EVALUATING INTER-COLUMN LOGICAL RELATIONSHIPS IN SYNTHETIC TABULAR DATA GENERATION

Yunbo Long[1], **Liming Xu**[1], **and Alexandra Brintrup**[1,2]
[1] Department of Engineering, University of Cambridge
[2] The Alan Turing Institute, London
{yl892, lx249, ab702}@cam.ac.uk

ABSTRACT

Current evaluations of synthetic tabular data mainly focus on how well joint distributions are modeled, often overlooking the assessment of their effectiveness in preserving realistic event sequences and coherent entity relationships across columns. This paper proposes three evaluation metrics designed to assess the preservation of logical relationships among columns in synthetic tabular data. We validate these metrics by assessing the performance of both classical and state-of-the-art generation methods on a real-world industrial dataset. Experimental results reveal that existing methods often fail to rigorously maintain logical consistency (e.g., hierarchical relationships in geography or organization) and dependencies (e.g., temporal sequences or mathematical relationships), which are crucial for preserving the fine-grained realism of real-world tabular data. Building on these insights, this study also discusses possible pathways to better capture logical relationships while modeling the distribution of synthetic tabular data.

1 INTRODUCTION

Tabular data is challenging to synthesize due to its heterogeneity, where columns can contain different variable types, exhibiting diverse distributions and complex interdependencies (Wang et al., 2024). These characteristics make it difficult to accurately model the joint distribution $P(X, Y)$ (Margeloiu et al., 2024) of tabular data, where X denotes the feature space and Y denotes the target variable(s). To address these challenges, Xu et al. (2019) proposed GTGANs, a variant of generative adversarial netowrk (GAN) that learns the joint distribution $P(X, Y)$ through a *minimax game* between a generator and a discriminator. While the generator produces synthetic data $\mathbf{x}_{\text{syn}} = [x_1, x_2, \ldots, x_n]$ by conditioning on a *latent* vector (Zhao et al., 2021), it often fails to explicitly capture specific conditional dependencies, such as $P(x_1 \mid x_2)$, because it does not directly model the logical relationships between features. Building on the success of diffusion models in image generation, recent work has adapted them for tabular data generation. For example, TabDDPM (Kotelnikov et al., 2023) treats continuous and categorical features in tables separately, while TabSyn (Zhang et al., 2023) embeds them as tokens together for diffusion processes. During training, those methods approximate the data distribution $P(X_1, X_2, \ldots, X_n, Y)$ by modeling the transition between noisy data at each step. In the forward process, noise is added *isotropically* to each feature according to $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$, where the identity matrix ($\mathbf{I}$) ensures isotropic noise addition (Lee et al., 2023). However, this isotropic noise assumption limits the model's ability to capture semantic dependencies among features. Unlike images, where adjacent pixels exhibit strong spatial correlations that aid denoising (Nichol & Dhariwal, 2021), tabular data consists of heterogeneous and non-linear feature relationships without a natural ordering (Ruan et al., 2024), making it challenging for diffusion models to preserve logic dependencies. Recently, autoregressive models, such as large language models (LLMs), have also been leveraged to approximate the joint distribution between features and target values through sequential modeling(Fang et al., 2024). These methods, such as GReaT (Borisov et al., 2022), transform tabular data into sequences of tokens, enabling the autoregressive models to predict the conditional probability of the j-th token given the preceding $j-1$ tokens in the permuted sequence $P(t_{i,j} \mid \Pi(t_i, k)_{1:j-1})$, where each sequence of tokens is represented as $t_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,m}]$, and the sequence is randomly reordered based on a permutation vector $k = [k_1, k_2, \ldots, k_m]$. While LLMs capture column dependencies

based on pretrained knowledge (Sui et al., 2024), they do not explicitly model the marginal distribution $P(t_{i,j})$, leading to *biased sampling* despite the introduction of feature permutations or data variability.

