Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TDAssure: a topological data analysis framework for assessing missing data imputation quality

Yiyang Ge
King's College London
Mathieu Carrière
Centre Inria de l'Université Paris-Saclay
Daniel Stahl
King's College London
Raquel Iniesta
raquel.iniesta@kcl.ac.uk

King's College London

Method Article

Keywords: Missing data imputation, Imputation quality assessment, Topological Data Analysis, Mapper

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11537886/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TDAssure: a topological data analysis framework for assessing missing data imputation quality

Yiyang Ge, Mathieu Carrière, Daniel Stahl, Raquel Iniesta*

(*) Corresponding author: raquel.iniesta@kcl.ac.uk

Abstract

Conventional measures of imputation quality assess how accurately missing values are reconstructed but provide limited information about preservation of the multivariate structure of the data. We introduce TDAssure, a topological data analysis framework for assessing structural preservation following missing-data imputation. The framework constructs Mapper graphs for complete and imputed cases, summarises their topological features using extended persistence diagrams, and quantifies differences using bottleneck distances. Permutation tests assess the statistical significance of these differences across a predefined Mapper parameter grid, with adjustment for multiple testing. We evaluated the framework on the Fashion-MNIST and Epileptic Seizure Recognition datasets using four imputation methods under missing completely at random, missing at random, and missing not at random mechanisms, with target missingness rates of 20% and 30%. Random imputation consistently produced the greatest reconstruction error and the largest number of Mapper configurations showing significant topological differences. The relative performance of k-nearest-neighbour, autoencoder, and mean imputation varied between datasets and assessment approaches. In particular, lower reconstruction error did not always correspond to fewer detectable topological differences, with disagreements most apparent when data were missing not at random. Interpretation in such settings requires caution because complete cases may provide an unrepresentative structural reference. An interactive. TDAssure, implemented in a Shinny application, provides a complementary assessment of imputation quality by examining structural preservation alongside pointwise reconstruction accuracy.

Keywords

Missing data imputation; Imputation quality assessment; Topological Data Analysis; Mapper;

Introduction

Missing data are a common problem in data analysis and can affect the validity and reliability of subsequent findings. Missing values may arise from data-entry errors, equipment malfunction, incomplete records, or limitations in data collection processes [1]. In health-related research, for example, missingness may occur when participants decline to provide information or do not attend follow-up assessments. Since inappropriate handling of missing data can introduce bias and reduce statistical power, both the choice of method for handling missingness and the evaluation of its performance are important components of data analysis.

Missing data are typically classified according to three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [2]. These mechanisms have different implications for statistical inference and for the suitability of approaches used to handle missing values. Methods for handling missing data broadly include deletion and imputation. Complete Case Analysis (CCA), or listwise deletion, is commonly used because of its simplicity, but its validity generally relies on assumptions that may be unrealistic in practice. Imputation methods instead estimate missing entries using information available in the observed data. These range from classical approaches such as mean, median, or mode substitution and Multiple Imputation by Chained Equations (MICE) [3], to methods such as k-nearest neighbors (KNN) [4] and missForest [5], which can capture more complex relationships in the data. The suitability and performance of an imputation method depend strongly on the characteristics of the dataset and the underlying missingness mechanism, making reliable evaluation of imputation quality essential.

However, evaluating the quality of missing-data imputation remains challenging. Conventional assessment approaches often introduce missingness artificially into complete

datasets and compare the imputed values with the known observations. The simulated missingness may not adequately reflect the actual missing-data pattern in the dataset, potentially affecting the validity of the evaluation [6]. Moreover, commonly used pointwise error measures primarily quantify the accuracy with which individual missing values are reconstructed and may provide limited information about whether the broader multivariate structure of the data has been preserved. An imputation method may therefore achieve low reconstruction error while still altering important relationships, clusters, or other structural characteristics of the original dataset. Assessing imputation quality may consequently require complementary approaches that evaluate structural preservation alongside conventional measures of reconstruction accuracy.

Topological Data Analysis (TDA) provides a mathematical framework for characterising the intrinsic structure, or "shape", of complex datasets [7]. Rooted in applied algebraic topology [8–10], TDA characterises structural features such as connectivity, clustering, loops, and voids in data and can examine how these features persist across different scales. This is particularly relevant for high-dimensional and heterogeneous data, in which important structural patterns may be difficult to characterise using conventional statistical approaches and simple geometric assumptions [11]. TDA has been applied across fields including neuroscience, genomics, and materials science; however, its potential for evaluating whether missing-data imputation preserves the underlying structure of a dataset remains comparatively underexplored. This provides an opportunity to complement value-level measures of imputation accuracy with an assessment of the extent to which the global organisation of the data is retained after imputation.

In this study, we introduce a topology-based framework for evaluating structural preservation following missing-data imputation. The framework uses the Mapper algorithm [12] to construct topological representations of the complete-case and imputed datasets and quantify differences in their structural organisation. Topological features are summarised using persistence diagrams, and structural differences are quantified using the bottleneck distance [13]. A key contribution of the framework is the integration of persistence-based topological comparison with permutation-based statistical inference,

allowing observed structural differences to be evaluated against an empirical null distribution. Together, these components provide complementary visual and quantitative evidence of structural preservation following imputation. We evaluate the framework using benchmark and real-world datasets and compare its assessments across different imputation methods. To facilitate practical application of the proposed framework, we implement it in TDAssure, an interactive Shiny for Python application designed to enable its use without requiring users to interact directly with the underlying code.

Related Work

Relatively few studies have examined the use of TDA specifically to assess structural changes associated with missing-data imputation. Arioli and colleagues [14] developed an interactive dashboard based on the Mapper algorithm to visualize missing-data distributions and the topological structures of complete and imputed datasets. Their approach allows users to inspect potential structural changes introduced by imputation and to evaluate acceptable levels of missingness. By presenting the results through Mapper visualizations, the method provides an intuitive way to explore how missing-data handling may affect the structure of the data. The approach, however, relies mainly on visual comparison of topological structures. This is useful for exploratory analysis, but subtle differences can be difficult to assess visually, particularly when many datasets or imputation settings must be compared. The method was also developed and evaluated primarily on proteomics data, so its applicability to other types of data has been less extensively examined. These limitations leave scope for approaches that complement visual inspection with quantitative measures of structural preservation following imputation.

Glushakov and colleagues [15] proposed Topology-Based Clinical Data Mining (TCDM), a TDA-based framework for evaluating missing-data imputation under MCAR and MAR assumptions. Their method uses Mapper graphs to represent the topological structure of clinical datasets and quantifies similarity between complete and imputed data using a pairwise-distance measure computed across multiple subsamples. Compared with the

approach of Arioli [14], TCDM extends the evaluation beyond visual inspection by introducing a quantitative measure of structural similarity.

Despite this advance, TCDM provides a descriptive measure of topological similarity but does not incorporate an inferential procedure for assessing whether observed structural differences exceed those expected under an appropriate reference distribution. In addition, the final assessment is based on averaging similarity measures across subsamples, which may mask variation among individual comparisons. This leaves an important methodological gap between quantifying topological differences and determining whether the observed differences provide evidence of meaningful structural change following imputation. Persistence-based representations provide a means of summarising topological features extracted from Mapper graphs, while bottleneck distance allows differences between these representations to be quantified [16]. Integrating these measures with statistical inference provides an approach to identify significant structural changes in imputed data.

