TARImpute: Task-Aware auto-Recommender System for Missing Value Imputation Algorithms with Clustering Case Studies

Xiaoou Ding, Yanshuo Liu, Zhounan Chen, Hongzhi Wang*
Harbin Institute of Technology, China
{dingxiaoou,wangzh}@hit.edu.cn

Chen Wang, Jianmin Wang
Tsinghua University, China
{wang_chen,jimwang}@tsinghua.edu.cn

ABSTRACT

Missing data prevalent in information systems impacts data diversity and fidelity, which systematically degrade clustering performance through biased similarity measures and unstable cluster boundaries. Current large-scale environments lack standardized imputation-clustering pipelines, as existing methods operate independently of downstream tasks without analyzing error propagation effects, leading to unreliable results. To address this, we propose TARImpute, a Task-Aware auto-Recommender system for missing value imputation for clustering. It owns three integrated features: Imputation Impact Profiler for quantitative evaluation of imputation-clustering interactions, Error Propagation Interpreter enabling explainable modeling of imputation error diffusion, and Adaptive Strategy Optimizer for dynamic selection of optimal imputation methods. TARImpute provides state-of-the-art imputation methods to evaluate their effects on clustering tasks. TARImpute also provides robust, interpretable solutions for low-quality data and shows extensibility to other analytical tasks.

PVLDB Reference Format:
Xiaoou Ding, Yanshuo Liu, Zhounan Chen, Hongzhi Wang and Chen Wang, Jianmin Wang. TARImpute: Task-Aware auto-Recommender System for Missing Value Imputation Algorithms with Clustering Case Studies. PVLDB, 18(12): 5343 - 5346, 2025.
doi:10.14778/3750601.3750667

1 INTRODUCTION

The proliferation of *data-driven* paradigms has exponentially increased demand for reliable data analysis, where *complete* and *high-quality* data is a fundamental requirement. However, the pervasive issue of systematic data missingness in real-world information systems has emerged as a bottleneck constraining analytical efficacy [3]. Such gaps not only cause sample information loss but also induce distribution distortions that fundamentally compromise similarity measurement reliability.

In clustering tasks, missing values compromise feature space integrity, biasing distance metrics and distorting cluster boundaries. High-dimensional spaces exacerbate these issues, causing feature collapse (>45% information loss) and metric distortion (2.7× error amplification), frequently leading to suboptimal convergence. This has driven imputation method development into three paradigms (see [7, 9] as recent surveys): (1) Statistical approaches (mean/median) offer efficiency but ignore feature interactions; (2) Machine learning methods (e.g., random forests) model complex dependencies for robust imputation; and (3) Deep generative models (VAEs, GANs) leverage expressive architectures for pattern learning, with VAEs optimizing explicit distributions and GANs learning implicit data manifolds. However, current imputation approaches exhibit two fundamental limitations: ❶ objective misalignment between imputation and downstream task losses, introducing cluster-distorting artifacts, and ❷ unquantified error propagation through feature interactions and distance metrics, causing compounding errors.

These limitations motivate our ***task-aware missing value imputation paradigm*** that emphasizes tight integration between imputation and downstream objectives. For clustering, developing an adaptive imputation framework faces two key challenges: (1) *Non-identifiable error propagation routes*: Traditional assumptions fail to capture how imputation errors nonlinearly affect clustering via feature coupling, distribution shifts, and noise. Existing methods lack systematic ways to track error paths and assess their impact on cluster stability. (2) *Limited adaptability*: Static imputation-clustering pairs (e.g., MICE+DBSCAN [10]) cannot model interactions between data traits, imputation strategies, and clustering goals (e.g., precision). Fixed strategies under varying missingness patterns often degrade performance, hindering cross-scenario adaptability.