To evaluate the *fidelity* of synthetic tabular data, numerous metrics have been proposed to assess accuracy and diversity, including both low-order statistics (e.g., Density Estimation and Correlation Score (Zhang et al., 2023), Average Coverage Scores (Zein & Urvoy, 2022)) and high-order statistics (e.g., α-Precision and β-Recall (Alaa et al., 2022)). However, these metrics operate at a high level and fail to evaluate whether synthetic data preserves logical relationships, such as hierarchical or semantic dependencies between features. This highlights the need for a more fine-grained, context-aware evaluation of multivariate dependencies. To address this, we propose three evaluation metrics: Hierarchical Consistency Score (HCS), Multivariate Dependency Index (MDI), and Distributional Similarity Index (DSI). To assess the effectiveness of these metrics in quantifying inter-column relationships, we select five representative tabular data generation methods from different categories for evaluation. Their performance is measured using both existing and our proposed metrics on a real-world dataset rich in logical consistency and dependency constraints. Experimental results validate the effectiveness of our proposed metrics and reveal the limitations of existing approaches in preserving logical relationships in synthetic tabular data. Additionally, we discuss potential pathways to better capture logical constraints within joint distributions, paying the way for future advancements in synthetic tabular data generation.

2 Inter-Column Relationships Evaluation Metrics

Logical relationships inherently capture both hierarchical consistency (e.g., city $\rightarrow$ country) and multivariate dependencies (e.g., temporal or mathematical relationships), reflecting structured inter-dependencies between columns. To evaluate hierarchical consistency, we define the **Hierarchical Consistency Score (HCS)**, which quantifies how well synthetic data preserves hierarchical relationships across columns. This metric is defined as:

$$\text{HCS} = \frac{1}{M \times N} \sum_{k=1}^{N} \sum_{j=1}^{M} \mathbb{1}\left((x_{i,j})_{i \in G_k} \in C_{k,j}\right), \tag{1}$$

where M is the number of rows and N is the number of consistency groups, and $x_{i,j}$ denotes the i-th attribute in the j-th row of the dataset. The set G_k refers to the k-th group of attributes (e.g., $G_1 = \{1, 2, 3\}$), while $C_{k,j}$ is a set of tuples, where each tuple represents a valid combination of attribute values for the group G_k in the j-th row. The indicator function $\mathbb{1}(\cdot)$ returns 1 if the tuple $(x_{i,j})_{i \in G_k}$ belongs to $C_{k,j}$, and 0 otherwise. To evaluate multivariate dependencies, we introduce the **Multivariate Dependency Index** (MDI), as follows:

$$\text{MDI} = \frac{1}{M \times N} \sum_{g=1}^{N} \sum_{j=1}^{M} \mathbb{1}\left(D_{g,j}\right), \quad D_{g,j} = \mathcal{F}(\{x_{i,j} \mid i \in G_g\}, x_{1,j}, x_{2,j} \ldots, x_{n,j}). \tag{2}$$

Where N is the number of dependency groups, M is the number of rows, and $D_{g,j}$ is the Boolean dependency condition for the g-th group G_g in the j-th row. For $i \in G_g$, $x_{i,j}$ must satisfy the dependency function $\mathcal{F}$ with respect to other n attributes in G_g in j-th row. To capture multivariate dependencies in non-linear relationships where explicit dependency rules are difficult to define, we propose the **Distributional Similarity Index** (DSI), which compares the log-likelihoods of Gaussian Mixture Models fitted to the each row in the synthetic dataset and the real dataset, as follows:

$$\text{DSI} = \frac{1}{K} \sum_{i=1}^{K} \left(1 - \frac{\left|\log \mathcal{L}(\hat{\mathbf{X}}^*_{\text{syn},i}) - \log \mathcal{L}(\hat{\mathbf{X}}^*_{\text{real}})\right|}{\left|\log \mathcal{L}(\hat{\mathbf{X}}^*_{\text{real}})\right|}\right), \tag{3}$$

where K is the number of rows, $\log \mathcal{L}(\hat{\mathbf{X}}^*_{\text{syn},i})$ and $\log \mathcal{L}(\hat{\mathbf{X}}^*_{\text{real}})$ are the log-likelihood of the GMM for the i-th synthetic and real datasets separately. The GMM log-likelihood for a dataset $\mathbf{X}$ is $\log \mathcal{L}(\mathbf{X}) = \sum_{j=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$, where π_k is the mixing coefficient, and $\mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density of the j-th data point under the k-th Gaussian component, with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. DSI reflects fine-grained differences between synthetic and real data.

Table 1: Results on inter-column logical relationship preservation. (**Note**: Higher values indicate better performance. The best and second-best results are highlighted in **bold** and **bold**, respectively.)

