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.

Metaobject representation of multimodal data with missing values for intelligent data analysis

Nikolay A. Ignatev
National University of Uzbekistan named after Mirzo Ulugbek
Shamsiddin Ramazonov
ramazonov377@gmail.com

National University of Uzbekistan named after Mirzo Ulugbek

Research Article

Keywords: Metaobject, Feature stability, Membership functions, Missing data, Mode comparison, Explainable artificial intelligence

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11409855/v1

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

Additional Declarations: No competing interests reported.

Metaobject representation of multimodal data with missing values for intelligent data analysis

Nikolay A. Ignatev[1] · Shamsiddin Ramazonov[1,*]

[1] Department of Software Engineering and Artificial Intelligence, Faculty of Applied Mathematics and Intelligent Technologies, National University of Uzbekistan named after Mirzo Ulugbek, 4 University Street, Tashkent 100174, Uzbekistan

[*] Corresponding author: Shamsiddin Ramazonov, ramazonov377@gmail.com

Nikolay A. Ignatev: n_ignatev@rambler.ru, ORCID https://orcid.org/0000-0002-7150-5837; Shamsiddin Ramazonov: ramazonov377@gmail.com, ORCID https://orcid.org/0009-0004-9774-6104

Abstract

We propose an approach to analysing data with missing values that dispenses with mandatory imputation and instead constructs a unified metadescription of a sample from stability indices of heterogeneous features. The central concept is the metaobject, a vector of aggregate training-sample characteristics whose components reflect the stability of discrimination between classes or modes using quantitative and nominal features. For quantitative features, stability is calculated through a minimum cover of values by non-overlapping intervals with membership functions; for nominal features, it is calculated from category proportions against a uniform reference associated with a symmetric Dirichlet model. Metaobjects yield a quality score for comparing modes, allowing mode pairs to be ordered, informative features to be identified, and interpretable explanations of differences to be obtained without reconstructing missing values. An experiment on 974,155 records of the North Sea International Bottom Trawl Survey (NS-IBTS), divided into 100 modes by catch weight, illustrates the approach for quantitative, nominal, and heterogeneous feature spaces. When up to 50% of the observed values were additionally removed at random or in a value-dependent way, the ordering of all 4,950 mode pairs remained highly stable (Spearman $\rho \geq 0.979$). Compared with eight conventional missing-data methods, the approach reproduced the true quality scores most accurately, with a 2–5 times smaller error than imputation methods, at a computational cost comparable to

the fastest simple methods. Large language models can subsequently verbalize the precomputed and verifiable metadescription in explainable artificial intelligence systems.

Keywords Metaobject · Feature stability · Membership functions · Missing data · Mode comparison · Explainable artificial intelligence

1 Introduction

Many established machine-learning procedures require complete feature vectors, whereas observational datasets frequently contain missing measurements. In clinical and cohort studies, missing values can arise from incomplete examinations and follow-up, among other causes [1, 2]. Medical data provide a motivating example here; the experiment in this paper uses fisheries survey data. The distinction between missing completely at random, missing at random, and missing not at random is essential when assessing the assumptions behind an analysis [3, 4]. Ignoring the missingness mechanism can bias inference, and imputation does not automatically guarantee valid statistical conclusions [4].

Missing-value recovery can involve conditional regression for quantitative variables and conditional classification for nominal variables, as in chained-equation approaches [5]. Random-forest-based methods also provide imputation procedures for mixed-type data [6, 7]. These procedures should be distinguished from analysing incomplete observations without reconstructing their missing entries. Comparisons in clinical data, metabolomics, and binary-variable simulations show that imputation performance depends on the data and evaluation setting [1, 8, 9].

Evaluation should distinguish reconstruction error from performance on the subsequent analytical task. A cohort-study comparison assessed both MAE/RMSE and predictive discrimination and obtained different rankings across evaluation criteria [2]. Distributional agreement and the preservation of relevant relationships therefore provide complementary evaluation objectives.

In this paper, missing values are handled without imputation. Throughout the paper, *multimodal data* means data whose objects are partitioned into several modes, that is, groups with different feature distributions; in the experiment, the modes are 100 groups of objects formed by catch weight. The term does not refer to data from several modalities, such as text, images, or audio. Each mode of a multimodal

sample is represented by a metaobject, a vector of feature stability indices computed only from observed values, and pairs of modes are compared in this unified space. The main contributions are as follows:

1. a stability-based metaobject representation for heterogeneous (quantitative and nominal) features that does not require reconstruction of missing values;
2. a quality score C(Q) for ordering pairs of modes and identifying the features that distinguish them, together with the number of observations on which each conclusion is based;
3. a computational experiment on NS-IBTS survey data in which 100 modes and all 4,950 mode pairs are analysed in quantitative, nominal, and heterogeneous feature spaces, together with a robustness study under up to 50% additional random (MCAR) and value-dependent (MNAR) missing values;
4. an experimental comparison with eight conventional missing-data methods, from complete-case analysis to MICE and missForest, against a known true ordering of the mode pairs;
5. a design for explainable artificial intelligence in which a language model only verbalizes precomputed and verifiable metaobject characteristics.

The paper is organized as follows. Section 2 reviews related work, and Sect. 3 states the research objective. Section 4 defines the stability index, the metaobject, and the quality score. Section 5 shows how metaobjects can support explainable artificial intelligence. Section 6 reports the experiment on NS-IBTS data, including a robustness study and a comparison with conventional missing-data methods. Section 7 discusses limitations, and Sect. 8 concludes the paper.

2 Related work

The background of the present study includes empirical-data learning [10] and feature selection with missing observations [11]. Earlier recovery approaches include curve-based modelling of incomplete tables [12] and hybrid extreme gradient boosting for precipitation data [13]. Deep-learning foundations are described in [14]; the present work focuses on feature-wise statistical descriptions rather than training a deep generative model. Chained-equation and random-forest imputation provide established methods for recovering incomplete observations [5, 7]. The present study instead constructs feature-wise descriptions for comparing groups. In Sect. 6.7, these

and other conventional methods are compared experimentally with the proposed approach.

The problem of accounting for missing values becomes even more important in Big Data settings. Here, missing-value handling is considered in the context of large-scale, multimodally distributed data. One possible solution is the notion of a metaobject [15–17], which is central to this paper. Its definition is constructive rather than purely axiomatic. The accompanying references provide context on metaknowledge and admissible object descriptions; the specific stability-based definition used here is given below.

Definition. A metaobject is a vector of aggregate characteristics of a training sample, whose components are stability indices of measurable or synthetic features calculated for that sample. A metaobject describes the entire training sample rather than an individual object.

Related work examines the choice of object-relation structures in metric classification [18] and the correspondence of class-relation structures across description manifolds [19]. These studies provide context for representing relations among samples.

From a theoretical perspective, a metaobject is a second-order object constructed not from original measurements but from statistics of relations among objects in a training sample. The prefix meta therefore denotes a transition to a higher level of data description. The construction of a training sample consisting of objects from two disjoint classes requires clarification. Two cases are considered:

- objects in both classes are represented by original measured features and belong to different categories;
- objects in one class have an original category (mode label), whereas objects in the opposite, alternative class are generated artificially.

The stability of heterogeneous features is calculated from the membership-function values of their observed values in the classes and is used to describe metaobjects in a unified space without missing values. In this space, features lose their explicit measurement units and their admissible values are reduced to a single fixed interval. For quantitative features, the representation is invariant to measurement scale and hence to the units of measurement. This invariance is achieved by partitioning quantitative feature values into non-overlapping intervals while taking object

categories into account. The original description of objects by measurable features may contain missing values.

3 Research objective

The central concept of this study is the metaobject, interpreted as an image of a training sample in a space of relations. Whereas an ordinary object is described by a system of features, a metaobject is described by a system of statistical characteristics. Thus, it provides a statistical description of a training sample. In machine learning, metaobjects enable a transition from learning on individual objects to learning on sample descriptions; the unit of analysis becomes the training sample as a whole.

Different metaobject representations can be obtained when the descriptions of one class are generated from hyperparameters of distribution laws. This is necessary for multimodal distributions. It makes it possible to obtain information about the actual feature distribution in each mode without imputing missing values. A metaobject is therefore a vector of aggregate sample characteristics based on feature stability indices that reflect neighborhood relations among objects.

In general, a metaobject may be formed from original measurable features or from synthetic features [20], for example those derived from k-nearest-neighbor estimates with category information. Synthetic features can support comparison of alternative data representations, metrics, normalization schemes, and feature-construction methods. They are not used in the present work. Missing values require an explicit policy for distance computation; mixed-data similarity measures can account for unavailable components, but the present construction uses feature-wise observed frequencies [21].

Existing regression- and classification-based imputation methods [22] are usually evaluated by their accuracy in reconstructing artificially removed values. MAE and RMSE are common for quantitative features, whereas classification accuracy, F1 score, and AUC are used for nominal features. However, accurate recovery of individual values does not ensure preservation of the statistical structure of the original sample or the quality of subsequent data analysis. Comparative studies therefore also assess the effects of imputation on downstream predictive tasks [2]. Real missing values are especially challenging because their true values are

unknown and reconstruction error cannot be evaluated directly. Analysis should consequently be supplemented by criteria that assess preservation of distributions, correlations, and other structural characteristics.

Feature stability is a combinatorial estimate calculated from real data samples. It is based on object membership-function values. For quantitative features, it is obtained by solving a minimum-cover problem in which the set of values is partitioned into non-overlapping intervals. In each interval, one of the two classes dominates in frequency, corresponding to a membership-function value greater than 0.5. The number of intervals in a minimum cover may differ across samples from the same population. The stability framework is related to the sample-relation approach in [15]; the stated convergence result requires the conditions of its original proof. This property provides the methodological basis for the proposed explanatory-system construction.