Methodology

The proposed framework evaluates whether the intrinsic topological structure of a dataset is preserved after missing data imputation. It consists of two main components: Mapper-based topological representation and permutation-based statistical validation. First, Mapper graphs are constructed for the complete-case dataset (the subset of observations containing no missing values) and the imputed dataset (comprising only previously incomplete observations whose missing values have been imputed) across a predefined parameter grid, and their topological features are summarized using persistence diagrams. The difference between the resulting topological representations is then quantified using the bottleneck distance. Second, a permutation test is used to determine whether the observed topological discrepancy between complete-case and imputed-datasets is significantly greater than expected under random permutation of group membership (complete-case vs imputed) while preserving the original group sizes. Together, these steps provide a quantitative procedure for assessing structural preservation following imputation.

Mapper construction

The Mapper algorithm constructs a graph representation of complex data by mapping high-dimensional observations through a filter function, covering the resulting filter space with overlapping intervals, and clustering observations within each interval. Each cluster forms a node in the Mapper graph, while nodes are connected when their corresponding clusters share observations. The resulting graph provides a simplified representation of the topological structure of the data and can reveal patterns of connectivity, branching, and clustering [11]. The structure of a Mapper graph depends on several choices, including the filter function, resolution, overlap, and clustering method [12,16,17]. In the present pipeline, the complete-case dataset and the imputed dataset are represented within a common coordinate and filter space, and the same Mapper construction procedure is applied to both. This provides a consistent basis for comparing their topological structures.

Mapper was selected because it provides an interpretable representation of the organisation of high-dimensional data, allowing structural patterns such as clusters, branches and connections to be visually inspected, while also providing a representation from which topological features can be quantified. This combination of visual interpretability and quantitative assessment is particularly important for TDAssure, which aims to provide users with complementary evidence of whether imputation has altered the underlying structure of their data. Mapper also provides a filter-guided representation of the data, allowing its organisation to be examined through selected aspects of the data rather than solely through pairwise distances. Extended persistence [18] is then computed from the resulting Mapper graphs, so that the quantitative assessment describes the same structural representation that users inspect visually. This provides a coherent framework in which changes following imputation can be both observed and quantified, rather than using a separate topological construction for the visual and quantitative components of the assessment.

The choice of distance metric depends on the type of data being analysed. The pipeline uses Euclidean distance for numerical data, whereas Gower distance may be considered when variables are of mixed types or include categorical information [19]. The framework allows the distance metric to be selected according to the characteristics of the dataset.

Mapper construction requires the specification of cover parameters, of which resolution and gain are the main parameters considered in this study. Resolution determines the number of intervals used to partition each dimension of the filter space. Higher resolution therefore produces a finer partition and can reveal more detailed local structure. Gain controls the overlap between adjacent intervals. Increasing gain increases the extent to which observations may be shared between neighbouring cover elements and generally produces greater connectivity between the resulting Mapper nodes. Figure 1 illustrates the process of Mapper construction.

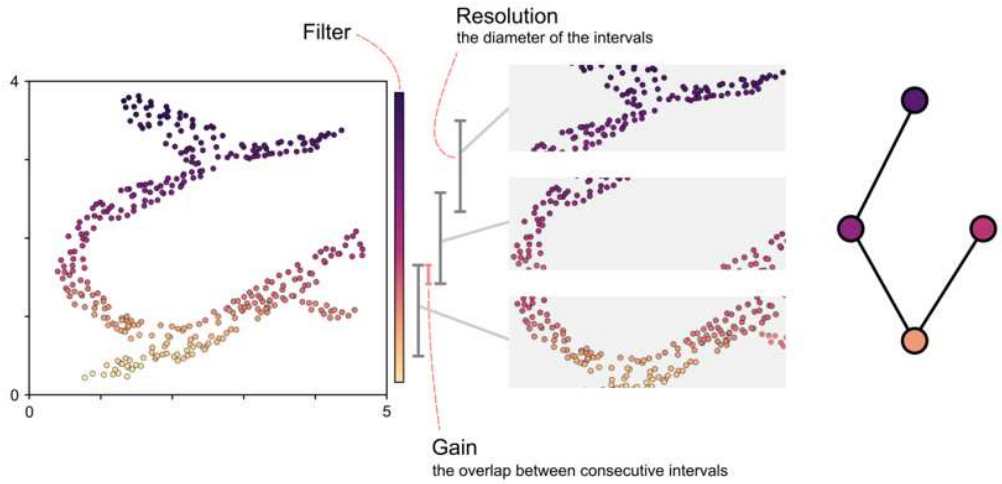


Fig. 1 The Mapper graph construction. Figure adapted from Munch [20].

Within each element of the Mapper cover, observations with similar filter values are further separated into local clusters. Each resulting cluster corresponds to a node in the Mapper graph, while shared observations between clusters determine the edges connecting those nodes. In this pipeline, K-Means is used for the clustering step. Its use is consistent with the Euclidean representation adopted in the present analysis and provides a computationally straightforward procedure for clustering observations within each local cover element.

Rather than fixing the same number of clusters for every cover element, the pipeline allows k to vary according to the number of observations available locally. Users may specify a

minimum cluster size, and for an interval containing *m* observations, the number of clusters is determined as follows:

$$k = \max\left(1, \min\left(\text{base_n_clusters}, \frac{m}{\text{min_cluster_size}}\right)\right)$$

Here, *base_n_clusters* specifies the maximum number of clusters allowed within each cover element, while *min_cluster_size* controls the number of clusters relative to the number of observations (m) in that element. This adjustment limits the number of clusters in sparsely populated cover elements but does not guarantee that every resulting cluster contains at least *min_cluster_size* observations.

Extended persistence and topological comparison

Once the Mapper graphs have been constructed, their topological features are extracted and represented using extended persistence diagrams. Persistence diagrams provide a two-dimensional representation of topological features through birth–death pairs [21]. A birth–death pair records the values at which a topological feature appears and disappears during a filtration. In ordinary persistence, the difference between these values represents the feature's persistence, with larger values generally indicating features that persist over a wider range of the filtration. In the context of Mapper, these features can be obtained directly from the graph through extended persistence. Under appropriate conditions, a Mapper graph can be viewed as a discrete approximation of its continuous analogue, the Reeb graph [16]. The Reeb graph describes changes in the level sets of a real-valued function over the underlying space, and these changes give rise to topological features that can be summarized by their birth and death values.

For each Mapper graph, we extract birth–death pairs corresponding to the topological events considered in the analysis, including connected components, down branches, up branches, and loops. Each event type produces a separate collection of persistence pairs. These persistence pairs provide representations of the corresponding categories of topological features across the complete-case and imputed datasets and are used in the subsequent topological comparison.

The resulting persistence diagrams are compared using the bottleneck distance. Each point in a persistence diagram represents a topological feature, with its coordinates corresponding to the feature's birth and death values. For two persistence diagrams, A and B, the bottleneck distance is defined over possible matchings between their points, with points also permitted to be matched to the diagonal when no corresponding feature is present in the other diagram [22]. For a given matching, the largest L^{∞} distance between paired points is identified, and the bottleneck distance is defined as the minimum, over all admissible matchings, of the maximum L^{∞} distance between matched points:

$$d_B(A, B) = \inf_{\gamma} \sup_{x \in A \cup \Delta} \|x - \gamma(x)\|_{\infty}$$

Here, γ denotes an admissible bijection between the diagrams augmented with diagonal points, and Δ denotes the diagonal.

Figure 2 illustrates this matching procedure. Different bijections produce different maximum pairwise distances, and the bottleneck distance corresponds to the smallest of these maximum distances. In the proposed pipeline, this comparison is performed separately for the different categories of topological features extracted from the Mapper graphs, providing the topological discrepancy measures used in the subsequent permutation test.