Metrics	Interpolation-based	Latent Space Representation-based			LLM-based
	SMOTE	CTGAN	TabDDPM	TabSyn	GReaT
Density Estimation	**98.02±0.02**	90.38±0.03	33.11±0.02	**96.38±0.04**	89.58±0.02
Correlation Score	**96.21±0.62**	74.41±0.14	36.78±0.01	**94.81±0.04**	71.00±1.16
Average Coverage Scores	**99.41±0.08**	82.27±0.14	76.23±0.23	**99.52±0.11**	92.34±0.15
α-Precision Scores	**93.54±0.01**	88.90±0.14	0.00±0.00	**98.48±0.10**	82.18±0.15
β-Recall Scores	**72.21±0.13**	1.71±0.01	0.00±0.00	22.62±0.12	**24.05±0.14**
HCS	**98.09±0.01**	39.23±0.03	16.07±0.01	71.63±0.08	**98.01±0.01**
MDI	**87.03±0.03**	38.87±0.05	59.08±0.05	68.34±0.08	**97.37±0.02**
DSI	77.46±0.01	11.10±0.00	68.38±0.01	**83.62±0.02**	**85.55±0.01**

3 Experiments and Discussion

We selected five representative synthetic tabular data generation methods for experiments. Originally introduced in early 2000s to address dataset imbalance (Chawla et al., 2002), the interpolation-based method—SMOTE—continues to outperform many generative models (Margeloiu et al., 2024). As such, we include it as a strong baseline, alongside the four state-of-the-art methods from different categories: CTGAN (Xu et al., 2019), TabDDPM (Kotelnikov et al., 2023), TabSyn (Zhang et al., 2023), and GReaT (Borisov et al., 2022). Each method generates around 140,000 samples from the DataCo dataset[1] (see more about this dataset in Section A.1). To ensure robustness, each evaluation was repeated ten times (see more experimental settings in Section A.2). We reported the mean and standard deviation of these ten runs in Table 1. As shown in Table 1, SMOTE and TabSyn achieve the best overall performance in terms of low-order statistical accuracies. With the highest accuracy in α-precision, TabSyn excels in accurately modeling the joint distribution among columns. Meanwhile, SMOTE significantly outperforms others in balancing data categories, achieving the highest β-recall scores. However, generative model-based methods, such as CTGAN, TabDDPM, and GReaT, struggle to accurately capture the distribution when synthesizing this complex and large-scale dataset, resulting in low values for low-order statistical metrics. Regarding inter-column relationship preservation, the results differ marginally. For data consistency, SMOTE and GReaT achieve the highest HCS of 98.09% and 98.01%, respectively, outperforming all other approaches. For data dependency, the MDI results indicate that GReaT effectively captures temporal and mathematical dependencies, achieving approximately 97.37% accuracy, considerably higher than the second-best method—SMOTE. Additionally, GReaT outperforms others in terms of DSI, however, it may generate unseen values for certain attributes, resulting in uncontrollable generation. In comparison, latent space-based generative models (CTGAN, TabDDPM, and TabSyn) are more reliable but struggle to effectively capture inter-column logical relationships in real-world tabular data. See the details and examples of generated data by each method in Section A.3.

4 Conclusion and Research Directions

We introduce three metrics—HCS, MDI, and DSI—for evaluating inter-column logical relationship in synthetic tabular data generation. Our experiments show that existing methods often fail to strictly maintain hierarchical consistency and multivariate dependencies—essential characteristics of real-world datasets. Our future work will focus on enhancing the preservation of inter-column logical relationships in synthetic tabular data generation. For LLM-based methods, the column serialization format and order are crucial for the model's ability to learn the joint distribution of logically related features. Knowledge graphs (Dong & Wang, 2024) or Bayesian networks (Ling et al., 2024) would be employed to reorder tokenization sequences or restructure the serialization of columns in natural language, leveraging prior knowledge to guide the synthetic tabular data generation. For latent space-based methods, LLM reasoning (Hegselmann et al., 2023; Dong & Wang, 2024) can be utilized to analyze column names and descriptions, identifying semantic or logical relationships without prior knowledge. Additionally, inspired by CTSyn (Lin et al., 2024), grouping data by log-

[1] https://data.mendeley.com/datasets/8gx2fvg2k6/3

ical relationships and embedding them into a shared latent space could potentially capture inherent structures, improving joint distribution modeling. Lastly, incorporating interpolation techniques like SMOTE may help can help balance data classes (Yang et al., 2024), particularly in learning minority logical relationships. These directions are *worthy* to explore for designing generative models that effectively capture inter-column logical relationships in synthetic tabular data generation.