The informativeness of a feature is indicated by the proximity of its stability value to 1 [23, 24]. Feature-space reduction is performed by ordering features according to stability. Here, stability is the feature-ranking criterion adopted in this study; the general feature-selection literature does not establish this particular formula. Stability vectors obtained for different pairs of classes are treated as metaobject representations in a unified space. The proposed approach does not regard missing-value imputation as a compulsory intermediate step; it maps incomplete source data directly to a unified metadescription of sample structure based on class relations. The choice of class combinations is determined by the research objective. An important task is to construct an interpretable explanation of interclass differences from the resulting metadescription.

4 Mathematical model

Before giving the formal definitions, we summarize the approach in plain terms. For two modes K_a and K_b:

1. for every feature, only the objects whose value of this feature is observed are used; missing values are skipped, not filled in;
2. the observed values of a quantitative feature are sorted and split into intervals in which one of the two modes dominates; for a nominal feature, every category is used as it is;

3. from the shares of the two modes in these intervals or categories, a stability index $g_i \in [0, 1]$ is computed; it shows how reliably the feature separates the two modes;
4. the indices of all features form the metaobject $Q = (g_1, \ldots, g_n)$, a description of the pair of modes without missing values;
5. the quality score $C(Q) = 1 - \text{mean}(g_i)$ summarizes the metaobject: a small $C(Q)$ means that the modes differ strongly, and a large $C(Q)$ means that they are similar. Ordering all pairs by $C(Q)$ and features by g_i gives the basis for the explanation.

The main notation is as follows: E_0 is the set of objects; $X(n) = (x_1, \ldots, x_n)$ is the set of features; $K_1, \ldots, K_l$ are the modes (classes); K_0 is an artificially generated alternative class; T_{jc} is the set of objects of K_j with an observed value of feature x_c; f_c is the membership function; g_c (or g_i) is the stability index of a feature; Q is a metaobject; and $C(Q)$ is the quality score.

Let $E_0 = \{S_1, \ldots, S_m\}$ be a set of objects described by n heterogeneous quantitative and nominal features: $X(n) = (x_1, \ldots, x_n)$. Missing (unmeasured) values are permitted. Assume that the set E_0 is partitioned into l modes according to a mode-defining feature. Each mode is assigned an index $j \in \{1, \ldots, l\}$ and is interpreted as the corresponding class K_j. The objectives are as follows:

- to construct a set TUP of descriptions of class K_0, which is an alternative to the set of object modes, from prescribed hyperparameters of a fixed set of distribution laws;
- to map every mode into a unified space as $\eta = |\text{TUP}|$ vectors corresponding to metaobjects;
- to identify distribution laws characteristic of every feature in every mode;
- to construct metaobjects for every pair of modes and compare them by a dedicated quality criterion;
- to construct an evidence-based metadescription of each pair of modes, identify features that provide their discriminability, assess the reliability of conclusions in view of the available observations, and formulate an interpretable explanation without imputing missing values.

Consider an alternative class K_0 whose object descriptions are generated from a distribution density $\chi(*) \in$ TUP. For quantitative features, hyperparameters are estimated from object descriptions in E_0. For every feature of class K_j, the degree of proximity between its empirical density and $\chi(*)$ is of interest.

Class-specific normalization permits comparison of samples of different sizes. However, finite-sample variation in a generated reference class can still affect interval boundaries and stability estimates. The size of K_0 and sensitivity to its generation should therefore be reported.

Feature values generated according to a prescribed distribution for K_0 are used only to construct an ordered sequence jointly with feature values of K_j, $j \in \{1, \ldots, l\}$. When interval boundaries for class dominance are calculated, the relevant quantity is not the absolute number of feature values but their class-specific proportion, taking missingness into account.

For a nominal feature with $z(c)$ categories, a symmetric Dirichlet model has equal expected category proportions, $1/z(c)$ [25]. Equation (5) uses these deterministic uniform reference proportions. A random draw from a symmetric Dirichlet distribution need not have exactly equal components.

4.1 Metaobject construction in the presence of missing values

Membership functions also have a classical role in fuzzy decision-making [26]. Here their interpretation is specified by the normalized class frequencies in Eq. (3); the cited general framework is not a source for the particular interval criterion or its convergence theorem.

Missing values complicate direct application of conventional complete-vector distances; appropriate missing-aware similarities are possible [21]. A further difficulty is that nominal and quantitative features cannot be jointly used in the alternative class K_0. Consequently, a tabular representation of K_0 can be constructed only in special cases.

Consider the set of class pairs $\{(K_0, K_i)\}$ used to form a set of metaobjects. Here K_i is the set of objects in a fixed mode, and K_0 is an artificially generated set of objects used as the alternative class in a binary problem. The term opposition for K_0 was

previously used in the calculation of generalized object estimates regarded as a numerical solution of a multicriteria problem with an initially unfixed set of criteria.

For feature values $x_c \in X(n)$ describing objects in $K_0 \cup K_j$, $j = 1, \ldots, l$, construct the nondecreasing ordered sequence

$$r_1, \ldots, r_i, \ldots, r_{h(c)}, \quad h(c) = \left|K_0 \cup T_{jc}\right| \tag{1}$$

Here T_{jc} is the set of objects in class K_j that have no missing value for feature x_c. To partition sequence (1) into p_c ($p_c \geq 2$) non-overlapping intervals $\{[r_u; r_v]^i\}$, $1 \leq u \leq v \leq h(c)$, $i = 1, \ldots, p_c$, the following interval criterion is applied:

$$\left|\frac{d_{0c}(u,v)}{|K_0|} - \frac{d_{jc}(u,v)}{|T_{jc}|}\right| \to \max \tag{2}$$

where $d_{0c}(u,v)$ and $d_{jc}(u,v)$ denote the numbers of objects from K_0 and K_j that fall into interval $[r_u; r_v]^i$. Values belonging to interval $[r_u; r_v]^i$ are treated as a separate category of a nominal feature. A one-to-one mapping of the categories to $\{1, \ldots, p_c\}$ is assumed. Applying the algorithm based on criterion (2) yields a minimum cover of all values in sequence (1) by non-overlapping intervals.

For brevity, $d_{0c}(u,v)$ and $d_{jc}(u,v)$ for interval $[r_u; r_v]^\mu$ of feature $x_c \in X(n)$ are denoted below by $d_{0c}(\mu)$ and $d_{jc}(\mu)$. The membership-function value $f_c(\mu)$ for class K_j in interval $[r_u; r_v]^\mu$ (category $\mu \in \{1, \ldots, p_c\}$) is computed as follows:

$$f_c(\mu) = \frac{d_{jc}(\mu)/|T_{jc}|}{d_{0c}(\mu)/|K_0| + d_{jc}(\mu)/|T_{jc}|} \tag{3}$$

Because nominal-feature categories are unordered, when calculating $f_c(\mu)$ for K_0 and category $\mu \in \{1, 2, \ldots, p_c\}$, $d_{0c}(\mu)$ and $d_{jc}(\mu)$ are the numbers of K_0 and K_j objects with value μ. If $f_c(\mu) = 0.5$ for $x_c \in X(n)$, the following condition holds for category $\mu \in \{1, 2, \ldots, p_c\}$:

$$\frac{d_{0c}(\mu)}{|K_0|} = \frac{d_{jc}(\mu)}{|T_{jc}|}$$

For the set $\{S_i \in K_0 \cup T_{jc}\}$, $i = 1, \ldots, h(c)$, $S_i = (a_{i1}, \ldots, a_{in})$, where $a_{ic} \in \{1, 2, \ldots, p_c\}$ is calculated by (3), define the stability index of feature $x_c \in X(n)$:

$$g_c = \frac{1}{h(c)} \sum_{i=1}^{h(c)} \begin{cases} f_c(a_{ic}), & f_c(a_{ic}) > 0.5, \\ 1 - f_c(a_{ic}), & f_c(a_{ic}) < 0.5, \\ 0, & f_c(a_{ic}) = 0.5. \end{cases} \tag{4}$$

Equation (4) retains $h(c)$ in the denominator and assigns zero to exact ties. Under this convention its general range is [0, 1]; it lies in (0.5, 1] when no ties occur. If tied

observations are excluded instead, the denominator must also be replaced by the number of retained observations, and the estimate is undefined if none remain. In practical terms, g_c close to 1 means that in almost every interval or category one of the two classes clearly dominates, so the feature separates the classes well, whereas g_c close to 0.5 means that both classes are represented in similar proportions everywhere, so the feature cannot tell them apart. The dependence of the stability index on the number of intervals used to partition the feature values is established by the following theorem.

Theorem 1 (Theorem 1 in [15]). As the number of objects in the sample E_0, partitioned into two disjoint classes, increases without bound, the feature-stability value converges in probability to a constant Φ, where $\Phi \in (0.5, 1]$. The proof in [15] invokes the law of large numbers and is stated for the feature-stability index defined in the cited work; its application to index (4) assumes the same conditions.

The quantity g_c in (4) is interpreted as the value of a metaobject feature in the unified feature space. The stability of class separation by feature x_i is g_i. The value $g_i = 1$ corresponds to maximum stability of class discrimination determined by observed-value frequencies in the compared classes. The value g_i characterizes discriminability by the feature; it is not, by itself, a measure of the statistical reliability of the conclusion. Reliability must also be assessed using the number of available observations for this feature in each compared class.

If feature $x_c \in X(n)$ has $z(c)$ categories, the membership function for class K_j is calculated for $\mu \in \{1, 2, \ldots, z(c)\}$ as follows:

$$f_c(\mu) = \frac{d_{jc}(\mu)/|T_{jc}|}{d_{jc}(\mu)/|T_{jc}| + 1/z(c)} \quad (5)$$

The quantity $1/z(c)$ is the proportion of K_0 objects under a uniform distribution across the feature categories. The specific category value μ is not material.

4.2 Quality assessment of class comparisons

The variability of metaobject descriptions is determined by the values of their components in the unified space. Each component reflects the degree of difference between two classes for the corresponding feature. Stability index (4) reaches its minimum when the empirical feature distributions of the compared classes are close.