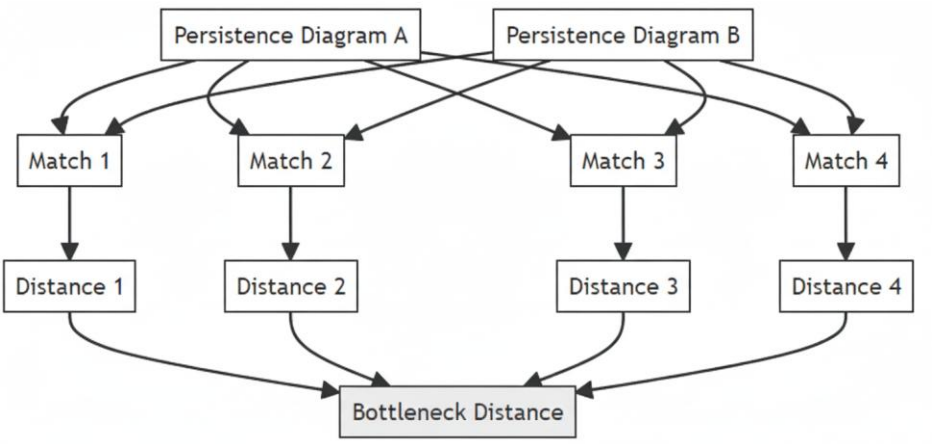


Fig. 2 "Match 1-4" represent possible bijections between the points of the two diagrams. "Distance 1-4" represent the maximum distances between matched points for each bijection. "Bottleneck Distance" is the minimum value of these maximum distances.

Permutation-based statistical inference

The permutation test is used to assess whether the observed topological difference between the complete case and imputed datasets is greater than would be expected under random group assignment**.** The complete-case and imputed observations are pooled, with preprocessing and filter construction performed in this common space. The observed test statistic is first calculated using the original group labels. Group labels are then repeatedly permuted while preserving the original group sizes. For each permutation, Mapper graphs and their corresponding persistence representations are reconstructed for the two permuted groups, and the same topological discrepancy statistic is recalculated. The resulting permutation statistics form an empirical null distribution against which the observed discrepancy is evaluated.

Stage 1: Hypothesis definition. The null hypothesis assumes that there is no systematic difference in topological structure between the complete-case and imputed datasets. Under the permutation null, group labels are assumed to be exchangeable. The alternative hypothesis is that their topological structures differ beyond that expected under random group assignment:

H_0: There is no systematic difference in topological structure between the complete-case and imputed datasets.

H_1: The topological structures of the complete-case and imputed datasets differ beyond that expected under random group assignment.

Under this formulation, the topological discrepancy constitutes the test statistic rather than the null hypothesis itself. A sufficiently large observed discrepancy relative to the permutation distribution provides evidence against the null hypothesis.

Stage 2: Compute the test statistic. For each Mapper parameter configuration, the topological difference between the complete case and imputed datasets is calculated from their extended persistence representations. The analysis considers the feature categories used in the pipeline, including connected components, down branches, up branches, and loops. The corresponding distances are calculated separately, and the test statistic T_{obs} is

defined as the maximum of these distances. This statistic captures the largest observed discrepancy across the topological feature categories considered in the analysis. The same definition is used for both the observed data and every subsequent permutation. This calculation is repeated across all 80 Mapper parameter configurations.

Stage 3: Permutation procedure. The complete case and imputed datasets are concatenated along the sample dimension to form a common data pool. Group labels are then randomly permuted while preserving the original numbers of observations assigned to the two groups. The pooled observations themselves remain unchanged. After each permutation, the observations are separated according to the permuted labels, and Mapper graphs are reconstructed for the two resulting groups using the same parameter configuration and analysis procedure as for the observed data. The corresponding persistence representations are then extracted and the test statistic is recalculated.

Under the null hypothesis, group membership is assumed to be exchangeable. Repeating this process generates the empirical null distribution of the test statistic for each Mapper configuration.

Stage 4: **Compute p-values and adjusted p-values**.

In the above stages, we first compute the observed statistic T_{obs} based on the original grouping (complete vs. imputed). All samples are then pooled together, and group labels are randomly permuted. For each permutation, a new Mapper graph is constructed according to the shuffled labels, from which the persistence diagrams are extracted and the corresponding topological distance $T^{(s)}$ is computed. Repeating this procedure N times yields an empirical null distribution of the test statistic. The p-value is defined as the proportion of permutation statistics that meet or exceed the observed value, using the unbiased estimator:

$$p = \frac{\#\{T^{(s)} \geq T_{\text{obs}}\} + 1}{N + 1}$$

Because the test is repeated across multiple parameter configurations (resolution × gain), multiple p-values are obtained. To account for multiple testing across the grid of

resolution–gain combinations, we adopt a max-T correction in the spirit of the Westfall–Young procedure. For each permutation, we record the maximum test statistic over all grid points and use these maxima to form a global null distribution. The adjusted p-value for a given observed statistic T_{obs} is

$$p_{max} = \frac{1 + \sum_{b=1}^{N} \mathrm{I}\{T_{max}^{(b)} \geq T_{\text{obs}}\}}{N + 1}$$

where $T_{max}^{(b)}$ denotes the maximal statistic in permutation b. This procedure controls the family-wise error rate by comparing each observed statistic to the most extreme values attainable under the joint null across the entire parameter grid. These adjusted p-values allow for a more robust determination of statistically significant topological differences across the parameter space.

Stage 5: Statistical interpretation

Statistical significance is assessed at a threshold of 0.05. An adjusted p-value below 0.05 indicates that the observed topological discrepancy is unlikely under the permutation null distribution and provides evidence of a difference between the topological structures of the complete case and imputed datasets. An adjusted p-value at or above 0.05 indicates that the null hypothesis cannot be rejected; under that Mapper configuration, the analysis does not detect a statistically significant topological difference between the two datasets. Across the parameter grid, the number and distribution of significant adjusted p-values are used to assess the extent to which an imputation method preserves the observed topological structure. The main pipeline steps are summarised in Figure 3.

