating characteristic area under the curve

SLR systematic literature review

SME small- and medium-sized enterprise

SSS stratified shuffle split

VAE variational autoencoder

References

- [1] Eversheim W. Organisation in der Produktionstechnik: [engl.: Organization in Production Engineering]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1996.

1611 [2] Juliano L, Bender JP, Russell R, Hughes D.: Managing Raw-Material Volatility.
1612 Boston Consulting Group. Accessed 4 August 2025. Available from: [https://](https://www.bcg.com/publications/2022/managing-raw-material-supply-volatility)
1613 www.bcg.com/publications/2022/managing-raw-material-supply-volatility.
1614
1615
1616
1617 [3] Kilpatrick J, Berckman L, Faver AD, Hardin K, Sloane M.: Optimizing the
1618 balance between performance and cost. Deloitte. Accessed 4 August 2025. Avail-
1619 able from: [https://www.deloitte.com/us/en/insights/industry/manufacturing-](https://www.deloitte.com/us/en/insights/industry/manufacturing-industrial-products/global-supply-chain-resilience-amid-disruptions.html)
1620 [industrial-products/global-supply-chain-resilience-amid-disruptions.html](https://www.deloitte.com/us/en/insights/industry/manufacturing-industrial-products/global-supply-chain-resilience-amid-disruptions.html).
1621
1622
1623
1624 [4] Iftikhar A, Ali I, Zhan Y, Stevenson M, Tarba SY. Firms' strategic responses to
1625 rising uncertainty amid ongoing geopolitical tensions: The synergistic mediating
1626 role of network capability and innovation ambidexterity. Transportation Research
1627 Part E: Logistics and Transportation Review. 2025;199:104146. [https://doi.org/](https://doi.org/10.1016/j.tre.2025.104146)
1628 [10.1016/j.tre.2025.104146](https://doi.org/10.1016/j.tre.2025.104146).
1629
1630
1631
1632
1633 [5] Ojochide PF. Investigating How Procurement Strategies Influence the Perform-
1634 mance of Small and Medium-Sized Enterprises (SMEs). International Journal
1635 of Advanced Research in Science and Technology. 2023;12(11):1160–1167. [https:](https://doi.org/10.62226/ijarst20231120)
1636 [//doi.org/10.62226/ijarst20231120](https://doi.org/10.62226/ijarst20231120).
1637
1638
1639
1640
1641 [6] Chawla D, Fletcher R, Tollens S.: How medium-size enterprises can better
1642 manage sourcing. McKinsey & Company. Accessed 4 August 2025. Avail-
1643 able from: [https://www.mckinsey.com/industries/industrials-and-electronics/](https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/how-medium-size-enterprises-can-better-manage-sourcing#/)
1644 [our-insights/how-medium-size-enterprises-can-better-manage-sourcing#](https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/how-medium-size-enterprises-can-better-manage-sourcing#/)/.
1645
1646
1647
1648 [7] Amar J, Rahimi S, Surak Z, von Bismarck N.: AI-driven operations forecast-
1649 ing in data-light environments. McKinsey & Company. Accessed 15 June
1650 2025. Available from: [https://www.mckinsey.com/capabilities/operations/our-](https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments#/)
1651 [insights/ai-driven-operations-forecasting-in-data-light-environments#](https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments#/)/.
1652
1653
1654
1655
1656

- [8] Dombrowski U, Wullbrandt J, Krenkel P. Industrie 4.0 in production ramp-up management. *Procedia Manufacturing*. 2018;17:1015–1022. <https://doi.org/10.1016/j.promfg.2018.10.085>. 1657
1658
1659
1660
1661
1662
- [9] Surbier L, Alpan G, Blanco E. A comparative study on production ramp-up: state-of-the-art and new challenges. *Production Planning & Control*. 2014;25(15):1264–1286. <https://doi.org/10.1080/09537287.2013.817624>. 1663
1664
1665
1666
1667
1668
- [10] House CH, Price RL.: The Return Map: Tracking Product Teams. *Harvard Business Review*. Accessed 2 September 2025. Available from: <https://hbr.org/1991/01/the-return-map-tracking-product-teams>. 1669
1670
1671
1672
1673
1674
- [11] Wirth R, Hipp J. CRISP-DM: Towards a Standard Process Model for Data Mining. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*. vol. 4. Blackpool, Lancashire; 2000. p. 22–39. Available from: <https://api.semanticscholar.org/CorpusID:1211505>. 1675
1676
1677
1678
1679
1680
1681
1682
1683
- [12] Döring N. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2023. 1684
1685
1686
1687
- [13] Dombrowski U, Hanke T. Lean Ramp-up. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*. 2009;104(10):877–883. <https://doi.org/10.3139/104.110161>. 1688
1689
1690
1691
- [14] Terwiesch C, Bohn R, Chea K. International product transfer and production ramp-up: a case study from the data storage industry. *R&D Management*. 2001;31(4):435–451. <https://doi.org/10.1111/1467-9310.00230>. 1692
1693
1694
1695
1696
1697
- [15] Stang J, Zaeh MF. A Concept For Data-Driven Decision-Making During The Production Ramp-Up To Increase Resilience In Value Networks. In: Herberger D, Hübner M, editors. *Proceedings of the Conference on Production Systems and* 1698
1699
1700
1701
1702

Logistics: CPSL 2024. vol. 1. Hannover: Hannover : publish-Ing; 2024. p. 204–214.

[16] Hilbe JM. Modeling Count Data. Cambridge University Press; 2014.