REFERENCES

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pp. 290–306. PMLR, 2022.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Haoyu Dong and Zhiruo Wang. Large language models for tabular data: Progresses and future directions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2997–3000, 2024.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Jane Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, Christos Faloutsos, et al. Large language models (llms) on tabular data: Prediction, generation, and understanding-a survey. 2024.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pp. 18940–18956. PMLR, 2023.

Xiaofeng Lin, Chenheng Xu, Matthew Yang, and Guang Cheng. Ctsyn: A foundational model for cross tabular data generation. *arXiv preprint arXiv:2406.04619*, 2024.

Yaobin Ling, Xiaoqian Jiang, and Yejin Kim. Mallm-gan: Multi-agent large language model as generative adversarial network for synthesizing tabular data. *arXiv preprint arXiv:2406.10521*, 2024.

Andrei Margeloiu, Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabebm: A tabular data augmentation method with distinct class-specific energy-based models. *arXiv preprint arXiv:2409.16118*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. Language modeling on tabular data: A survey of foundations, techniques and evolution. *arXiv preprint arXiv:2408.10548*, 2024.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024.

Alex X Wang, Stefanka S Chukova, Colin R Simpson, and Binh P Nguyen. Challenges and opportunities of generative models on tabular data. *Applied Soft Computing*, pp. 112223, 2024.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Zeyu Yang, Peikun Guo, Khadija Zanna, and Akane Sano. Balanced mixed-type tabular data synthesis with diffusion models. *arXiv preprint arXiv:2404.08254*, 2024.

EL Hacen Zein and Tanguy Urvoy. Tabular data generation: Can we fool xgboost? In *NeurIPS 2022 First Table Representation Workshop*, 2022.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pp. 97–112. PMLR, 2021.

A APPENDIX

A.1 DATASET

DataCo is a large real-world dataset that includes complex, high-dimensional features and a wide variety of values for each feature. A summary of this complex dataset's characteristics is presented in Table 2. The dataset contains around 140,000 samples designed for model training, and each method generate same number of samples for synthetic tabular data evaluation. Notably, certain categorical columns contain over 3,000 unique values, reflecting the dataset's complexity.

This dataset represents a series of events related to purchasing, production, sales, and commercial distribution for an e-commerce company operating in global markets. This dataset is crucial and representative for synthetic tabular data generation in because its high-dimensional features, diverse inter-column relationships, and extensive unique values in categorical columns provide a realistic and challenging benchmark for evaluating the ability of synthetic data models to capture complex real-world patterns.

Table 2: Statistics of the Dataco dataset. "Numerical" represents the number of numerical columns, and "Categorical" stands for the number of categorical columns.

Dataset	Rows	Numerical	Categorical	Train	Validation	Test
Dataco	172,766	26	15	138,213	17,277	17,277

A.1.1 HIERARCHICAL CONSISTENCY

For hierarchical consistency in the Dataco dataset, there are three sets of tuples, C_1, C_2, and C_3, which represent the sets of valid tuples for specific attribute groups in the dataset. Here, i represents the attribute index (e.g., the i-th attribute in the dataset), and j represents the row index (e.g., the j-th row in the dataset). The tuple $((x_{i,j})_{i\in G_1})$ representing the **geographical information of orders** in each row j includes the following features :

- $x_{1,j}$ (order city),
- $x_{2,j}$ (order state),
- $x_{3,j}$ (order country),
- $x_{4,j}$ (order region),
- $x_{5,j}$ (order market).

The tuple $((x_{i,j})_{i\in G_2})$ representing the **product information** in each row j includes the following features :

- $x_{6,j}$ (category ID),
- $x_{7,j}$ (category name),
- $x_{8,j}$ (department ID),
- $x_{9,j}$ (department name),
- $x_{10,j}$ (product card ID),
- $x_{11,j}$ (product category ID),
- $x_{12,j}$ (product name).

The tuple $((x_{i,j})_{i\in G_3})$ representing the **geographical information of customers** in each row j includes the following features :

- $x_{13,j}$ (customer city),
- $x_{14,j}$ (customer state),
- $x_{15,j}$ (customer country).