Building upon our previous research in data cleaning and missing value imputation [1, 2, 4], we propose TARImpute, a task-aware adaptive imputation system that co-optimizes imputation strategies with clustering tasks. It significantly improves clustering stability on incomplete data while demonstrating extensibility to other analytical tasks (e.g., time series classification and prediction). The contributions of TARImpute are threefold:

❶ **Modeling imputation's impact on clustering result**: By reformulating clustering algorithms as differentiable operators, we enable end-to-end error topology modeling through gradient pathways. DAGs explicitly map imputation error propagation, quantifying path-specific contributions via feature interactions and shifts. This reveals error network properties and provides topological criteria for strategy selection, giving TARImpute the theoretical capacity to interpret imputation effects.

❷ **Quantitative tolerance analysis of clustering methods**: Using the propagation model, we derive Lipschitz continuity conditions for imputation, setting upper bounds on cluster quality degradation. Quantifies error sensitivity and guides decision making through tolerance thresholds. Experiments show its efficacy in predicting stability limits across MCAR/MAR/MNAR scenarios, preventing performance drops from poor imputation choices.

*Corresponding author.

Proceedings of the VLDB Endowment, Vol. 18, No. 12 ISSN 2150-8097.
doi:10.14778/3750601.3750667

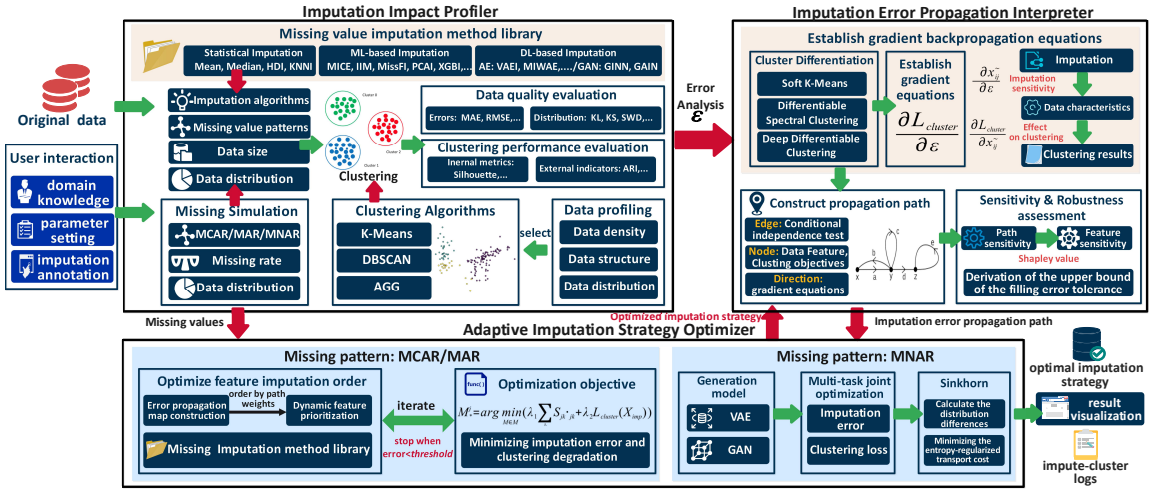

Figure 1: TARImpute system overview.

❸ **Task-aware adaptive recommendation**: TARImpute integrates a library of SOTA imputation methods (e.g., DeepMVI, IIM, MIWAE [9]) with several clustering algorithms (K-Means, DBSCAN, Agglomerative) for dynamic recommendations. It evaluates propagation paths to create an adaptive space using task-specific metrics, employing dynamic tuning for MCAR/MAR and deep learning (e.g., GANs, VAEs) for MNAR to automate optimal pair selection.

TARImpute revolutionizes missing imputation for clustering by integrating feature awareness, mechanism insight, and optimal recommendations. It offers robust, interpretable solutions for low-quality data and exhibits modular scalability for task expansion.