Let $Q = (g_1, \ldots, g_n)$ be a vector of aggregate characteristics whose component g_i reflects the stable separation of two classes by feature x_i. The quality of mapping to the unified space is defined by

$$C(Q) = 1 - \frac{1}{n}\sum_{i=1}^{n} g_i \quad (6)$$

Score (6) can also be computed for subsets of $X(n)$, for example, separately for quantitative and nominal features, as is done in Sect. 6. Because $C(Q)$ is one minus the mean stability, in the tie-free case it ranges from 0, when every feature separates the two modes perfectly, to 0.5, when no feature separates them. The following linguistic interpretation of the quality score $C(Q)$ is proposed:

- $C(Q) \in [0, 0.25]$: strong separation under the proposed score;
- $C(Q) \in (0.25, 0.5)$: weak separation when all $g_i \in (0.5, 1]$.

In this tie-free case, $C(Q) \in [0, 0.5)$. With zero contributions for ties and the original denominator, the general bound is instead $C(Q) \in [0, 1]$; the interpretation then also depends on the proportion of ties. The threshold 0.25 is a descriptive convention, not a significance level. These descriptive comparisons should not be interpreted as formal distributional hypothesis tests.

One possible analysis is to determine, for the i-th feature of class (mode) K_j, the minimum value of the index defined by (4) relative to the alternative class K_0. In this case, when comparing two modes K_a and K_b, it is useful to report not only the stability value for the i-th feature but also the distribution law that best matches the distribution of that feature in each of the two modes.

Rather than imputing missing values, feature characteristics are constructed exclusively from available observations and represented in the space of relations between classes. The task is therefore not recovery of a missing value as such, but generation of a stable description of data structure suitable for subsequent intelligent analysis under missingness. Pairwise mode comparisons are also metaobjects; the set of $I(I - 1)/2$ metaobjects can be studied using established data-mining methods.

5 Metaobject-based explainable intelligence

The proposed method for processing data with missing values is directly connected to explainable artificial intelligence. Unlike local model-explanation methods [27, 28],

which interpret a model decision for an individual object, this approach explains the difference between two modes of a data distribution.

Each mode K_j corresponds to a set of objects in the original sample, whose descriptions may contain missing values. Avoiding mandatory imputation prevents reconstructed entries from being treated as observations. It does not eliminate selection bias caused by informative missingness [3, 4].

For every pair of modes K_a and K_b, a metaobject $Q_{ab} = (g_1, \ldots, g_n)$ is constructed, where g_i characterizes the stable discrimination of the modes by feature x_i. Metaobject components are therefore both data characteristics and elements of the evidence base for explaining differences between modes.

An important property of this representation is that feature discriminability can be separated from the completeness of data used to estimate it. Therefore, together with g_i, the number of available observations N_i and the relative proportion of observed values $R_i = N_i/(|K_a| + |K_b|)$ should be retained. Consequently, an explanation can state not only which feature most strongly distinguishes two modes but also the number of observations on which the conclusion is based. This is especially important for real data, where the mechanism of missingness may be related to the structure of the population under study.

An explanation can be constructed by ranking features according to their g_i values. For each feature included in the explanation, the characteristic distribution law, the number of available observations, and other computed characteristics may also be reported. This yields a structured explanation of the following form: “Modes K_a and K_b differ primarily in features x_i, x_j, …, with the highest stability observed for x_i; the estimate is based on N_i observations.” This representation allows the metaobject to serve as a basis for an evidence-based explanation. Unlike a generative explanation produced directly from the original data table, each element of the textual conclusion can be linked to a specific component of the metaobject and the procedure used to compute it.

On this basis, large language models (LLMs) can be used to verbalize analysis results automatically. The language model must not determine the existence of a statistical regularity or invent quantitative characteristics absent from the data. Its role is to transform a precomputed, structured metadescription into domain-language text.

This is a design requirement for grounding generated text in verifiable quantities. Because language generation can produce unsupported statements [29], the output still requires factual checking; no LLM evaluation is reported here.

Thus, the proposed technology shifts attention from processing missing values to explaining the structure of incomplete data. Instead of recovering unknown values, it forms a unified metadescription, while stability indices serve simultaneously as descriptors of mode discriminability and evidence for explanation. The metaobject is an intermediate level between the source data and an explainable artificial intelligence system.

An LLM verbalizes evidence rather than generating it. A conventional machine-learning model answers the question: which class does an object belong to? The proposed approach addresses a different question: how do two groups of objects differ, and which data properties support that difference? It explains the structure of the difference between two modes rather than an individual model decision. It differs from SHAP and LIME as follows:

- SHAP [30] explains a model decision for an object;
- LIME [31] constructs a local approximation of a model;
- a metaobject explains a difference between two samples;
- missing values are not replaced by assumed values;
- the explanation is based only on actually observed data;
- every claim can be linked to a specific feature, its stability, the number of observations, and a characteristic distribution.

When explaining differences between modes using metaobjects, an LLM should receive only computed and verifiable facts about the metaobject. It must not independently introduce quantitative claims that are absent from its metadescription. Developing explainable intelligence requires a conceptual transition to the following representation:

$$Q_{ab} = \{(g_i, N_i, R_i, D_i, U_i)\}_{i\in\{1,\dots,n\}}$$

where g_i is the stability of discrimination between modes based on feature x_i; N_i is the number of observations used; R_i is the proportion of available observations; D_i is the most characteristic distribution law; and U_i is a measure of confidence in, or reliability of, the estimate.

The vector $\{g_i\}$, $i \in \{1, \ldots, n\}$, remains the primary unified representation, whereas the remaining quantities serve as explanatory attributes of the metaobject. The uncertainty attribute U_i is a proposed extension, not a computed result in the current experiment. Counts and observed proportions describe data support but are not themselves confidence intervals.

In other words, a meta-feature answers the question “How well does the feature distinguish between modes?”, whereas the explanatory attributes address the questions “What data support the conclusion?”, “What distribution is observed?”, and “How reliable is the conclusion?”. The explanation process should proceed without reconstructing unknown values and should rely on three concepts:

1. discriminability;
2. the contribution of a feature to the explanation;
3. the reliability of the explanation.

The first two concepts are determined by g_i, whereas the reliability of the explanation is primarily related to the number of observations N_i and, possibly, to the proportion of available observations R_i.

6 Computational experiment

This section first describes the data and the experimental protocol (Sects. 6.1 and 6.2). Sections 6.3–6.5 report the ordering of mode pairs for quantitative, nominal, and heterogeneous features. Section 6.6 examines how stable this ordering is when additional values are missing, and Sect. 6.7 compares the approach with conventional ways of handling missing values.

6.1 Data and preprocessing

The experiment used data from the North Sea International Bottom Trawl Survey (NS-IBTS), obtained from the DATRAS database of the International Council for the Exploration of the Sea (ICES) [32]. The dataset was compiled from HL (length-based) records of surveys conducted in the first and third quarters of 2015–2024. In total, it contained 974,155 records described by 29 fields of the DATRAS HL exchange format. Not all of these fields were used in the computational experiment (Table 1). Three fields, RecordType, Survey, and SpecCodeType, are service attributes of the exchange format, and each takes a single constant value across all records (HL, NS-

IBTS, and W, respectively). Since a constant feature carries no information about differences between objects, its stability value cannot reflect any distinction between modes, and these fields were therefore excluded.

Table 1 Fields of the DATRAS HL format that were not used as descriptive features

Field	Description	Reason for exclusion
RecordType	Record type of the exchange format	Constant value (HL) in all records
Survey	Survey name	Constant value (NS-IBTS) in all records
SpecCodeType	Species coding system	Constant value (W, WoRMS) in all records
CatCatchWgt	Catch weight of a species category in a haul	Used as the modal attribute defining K1–K100

The CatCatchWgt feature, which gives the catch weight of a species category in a haul, was not treated as a descriptive feature, because it was used as the modal attribute for forming the classes. It was selected for this role because it represents the key outcome of the survey and varies over a wide continuous range. Including it in the feature space would have made the comparison of modes trivial, since the modes are defined by its values. In addition, 13,025 records with a missing CatCatchWgt value could not be assigned to any mode and were removed, which left a working sample of 961,130 objects. The remaining observations were sorted in ascending order of CatCatchWgt and partitioned into 100 consecutive groups of approximately equal size. Each group was treated as a separate mode, denoted by K_1, K_2, …, K_{100}. The resulting mode labels were stored in CatCatchWgt_class and subsequently used to construct and compare metaobjects. Thus, the metaobjects were constructed from the remaining 25 features: 8 quantitative (SweepLngt, HaulNo, TotalNo, NoMeas, SubFactor, SubWgt, LngtClass, and HLNoAtLngt) and 17 nominal (Quarter, Country, Ship, Gear, GearEx, DoorType, StNo, Year, SpecCode, SpecVal, Sex, CatIdentifier, LngtCode, DevStage, LenMeasType, DateofCalculation, and Valid_Aphia). Missing values in these features were retained and handled without imputation.

6.2 Experimental protocol

To identify patterns related to the proximity of quantitative-feature distributions to known laws, twelve distributions were analysed for every mode: normal, uniform,

exponential, gamma, lognormal, Weibull, chi-square, Student's t, beta, Laplace, logistic, and Pareto. For every feature, the parameters of the corresponding distribution were estimated and its stability was calculated. Three linguistic labels describe agreement between the selected distribution law and the empirical data according to stability values:

- (0.50, 0.70]: close agreement;
- (0.70, 0.83]: satisfactory agreement;
- (0.83, 1.00]: weak agreement.

All $I(I-1)/2 = 4{,}950$ unordered pairs of the 100 modes were then evaluated using (6), separately for the quantitative, nominal, and heterogeneous feature sets. For visualization, the 10 pairs with the lowest and the 10 pairs with the highest values of $C(Q)$ were projected onto the first two principal components. Principal component analysis (PCA) was used only for visualization: missing components were replaced by feature medians, features were standardized, and two components were extracted [33]. Finally, the robustness of the ordering to additional missing values was assessed by masking experiments (Sect. 6.6).