To ensure hierarchical consistency, we verify whether each tuple $((x_{i,j})_{i\in G_k})$ belongs to its corresponding valid set of tuples $C_{k,j}$, where $k \in \mathbb{N}^+$ and j represents the row index (i.e., the j-th row in the dataset). For example, the tuple $((x_{i,1})_{i\in G_3})$, where $G_3 = \{13, 14, 15\}$ represents the combination of geographical information of customers, must belong to $C_{3,1}$ in the first row. This validation ensures that all attribute values are consistent with their hierarchical relationships.

A.1.2 TEMPORAL DEPENDENCY

Temporal dependencies are evident in the order information, where order dates and delivery times are sequentially linked. Let G_1 represent the **temporal group**, which includes x_1 as the order date and x_2 as the delivery time. The Boolean dependency condition $D_{1,j}$ for the j-th row in G_1 is defined as:

$$D_{1,j} : x_{1,j} < x_{2,j},$$

where $x_{1,j}$ (order date) must be earlier than $x_{2,j}$ (delivery time) to satisfy the temporal dependency.

A.1.3 MATHEMATICAL DEPENDENCY

Mathematical dependencies are presented in financial data. Let G_2 represent the **financial group**, which includes x_1 as the products quantity, x_2 as the products price, x_3 as the discount rate, x_4 as the discount value, x_5 as the original price, and x_6 as the sales price. The Boolean dependency conditions $D_{1,j}$, $D_{2,j}$, and $D_{3,j}$ for the j-th row in G_2 are defined as:

$$D_{1,j} : x_{5,j} = x_{1,j} \times x_{2,j},$$
$$D_{2,j} : x_{5j} = x_{1,j} \times x_{2,j} \times x_{3,j},$$
$$D_{3,j} : x_{4,j} = x_{7,j} - x_{5,j},$$

where the discount value should be equal to the product of the product price and the discount rate. The sales price should be calculated as the original price minus the discount value. Additionally, the original price should be determined by multiplying the product's quantity by its unit price.

A.2 EXPERIMENTS SETTINGS

To ensure reproducibility and fairness, all experiments are conducted on an open-source industrial dataset—DataCo. For SMOTE, synthetic data is sampled ten times to ensure robustness and account for variability in the interpolation-based generation process. For other methods (CTGAN, TabDDPM, TabSyn, and GReaT), models are trained on 80% of the dataset, with 10% used for validation and 10% for testing, following the default hyperparameter settings from their respective papers. The trained models are then used to generate synthetic tabular data. Each method is evaluated ten times, and the results are reported as the mean and standard deviation to account for variability.

To comprehensively evaluate the quality of synthetic data, each method is assessed using a set of metrics designed to measure different aspects of data fidelity and utility. These metrics include:

- **Hierarchical Consistency Score (HCS)**: Measures the preservation of hierarchical relationships in the data.
- **Multivariate Dependency Index (MDI)**: Quantifies the preservation of multivariate dependencies between features.
- **Distributional Similarity Index (DSI)**: Evaluates the overall similarity between the synthetic and real data distributions, where we embed both categorical and numerical features into a shared continuous space.

Additionally, we also adopted existing statistical metrics for evaluation:

- **Density Estimation**: Measures how well the synthetic data matches the probability density of the real data, with higher values indicating better alignment (Zhang et al., 2023).
- **Correlation Score**: Quantifies the preservation of pairwise correlations between features, with higher scores reflecting better retention of feature relationships (Zhang et al., 2023).
- **Average Coverage Score**: Evaluates the extent to which the synthetic data covers the range of values in the real data, with higher scores indicating better diversity (Zein & Urvoy, 2022).
- α**-Precision Score**: Evaluates the fidelity of synthetic data by determining whether each synthetic example originates from the real-data distribution, providing a measure of how well the synthetic data aligns with the true underlying data structure.(Alaa et al., 2022).
- β**-Recall Score**: Evaluates how well the synthetic data captures the real data distribution, with higher values indicating better representation (Alaa et al., 2022).

As current experiments were conducted over a single dataset, our future work will validate these three metrics (HCS, MDI, and DSI) on broader datasets to ensure their generalizability and robustness across different domains.

A.3 ADDITIONAL RESULTS ON LOGICAL RELATIONSHIPS PRESERVATION

To further examine the generated tabular data, we randomly select 10 rows from the generated dataset produced by each method to illustrate how effectively they preserve logical relationships. Specifically, we examine three types of relationships: mathematical dependencies, geographical hierarchies, and temporal sequences.