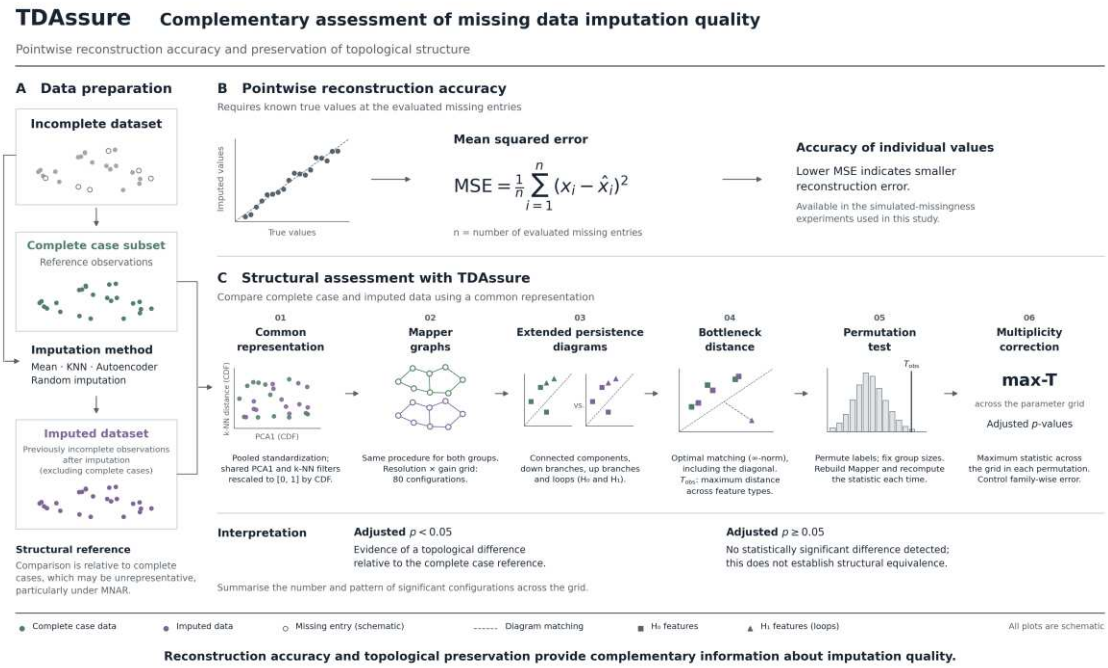


Fig. 3 TDAssure framework for complementary assessment of imputation quality. Conventional evaluation assesses pointwise reconstruction accuracy by comparing imputed and true values using measures such as mean squared error (MSE). TDAssure complements this assessment by evaluating whether an imputation method preserves the topological structure of the data relative to complete cases. Using a common filter representation, Mapper graphs are constructed across multiple parameter settings and summarised using extended persistence diagrams. Bottleneck distances quantify topological differences between complete-case and imputed data, with permutation testing used to assess structural preservation and p-values adjusted across the parameter grid. Together, these complementary perspectives distinguish reconstruction accuracy at the level of individual values from preservation of global data structure.

Datasets and Experimental design

The experiments were designed to evaluate the proposed topology-based framework across different datasets, missingness mechanisms, and imputation methods. Two datasets were analysed using four imputation techniques under three missingness mechanisms (MCAR, MAR, and MNAR) and two target missingness rates (20% and 30%). To provide an

additional comparison, the results were also evaluated against an established imputation assessment method [23]. For each experimental setting, Mapper analyses were performed over a predefined parameter grid, producing adjusted p-values that were used to assess structural preservation following imputation. This design allows imputation performance to be compared across different missingness conditions and provides a consistent basis for evaluating the proposed framework. The datasets used as the starting point for the experiments were complete, containing no missing values; missingness was subsequently introduced according to the experimental conditions described below.

Datasets and Preprocessing

The first dataset is Fashion-MNIST [24], a widely used benchmark consisting of 60,000 training images and 10,000 test images of Zalando clothing items. Each image is 28 × 28 pixels and is represented as a 784-dimensional vector, with pixel intensities taking integer values from 0 to 255. The dataset contains ten clothing categories, including T-shirt/top, coat, sandal, sneaker, and ankle boot, with 6,000 training images in each category.

For this study, the original training and test sets were combined and represented in tabular form, with each image corresponding to a 784-dimensional feature vector and its associated class label. Because the proposed analysis is computationally intensive, a balanced subset was obtained by stratified sampling, with equal numbers of observations selected from each clothing category. The resulting dataset contained 7,000 observations and 784 features and was used for subsequent missingness simulation, imputation, and topological analysis.

The second dataset is the Epileptic Seizure Recognition (ESR) dataset [25], which is commonly used in neurological and biomedical data analysis. It contains 11,500 EEG segments, each represented by 178 continuous features corresponding to uniformly sampled points from a one-second EEG recording. The dataset also provides labels indicating different EEG recording conditions, including seizure activity. Compared with Fashion-MNIST, ESR provides a different numerical data structure, with continuous measurements derived from EEG recordings, and therefore serves as a complementary dataset for evaluating the framework across different data characteristics.

Both datasets are publicly available. The versions used in this study are complete and contain no missing values.

The two datasets consist entirely of numerical variables and are represented as finite point clouds in $\mathbb{R}^d$, where each observation corresponds to a point and the variables define the coordinate dimensions. Euclidean distance is used to measure the distance between observations**.** The input variables are standardized to zero mean and unit variance so that variables measured on larger scales do not disproportionately influence subsequent analyses, including PCA and clustering. Standardization is performed within the common pooled data space to ensure that complete-case and imputed observations are represented on the same scale**.**

For the Mapper construction, we use two complementary filter functions, the first principal component (PCA1) [26] and the k-nearest neighbour distance (k-NN distance), to define a two-dimensional filter space for Mapper (Figure 4). PCA1 summarizes the dominant direction of variation in the data and therefore provides information about its global structure. The k-NN distance, by contrast, reflects local density and allows Mapper to capture variation in local neighbourhood structure. Together, the two filters provide complementary information about global variation and local density. Each observation is represented by a two-dimensional filter vector:

$$\text{fil} = \begin{bmatrix} PCA1_1 & k-NN\ dist_1 \\ PCA1_2 & k-NN\ dist_2 \\ \vdots & \vdots \end{bmatrix}$$

Fig. 4 Construction of the two-dimensional filtering matrix used in the Mapper analysis. The filter combines the first principal component (PCA1), representing global variation in the data, and the k-nearest-neighbour (k-NN) distance, representing local density.

Both filter values are transformed using their cumulative distribution functions (CDFs), rescaling them to the interval [0, 1]. This places the two filter dimensions on a common scale and allows the same Mapper cover specification to be applied consistently across the complete case and imputed datasets.

Because the topology of a Mapper graph can vary with the choice of resolution and gain, the analysis is performed over a predefined parameter grid rather than relying on a single configuration. The same grid is applied to all dataset–imputation comparisons, providing a consistent basis for comparison across experimental settings. In total, 80 combinations are evaluated (Table 1).

Table 1 Mapper parameter grid used in the analysis

Parameter	**Options**
Resolution	5, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35
Gain	0.1, 0.2, 0.3, 0.4

Missing Values Simulation

Three missingness mechanisms — MCAR, MAR, and MNAR — were simulated for both datasets. Following Sousa and Mayer [27], each mechanism was generated algorithmically from the fully observed data, allowing the imputed values to be evaluated against known ground truth.

For each dataset, missingness mechanism, and target missingness rate (20% or 30%), a missingness simulation procedure was applied to generate an incomplete dataset. Under MCAR, entries were removed independently with a constant probability calibrated to achieve the target missingness rate, so that missingness was unrelated to either observed or unobserved values. Under MAR, the probability of missingness depended on a weighted combination of observed variables, making missingness conditionally dependent on information available in the dataset. Under MNAR, the probability of missingness depended on the value of the variable itself, producing non-ignorable missingness. These procedures allowed the three missingness mechanisms to be simulated while retaining the original complete data as a reference for evaluating imputation performance.

Because the proposed framework compares the topological structure of complete case and imputed data, a proportion of observations had to remain fully observed. To provide a

complete-case reference group for the topological comparison, 40% of the observations in each dataset were randomly selected and retained without modification. Missingness was then introduced into the remaining 60% of observations, followed by imputation. This proportion was kept fixed across all experimental settings to ensure a consistent comparison. The target missingness rates of 20% and 30% were defined relative to the entire dataset rather than only to the imputed subset. Accordingly, the missingness simulation procedure within the 60% subset was calibrated to achieve the corresponding target rate across the full dataset. Dataset characteristics are summarized in Table 2, and the resulting missingness patterns for each experimental setting are reported in Table 3. For MAR and MNAR, the simulation procedure was calibrated to approximate the target missingness rate while preserving the dependence structure associated with each mechanism.