6.3 Results for quantitative features

Table 2 presents the results for mode 9. Ordering the 4,950 pairs by the resulting score makes it possible to identify latent patterns while accounting for missing values. Table 3 presents the 10 pairs with the smallest and largest quality scores for quantitative features.

Table 2 Distribution characteristics and stability of features in mode 9

No.	Feature	Type	Distribution	Stability	Agreement
1	SweepLngt	Quantitative	Pareto	0.848	Weak
2	HaulNo	Quantitative	Uniform	0.849	Weak
3	TotalNo	Quantitative	Chi-square	0.829	Satisfactory
4	NoMeas	Quantitative	Lognormal	0.858	Weak
5	SubFactor	Quantitative	Chi-square	0.873	Weak
6	SubWgt	Quantitative	Chi-square	0.743	Satisfactory
7	LngtClass	Quantitative	Weibull	0.841	Weak
8	HLNoAtLngt	Quantitative	Lognormal	0.870	Weak
9	Quarter	Nominal	Dirichlet	0.522	Close
10	Country	Nominal	Dirichlet	0.564	Close
11	Ship	Nominal	Dirichlet	0.651	Close
12	Gear	Nominal	Dirichlet	0.666	Close
13	GearEx	Nominal	Dirichlet	0.643	Close

No.	Feature	Type	Distribution	Stability	Agreement
14	DoorType	Nominal	Dirichlet	0.642	Close
15	StNo	Nominal	Dirichlet	0.875	Weak
16	Year	Nominal	Dirichlet	0.521	Close
17	SpecCode	Nominal	Dirichlet	0.764	Satisfactory
18	SpecVal	Nominal	Dirichlet	0.834	Weak
19	Sex	Nominal	Dirichlet	0.554	Close
20	CatIdentifier	Nominal	Dirichlet	0.792	Satisfactory
21	LngtCode	Nominal	Dirichlet	0.597	Close
22	DevStage	Nominal	Dirichlet	0.666	Close
23	LenMeasType	Nominal	Dirichlet	0.826	Satisfactory
24	DateofCalculation	Nominal	Dirichlet	0.727	Satisfactory
25	Valid_Aphia	Nominal	Dirichlet	0.771	Satisfactory

Table 3 Mode pairs with the lowest and highest quality scores for quantitative features

No.	Mode pair	Number of objects	Quality score C(Q)	Separation
1	(1, 100)	19272	0.1598	Strong
2	(1, 99)	19360	0.1624	Strong
3	(1, 98)	19304	0.1687	Strong
4	(1, 94)	19289	0.1700	Strong
5	(1, 97)	19333	0.1705	Strong
6	(2, 100)	19849	0.1713	Strong
7	(1, 96)	19301	0.1728	Strong
8	(2, 99)	19937	0.1731	Strong
9	(1, 95)	19313	0.1742	Strong
10	(3, 100)	18903	0.1759	Strong
…	…	…	…	…
4941	(5, 7)	18924	0.3970	Weak
4942	(6, 8)	19421	0.3973	Weak
4943	(9, 10)	19110	0.3988	Weak
4944	(3, 4)	18711	0.3989	Weak
4945	(2, 3)	19620	0.3993	Weak
4946	(6, 7)	19276	0.4014	Weak
4947	(8, 9)	19106	0.4016	Weak
4948	(4, 5)	18853	0.4030	Weak
4949	(7, 8)	19035	0.4034	Weak
4950	(5, 6)	19310	0.4056	Weak

Using the ordering in Table 3, the first pair (modes 1 and 100) and the last, 4,950th pair (modes 5 and 6) were analysed. The results are shown in Tables 4 and 5.

Table 4 Distributions and stability of quantitative features in modes 1 and 100

No.	Feature	Distribution (mode 1)	Distribution (mode 100)	Stability (mode 1)	Stability (mode 100)
1	SweepLngt	Pareto	Pareto	0.850	0.849

No.	Feature	Distribution (mode 1)	Distribution (mode 100)	Stability (mode 1)	Stability (mode 100)
2	HaulNo	Uniform	Normal	0.842	0.860
3	TotalNo	Chi-square	Normal	0.829	0.877
4	NoMeas	Lognormal	Exponential	0.864	0.867
5	SubFactor	Chi-square	Normal	0.858	0.897
6	SubWgt	Chi-square	Chi-square	0.735	0.831
7	LngtClass	Exponential	Weibull	0.826	0.856
8	HLNoAtLngt	Lognormal	Student's t	0.844	0.866

Table 5 Distributions and stability of quantitative features in modes 5 and 6

No.	Feature	Distribution (mode 5)	Distribution (mode 6)	Stability (mode 5)	Stability (mode 6)
1	SweepLngt	Pareto	Pareto	0.848	0.851
2	HaulNo	Uniform	Uniform	0.846	0.848
3	TotalNo	Chi-square	Chi-square	0.871	0.831
4	NoMeas	Lognormal	Lognormal	0.856	0.859
5	SubFactor	Exponential	Exponential	0.870	0.878
6	SubWgt	Chi-square	Chi-square	0.740	0.742
7	LngtClass	Exponential	Exponential	0.837	0.842
8	HLNoAtLngt	Lognormal	Lognormal	0.867	0.871

The reported examples associate similar scores with some common selected distribution families. Selecting the same family does not establish equality of fitted parameters or distribution densities. To visualize differences between modes in the sequence ordered by score (6), Fig. 1 presents a principal-component visualization of metaobjects based on the eight quantitative features. PCA summarizes variance along orthogonal directions [33]. Ten metaobjects at the beginning of the sequence and ten at the end were selected for the analysis. The two groups are separated along PC1, which explains 91.18% of the variance.

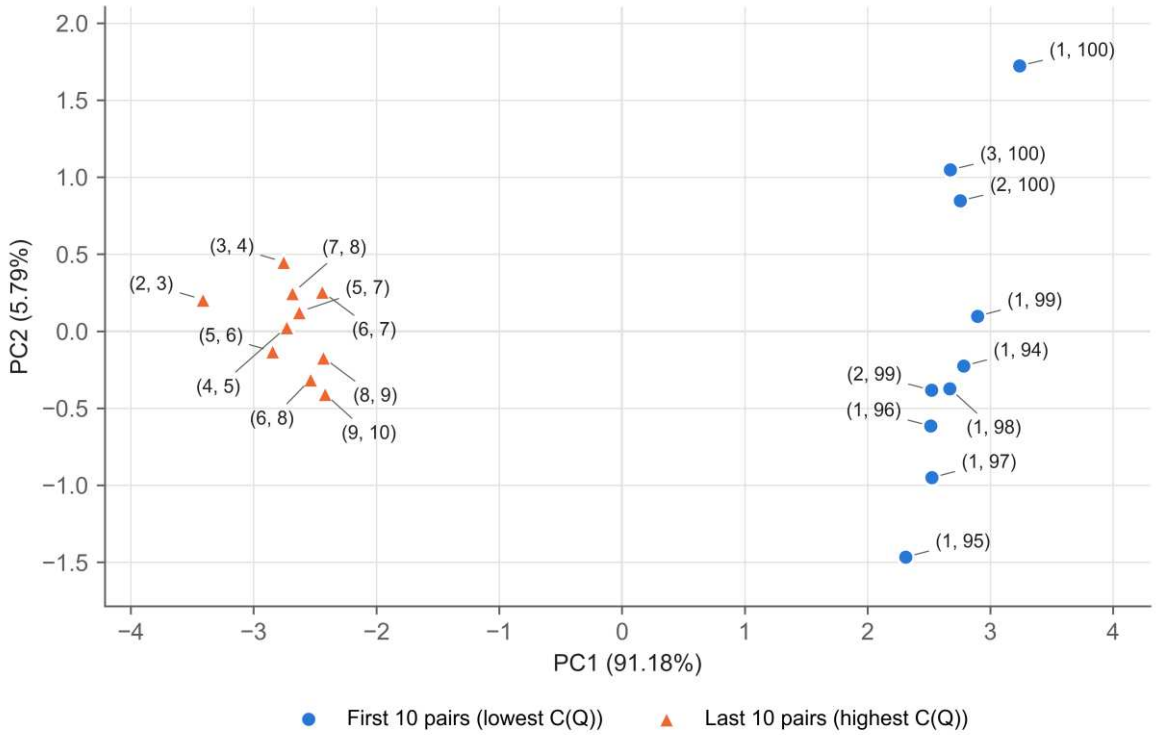


Fig. 1 PCA projection of the first 10 (lowest C(Q)) and last 10 (highest C(Q)) mode-pair metaobjects built from 8 quantitative features. PC1 explains 91.18% and PC2 explains 5.79% of the variance (96.97% in total). Circles denote the first 10 pairs and triangles denote the last 10 pairs

6.4 Results for nominal features

Quality score (6) was also used to analyse patterns in mode pairs described by nominal features. Analogously to Table 3, Table 6 presents the ordered results for all 4,950 mode pairs.

Table 6 Mode pairs with the lowest and highest quality scores for nominal features

No.	Mode pair	Number of objects	Quality score C(Q)	Separation
1	(1, 100)	19272	0.3478	Weak
2	(1, 99)	19360	0.3610	Weak
3	(2, 100)	19849	0.3615	Weak
4	(1, 94)	19289	0.3637	Weak
5	(1, 97)	19333	0.3642	Weak
6	(1, 98)	19304	0.3663	Weak
7	(1, 95)	19313	0.3672	Weak
8	(1, 96)	19301	0.3672	Weak
9	(3, 100)	18903	0.3691	Weak
10	(1, 91)	19322	0.3692	Weak
…	…	…	…	…

No.	Mode pair	Number of objects	Quality score C(Q)	Separation
4941	(8, 9)	19106	0.4739	Weak
4942	(7, 9)	18961	0.4739	Weak
4943	(11, 12)	19428	0.4742	Weak
4944	(7, 8)	19035	0.4746	Weak
4945	(5, 6)	19310	0.4748	Weak
4946	(15, 16)	19260	0.4749	Weak
4947	(10, 12)	19343	0.4750	Weak
4948	(3, 4)	18711	0.4750	Weak
4949	(4, 5)	18853	0.4757	Weak
4950	(6, 7)	19276	0.4761	Weak

Analysis of Tables 3 and 6 using quality score (6) reveals substantial differences in mode descriptions based on quantitative and nominal features. The following feature-wise reference profiles use the uniform nominal baseline $1/z(c)$. These profiles must be distinguished from stability estimates computed directly between two real modes. For Tables 7 and 8, the absolute difference between the two reference-profile components is

$$\Delta g_i = \left| g_i^{(a)} - g_i^{(b)} \right|$$

The descriptive label is Weak when $\Delta g_i \leq 0.01$ and Strong otherwise. This difference is not the stability index obtained by directly comparing K_a and K_b and is not a statistical significance test. Results for modes 1 and 100 are presented in Table 7, and results for modes 6 and 7 are presented in Table 8.