Mathematical Dependencies For mathematical dependency, as shown in Table 3, it is evident that CTGAN and TabDDPM produce numerous errors in maintaining these relationships, with none of randomly selected samples preserve mathematic dependency. In contrast, TabSyn successfully captures the mathematical relationships for all sampled rows. GReaT performs fairly good, preserving most of the mathematical relationships completely across the ten samples, followed closely by SMOTE.

Geographical Hierarchies For geographical relationship preservation, as shown in Table 4, CTGAN and TabDDPM fail to maintain any correct consistency in the logical chains. TabSyn, however, preserves the hierarchical consistency for some samples. GReaT and SMOTE achieve nearly 100% accuracy in maintaining geographical consistency, though GReaT exhibits a small number of errors, including generating incorrect names for order cities.

Temporal Relationships For temporal relationship(see Table 5), latent space-based methods (e.g., CTGAN, TabDDPM) preserve only about half of the correct temporal relationships, with results appearing somewhat random. In contrast, GReaT and SMOTE achieve near-perfect performance in preserving temporal relationships, although minor errors occur when the delivery and order dates are very close.

These results demonstrate that GReaT and SMOTE outperform other methods in preserving logical relationships, with GReaT showing slight inconsistencies in geographical and mathematical relationships. TabSyn performs moderately well in capturing temporal dependencies but struggles with geographical and mathematical consistency. In contrast, CTGAN and TabDDPM exhibit significant limitations across all types of logical relationships.

Table 3: Mathmatical dependency preservation in synthetic tabular data

Tabular Data Generation Methods	Group 1(Financial Data)						Preserved
	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	
	Quantity	Product Price	Discount Rate	Discount Values	Original Price	Sales Price	
Original Tables	5	49.98	0.09	22.49	249.90	227.41	✓
	2	49.98	0.04	4.00	99.96	95.96	✓
	1	129.99	0.05	6.50	129.99	123.49	✓
CTGAN	5	48.40	0.02	2.53	191.56	463.84	✗
	1	199.83	0.04	8.63	201.26	204.72	✗
	4	297.48	0.15	46.76	200.08	198.92	✗
	4	57.67	0.02	0.20	249.58	192.38	✗
	5	46.77	0.15	20.41	254.27	239.82	✗
	1	302.36	0.17	88.55	301.53	292.56	✗
	1	129.02	0.04	0.57	131.43	100.26	✗
	1	402.32	0.03	7.43	200.15	378.42	✗
	2	52.56	0.16	3.34	100.02	100.01	✗
	2	53.91	0.10	8.71	23.39	151.86	✗
TabDDPM	1	1999.99	0	500	1999.99	1939.99	✗
	5	9.99	0	500	9.99	7.99	✗
	5	1999.99	0.25	500	1999.99	1939.99	✗
	5	1999.99	0	500	9.99	1939.99	✗
	1	9.99	0	500	1999.99	1939.99	✗
	1	9.99	0	500	1999.99	1939.99	✗
	1	9.99	0	0	9.99	1939.99	✗
	1	1999.99	0	500	1999.99	7.99	✗
	5	1999.99	0	500	9.99	7.99	✗
	1	9.99	0	500	9.99	7.99	✗
TabSyn	1	299.98	0.01	3	299.98	299.96	✗
	1	129.99	0.06	7.2	129.99	123.66	✗
	1	199.99	0.04	8	199.99	191.97	✓
	1	251.37	0.04	10.72	299.95	254.67	✗
	4	99.99	0.03	10.82	399.96	387.98	✓
	3	50	0.05	7.46	179.96	139.50	✗
	5	99.99	0.13	70.68	499.95	427.09	✗
	5	50	0.05	12.5	250	239.98	✗
	2	50	0.25	24.60	100	73.19	✗
	1	50	0.09	9	100	92.93	✗
GReaT	3	99.99	0.03	9	299.97	290.97	✓
	4	50	0.15	30	200	170	✓
	1	199.99	0.09	18	199.99	181.99	✓
	1	129.99	0.13	16.90	129.99	113.09	✓
	1	50	0.16	8	50	42	✓
	1	199.99	0.04	8	199.99	191.99	✓
	4	49.98	0.10	19.99	199.92	179.93	✓
	2	49.98	0.04	12	99.96	95.96	✗
	1	399.98	0.03	12	399.98	387.98	✓
	2	50	0.07	7	100	93	✓
SMOTE	1	129.99	0.04	5.40	129.99	124.22	✗
	1	50	0.16	8	50	42	✓
	1	49.98	0.13	4.80	39.99	34.84	✗
	5	17.99	0.02	1.60	89.95	87.97	✗
	1	199.99	0	0	199.99	199.99	✓
	1	24.99	0.09	2	24.99	21.65	✗
	1	210.85	0	0	210.85	210.85	✓
	1	50	0.25	12.5	50	37.5	✓
	3	79.99	0.13	30	249.90	217.42	✗
	1	129.99	0.03	3.53	129.99	126.09	✗

