Set-Based Domain Analysis for Missing Value Imputation in GIS Data: A Clustering-Driven Approach

Khyari Hamza[1] and Benferhat Salem[2]

[1] LSIA, USMBA University, Fez, 2202, Morocco
[2] CRIL, Univ. Artois & CNRS, UMR 8188, Lens, 62300, France
hamza.khyari@usmba.ac.ma, benferhat@cril.fr

Abstract. Missing values in Geographic Information Systems (GIS) datasets present significant challenges for spatial analysis and decision-making. This paper introduces a set-based domain analysis approach for handling missing values in spatial datasets. Our method combines a value-based representation compatible with its domain, and use clustering-based analysis to select optimal imputation strategies. By representing missing values as complete domain sets and systematically evaluating different value selection strategies through clustering quality metrics, we maintain data integrity while optimizing the overall dataset structure. Experimental results on real-world GIS datasets demonstrate that our approach outperforms traditional imputation methods in terms of clustering quality and domain constraint preservation. The proposed framework provides a robust solution for missing value imputation in spatial datasets while maintaining their inherent characteristics and relationships.

Keywords: Missing Value Imputation· GIS Data· Set-Based Domain· Clustering Analysis

1 Introduction

Geographic Information Systems (GIS) [8] datasets frequently suffer from missing or incomplete data due to various factors including sensor failures, data collection errors, and integration challenges [6, 15]. These missing values pose significant challenges for spatial analysis, as they can lead to biased results and unreliable conclusions. While numerous methods exist for handling missing values in general datasets [16], the spatial nature and domain constraints of GIS data require specialized approaches that preserve both the statistical properties and spatial relationships of the data.

Existing missing value imputation methods, such as mean substitution [11], k-nearest neighbors [10], Expectation-Maximization (EM) [4], or multiple imputation [2], often fail to account for the complex domain constraints and spatial dependencies inherent in GIS data.

For instance, certain attributes may only accept specific categorical values, while others must fall within valid numerical ranges determined by physical or administrative constraints. Moreover, these methods typically apply a single imputation strategy across all missing values, disregarding the possibility that different strategies might be optimal for different attributes or spatial contexts.

To address these limitations, we propose a set-based domain optimization framework for handling missing values in GIS datasets. Our approach consists of three key elements:

1. **Domain-Aware Representation:** Instead of immediately imputing missing values with single estimates, we first replace them with complete domain sets that represent all possible valid values. This preserves the domain constraints and maintains the solution space for subsequent optimization.
2. **Selection-based strategies:** Due to the computational intractability of evaluating all possible value combinations, we systematically generate different dataset versions using key statistical measures: minimum, maximum (mode for categorical data), and median, from both the complete domain and spatial neighbors. This strategic sampling approach enables efficient exploration of the solution space while capturing the essential characteristics of the data distribution, reducing the computational complexity from exponential to manageable levels.
3. **Clustering-Driven Analysis:** We evaluate each dataset version using clustering quality metrics, leveraging the assumption that optimal imputation should enhance, rather than disrupt, the natural groupings and patterns in the data.

This approach offers several advantages over existing methods. First, it ensures that imputed values always respect domain constraints, as they are selected from pre-validated domain sets. Second, it allows different imputation strategies to be applied to different attributes, acknowledging that optimal strategies may vary across the dataset. Third, by using clustering quality as an optimization criterion, it maintains the global structure and relationships within the data.

Our main contributions can be summarized as follows. We propose a set-based framework aimed at representing and managing missing values in spatial datasets in a structured manner. In addition, we introduce a methodical approach for generating and evaluating various imputation strategies. A clustering-based optimization process is also explored to guide the selection of suitable imputation combinations. Finally, we present an implementation architecture designed to support the processing of large-scale GIS datasets efficiently.

The remainder of this paper is organized as follows. Section 2 reviews related work in missing value imputation and spatial data analysis. Section 3 presents our methodology, including the mathematical formulation and algorithmic details. Section 4 describes the implementation architecture and computational considerations. Section 5 presents experimental results on real-world GIS datasets. Section 6 discusses the advantages and limitations of our approach, concludes the paper, and suggests directions for future research.