Table 7 Reference stability profiles of nominal features in modes 1 and 100

No.	Feature	Stability (mode 1)	Stability (mode 100)	Δg_i	Difference
1	Quarter	0.522257	0.521788	0.000469	Weak
2	Country	0.567070	0.549447	0.017623	Strong
3	Ship	0.651517	0.648848	0.002669	Weak
4	Gear	0.666830	0.667163	0.000333	Weak
5	GearEx	0.638715	0.651732	0.013017	Strong
6	DoorType	0.640595	0.661978	0.021383	Strong
7	StNo	0.902804	0.774222	0.128582	Strong
8	Year	0.520520	0.522928	0.002408	Weak
9	SpecCode	0.852875	0.831029	0.021846	Strong
10	SpecVal	0.878030	0.800224	0.077806	Strong
11	Sex	0.572144	0.556221	0.015923	Strong
12	CatIdentifier	0.811489	0.828112	0.016623	Strong
13	LngtCode	0.588842	0.652787	0.063945	Strong
14	DevStage	0.666302	0.666261	0.000041	Weak
15	LenMeasType	0.832925	0.817016	0.015909	Strong
16	DateofCalculation	0.740575	0.696250	0.044325	Strong

No.	Feature	Stability (mode 1)	Stability (mode 100)	Δg_i	Difference
17	Valid_Aphia	0.842101	0.826052	0.016049	Strong

Table 8 Reference stability profiles of nominal features in modes 6 and 7

No.	Feature	Stability (mode 6)	Stability (mode 7)	Δg_i	Difference
1	Quarter	0.522120	0.522100	0.000020	Weak
2	Country	0.561762	0.562795	0.001033	Weak
3	Ship	0.649741	0.650040	0.000299	Weak
4	Gear	0.666828	0.667102	0.000274	Weak
5	GearEx	0.641699	0.643808	0.002109	Weak
6	DoorType	0.640330	0.640983	0.000653	Weak
7	StNo	0.880804	0.872869	0.007935	Weak
8	Year	0.521328	0.521555	0.000227	Weak
9	SpecCode	0.769009	0.766635	0.002374	Weak
10	SpecVal	0.859551	0.835427	0.024124	Strong
11	Sex	0.558698	0.557341	0.001357	Weak
12	CatIdentifier	0.810156	0.792405	0.017751	Strong
13	LngtCode	0.598938	0.597840	0.001098	Weak
14	DevStage	0.666302	0.666968	0.000666	Weak
15	LenMeasType	0.822419	0.840751	0.018332	Strong
16	DateofCalculation	0.725685	0.724773	0.000912	Weak
17	Valid_Aphia	0.776386	0.769355	0.007031	Weak

For a visual representation of mode differences in the sequence ordered by score (6), Fig. 2 presents a principal-component visualization of metaobjects based on the 17 nominal features.

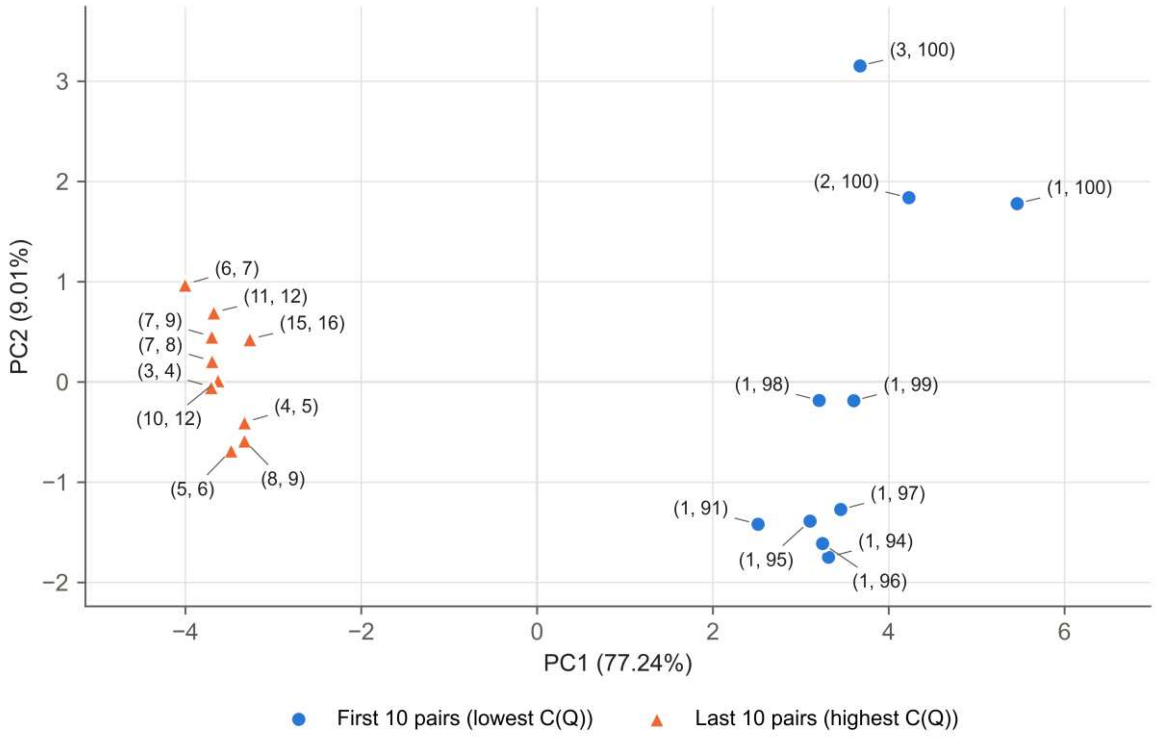


Fig. 2 PCA projection of the first 10 and last 10 mode-pair metaobjects built from 17 nominal features. PC1 explains 77.24% and PC2 explains 9.01% of the variance (86.25% in total). Circles denote the first 10 pairs and triangles denote the last 10 pairs

6.5 Joint analysis of heterogeneous features

The ordering of metaobjects by heterogeneous features is presented in Table 9. Figure 3 presents the corresponding visualization. The first two components account for 88.61% of the variance, compared with 86.25% for nominal features. These projections describe selected extreme-ranked pairs; they do not establish separation of all 4,950 pairs or robustness to additional missingness.

Table 9 Mode pairs with the lowest and highest quality scores for heterogeneous features

No.	Mode pair	Number of objects	Quality score C(Q)	Separation
1	(1, 100)	19272	0.2876	Weak
2	(1, 99)	19360	0.2975	Weak
3	(2, 100)	19849	0.3006	Weak
4	(1, 94)	19289	0.3017	Weak
5	(1, 97)	19333	0.3022	Weak
6	(1, 98)	19304	0.3031	Weak

No.	Mode pair	Number of objects	Quality score C(Q)	Separation
7	(1, 96)	19301	0.3050	Weak
8	(1, 95)	19313	0.3054	Weak
9	(3, 100)	18903	0.3073	Weak
10	(2, 99)	19937	0.3081	Weak
…	…	…	…	…
4941	(11, 12)	19428	0.4486	Weak
4942	(7, 9)	18961	0.4488	Weak
4943	(10, 12)	19343	0.4490	Weak
4944	(2, 3)	19620	0.4500	Weak
4945	(3, 4)	18711	0.4507	Weak
4946	(8, 9)	19106	0.4508	Weak
4947	(7, 8)	19035	0.4518	Weak
4948	(6, 7)	19276	0.4522	Weak
4949	(4, 5)	18853	0.4525	Weak
4950	(5, 6)	19310	0.4526	Weak