[17] Cragg JG. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*. 1971;39(5):829. <https://doi.org/10.2307/1909582>.

[18] Zhang C, Barbano R, Jin B. Conditional Variational Autoencoder for Learned Image Reconstruction. *arXiv*. 2021;p. 1–22. <https://doi.org/10.48550/arXiv.2110.11681>.

[19] Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: Yoshua Bengio, Yann LeCun, editors. 2nd International Conference on Learning Representations, ICLR 2014; 2014. p. 1–14.

[20] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed)*. 2021;372:n71. <https://doi.org/10.1136/bmj.n71>.

[21] van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*. 2021;3(2):125–133. <https://doi.org/10.1038/s42256-020-00287-7>.

[22] Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. *JBIC Evidence Implementation*. 2015;13(3):179–187.

- [23] Shon HS, Batbaatar E, Cho WS, Choi SG. Unsupervised Pre-Training of Imbalanced Data for Identification of Wafer Map Defect Patterns. IEEE Access. 2021;9:52352–52363. <https://doi.org/10.1109/ACCESS.2021.3068378>.
- [24] Karamti H, Lashin MMA, Alrowais FM, Mahmoud AM. A New Deep Stacked Architecture for Multi-Fault Machinery Identification With Imbalanced Samples. IEEE Access. 2021;9:58838–58851. <https://doi.org/10.1109/ACCESS.2021.3071796>.
- [25] Zhang X, Kano M, Tani M, Mori J, Ise J, Harada K. Hurdle Modeling for Defect Data with Excess Zeros in Steel Manufacturing Process. IFAC-PapersOnLine. 2018;51(18):375–380. <https://doi.org/10.1016/j.ifacol.2018.09.329>.
- [26] Workie MS, Azene AG. Bayesian zero-inflated regression model with application to under-five child mortality. Journal of Big Data. 2021;8(1). <https://doi.org/10.1186/s40537-020-00389-4>.
- [27] Aragaw AM, Azene AG, Workie MS. Poisson logit hurdle model with associated factors of perinatal mortality in Ethiopia. Journal of Big Data. 2022;9(1). <https://doi.org/10.1186/s40537-022-00567-6>.
- [28] Akkas A, Gaur V, Simchi-Levi D. Drivers of Product Expiration in Consumer Packaged Goods Retailing. Management Science. 2018;65:1–17. <https://doi.org/10.1287/mnsc.2018.3051>.
- [29] Tang P, Peng K, Dong J, Zhang K, Zhao S. Monitoring of Nonlinear Processes With Multiple Operating Modes Through a Novel Gaussian Mixture Variational Autoencoder Model. IEEE Access. 2020;8:114487–114500. <https://doi.org/10.1109/ACCESS.2020.3003095>.

- 1795 [30] Ibrahim M, Zhang C, Mahmood T. Surveillance of high-yield processes
1796 using deep learning models. *Quality and Reliability Engineering International*.
1797 2024;40(8):4365–4393. <https://doi.org/10.1002/qre.3635>.
1798
1799
1800
1801 [31] van Rossum G.: Python. Python Software Foundation. Available from: <https://www.python.org/>.
1802
1803
1804
1805 [32] Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al.:
1806 JAX: composable transformations of Python+NumPy programs. Available from:
1807 <http://github.com/google/jax>.
1808
1809
1810
1811 [33] Heek J, Levskaya A, Oliver A, Ritter M, Rondepierre B, Steiner A, et al.: Flax:
1812 A neural network library and ecosystem for JAX. Google. Available from: <http://github.com/google/flax>.
1813
1814
1815
1816 [34] Schuh G, Stich V. Produktionsplanung und -steuerung 1: Grundlagen der PPS:
1817 [engl.: Production Planning and Control 1: Fundamentals of PPC]. Berlin:
1818 Springer; 2012. Available from: [https://link.springer.com/book/10.1007/978-3-](https://link.springer.com/book/10.1007/978-3-642-25423-9)
1819
1820
1821
1822
1823
1824 [35] Miller J. Reaction time analysis with outlier exclusion: bias varies with sample
1825 size. *The Quarterly journal of experimental psychology A, Human experimental*
1826
1827
1828
1829
1830 [36] Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer-Verlag;
1831 2002.
1832
1833
1834 [37] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation
1835 Hyperparameter Optimization Framework. In: Teredesai A, Kumar V, Li Y,
1836 Rosales R, Terzi E, Karypis G, editors. *Proceedings of the 25th ACM SIGKDD*
1837
1838
1839
1840

NY, USA: ACM; 2019. p. 2623–2631.	1841
	1842
[38] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: Yoshua	1843
Bengio, Yann LeCun, editors. 7th International Conference on Learning Repre-	1844
sentations, ICLR 2019; 2019. p. 1–19.	1845
	1846
	1847
	1848
[39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.	1849
Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.	1850
2011;12:2825–2830.	1851
	1852
	1853
	1854
[40] Ramdas A, Trillos N, Cuturi M. On Wasserstein Two-Sample Testing and	1855
Related Families of Nonparametric Tests. Entropy. 2017;19(2):47. https://	1856
doi.org/10.3390/e19020047 .	1857
	1858
	1859
	1860
	1861
	1862
	1863
	1864
	1865
	1866
	1867
	1868
	1869
	1870
	1871
	1872
	1873
	1874
	1875
	1876
	1877
	1878
	1879
	1880
	1881
	1882
	1883
	1884
	1885
	1886