Table 2 Dataset summary

Dataset	Sample size	Number of features	Total cells
Fashion-MNIST	7000	784	5488000
Epileptic Seizure Recognition	11500	178	2047000

Table 3 Missingness simulation

	MCAR			MAR			MNAR		
	TMR	AMR	AMC	TMR	AMR	AMC	TMR	AMR	AMC
Fashion-MNIST	20%	20%	1097600	20%	18.76%	1029557	20%	19.93%	1093684
	30%	30%	1646400	30%	28.50%	1564230	30%	29.83%	1637276
Epileptic Seizure Recognition	20%	20%	409400	20%	18.58%	380367	20%	19.67%	402745
	30%	30%	614100	30%	29.18%	597219	30%	29.98%	613694

Note: TMR = Target missingness rate, AMR = Actual missingness rate, AMC = Actual missing count

Imputation Techniques

Four imputation methods were selected to provide a range of performance characteristics for evaluating the proposed framework. KNN imputation and an autoencoder approach [28]

were included as methods capable of using multivariate structure in the observed data. Mean imputation was included as a simple conventional baseline, while random imputation was used as an intentionally weak comparator because it does not use structural information when replacing missing values. Together, these methods provide contrasting levels of reconstruction quality for assessing whether the proposed evaluation procedure can distinguish differences in imputation performance. Because missingness was simulated on originally complete datasets, the true values at all missing positions were known. The mean squared error (MSE) was therefore calculated for each imputation method and used as a reference measure of reconstruction accuracy.

K-Nearest Neighbour (KNN) imputation. KNN imputation replaces each missing value using information from the (k) most similar observations. To maintain the same setting across all experiments, (k) was fixed at 10.

Autoencoder imputation. Autoencoder imputation was implemented as a deep learning approach in PyTorch using a feed-forward architecture with hidden layers of 256, 128, and 64 units, with the 64-unit layer serving as the bottleneck representation. The decoder mirrored this structure to reconstruct the original feature space, and ReLU activation functions were used between hidden layers. The model was trained for 200 epochs with a batch size of 128.

Mean imputation. Mean imputation replaces each missing value with the arithmetic mean of the observed values for the corresponding feature. Because this procedure does not account for relationships among observations or variables, it may reduce variability and alter the structure of the data, making it a useful conventional baseline for comparison.

Random imputation. Random imputation replaces missing values with randomly generated values drawn from the empirical distribution of the corresponding feature. Since the replacement values are generated without using relationships among observations, this method serves as a deliberately weak baseline.

To provide an additional external assessment, we also applied the method proposed by Bak and Hansen [23] to estimate imputation error. Their approach uses the missingness patterns observed in the input data and reproduces these patterns within the complete case subset.

The artificially incomplete complete case data are then imputed using the method being evaluated. Because the original values in the complete case subset are known, the resulting imputation error can be calculated directly. This observed error, together with the corresponding imputed values, is subsequently used by the Bak and Hansen procedure to estimate imputation error for the full dataset under different imputation methods. As with the proposed framework, the reliability of this approach depends in part on how well the complete-case observations represent the data relevant to the imputation assessment. Further methodological details are provided in Bak and Hansen [23].

Results

The experimental results are summarized in Tables 4–7. For each dataset, results are presented separately for the 20% and 30% missingness scenarios and include all four imputation methods evaluated under the three missingness mechanisms (MCAR, MAR, and MNAR). The primary outcome reported in the main text is the number of adjusted p-values below the 0.05 significance threshold across the 80 parameter combinations examined for each experimental setting.

These counts provide a descriptive summary of the number of parameter configurations showing statistically detectable differences, rather than a direct measure of the magnitude of structural disruption. As neighbouring Mapper configurations may produce correlated results, small differences in counts should be interpreted cautiously.

Table 4 Fashion-MNIST, 20% missingness level

Mechanisms	Imputation Techniques	MSE	Number of p-values < 0.05
MCAR	Autoencoder	1734.1660	5
	KNN	1195.0549	3
	Mean	5689.4817	9
	Random	16517.9702	13
MAR	Autoencoder	2399.5121	5
	KNN	1395.3268	3
	Mean	6076.0995	8
	Random	16115.1501	12
MNAR	Autoencoder	4116.3416	6

	KNN	1997.6903	7
	Mean	7908.2961	7
	Random	15232.7948	12

Table 5 Fashion-MNIST, 30% missingness level

Mechanisms	Imputation Techniques	MSE	Number of p-values <0.05
MCAR	Autoencoder	2427.1969	5
	KNN	1245.2967	4
	Mean	5684.1827	7
	Random	16529.8373	9
MAR	Autoencoder	2982.3932	6
	KNN	1441.4452	2
	Mean	5911.9627	9
	Random	16213.1559	10
MNAR	Autoencoder	4497.8716	8
	KNN	2059.7996	8
	Mean	7156.2275	9
	Random	15535.5940	10

Table 6 Epileptic Seizure Recognition, 20% missingness level

Mechanisms	Imputation Techniques	MSE	Number of p-values < 0.05
MCAR	Autoencoder	3509.8869	3
	KNN	18647.9205	5
	Mean	26662.3001	6
	Random	53685.7186	13
MAR	Autoencoder	7825.2321	1
	KNN	16970.5763	5
	Mean	23334.8121	6
	Random	51275.1658	12
MNAR	Autoencoder	16596.2590	4
	KNN	29575.3296	3
	Mean	32997.8865	1
	Random	58689.8949	16

Table 7 Epileptic Seizure Recognition, 30% missingness level

Mechanisms	Imputation Techniques	MSE	Number of p-values < 0.05
MCAR	Autoencoder	5315.9269	1
	KNN	20880.0742	5
	Mean	26778.5027	6
	Random	54074.6057	12
MAR	Autoencoder	10191.3425	3
	KNN	18663.1126	5
	Mean	23581.6242	7
	Random	51817.0354	13
MNAR	Autoencoder	15743.9170	3
	KNN	27378.0148	3
	Mean	27044.7483	1
	Random	54636.7355	28

MSE results

Because the missing values were simulated from originally complete datasets, MSE could be calculated directly against the known true values and therefore provides a reference measure of reconstruction accuracy. For the Epileptic Seizure Recognition (ESR) dataset, the autoencoder achieves the lowest MSE across all missingness mechanisms and both missingness rates, while random imputation consistently produces the highest values. KNN generally performed better than mean imputation, with the exception of the 30% MNAR setting, in which mean imputation produced a slightly lower MSE than KNN (27,044.75 vs. 27,378.01). Overall, reconstruction error tended to increase with the missingness rate, although the magnitude of this effect varied across imputation methods and missingness mechanisms.

For Fashion-MNIST, KNN consistently produces the lowest MSE across all missingness mechanisms and missingness rates, followed by the autoencoder, mean imputation, and random imputation. The ranking of the four methods was therefore highly consistent across the Fashion-MNIST experimental conditions. KNN remained the best-performing method

under both 20% and 30% missingness, whereas random imputation produced substantially larger reconstruction errors than the other methods across all three missingness mechanisms.

TDA results

The TDA results broadly follow the patterns observed in the MSE results. Across both datasets, autoencoder, KNN, and mean imputation generally produce fewer adjusted p-values below 0.05 under MCAR and MAR, whereas random imputation consistently produces larger counts. The relative ordering of the methods is therefore broadly consistent with that indicated by the true MSE results, particularly in distinguishing random imputation from the other methods.

The overall number of significant adjusted p-values is relatively small across most experimental settings. Each significant adjusted p-value represents a Mapper parameter configuration for which a statistically detectable topological difference between the complete-case and imputed datasets remains after correction for multiple testing. The number of significant configurations is therefore interpreted comparatively across imputation methods within the same dataset, missingness mechanism, and missingness rate, rather than as an absolute measure of imputation quality.