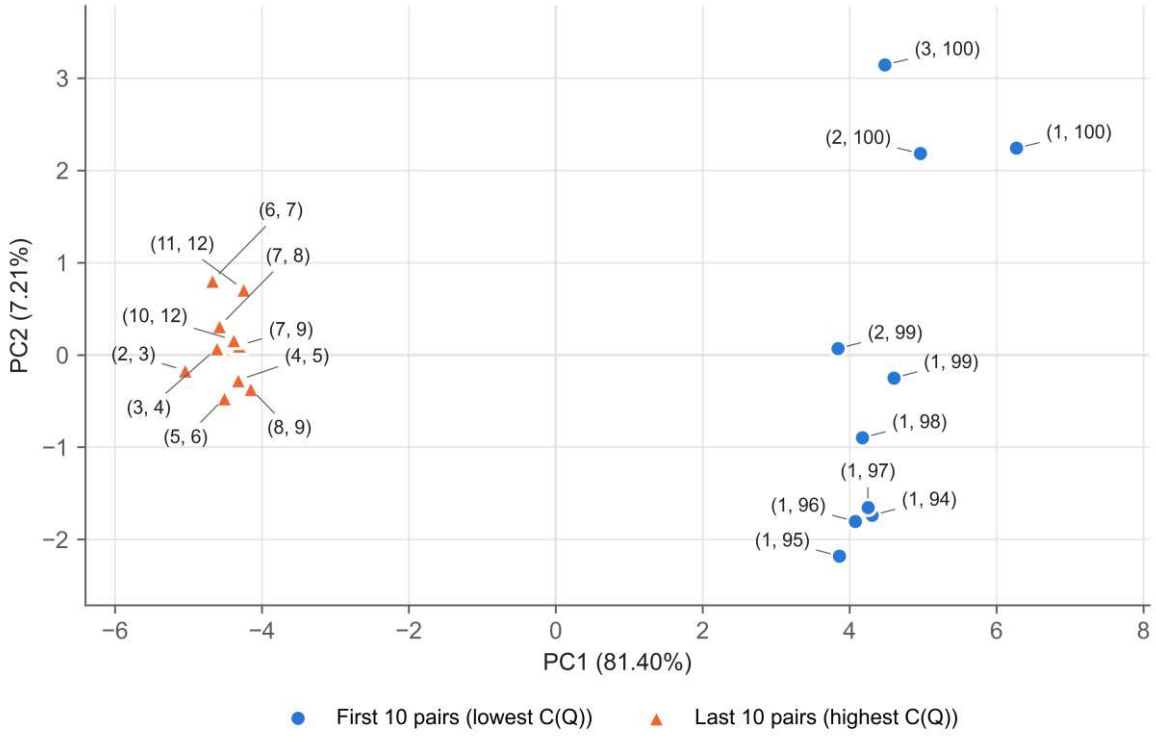


Fig. 3 PCA projection of the first 10 and last 10 mode-pair metaobjects built from 25 heterogeneous (8 quantitative and 17 nominal) features. PC1 explains 81.40% and PC2 explains 7.21% of the variance (88.61% in total). Circles denote the first 10 pairs and triangles denote the last 10 pairs

Tables 10 and 11 report feature stability and the number of available observations for modes 1 and 100 and modes 5 and 6, respectively. For modes 1 and 100, the most informative features are SubWgt, TotalNo, and NoMeas, whereas the proximity of

feature-stability values to 0.5 (Table 11) indicates weak discriminability of objects in modes 5 and 6. The numbers of objects also show how strongly the available evidence differs between features; for example, SubWgt has the highest stability for both pairs but is observed for fewer than one third of the objects.

Table 10 Feature stability and number of available observations for modes 1 and 100, ranked by stability

No.	Feature	Type	Stability	Number of objects
1	SubWgt	Quantitative	1.000000	5760
2	TotalNo	Quantitative	0.999896	19076
3	NoMeas	Quantitative	0.985528	17477
4	SpecCode	Nominal	0.961858	19272
5	Valid_Aphia	Nominal	0.961858	19272
6	LngtClass	Quantitative	0.911153	17668
7	HLNoAtLngt	Quantitative	0.889927	17668
8	StNo	Nominal	0.796907	19272
9	Sex	Nominal	0.727907	1841
10	SubFactor	Quantitative	0.711263	18890
11	LngtCode	Nominal	0.675113	17668
12	HaulNo	Quantitative	0.658526	19272
13	DateofCalculation	Nominal	0.637486	19272
14	Quarter	Nominal	0.627193	19272
15	Country	Nominal	0.624496	19272
16	Ship	Nominal	0.618410	19272
17	LenMeasType	Nominal	0.610490	6044
18	DoorType	Nominal	0.588335	19067
19	SpecVal	Nominal	0.582694	19272
20	Year	Nominal	0.577523	19272
21	SweepLngt	Quantitative	0.565567	19272
22	CatIdentifier	Nominal	0.558624	19272
23	GearEx	Nominal	0.537923	8431
24	Gear	Nominal	0.500525	19272
25	DevStage	Nominal	0.500064	15703

Table 11 Feature stability and number of available observations for modes 5 and 6, ranked by stability

No.	Feature	Type	Stability	Number of objects
1	SubWgt	Quantitative	0.989098	5801
2	StNo	Nominal	0.667141	19310
3	HaulNo	Quantitative	0.585931	19310
4	TotalNo	Quantitative	0.559928	19263
5	SpecCode	Nominal	0.554104	19310
6	Valid_Aphia	Nominal	0.553698	19310
7	NoMeas	Quantitative	0.545080	18300

No.	Feature	Type	Stability	Number of objects
8	Sex	Nominal	0.542594	2193
9	LngtClass	Quantitative	0.540392	18852
10	Year	Nominal	0.520726	19310
11	HLNoAtLngt	Quantitative	0.519726	18852
12	GearEx	Nominal	0.516409	9281
13	DateofCalculation	Nominal	0.515379	19310
14	Country	Nominal	0.514648	19310
15	LngtCode	Nominal	0.511039	18852
16	Ship	Nominal	0.510648	19310
17	SubFactor	Quantitative	0.508325	19206
18	SweepLngt	Quantitative	0.506591	19310
19	CatIdentifier	Nominal	0.506067	19310
20	LenMeasType	Nominal	0.505753	4944
21	Quarter	Nominal	0.504929	19310
22	DoorType	Nominal	0.503190	19102
23	SpecVal	Nominal	0.502664	19310
24	DevStage	Nominal	0.500136	15140
25	Gear	Nominal	0.500115	19310

6.6 Robustness to additional missing values

Because the approach is intended for incomplete data, we examined how the ordering of mode pairs changes when more values are missing. Starting from the working sample, a proportion $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ of the observed values of each of the 25 features was removed, and the $C(Q)$ values of all 4,950 mode pairs were recomputed by Eqs. (1)–(6) without imputation. Two mechanisms were used. Under MCAR masking, every observed value was removed with probability r. Under value-dependent (MNAR) masking, the removal probability of a quantitative value was proportional to its percentile rank, so that large values were removed more often, and every category of a nominal feature received a random weight drawn from U(0.25, 1.75); in both cases the mean removal probability was equal to r. Each setting was repeated 10 times with different random seeds. The recomputed scores were compared with the baseline scores by the Spearman and Kendall rank correlations, by the proportion of the baseline first 10 and last 10 pairs that remained in these positions, and by the mean absolute change in $C(Q)$.

Table 12 and Fig. 4 summarize the results. Under MCAR masking, the ordering was almost unchanged: even with 50% additional missing values, the mean Kendall τ was at least 0.973 in all feature spaces, and the mean absolute change in $C(Q)$ did not exceed 0.0025. Value-dependent masking had a stronger effect, especially for

quantitative features, for which large values were removed preferentially; at 50% additional MNAR missingness, Kendall τ decreased to 0.885 and the mean absolute change in $C(Q)$ reached 0.040. The nominal and heterogeneous feature spaces were less affected (τ = 0.948 and 0.934 at 50%), and the heterogeneous space retained the largest share of the extreme pairs (0.84 of the first and last 10 pairs at 50% MNAR masking). In all 300 runs, the most distinct pair (1, 100) remained first in every feature space, and the closest pair (5, 6) never left the last 27 positions.

The feature-level explanations were also stable. For the pair (1, 100), the five most informative features of the heterogeneous space were identical to the baseline in every run, and the Spearman correlation between the baseline and recomputed stability vectors was at least 0.998 under MCAR masking. For the weakly separated pair (5, 6), whose stability values are close to 0.5, the top-five features changed more often (on average 4 of 5 were retained at 50% masking under both mechanisms), which is consistent with the weak separation of this pair.

Table 12 Agreement between the baseline ordering of the 4,950 mode pairs and the ordering obtained after additional missing values were introduced (means over 10 repetitions; ± standard deviation)

Feature space	Mechanism	Added missing	Spearman ρ	Kendall τ	Top/bottom-10 retained	Mean |ΔC(Q)|	Lowest rank of (5, 6)
Quantitative	MCAR	10%	1.0000	0.995 ± 0.000	1.00	0.0004	4950
Quantitative	MCAR	30%	0.9999	0.990 ± 0.000	0.96	0.0011	4950
Quantitative	MCAR	50%	0.9996	0.983 ± 0.001	0.94	0.0025	4949
Quantitative	MNAR	10%	0.9999	0.992 ± 0.000	0.96	0.0036	4950
Quantitative	MNAR	30%	0.9985	0.968 ± 0.001	0.89	0.0147	4950
Quantitative	MNAR	50%	0.9793	0.885 ± 0.001	0.74	0.0399	4950
Nominal	MCAR	10%	0.9999	0.991 ± 0.000	0.95	0.0002	4943
Nominal	MCAR	30%	0.9996	0.982 ± 0.001	0.95	0.0005	4943
Nominal	MCAR	50%	0.9990	0.974 ± 0.002	0.83	0.0010	4931
Nominal	MNAR	10%	0.9999	0.990 ± 0.000	0.96	0.0003	4944
Nominal	MNAR	30%	0.9991	0.974 ± 0.002	0.90	0.0009	4940
Nominal	MNAR	50%	0.9961	0.948 ±	0.81	0.0022	4924

Feature space	Mechanism	Added missing	Spearman ρ	Kendall τ	Top/bottom-10 retained	Mean |ΔC(Q)|	Lowest rank of (5, 6)
				0.016			
Heterogeneous	MCAR	10%	1.0000	0.995 ± 0.000	0.98	0.0002	4949
Heterogeneous	MCAR	30%	0.9999	0.990 ± 0.000	0.98	0.0006	4947
Heterogeneous	MCAR	50%	0.9997	0.986 ± 0.001	0.92	0.0014	4947
Heterogeneous	MNAR	10%	0.9999	0.993 ± 0.001	0.97	0.0012	4948
Heterogeneous	MNAR	30%	0.9993	0.979 ± 0.002	0.90	0.0049	4948
Heterogeneous	MNAR	50%	0.9932	0.934 ± 0.009	0.83	0.0132	4947