2 SYSTEM OVERVIEW

As illustrated in Figure 1, TARImpute comprises four integrated modules: (1) `Imputation Impact Profiler` for quantitative evaluation of imputation-clustering interactions, (2) `Error Propagation Interpreter` enabling explainable modeling of imputation error diffusion, (3) `Adaptive Strategy Optimizer` for dynamic selection of optimal imputation methods, and (4) an Interactive User Interface facilitating system operation.

• **`Imputation Impact Profiler`**. It offers a framework to evaluate imputation-clustering performance via configurable missing data simulations and multi-faceted assessments. It features a multi-mechanism missingness injector for dynamic configuration of missingness types, rates (0-90%), scales, and distributions (normal/non-normal) on clean data. Integrating SOTA imputation methods and classical clustering algorithms, it auto-selects suitable clustering methods based on data characteristics, generating interactive reports with metrics like silhouette coefficient and adjusted rand index (ARI). Users can adjust settings to view performance heatmaps, identify sensitive features, and critical thresholds (e.g., KNN imputation reduces DBSCAN silhouette by 42% over 35% missingness in MNAR), establishing a baseline for subsequent modules.

• **`Imputation Error Propagation Interpreter`**. This module analyzes how data imputation affects clustering by modeling its impact mechanistically. It reformulates pre-selected clustering methods as differentiable operators for gradient backpropagation, assessing factors driving performance changes. Conditional independence tests identify key error propagation pathways, integrated with an enhanced PC algorithm [6] to determine pathway directionality, forming a DAG (*features-errors-clustering performance*) for gradient and error analysis. The module derives the upper bound of imputation error propagation under Lipschitz continuity [8], evaluating clustering robustness and sensitivity to imputation errors. This theoretical framework supports optimization modules and clarifies clustering method against imputation errors.

• **`Adaptive Imputation Strategy Optimizer`**. This module optimizes imputation-clustering strategies by dynamically adjusting methods based on missingness mechanisms and sensitive feature weights. For MCAR/MAR patterns, it uses HyperImpute to minimize errors and clustering degradation. For MNAR patterns, it employs MIRACLE causal inference with Sinkhorn optimal transport. TARImpute constructs an imputation-clustering strategy library, and, through an adaptive adjustment mechanism, recommends the optimal imputation strategy, providing users with flexible imputation method options for clustering tasks in complex data scenarios.

3 IMPLEMENTATION DETAILS

This section details the technical implementation of TARImpute.

3.1 Interpretable evaluation of imputation impact on clustering

In clustering tasks with missing values, imputation errors propagate and accumulate, impacting result stability and effectiveness. Traditional methods overlook error diffusion during clustering, worsening similarity and partitioning errors. Thus, quantifying error pathways is key to optimizing imputation and improving clustering robustness and accuracy, central to TARImpute.

(1) Establishment of gradient backpropagation equations. Classic clustering methods employ hard assignments causing discontinuous gradients unsuitable for backpropagation. TARImpute makes them differentiable by introducing a temperature coefficient β to soften the assignment, defining the probability distribution of a data point x_i belonging to cluster j as $\Pr_{ij} = \frac{\exp(-\beta||x_i-\mu_j||^2)}{\sum_k \exp(-\beta||x_i-\mu_j||^2)}$. This allows cluster centers $\mu_j = \frac{\sum_i \Pr_{ij} x_i}{\sum_i \Pr_{ij}}$ to be expressed as weighted averages of data points, constructing a differentiable relationship between data points and cluster centers. We design a multi-dimensional

imputation error function

$$\epsilon = \gamma_1 Err_{\text{RMSE}} + \gamma_2 Err_{\text{KL}} + \gamma_3 Err_{\text{SWD}}$$

to capture error propagation effects, where Err_{RMSE} assesses numerical imputation error, Err_{KL} constrains imputed data distribution, Err_{SWD} quantifies the distribution differences.