The MNAR results show a different pattern. Although the true MSE values continue to distinguish among the imputation methods, the TDA results provide less separation between autoencoder, KNN, and mean imputation. This is particularly evident for the ESR dataset, where mean imputation produces fewer significant adjusted p-values than the other non-random methods despite not consistently achieving the lowest MSE. Random imputation, however, remains clearly distinguishable and produces the largest number of significant results under MNAR in both datasets. These findings indicate that pointwise reconstruction accuracy and topological preservation do not necessarily provide identical rankings of imputation methods, particularly under MNAR.

Across the experimental conditions, random imputation produced the most consistent evidence of topological disruption, with the largest number of significant adjusted p-values in almost all settings. By contrast, differences among autoencoder, KNN, and mean

imputation were smaller and varied according to the dataset and missingness mechanism. Overall, the TDA results broadly corresponded with MSE when distinguishing strongly performing from deliberately weak imputation, while also capturing differences in structural preservation that were not directly reflected by pointwise reconstruction error.

Comparison with an existing imputation assessment method

To further examine the consistency of the proposed framework, the experimental results were compared with those obtained using the evaluation method of Bak and Hansen. Because this method could not be directly applied to the autoencoder procedure used in the present study, the comparison was restricted to KNN, mean, and random imputation. The estimated imputation errors across the two datasets, missingness mechanisms, and missingness rates are summarized in Table 8.

Table 8 Imputation error estimates obtained using the Bak and Hansen method

Dataset	Missingness (%)	Mechanism	KNN	MEAN	RANDOM
Fashion-MNIST	20	MCAR	0.6801	5.1624	8.9039
Fashion-MNIST	20	MAR	0.6129	3.8682	5.9442
Fashion-MNIST	20	MNAR	0.6529	4.6706	7.3610
Fashion-MNIST	30	MCAR	0.9931	8.0439	14.0252
Fashion-MNIST	30	MAR	1.1756	7.3365	12.0521
Fashion-MNIST	30	MNAR	1.2321	8.0579	12.8659
Epileptic Seizure Recognition	20	MCAR	0.3938	0.5109	4.3921
Epileptic Seizure Recognition	20	MAR	0.2864	0.3615	2.1740
Epileptic Seizure Recognition	20	MNAR	0.3905	0.5100	4.1209
Epileptic Seizure Recognition	30	MCAR	0.5728	0.7761	5.8325
Epileptic Seizure Recognition	30	MAR	0.5656	0.6482	5.4147
Epileptic Seizure Recognition	30	MNAR	0.5642	0.7658	5.5536

Across the experimental settings, the Bak and Hansen method produces a consistent ordering of the three imputation methods. KNN yields the lowest estimated error in all settings, followed by mean imputation, while random imputation consistently produces the highest error. For Fashion-MNIST, the estimated errors for KNN range from 0.6129 to 1.2321, compared with 3.8682–8.0579 for mean imputation and 5.9442–14.0252 for

random imputation. A similar separation is observed for ESR, where KNN again produces lower estimated errors than mean and random imputation across all conditions.

This ordering is broadly consistent with the patterns observed in the true MSE and TDA results, particularly in identifying random imputation as the weakest-performing method. KNN also consistently shows lower estimated error than mean imputation using the Bak and Hansen method, in agreement with the true MSE results in most experimental settings. The TDA results show broader agreement in distinguishing random imputation but do not always reproduce the same ordering between KNN and mean imputation, particularly under MNAR. This is consistent with the different aspects of imputation quality captured by the approaches: whereas MSE-based methods assess reconstruction error, the proposed TDA framework evaluates preservation of topological structure.

Practical implementation: TDAssure

To support practical use of the proposed methodology, we implemented the analysis pipeline as an interactive Shiny for Python application, TDAssure. The application allows users to upload datasets, specify Mapper parameters, run the permutation procedure, and inspect the resulting topological and statistical outputs through a graphical interface.

Figures 5 and 6 illustrate the main components of the application. Figure 4 shows the Mapper graph visualization generated during the analysis, while Figure 5 presents the pipeline monitoring interface and statistical results. TDAssure integrates the components of the proposed framework within a single workflow, enabling users to examine Mapper representations alongside the corresponding statistical assessment of topological differences. Adjusted p-values are presented across the Mapper parameter grid, allowing users to identify the configurations for which statistically significant topological differences are detected.

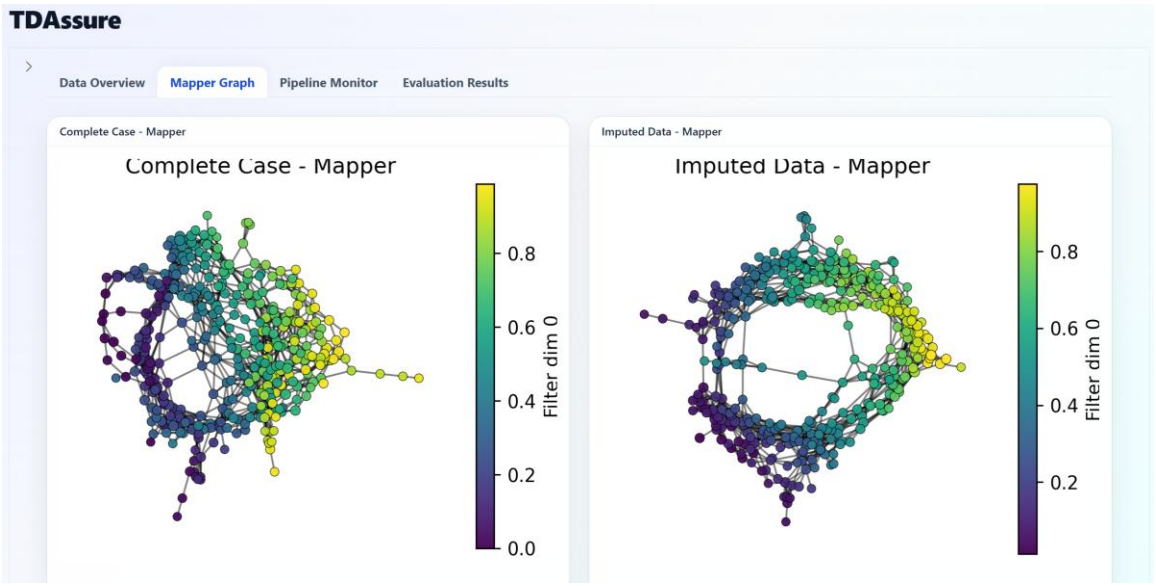


Fig. 5 Mapper graph visualization generated using TDAssure.