2 Related Work

Research on missing value imputation in spatial datasets spans multiple domains, from traditional statistical methods to modern machine learning approaches. We organize this review around three key areas: imputation methods, spatial-aware approaches, and domain-constrained imputation.

Missing Value Imputation: The foundation of missing value imputation was established by Rubin's seminal work [12] on missing data mechanisms, which introduced the fundamental concepts of Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). This theoretical framework has guided the development of numerous statistical approaches, comprehensively documented by Little and Rubin [7], including mean substitution, hot-deck imputation, and multiple imputation methods. Recent advances in machine learning have introduced more sophisticated imputation techniques. Notably, Stekhoven and Bühlmann [14] proposed MissForest, a non-parametric imputation method using random forest models that can handle mixed-type data and complex interactions between variables. Deep learning approaches have further expanded the field, with Yoon et al. [18] introducing GAIN (Generative Adversarial Imputation Networks), which demonstrates remarkable capability in handling complex missing data patterns by leveraging adversarial training. However, these methods, while sophisticated, often fail to account for the spatial relationships and domain constraints inherent in GIS data.

Spatial-Based Imputation Methods: The unique characteristics of spatial data, defined in GIS context as data with explicit geographic coordinates and topological relationships, have led to the development of specialized imputation methods that incorporate these spatial relationships. Cressie [3] provides a comprehensive framework for spatial statistics and interpolation methods, including kriging-based approaches that have become fundamental in spatial data analysis. These methods explicitly account for spatial autocorrelation and geographical dependencies in the imputation process.

Domain-Constrained and Hybrid Approaches: Recent research has increasingly focused on incorporating domain knowledge and spatial constraints into data analysis methods for GIS. Cao et al. [1] developed a comprehensive framework for spatiotemporal data analysis that effectively handles missing and incomplete data while considering spatial and temporal dependencies. Their work demonstrates how domain-specific constraints and spatial relationships can be integrated into data processing pipelines, particularly for large-scale geographic datasets with complex topological and proximity relationships.

Although significant progress has been made, several challenges remain in the current literature on missing value imputation in spatial datasets. Integrated approaches that simultaneously account for both spatial relationships and domain-specific constraints are still relatively uncommon, as many existing methods tend to emphasize one aspect over the other. Additionally, the widespread use of a single imputation strategy across all missing values may overlook opportunities for more tailored approaches better suited to specific attributes. Furthermore, some techniques prioritize local optimization, which can limit their ability to maintain the overall structure of the dataset. Lastly, balancing computational efficiency with imputation accuracy continues to be a concern, particularly when dealing with large-scale spatial data.

Our research addresses these gaps by introducing a framework that combines set-based domain representation with clustering-driven optimization. Unlike existing approaches that focus on single-value imputation strategies, our method maintains the complete domain of possible values throughout the optimization process. This allows for a more comprehensive exploration of the solution space. By incorporating clustering-based evaluation, we ensure that the imputed values maintain both local consistency and global dataset structure, while our set-based approach naturally preserves domain constraints.

Problem Statement

Let us formally define the problem of missing value imputation in spatial datasets. Given a dataset D with n objects and m features, where some values are missing, we denote the dataset as a matrix X where x_{ij} represents the value of feature j for object i. We define a mask matrix $M \in \{0,1\}^{n \times m}$ where $M_{ij} = 1$ if x_{ij} is observed and $M_{ij} = 0$ if x_{ij} is missing.