Accordingly, to build an end-to-end differentiable computational framework for *data imputation-data clustering*, we systematically establish a backward gradient propagation link from clustering decisions to data imputation: $\frac{\partial L_{\text{cluster}}}{\partial \epsilon} = \frac{\partial L}{\partial \tilde{x_{ij}}} \cdot \frac{\partial \tilde{x_{ij}}}{\partial \epsilon}$, where $\tilde{x_{ij}}$ represents the imputed data point, L_{cluster} reflects changes in clustering structure post-imputation, $\frac{\partial L_{\text{cluster}}}{\partial \epsilon}$ reflects the influence of changes in data point positions on clustering results, and $\frac{\partial L}{\partial \tilde{x_{ij}}}$ is imputation sensitivity, i.e., the impact of the imputation method on imputation errors. This transforms the "*imputation-feature space-clustering*" causality into a computable equation, enabling transparent decision-making in the imputation-clustering.

(2) Propagation path graph construction. To visualize how imputation errors affect clustering, we construct a directed acyclic graph (DAG) where nodes represent data features X_i, and edges represent the direction and strength of error influence. First, conditional independence tests identify critical error paths: edges between features (X_i, X_j) are removed if $\rho > \alpha$ (significance level), forming an undirected graph. Next, we determine the direction of the edges by improving the traditional PC algorithm in conjunction with gradient relationships. If $\frac{|\partial L_{\text{cluster}}/\partial X_i|}{|\partial L_{\text{cluster}}/\partial X_j|} > 2$, the direction $X_i \rightarrow X_j$ is established. For complex structures like $X_i - X_k - X_j$, if the gradient propagation paths of X_i and X_j independently affect X_k, the structure $X_i \rightarrow X_k \leftarrow X_j$ is formed. Additionally, edges with path weights exceeding a threshold are retained. The path weight is defined as: $\omega_{ij} = \frac{|\partial L_{\text{cluster}}/\partial X_j|}{\sum_k |\partial L_{\text{cluster}}/\partial X_k|}$ (k represents all features). This constructs a DAG of the error propagation paths.

(3) **Path sensitivity quantification:** For each propagation path p, we can calculate its contribution as

$$\text{Contribution}(p) = \prod_{(i \rightarrow j) \in p} \left(\phi_i \cdot \frac{\partial L_{\text{cluster}}}{\partial X_j} \right),$$

where ϕ_i is the Shapley value of each node i quantifying its influence on the model output. Using the weights, we derive the global sensitivity of imputed feature X_i, measuring its importance. This quantifies feature sensitivity, identifies critical features impacting results, and optimizes model performance.

(4) **Upper bound derivation for imputation error tolerance:** To gauge the maximal impact of imputation errors on clustering, we derive an upper bound for clustering tasks' tolerance to errors. This bound offers a rigorous quantitative framework for analyzing error propagation and assessing clustering algorithm robustness. It identifies the tolerance range, guiding adaptive imputation optimization to reduce performance degradation and ensure robust clustering. Given an incomplete data matrix $X \in \mathbb{R}^{n \times d}$ with complete ground truth X_{com}, and an imputation function $f: \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d}$, we define the **imputation error** as $\epsilon = \|f(X) - X_{\text{com}}\|$. Our objective is to derive the upper bound for clustering quality degradation $\Delta Q = \left|Q(f(X)) - Q(X_{\text{com}})\right|$, where $Q(\cdot)$ is a clustering quality metric (e.g., silhouette, ARI). Due to the non-differentiability of metrics like ARI and silhouette, we approximate error propagation via:

$$\Delta Q = \sup_{\varepsilon} \left|Q(f(X)) - Q(X_{\text{com}})\right|.$$

This seeks the worst-case upper bound of clustering quality change due to imputation errors. We analyze imputation errors' impact on clustering quality, focusing on the worst-case upper bound. Experiments characterize the error-quality relationship, fit the data, explore tolerance, and provide an approximate upper bound.