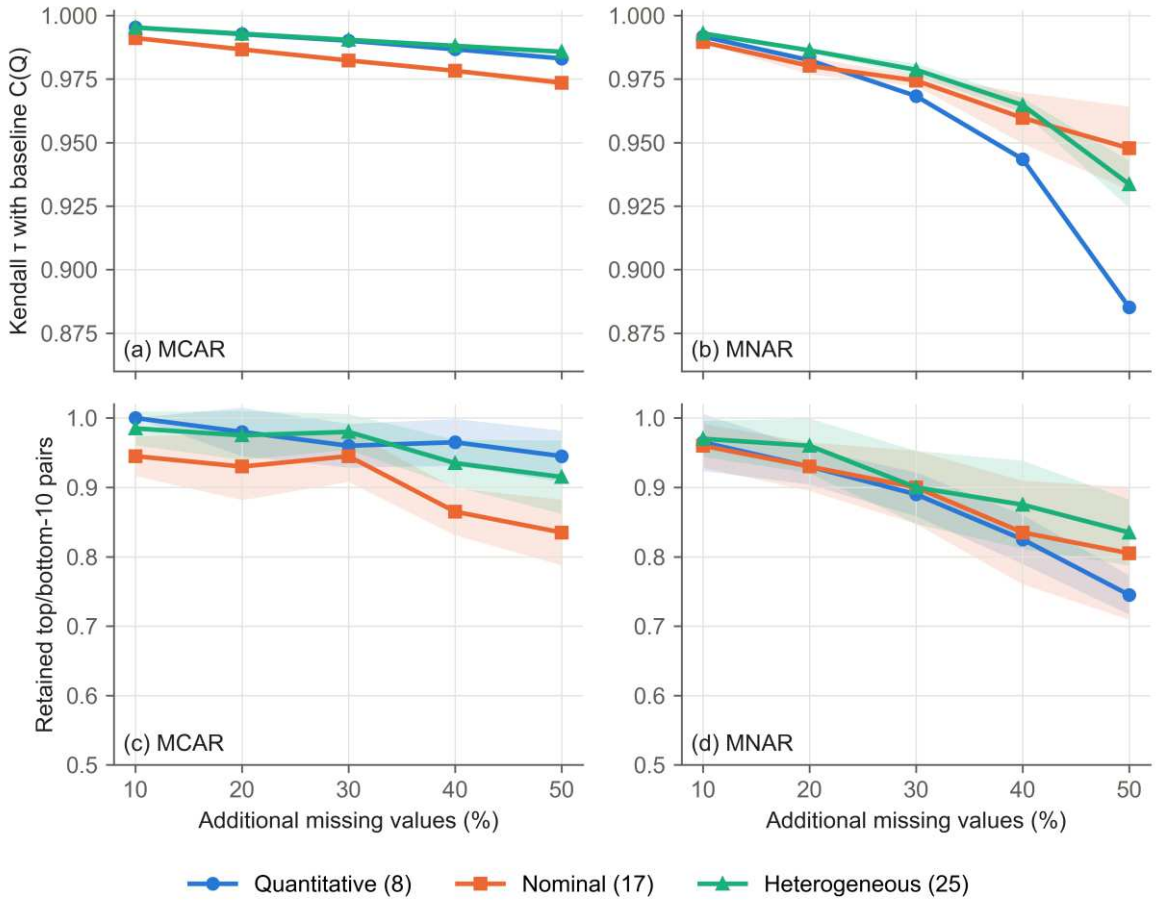


Fig. 4 Robustness of the mode-pair ordering to additional missing values. (a, b) Kendall rank correlation between the baseline C(Q) values of the 4,950 mode pairs and those recomputed after masking; (c, d) proportion of the baseline first 10 and last 10 pairs that remain in these positions. Panels (a, c) correspond to MCAR masking

and panels (b, d) to value-dependent (MNAR) masking. Lines show means over 10 repetitions and shaded bands show ± one standard deviation

6.7 Comparison with conventional missing-data methods

The masking experiments above show that the proposed ordering is stable, but they do not show how it compares with the usual practice of deleting incomplete objects or filling in the missing values before analysis. Because the true values of real missing entries are unknown, we followed the standard benchmark design for imputation methods: the true answer was computed on complete data, values were then removed artificially, and each method was judged by how well it recovered the true answer. We selected the 20 features with at most 6% missing values in the working sample (7 quantitative and 13 nominal; GearEx, Sex, SubWgt, DevStage, and LenMeasType were excluded) and the 897,825 objects that are complete on these features. To make the slower imputation methods feasible, 200 objects were drawn at random from each mode (20,000 objects). The $C(Q)$ values of the 4,950 mode pairs computed on these complete data were taken as the true values.

Then 10%, 30%, or 50% of the values of every feature were removed by the MCAR or MNAR mechanism described above, with three repetitions per setting. Eight conventional methods were compared with the proposed approach: complete-case analysis [4]; mean/mode and median/mode imputation; within-mode median/mode imputation, which uses the mode labels; random hot-deck imputation [34]; k-nearest-neighbour imputation with $k = 5$ [35]; MICE with Bayesian ridge regression [5]; and missForest with 30 trees and three iterations [7]. Nominal values produced by kNN and MICE were rounded to the nearest observed category. After imputation, $C(Q)$ was computed by the same Eqs. (1)–(6), so that the methods differ only in how they treat missing values. Each method was evaluated by the Kendall correlation with the true ordering, by the share of the true first 10 and last 10 pairs that it recovered, by the mean absolute error of $C(Q)$, and by its running time.

Table 13 and Fig. 5 summarize the results for the heterogeneous feature space. The proposed approach ranked first overall. It reproduced the true $C(Q)$ values most accurately: its mean absolute error (0.0071) was 2.1–5.2 times smaller than that of any imputation method, because filling in values shifts the distributions from which the stability indices are computed. In terms of ordering, only missForest performed

slightly better (mean τ = 0.929 versus 0.920; at 50% MNAR, 0.885 versus 0.861), but at the cost of a 2.4 times larger error in $C(Q)$ and a 28 times longer running time. kNN imputation was comparable at 10% missing values but fell behind at 30% and 50%, and it was the slowest method. Simple imputation by the mean, median, or a random observed value was as fast as the proposed approach but distorted the ordering, especially under MNAR missingness (τ = 0.56–0.67 at 50%). Within-mode imputation, which fills values with statistics of the same mode, distorted the differences between modes and gave the worst ordering among the imputation methods. Complete-case analysis failed: with 20 features, only about 12% of the objects remained complete at 10% missing values, and none remained at 50%.

Two qualifications should be noted. First, in the nominal feature space taken alone, MICE, missForest, and kNN reproduced $C(Q)$ slightly more accurately than the proposed approach (mean error 0.005–0.006 versus 0.007), and missForest also gave a better nominal ordering; the advantage of the proposed approach comes mainly from the quantitative features, for which its error was 2.4–6.2 times smaller. Second, the comparison was carried out on a complete subsample of 20 features; the five features with heavy real missingness could not be included, because their true values are unknown.

Table 13 Comparison of the proposed approach with eight conventional ways of handling missing values. Each method had to reproduce the C(Q) values of the 4,950 mode pairs computed on complete data after 10%, 30%, or 50% of the values were removed (means over three repetitions). Cell colour: green = better, red = worse within each column; the best value in each column is bold

Place	Method	How missing values are handled	Agreement with the true ordering, Kendall τ: random missing (MCAR)			Agreement with the true ordering, Kendall τ: value-dependent missing (MNAR)			Extreme pairs kept (share of first/last 10)	Mean error of C(Q)	Time per run (s)	Verdict
			10%	30%	50%	10%	30%	50%				
1	**Proposed metaobject approach (no imputation)**	Uses only the observed values; nothing is filled in	0.963	0.929	0.884	0.963	0.923	0.861	0.800	**0.0071**	0.77	Best overall: lowest C(Q) error (2–5 times lower than imputation) and near-best ordering at the lowest cost
2	missForest	Predicts each feature from the others with random forests	**0.971**	**0.936**	**0.887**	**0.969**	**0.924**	**0.885**	**0.819**	0.0173	21.7	Slightly better ordering (τ +0.008), but 2.4 times larger C(Q) error and 28 times slower
3	kNN imputation (k = 5)	Average of the 5 most similar objects	0.962	0.900	0.825	0.960	0.886	0.830	0.711	0.0150	40.7	Falls behind at ≥ 30% missing; 2.1 times larger error; slowest (53 times)
4	Random hot-deck imputation	Random observed value of the same feature	0.943	0.882	0.792	0.942	0.868	0.674	0.653	0.0182	0.76	Fast, but the ordering degrades strongly under MNAR
5	MICE	Chained regressions of each feature on the others	0.957	0.910	0.844	0.956	0.898	0.813	0.733	0.0289	3.9	Moderate ordering; 4.1 times larger C(Q) error
6	Mean/mode imputation	Feature mean, or most frequent category	0.950	0.899	0.829	0.932	0.798	0.593	0.669	0.0327	0.77	Shrinks distributions towards one value; poor under MNAR
7	Median/mode imputation	Feature median, or most frequent category	0.951	0.902	0.836	0.932	0.792	0.560	0.672	0.0350	0.75	Shrinks distributions towards one value; poor under MNAR
8	Within-mode median/mode imputation	Median or most frequent category within the same mode	0.900	0.732	0.530	0.905	0.754	0.529	0.375	0.0369	1.0	Distorts the differences between modes; worst ordering among imputation methods
9	Complete-case analysis	Deletes every object with any missing value	0.498	0.008	n/a[a]	0.557	−0.030	n/a[a]	0.100	0.1074	**0.22**	Fails: only 12% of objects remain at 10% and none at 50% missing

[a] Not computable: no object without missing values remained. Extreme pairs kept, mean error of C(Q), and time are averaged over all six settings; for complete-case analysis they are averaged over the computable settings only. Place is the mean rank over the six settings by Kendall τ and by the error of C(Q)