Each feature j has an associated domain D_j that defines the set of valid values for that feature. Additionally, each object i has geographical coordinates (lat_i, lon_i) that define its spatial location. For any object i, we define its spatial neighbors (NBR) $N(i,r)$ as the set of objects within radius r of object i:

$$N(i,r) = \{k \mid \text{distance}((lat_i, lon_i), (lat_k, lon_k)) \leq r, k \neq i\} \tag{1}$$

The objective is to find an imputation function f that produces a complete dataset $\hat{X}$ such that:

$$\hat{x}_{ij} = \begin{cases} x_{ij} & \text{if } M_{ij} = 1 \\ f(i, j, X, M, \{D_j\}_{j=1}^{m}, \{N(i,r)\}_{i=1}^{n}) & \text{if } M_{ij} = 0 \end{cases} \tag{2}$$

subject to the constraints:

$$\hat{x}_{ij} \in D_j \text{ for all } i, j \text{ where } M_{ij} = 0 \tag{3}$$

and optimized according to clustering quality:

$$\text{maximize } Q(\hat{X}) \tag{4}$$

where $Q(\hat{X})$ is a clustering quality function that measures how well the imputed dataset $\hat{X}$ can be clustered. The intuition behind this optimization objective is that good imputations should maintain or enhance the natural structure of the data, which is reflected in high-quality clustering results.

Table 1. Example dataset with missing values

ID	Latitude	Longitude	Diameter (mm)	Material	Install Date
1	43.6101	3.8677	100	Steel	2015-03-15
2	43.6120	3.8701	-	Steel	2016-07-22
3	43.6098	3.8690	150	-	2014-09-10
4	43.6110	3.8730	100	Plastic	-
5	43.6089	3.8710	150	Plastic	2017-11-05

To illustrate this problem with a concrete example, consider a small GIS dataset containing information about fire hydrants as shown in Table 1. The dataset contains five instances with three attributes: diameter (numerical), material (categorical), and installation date (temporal). Some values are missing (marked as '-').

Based on this dataset, we extract the following domains:

- Diameter domain: $D_{diameter} = [100, 150]$
- Material domain: $D_{material} = \{\text{Steel}, \text{Plastic}\}$
- Installation date domain: $D_{date} = [\text{2014-09-10}, \text{2017-11-05}]$

For spatial relationships, if we use a radius of 500 meters (chosen based on the typical urban infrastructure density in the study area), the spatial neighborhood of instance 2 would include instances 1, 3, 4, and 5, since all are within the specified radius.

Given these domains and spatial relationships, our goal is to generate multiple versions of the dataset by applying different selection strategies to the missing values, evaluate each version based on clustering quality, and select the optimal version (the one with the highest clustering score). The imputation is successful if:

1. All imputed values respect their domain constraints
2. The resulting dataset produces high-quality clusters that maintain the inherent structure of the data
3. Spatial relationships and dependencies are preserved

The outcome of our approach is a completely imputed dataset where missing values are replaced by optimal values selected through the domain-based, clustering-driven process. This optimized dataset can then be used for further spatial analysis, decision-making, and modeling purposes.

3 Set-Based Domain Optimization for Missing Value Imputation

3.1 Methodology

We propose a method for handling missing values in spatial datasets by combining set-based domain representation with clustering-driven optimization. It comprises four main phases: domain extraction, set-based value replacement, strategy-based value selection, and clustering-based evaluation. Our proposed framework, detailed in Algorithm 1, consists of four main steps:

1. Domain Extraction: Identifying valid value ranges for each data type
2. Strategy-based Value Selection: Generating dataset versions using different selection strategies
3. Clustering-based Evaluation: Evaluating and selecting the optimal version
4. Final Selection: Choosing the best dataset version for imputation

Algorithm 1 Set-Based Domain Optimization for Missing Value Imputation

```
Require:
  $D$: Original dataset with missing values
  $S$: Set of selection strategies
  $C_{num}$, $C_{cat}$, $C_{date}$: Numerical, categorical, and date columns
Ensure: Optimal dataset version with imputed values
  Phase 1: Domain Extraction
for each column $c$ in $D$ do
    if $c \in C_{num}$ then
        $D_c \leftarrow [\min(c), \max(c)]$                ▷ Numerical range
    else if $c \in C_{cat}$ then
        $D_c \leftarrow \{v_1, v_2, \ldots, v_k\}$          ▷ Unique values
    else if $c \in C_{date}$ then
        $D_c \leftarrow [t_{\min}, t_{\max}]$              ▷ Date range
    end if
end forPhase 2: Version Generation
$V \leftarrow \emptyset$                                   ▷ Set of dataset versions
for each strategy combination $s \in S^n$ do              ▷ $n$ is the number of features
    $V_s \leftarrow$ Apply strategies $s$ to $D$
    $V \leftarrow V \cup \{V_s\}$
end for
  Phase 3: Version Evaluation
$scores \leftarrow \emptyset$
for each version $v \in V$ do
    $score_v \leftarrow Q(v)$                              ▷ Clustering quality metrics
    $scores \leftarrow score_v \cup \{score_v\}$
end for
  Phase 4: Final Selection
$v_{best} \leftarrow \arg\max_{v \in V} scores$
return $v_{best}$
```