3.2 Adaptive imputation method selection

(1) **MAR and MCAR patterns**. For MCAR and MAR, where missingness depends on observed data. We adopt HyperImpute's column-wise iterative imputation, selecting appropriate strategies and dynamically assessing imputation adaptability to minimize errors and clustering degradation.

HyperImpute imputes features column by column, considering inter-feature relationships for accuracy. However, its reliance on RMSE as a static threshold limits adaptability to complex patterns. We extend its strategies for flexibility in MCAR and MAR scenarios. We incorporate clustering metrics (e.g., silhouette coefficient, ARI) alongside imputation errors (RMSE, MAE) to ensure imputed data benefits clustering. Using HyperImpute's column-wise approach, we evaluate imputation errors and clustering performance after each round t. If the evaluation score falls below a threshold θ, we adjust the strategy $f_{t+1} = Adjust(f_t)$. After each imputation round, the proposed `TARImpute` uses a new threshold function

$$M_j^* = \arg\min_{M \in \mathcal{M}} (\lambda_1 \sum_k S_{jk} \cdot \epsilon_{jk} + \lambda_2 L_{\text{cluster}}(X_{imp}))$$

to check for negative impacts on clustering. If errors accumulate on key features or degrade performance, it adaptively adjusts strategies. This mechanism allows flexible handling of complex missing patterns and data characteristics in MCAR and MAR scenarios.

(2) **MNAR patterns**. For MNAR, where missingness depends on unobserved variables or the missing components themselves, imputation is more challenging. Traditional methods struggle to recover MNAR characteristics accurately, potentially causing significant imputation errors and clustering degradation. Thus, complex imputation methods and dynamic adaptive strategies are needed in MNAR scenarios to improve imputation and reduce error impacts on clustering. We apply MIRACLE framework [5] to address MNAR distribution shift and error propagation through a three-stage optimization: generative models, multi-task learning, and Sinkhorn regularization, enabling adaptive MNAR imputation. A VAE models the underlying data distribution in its latent space, supporting probabilistic MNAR modeling. Jointly optimizing reconstruction and clustering losses balances imputation quality and task performance:

$$\mathcal{L} = \gamma \mathcal{L}_{\text{rec}}(\hat{X}, X) + \delta \mathcal{L}_{\text{clu}}(f(\hat{X}), Y),$$

where $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{clu}}$ are reconstruction and clustering losses, and γ and δ adjust their weights. Sinkhorn optimal transport regularizes global similarity between imputed and real data, mitigating MNAR distribution shift by minimizing entropy-regularized transport costs. These components form a closed loop: the generative model imputes initially, multi-task learning balances objectives, and Sinkhorn corrects distribution, enhancing imputation precision and clustering robustness. This provides an end-to-end MNAR solution from local to global and generation to optimization.

Figure 2: Pages demonstration of TARImpute.

4 DEMONSTRATION

As illustrated in Figure 2, TARImpute offers an operationally concise and user-friendly demonstration interface. The process primarily consists of the following steps:

Step 1: Data uploading and preprocessing. Users upload raw datasets to TARImpute, which offers a dual-mode analysis framework: (1) **Simulated Missingness**: The system analyzes global features and recommends clustering algorithms based on data distribution.(2) **Real Missing Data**: This mode skips missing value generation and directly analyzes feature deviations. Users adjust parameters via an interactive interface: a slider sets missingness rates, and a numeric input specifies dataset size. Clicking "*Start Simulation*" generates a statistical atlas of missingness patterns, quantifying impacts on covariance matrices and feature correlations to help users understand data characteristics efficiently.

Step 2: Multi-dimensional evaluation and visualization of clustering strategies. Upon clicking "*Perform Imputation*",our system imputes missing data and quantifies errors using RMSE/MAE metrics. A dynamic table compares imputation algorithm performance across scenarios (e.g., 20% MAR vs. 10% MNAR). During "*Perform Clustering*", it evaluates combined imputation-clustering outcomes using silhouette coefficient and ARI. The report recommends optimal cluster counts and uses heatmaps to visualize strategy effectiveness variability. An interactive dashboard allows users to filter parameters for real-time performance analysis, aiding in imputation method suitability assessment.