Table 4: Hierachical consistency (geographical relationships) preservation in synthetic tabular data

Methods	Group 1(Geographical Data)					Preserved
	Attribute 1 Order City	Attribute 2 Order State	Attribute 3 Order Country	Attribute 4 Order Region	Attribute 5 Order Market	
Original Tables	Providence	Rhode Island	United States	East of USA	USCA	✓
	Porirua	Wellington	New Zealand	Oceania	Pacific Asia	✓
	Tegucigalpa	Francisco Morazan	Honduras	Central America	LATAM	✓
CTGAN	Aurangabad	Yalova	Nicaragua	Central America	LATAM	✗
	Rustenburg	Sao Paulo	Netherlands	Western Europe	Europe	✗
	Rasht	Maluku	Mexico	Central America	LATAM	✗
	San Pedro	Kano	Morocco	West Africa	Africa	✗
	Medan	Victoria	Australia	Oceania	Pacific Asia	✗
	Birobidzhan	Alsace-Champagene-Ardenne	France	Western Europe	Europe	✗
	Lagos	Kinshasa	Turkmenistan	East Africa	Africa	✗
	Yakarta	Punjab	India	Southeast Asia	Pacific Asia	✗
	Roermond	Mersin	Turkey	West Asia	Pacific Asia	✗
	Houston	Auckland	Mexico	Central America	LATAM	✗
TabDDPM	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
	Detmold	Zagrebacka	Detmold	Southeast Asia	LATAM	✗
TabSyn	Manila	Capital Nacional	Filipinas	Southeast Asia	Pacific Asia	✓
	Tokio	Tokyo	Japan	Eastern Asia	Pacific Asia	✓
	Concord	New Hampshire	United States	East of USA	USCA	✓
	Piedecuesta	Henan	China	Eastern Asia	Pacific Asia	✗
	Miguel Hidalgo	New Mexico	Mexico	Central America	LATAM	✓
	Lakeville	Mecklenburg-Western Pomerania	China	Eastern Asia	Pacific Asia	✗
	Baguio City	Osjecko-Baranjska	Norway	Northern Europe	Europe	✗
	Gaziemir	Vilnius	Iraq	Southeast Asia	Pacific Asia	✗
	Yucaipa	Guangdong	China	Eastern Asia	Pacific Asia	✗
	Erftstadt	Uusimaa	Italy	Southern Europe	Europe	✗
GReaT	Puebla	Puebla	Mexico	Central America	LATAM	✓
	Zhuzhou	Liaoning	China	Eastern Asia	Pacific Asia	✗
	Nicolas Romero	Mexico	Mexico	Central America	LATAM	✓
	Munich	Bavaria	Germany	Western Europe	Europe	✓
	Seattle	Washington	United States	West of USA	USCA	✓
	Culiacan	Sinaloa	Mexico	Central America	LATAM	✓
	Vitoria	Basque Country	Spain	Southern Europe	Europe	✓
	Reims	Alsace-Champagne-Ardenne-Lorraine	France	Western Europe	Europe	✓
	Cuneo	Piedmont	Italy	Southern Europe	Europe	✓
	13551	North Rhine-Westphalia	Germany	Western Europe	Europe	✗
SMOTE	Coyoacan	Federal District	Mexico	Central America	LATAM	✓
	Lancaster	California	United States	West of USA	USCA	✓
	Wiesbaden	Hesse	Germany	Western Europe	Europe	✓
	Ciego de Avila	Ciego de Avila	Cuba	Caribbean	LATAM	✓
	Grodno	Grodno	Belarus	Eastern Europe	Europe	✓
	Bugia	Buja	Argelia	North Africa	Africa	✓
	Nagpur	Maharashtra	India	South Asia	Pacific Asia	✓
	Nacka	Stockholm	Sweden	Northern Europe	Europe	✓
	Aachen	North Rhine-Westphalia	Germany	Western Europe	Europe	✓
	Oyonnax	Auvergne-Rhone-Alpes	France	Western Europe	Europe	✓