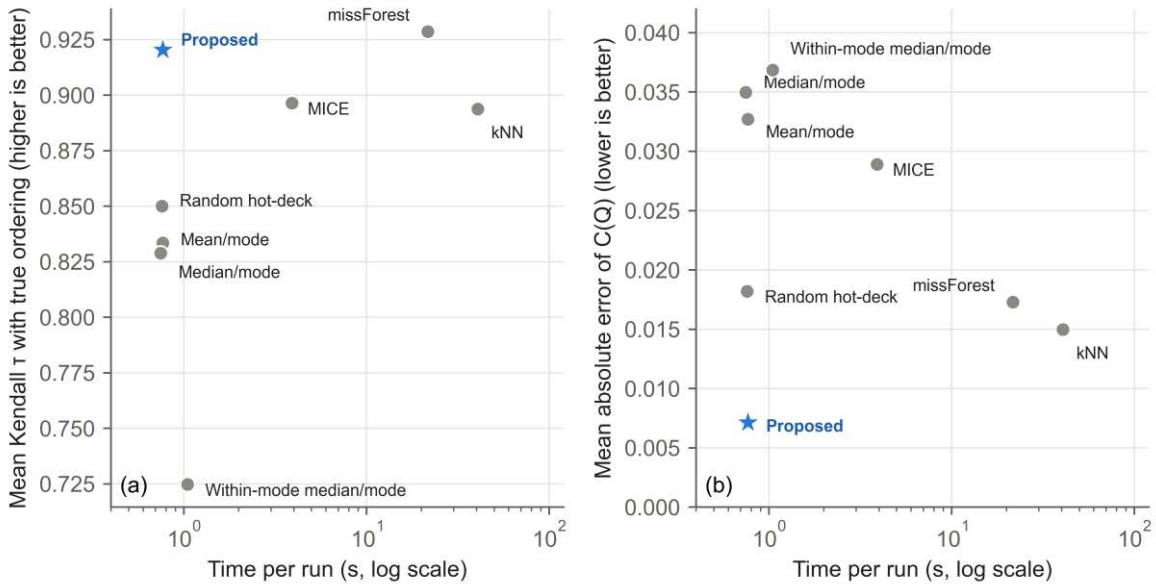


Fig. 5 Accuracy and cost of the compared methods. (a) Mean Kendall correlation between the true ordering of the 4,950 mode pairs and the ordering obtained after masking; (b) mean absolute error of C(Q). Values are averaged over both mechanisms and all masking levels. The star denotes the proposed approach. Complete-case analysis is not shown because it could not be computed at 50% missing values

7 Discussion and limitations

The experiment describes the observed portion of one survey dataset. Feature-wise exclusion avoids complete-case deletion across all variables but does not guarantee recovery of the unobserved population distribution, particularly under informative missingness [3, 4]. The masking experiments in Sect. 6.6 show that the ordering of mode pairs is robust to additional random missing values and degrades gradually under value-dependent missingness; however, they were performed on a single dataset, and the simulated MNAR mechanism is only one of many possible informative mechanisms. Future evaluation should also quantify sensitivity to the reference-class size and its random generation. Bootstrap resampling offers one possible framework for estimating sampling variability [36], with the resampling unit chosen to reflect dependence within the survey. No bootstrap intervals are claimed here.

The PCA figures are exploratory summaries. The selected distribution families and the 0.25 and 0.01 thresholds are descriptive choices. The comparison with

conventional methods (Sect. 6.7) used a complete subsample of 20,000 objects and 20 features and three repetitions per setting; larger benchmarks on other datasets, uncertainty estimation, and evaluation of LLM-generated explanations remain future work.

8 Conclusion

This study proposes a metaobject approach to analysing multimodal data with missing values by representing samples in a unified space of stability indices for heterogeneous features. Unlike methods requiring preliminary imputation, the approach uses available observations directly to construct metaobjects and to compare modes pairwise, while preserving quantitative assessment of each feature contribution to distribution discriminability. The NS-IBTS experiment illustrates how the resulting metaobjects rank mode pairs and identify features associated with substantial observed differences. Masking experiments with up to 50% additional missing values showed that this ranking is highly stable under random missingness and degrades gradually under value-dependent missingness. Compared with deletion and with seven imputation methods, the approach reproduced the true quality scores most accurately and was among the fastest, while only missForest gave a slightly better ordering at a much higher cost; robustness to reference-sample generation requires further assessment. A key result is the ability to supplement discriminability with the number of available observations, providing a basis for verifiable explanations and separating the degree of distributional difference from completeness of the original data description. The findings support metaobject representation as an intermediate layer between incomplete source data and explainable artificial intelligence systems, in which large language models can verbalize precomputed, verifiable characteristics without independently generating statistical conclusions.

Declarations

Acknowledgements The authors thank the International Council for the Exploration of the Sea (ICES) and the national institutes participating in the North Sea International Bottom Trawl Survey for collecting the survey data and making them publicly available through the DATRAS database.

Author contributions N.A.I.: conceptualization, methodology, formal analysis, supervision, writing – review and editing. S.R.: data curation, software, investigation, visualization, writing – original draft. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability The NS-IBTS HL data analysed in this study are publicly available from the ICES DATRAS database (https://datras.ices.dk) [32]. The derived datasets (mode labels, metaobjects, and pairwise quality scores) are available from the corresponding author on reasonable request.

Code availability The Python code used to form the modes, compute the metaobjects and quality scores, and produce the figures is available from the corresponding author on reasonable request.

Ethics approval and consent to participate Not applicable. The study uses publicly available fisheries survey data and involves no human participants or animal experiments.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Clinical trial number: not applicable.

References

1. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179(6):764–74. https://doi.org/10.1093/aje/kwt312.
2. Li J, Guo S, Ma R, He J, Zhang X, Rui D, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Med Res Methodol. 2024;24:41. https://doi.org/10.1186/s12874-024-02173-x.
3. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92. https://doi.org/10.1093/biomet/63.3.581.
4. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147–77. https://doi.org/10.1037/1082-989X.7.2.147.

5. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
6. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20:199. https://doi.org/10.1186/s12874-020-01080-1.
7. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8. https://doi.org/10.1093/bioinformatics/btr597.
8. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics. 2019;20:492. https://doi.org/10.1186/s12859-019-3110-0.
9. Ge Y, Li Z, Zhang J. A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods. Sci Rep. 2023;13:9432. https://doi.org/10.1038/s41598-023-36509-2.
10. Vapnik VN. Estimation of dependences based on empirical data. Moscow: Nauka; 1979. (In Russian).
11. Ignat'ev NA, Rakhimova MA, Lolaev MY. Specifics of selecting informative feature sets based on data with missings. Problems of Computational and Applied Mathematics. 2021;6/1(37):113–22.
12. Rossiev AA. Modeling data using curves to restore gaps in tables. In: Gorban AN, editor. Methods of neuroinformatics. Krasnoyarsk: KSTU; 1998. (In Russian).
13. Gorshenin AK, Martynov OP. Hybrid extreme gradient boosting models to impute the missing data in precipitation records. Informatics and Applications. 2019;13(3):34–40. https://doi.org/10.14357/19922264190306. (In Russian).
14. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
15. Ignatyev NA, Akbarov BKh. Estimation of the proximity of structures of relations of objects of the training sample on manifolds of sets of latent features. Tomsk State University Journal of Control and Computer Science. 2023;(65):69–78. (In Russian).
16. Brazdil P, Giraud-Carrier C, Soares C, Vilalta R. Metalearning: applications to data mining. Berlin: Springer; 2009. https://doi.org/10.1007/978-3-540-73263-1.
17. Lolaev M, Madrakhimov S, Makharov K, Saidov D. Choosing allowability boundaries for describing objects in subject areas. IAES Int J Artif Intell. 2024;13(1):329–36. https://doi.org/10.11591/ijai.v13.i1.pp329-336.
18. Ignatyev NA. Structure choice for relations between objects in metric classification algorithms. Pattern Recognit Image Anal. 2018;28(4):695–702. https://doi.org/10.1134/S1054661818040132.

19. Ignat'ev NA, Lolaev MY. An analysis of the compliance of the structures of relations of objects of classes on the varieties of their descriptions. Informatsionnye Tekhnologii. 2021;27(1):18–24. https://doi.org/10.17587/it.27.18-24. (In Russian).
20. Torshin IYu. On the generation of synthetic features based on support chains and arbitrary metrics within a topological approach to data analysis. Part 1. Inclusion of empirical distance functions into the formalism. Informatics and Applications. 2024;18(1):71–7. https://doi.org/10.14357/19922264240110. (In Russian).
21. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27(4):857–71. https://doi.org/10.2307/2528823.
22. Delfino M, Rapallo F. Not all missing data are equal: choosing the right imputation method for binary datasets. Qual Reliab Eng Int. 2026. https://doi.org/10.1002/qre.70342.
23. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
24. Madrakhimov S, Makharov K. Informativeness of feature sets in data with missing values. In: 12th World Conference "Intelligent System for Industrial Automation" (WCIS-2022). Lecture Notes in Networks and Systems, vol 718. Cham: Springer; 2024. p. 30–9. https://doi.org/10.1007/978-3-031-51521-7_6.
25. Minka TP. Estimating a Dirichlet distribution. Technical report; 2000 (revised 2012). https://tminka.github.io/papers/dirichlet/. Accessed 14 Oct 2026.
26. Bellman RE, Zadeh LA. Decision-making in a fuzzy environment. Manage Sci. 1970;17(4):B141–64. https://doi.org/10.1287/mnsc.17.4.B141.
27. Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2nd ed. 2022. https://christophm.github.io/interpretable-ml-book/. Accessed 14 Oct 2026.
28. Madrakhimov S, Makharov K, Khurramov A. On the transparency of decision-making in classification by precedents with fuzzy descriptions. IEEE Access. 2025;13:173656–64. https://doi.org/10.1109/ACCESS.2025.3616052.
29. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):248. https://doi.org/10.1145/3571730.
30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. p. 4765–74.
31. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 1135–44. https://doi.org/10.1145/2939672.2939778.

32. International Council for the Exploration of the Sea (ICES). DATRAS web services. https://datras.ices.dk/WebServices/Webservices.aspx. Accessed 14 Oct 2026.
33. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. https://doi.org/10.1098/rsta.2015.0202.
34. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. Int Stat Rev. 2010;78(1):40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x.
35. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5. https://doi.org/10.1093/bioinformatics/17.6.520.
36. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7(1):1–26. https://doi.org/10.1214/aos/1176344552.