Step 3: Construction of imputation error propagation pathway diagrams. To elucidate how imputation errors affect clustering outcomes, TARImpute introduces error propagation pathway diagrams. Using the error matrix, it builds a feature-level network, where directed edge weights (error amplification factors) and node color gradients (error accumulation) visualize error propagation, cumulative effects, and potential impacts on clustering. Users can view error contribution rates and downstream influences, enabling analysis of error dissemination mechanisms. For sensitive features, TARImpute allows manual prioritization. If users identify critical features via the diagram, it adjusts imputation strategies based on user annotations to improve clustering accuracy and stability.

Step 4: Optimization and dynamic adjustment of imputation-clustering combinations. TARImpute's adaptive imputation optimization module automatically adjusts imputation strategies to minimize both imputation errors and clustering degradation. Leveraging the imputation error propagation pathway diagram, users can flexibly modify imputation approaches based on the weights of sensitive features and error propagation patterns, prioritizing corrections for critical features and mitigating error impacts.

Upon clicking "*Data Imputation*", the proposed TARImpute recommends optimal imputation strategies for each feature. Clicking "*Performance Analysis*" activates the imputation-clustering optimization module, displaying the best combination strategy and enables real-time observation of strategy on clustering outcomes. This dynamic adjustment lets users promptly refine strategies based on real conditions, achieving superior clustering performance.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2023YFB3308003), National Natural Science Foundation of China (NSFC) (62202126, 62232005, 92267203), and Natural Science Foundation Project of Heilongjiang Province of China (YQ2024F005).

REFERENCES

[1] Xiaoou Ding, Yingze Li, Hongzhi Wang, Chen Wang, Yida Liu, and Jianmin Wang. 2024. TSDDISCOVER: Discovering Data Dependency for Time Series Data. IEEE ICDE, 3668–3681.

[2] Xiaoou Ding, Yichen Song, Hongzhi Wang, Donghua Yang, Chen Wang, and Jianmin Wang. 2024. Clean4TSDB: A Data Cleaning Tool for Time Series Databases. *VLDB* 17, 12 (2024), 4377–4380.

[3] Xiaoou Ding, Hongzhi Wang, Genglong Li, Haoxuan Li, Yingze Li, and Yida Liu. 2022. IoT data cleaning techniques: A survey. *Intell. Converged Networks* 3, 4 (2022), 325–339.

[4] Xiaoou Ding, Hongzhi Wang, Jiaxuan Su, Zijue Li, Jianzhong Li, and Hong Gao. 2019. Cleanits: A Data Cleaning System for Industrial Time Series. *VLDB* 12, 12 (2019), 1786–1789.

[5] Daniel Jarrett, Bogdan Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. 2022. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. In *ICML*, Vol. 162. 9916–9937.

[6] Markus Kalisch and Peter Bühlmann. 2007. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *J. Mach. Learn. Res.* 8 (2007), 613–636.

[7] Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. 2020. Mind the Gap: An Experimental Evaluation of Imputation of Missing Values Techniques in Time Series. *VLDB* 13, 5 (2020), 768–782.

[8] Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. 2023. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *CoRR* abs/2301.04431 (2023).

[9] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. 2023. An Experimental Survey of Missing Data Imputation Algorithms. *IEEE Trans. Knowl. Data Eng.* 35, 7 (2023), 6630–6650.

[10] Keyu Yang, Yunjun Gao, Rui Ma, Lu Chen, Sai Wu, and Gang Chen. 2019. DBSCAN-MS: Distributed Density-Based Clustering in Metric Spaces. In *ICDE*. 1346–1357.