Table 5: Temporal dependency preservation in synthetic tabular data

Methods	Group 3 (Temporal Data)		Preserved
	Attribute 1	Attribute 2	
	Order Date	Delivery Date	
Original Tables	19/08/2015 12:59:00	24/08/2015 12:59:00	✓
	16/06/2015 13:46:00	19/06/2015 13:46:00	✓
	14/04/2016 23:41:00	15/04/2016 11:41:00	✓
CTGAN	10/07/2016 07:38:18	04/03/2015 04:52:46	✗
	14/03/2017 08:57:22	24/02/2015 03:31:49	✗
	02/08/2017 12:26:03	08/11/2015 04:32:00	✗
	14/01/2015 17:23:01	30/01/2015 21:37:26	✓
	09/03/2015 03:13:50	21/04/2016 01:04:10	✓
	24/08/2015 21:04:44	29/10/2016 00:46:55	✓
	15/01/2017 21:17:56	21/05/2015 17:55:06	✗
	24/08/2015 12:22:23	13/02/2016 09:49:48	✓
	20/03/2017 02:38:29	01/03/2017 19:23:58	✗
	26/09/2016 10:14:31	02/12/2015 15:52:12	✗
TabDDPM	31/01/2018 23:36:58	03/01/2015 00:00:00	✗
	01/01/2015 00:00:00	03/01/2015 00:00:00	✓
	01/01/2015 00:00:00	03/01/2015 00:00:00	✓
	01/01/2015 00:00:00	03/01/2015 00:00:00	✓
	01/01/2015 00:00:00	06/02/2018 18:43:12	✓
	31/01/2018 23:36:58	03/01/2015 00:00:00	✗
	01/01/2015 00:00:00	06/02/2018 18:43:12	✓
	31/01/2018 23:36:58	06/02/2018 18:43:12	✓
	31/01/2018 23:36:58	03/01/2015 00:00:00	✗
	31/01/2018 23:36:58	06/02/2018 18:43:12	✓
TabSyn	25/08/2017 02:09:36	31/08/2017 06:43:12	✓
	06/06/2015 04:24:58	13/06/2015 00:43:12	✓
	22/10/2015 06:43:12	27/10/2015 05:54:14	✓
	01/12/2016 03:00:00	01/12/2016 10:30:43	✗
	19/12/2015 08:24:00	16/12/2015 18:14:24	✗
	01/02/2017 08:42:43	06/02/2017 01:29:17	✓
	17/07/2016 16:19:12	18/07/2016 23:16:48	✓
	13/08/2016 11:48:29	10/08/2016 12:57:36	✗
	05/07/2015 22:33:36	12/07/2015 04:33:36	✓
	07/10/2016 14:03:50	15/10/2016 15:28:48	✓
GReaT	21/07/2017 11:18:00	23/07/2017 11:18:00	✓
	14/06/2016 15:22:00	16/06/2016 15:22:00	✓
	23/06/2017 01:00:00	28/06/2017 01:00:00	✓
	31/01/2015 10:49:00	02/02/2015 10:49:00	✓
	16/10/2016 00:50:00	19/10/2016 00:50:00	✓
	16/08/2017 04:36:00	22/08/2017 04:36:00	✓
	03/08/2016 11:47:00	09/08/2016 11:47:00	✓
	10/03/2015 22:40:00	13/03/2015 22:40:00	✓
	16/01/2015 22:19:11	16/01/2015 10:19:00	✗
	21/03/2016 21:16:00	25/03/2016 21:16:00	✓
SMOTE	31/05/2016 01:07:42	04/06/2016 13:20:19	✓
	21/07/2016 11:38:00	23/07/2016 11:38:00	✓
	06/05/2017 01:00:54	08/05/2017 16:36:31	✓
	15/10/2016 00:37:28	16/10/2016 22:47:19	✓
	22/07/2017 04:52:12	24/07/2017 03:10:12	✓
	10/02/2016 05:15:24	15/02/2016 14:55:47	✓
	10/01/2015 00:27:21	14/01/2015 04:51:29	✓
	16/01/2016 00:38:18	19/01/2016 09:41:35	✓
	22/05/2015 11:33:20	25/05/2015 11:19:17	✓
	17/03/2015 19:08:39	19/03/2015 08:27:54	✓