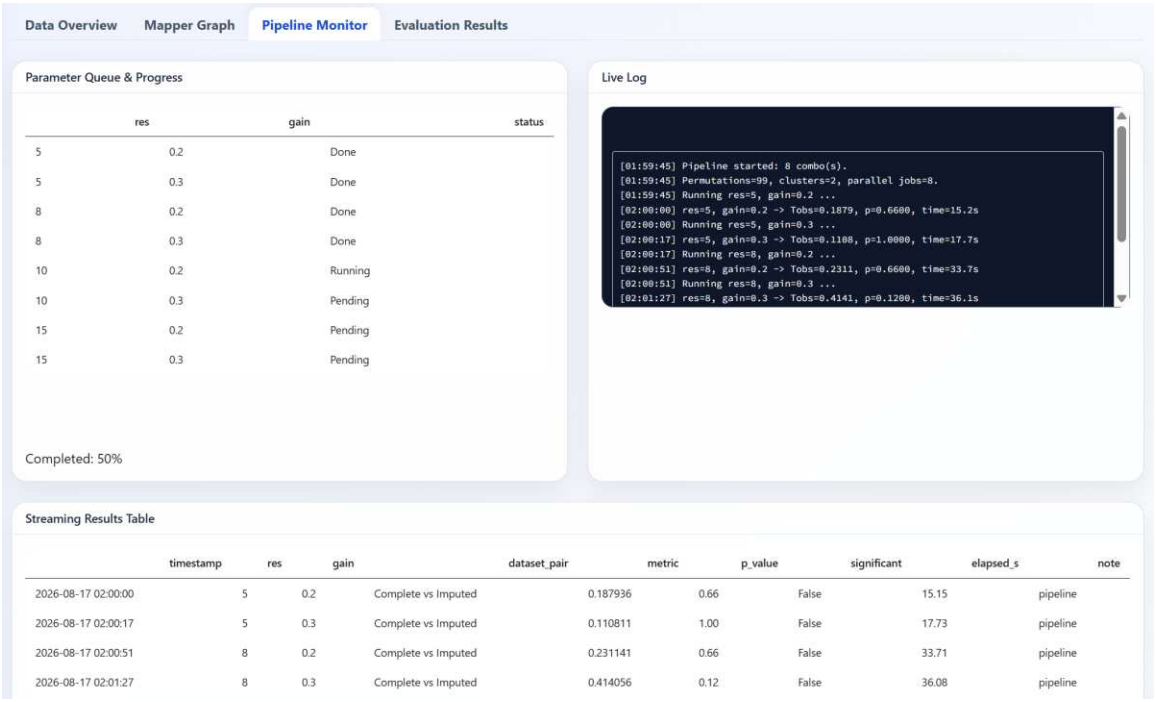


Fig. 6 Pipeline monitoring interface of TDAssure, showing an illustrative run with a reduced Mapper parameter grid.

The source code and user documentation for TDAssure are publicly available at https://github.com/GeYiyang/tda-imputation-quality-evaluator.

Discussion

This study introduces TDAssure, a topology-based framework for assessing whether missing-data imputation preserves the structural characteristics of a dataset as a measurement of imputation quality. Across two datasets, three missingness mechanisms, two missingness rates, and four imputation methods, the framework identified differences in topological preservation that were broadly consistent with conventional measures of reconstruction accuracy while providing complementary information about changes in data structure. In particular, random imputation, included as an intentionally weak comparator, produced the greatest reconstruction error and the most consistent evidence of topological disruption across experimental settings. By contrast, methods that used information from the observed data generally resulted in fewer but still statistically detectable topological differences between the complete-case and imputed datasets.

An important finding is that pointwise reconstruction accuracy and topological preservation were related but not equivalent. Under MCAR and MAR, the patterns identified by TDAssure broadly corresponded with those observed using true MSE and the Bak and Hansen approach, particularly in distinguishing random imputation from better-performing methods [23]. However, the relative ordering of autoencoder, KNN, and mean imputation was not always preserved across evaluation approaches. This distinction is important because conventional error measures evaluate how accurately individual missing values are reconstructed, whereas TDAssure evaluates whether imputation alters higher-level structural characteristics of the data. This distinction is consistent with previous work highlighting the potential complementary value of topological representations for examining structural changes associated with missingness and imputation [14,15,29]. An imputation method may therefore achieve relatively low reconstruction error without necessarily providing the greatest preservation of topological structure, or conversely may preserve aspects of global structure despite less accurate reconstruction of individual values. The proposed framework should consequently be viewed as complementary to, rather than a replacement for, conventional measures of imputation accuracy.

The experimental results illustrate how these two perspectives can lead to different conclusions. For example, under 20% MCAR in Fashion-MNIST, the two assessments were closely aligned: KNN produced the lowest MSE (1195.05) and the fewest significant adjusted p-values (3), followed by the autoencoder, mean imputation, and random imputation. In contrast, under 30% MNAR in ESR, the autoencoder achieved the lowest MSE (15,743.92), whereas mean imputation produced the fewest significant adjusted p-values (1, compared with 3 for both the autoencoder and KNN). The latter result should not be interpreted as evidence that mean imputation is universally preferable. Rather, it shows that pointwise reconstruction accuracy and preservation of the topological structure represented by Mapper can differ, particularly under MNAR. In practice, disagreement between the two assessments can therefore be used to identify cases in which an imputation method reconstructs individual values accurately but alters structural characteristics of the data, or conversely preserves aspects of structure despite larger pointwise errors. Such cases warrant inspection of both the quantitative results and the corresponding Mapper representations rather than selection of a method from a single ranking alone.

Differences between pointwise and topological assessment were particularly apparent under MNAR, where TDAssure provided less separation among autoencoder, KNN, and mean imputation despite differences in their MSE. In naturally incomplete datasets, MNAR mechanisms may cause complete cases to differ systematically from observations containing missing values [2,6]. In the present simulations, however, the complete-case reference was randomly selected before missingness was introduced. Therefore, selection of complete cases by the missingness mechanism does not directly explain the differences observed under MNAR. Nevertheless, the potential lack of representativeness of complete cases remains an important limitation when applying the framework to naturally incomplete datasets. Random imputation remained consistently distinguishable, suggesting that the framework can detect substantial structural disruption even under more challenging missingness conditions.

The results also varied according to the characteristics of the dataset. KNN achieved the lowest reconstruction error across all Fashion-MNIST conditions, whereas the autoencoder performed best on ESR. These differences illustrate that imputation performance is data

dependent and reinforce the value of evaluating imputation within the structural context of the dataset being analysed rather than assuming that a single method will perform optimally across settings. More importantly for the proposed framework, TDAssure was able to evaluate structural preservation across two datasets with substantially different dimensionality and underlying characteristics, suggesting potential applicability beyond a single data domain.

TDAssure extends previous applications of TDA to missing-data assessment by combining Mapper-based representation with quantitative topological comparison and permutation-based statistical inference. Earlier approaches demonstrated the value of Mapper for visualising structural changes associated with missingness and imputation [14], while subsequent work introduced quantitative measures of similarity between topological representations [15,29]. The present framework builds on these developments by evaluating the statistical significance of observed topological discrepancies against an empirical null distribution as a novelty, and accounting for the multiple Mapper parameter configurations considered in the analysis. This is particularly relevant because Mapper representations depend on parameter choices such as resolution and gain [11,16]. Evaluating a predefined parameter grid and correcting across these configurations reduces reliance on conclusions drawn from a single Mapper specification and provides a more systematic assessment of structural preservation.

Several limitations should be considered. First, the experimental evaluation was conducted using two fully observed benchmark datasets in which missingness was introduced algorithmically. This design was necessary to provide known ground-truth values for comparison using MSE, but simulated missingness cannot reproduce all of the complexities encountered in real-world incomplete datasets [6]. The framework also evaluates topological preservation relative to the complete-case subset, which may not fully represent the underlying structure of the data, particularly when complete-case status is related to observed or unobserved values. This limitation is especially relevant under MNAR, where complete cases may differ systematically from observations containing missing values [2,6]. In addition, the framework requires a sufficient number of complete cases to provide a meaningful reference for topological comparison, which may limit its

applicability when most observations contain missing values. Further evaluation using naturally occurring missing data across different domains and data structures is therefore warranted. Second, the present experiments focused on numerical data and used Euclidean distance, although the framework permits mixed or categorical data [19]. Third, the validity of permutation-based inference relies on the exchangeability assumption. In naturally incomplete datasets, this assumption may be violated when the missingness mechanism induces systematic differences between complete-case and incomplete observations, particularly under MAR or MNAR. Consequently, statistically significant topological differences should not necessarily be attributed to imputation alone. Fourth, Mapper representations remain dependent on methodological choices including filter functions, cover parameters, and clustering procedures [11,16]. Although the parameter-grid approach reduces dependence on a single resolution and gain combination, it does not imply that all configurations are equally informative. Extreme combinations of Mapper parameters may produce highly fragmented or overly connected graphs that provide limited structural information. The max-T adjustment accounts for multiple testing across the parameter grid, but it does not by itself determine whether an individual Mapper configuration is topologically meaningful. Significant results arising from isolated configurations should therefore not be interpreted independently of the broader pattern across the grid. Future work should incorporate stability or graph-quality criteria to identify poorly informative parameter settings before statistical comparison.

The implementation of the framework as the interactive TDAssure application is intended to facilitate its use without requiring users to implement the underlying topological procedures directly. By integrating Mapper visualisation, persistence-based comparison, and permutation inference, the application enables structural changes following imputation to be examined alongside conventional assessments of imputation quality. Further work should evaluate the framework across a broader range of datasets, imputation approaches, and naturally occurring missingness patterns. Automated or stability-guided selection of Mapper parameters could reduce reliance on manual parameter specification, while sensitivity analyses or auxiliary modelling strategies may help address settings in which complete cases provide a less representative structural reference, particularly under MNAR. Computational efficiency could also be improved through parallel computation,

approximate algorithms, or adaptive parameter grids. More broadly, these results support the use of structural preservation as an additional dimension of imputation quality, complementing pointwise reconstruction accuracy when evaluating the consequences of missing-data handling.

Conclusion

This study presents TDAssure, a topology-based framework for evaluating whether missing-data imputation preserves the structural characteristics of a dataset. By combining Mapper-based representation, persistence-based topological comparison, and permutation-based statistical inference, TDAssure provides an assessment of imputation quality that complements conventional measures of pointwise reconstruction accuracy. Across the experimental settings examined, the framework broadly distinguished methods associated with greater structural disruption while also demonstrating that reconstruction accuracy and topological preservation do not necessarily provide identical assessments of imputation performance. By providing both a methodological framework and an interactive implementation, TDAssure offers a practical approach for incorporating structural preservation into the evaluation of missing-data imputation.

Abbreviations

AMC	Actual missing count
AMR	Actual missingness rate
CCA	Complete case analysis
CDF	Cumulative distribution function
EEG	Electroencephalogram
ESR	Epileptic Seizure Recognition
KNN	k-nearest neighbour
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputation by chained equations
MNAR	Missing not at random
MSE	Mean squared error
PCA	Principal component analysis
PCA1	First principal component
ReLU	Rectified linear unit
TCDM	Topology-based clinical data mining

TDA	Topological data analysis
TMR	Target missingness rate

Acknowledgements

This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Author contributions

Y.G. developed the methodology, performed the data analysis, and drafted and revised the manuscript. M.C. contributed to the methodology development and critically reviewed the manuscript. D.S. co-supervised the research, and critically reviewed the manuscript. R.I. conceived the study, supervised the research, contributed to the methodology development, and critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Funding

Yiyang Ge and Raquel Iniesta were funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London grant NIHR203318.

Availability of data and materials

Datasets are freely available. Code has been uploaded: https://github.com/GeYiyang/TDA-missing-value-imputation-quality-evaluation

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence. 2003;17:519–33. https://doi.org/10.1080/713827181

2. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley; 2002.

3. Buuren S van, Groothuis-Oudshoorn K. **mice** : Multivariate Imputation by Chained Equations in *R*. J Stat Softw. 2011;45. https://doi.org/10.18637/jss.v045.i03

4. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat. 1992;46:175–85. https://doi.org/10.1080/00031305.1992.10475879

5. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28:112–8. https://doi.org/10.1093/bioinformatics/btr597

6. van Buuren S. Flexible Imputation of Missing Data, Second Edition. Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC; 2018. https://doi.org/10.1201/9780429492259

7. Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009;46:255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

8. Edelsbrunner H. Geometry and Topology for Mesh Generation. Cambridge University Press; 2001. https://doi.org/10.1017/CBO9780511530067

9. Zomorodian A, Carlsson G. Computing Persistent Homology. Discrete Comput Geom. 2005;33:249–74. https://doi.org/10.1007/s00454-004-1146-y

10. Ghrist R. Homological algebra and data. 2018. p. 273–325. https://doi.org/10.1090/pcms/025/06

11. Iniesta R, Carr E, Carriere M, Yerolemou N, Bertrand M, Chazal F. Topological Data Analysis and its usefulness for precision medicine studies. SORT. 2022;115–36.

12. Singh G, Mémoli F, Carlsson G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. Eurographics Symposium on Point-Based Graphics 2007. 2007. p. 91–100. https://doi.org/10.2312/SPBG/SPBG07/091-100

13. Chazal F, Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. 2017;1–38. http://arxiv.org/abs/1710.04019

14. Arioli A, Dagliati A, Geary B, Peek N, Kalra PA, Whetton AD, et al. OptiMissP: A dashboard to assess missingness in proteomic data-independent acquisition mass spectrometry. PLoS One. 2021;16:e0249771. https://doi.org/10.1371/journal.pone.0249771

15. Glushakov S, Kotenko I. Handling Missing Data in Clinical Trials Using Topological Data Analysis [Internet]. 2018. https://www.semanticscholar.org/paper/Handling-Missing-Data-in-Clinical-Trials-Using-Data-Glushakov-Kotenko/96f4411017ba2a80553927d0a1d2492a119874d8

16. Carrière M, Oudot S. Structure and Stability of the One-Dimensional Mapper. Foundations of Computational Mathematics. 2018;18:1333–96. https://doi.org/10.1007/s10208-017-9370-z

17. Chazal F. High-Dimensional Topological Data Analysis. 2016.

18. Cohen-Steiner D, Edelsbrunner H, Harer J. Extending Persistence Using Poincaré and Lefschetz Duality. Foundations of Computational Mathematics. 2009;9:79–103. https://doi.org/10.1007/s10208-008-9027-z

19. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;27:857. https://doi.org/10.2307/2528823

20. Munch E. A User's Guide to Topological Data Analysis. Journal of Learning Analytics. 2017; 4(2):47–61. https://doi.org/10.18608/jla.2017.42.6

21. Edelsbrunner H, Letscher D, Zomorodian A. Topological Persistence and Simplification. Discrete Comput Geom. 2002;28:511–33. https://doi.org/10.1007/s00454-002-2885-2

22. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of Persistence Diagrams. Discrete Comput Geom. 2007;37:103–20. https://doi.org/10.1007/s00454-006-1276-5

23. Bak N, Hansen LK. Data Driven Estimation of Imputation Error—A Strategy for Imputation with a Reject Option. PLoS One. 2016;11:e0164464. https://doi.org/10.1371/journal.pone.0164464

24. Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. 2017;

25. Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys Rev E. 2001;64:061907. https://doi.org/10.1103/PhysRevE.64.061907

26. Jolliffe IT. Principal component analysis. Springer; 2004.

27. de Sousa TA, Mayer I. How to generate missing values? R-miss-tastic. 2021. Available from: https://rmisstastic.netlify.app/how-to/generate/misssimul

28. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 2006;313:504–7. https://doi.org/10.1126/science.1127647

29. Glushakov S, Group I, Kotenko I, Rekalo AM. Topology-based Clinical Data Mining for Discovery of Hidden Patterns in Multidimensional Data. 2017. https://api.semanticscholar.org/CorpusID:34420073