The workflow begins with an original dataset containing missing values. Each missing value undergoes domain extraction based on its data type (numerical ranges, categorical values, or date ranges). These domains are then used to create a set-based representation where each missing value is replaced with its complete domain set. Various selection strategies (S_1 to S_n) are applied to generate multiple dataset versions, with each version representing a unique combination of selection strategies across features. Finally, clustering analysis determines the optimal dataset version based on quality metrics. The following sections detail each of these steps.

The algorithm takes as input the original dataset D, a set of selection strategies S, and the types of data columns (C_{num}, C_{cat}, C_{date}). It outputs the optimal dataset version with imputed values. The four phases are executed sequentially, with each phase building on the results of the previous one to generate the final imputed dataset. It has a time complexity of $O(S^f \cdot n)$, where S is the number of selection strategies, f is the number of features with missing values, and n is the number of data points.

Domain Extraction and Set-Based Representation In the GIS context, we deal with three main types of data: numerical (e.g., measurements), categorical (e.g., material types), and temporal (e.g., dates). For each feature containing missing values, we first extract the domain of possible values. Let D_i represent the domain for feature i, where:

$$D_i = \begin{cases} [\min(x_i), \max(x_i)] & \text{if feature } i \text{ is numerical or temporal} \\ \{v_1, v_2, \ldots, v_k\} & \text{if feature } i \text{ is categorical} \end{cases} \tag{5}$$

Where x_i represents the non-null values in numerical or date features, $[\min(x_i), \max(x_i)]$ represents the valid range of values, and $\{v_1, v_2, \ldots, v_k\}$ represents the unique values in categorical features. This domain extraction process is crucial for ensuring that imputed values respect the inherent constraints of the dataset.

Strategy-Based Value Selection For value selection, we define six strategies $S = \{s_1, \ldots, s_6\}$:

- Feature-level Domain strategies:
  - s_1: Minimum value from feature-level domain
  - s_2: Majority value from feature-level domain
  - s_3: Median value from feature-level domain
- Spatial-based Domain strategies (based on spatial neighbors):
  - s_4: Minimum value from spatial domain
  - s_5: Majority value from spatial domain
  - s_6: Median value from spatial domain

For categorical data, where there is no natural ordering, the selection strategies are based on frequency:

- Minimum: The least frequent value in the domain

- Majority: The most frequent value in the domain
- Median: The value with median frequency in the domain

This frequency-based approach ensures that selection strategies can be applied consistently across all data types, including categorical attributes where traditional min/max concepts don't apply naturally.

Given the computational challenge of exploring all S^f possible strategy combinations (where f is the number of features with missing values), we implement a practical approximation approach. Rather than exhaustively evaluating every combination, we randomly sample $S^f/10$ representative strategy combinations from the solution space. This sampling ratio was chosen empirically to balance computational efficiency with solution quality coverage. Each selected combination generates a unique dataset version by applying different imputation strategies across features. Formally, for any missing value at position (i, j) in dataset version v, the imputed value is determined by:

$$v_{ij} = s_k(D_i) \text{ where } s_k \in S \text{ is the selected strategy} \tag{6}$$

This sampling approach maintains solution diversity while making the computational complexity manageable for large datasets with multiple features containing missing values.

Clustering-Based Evaluation At this stage, we have generated multiple dataset versions, each with a unique combination of imputation strategies. The question becomes: how do we determine which version provides the optimal imputation? We propose evaluating each version using clustering analysis and selecting the one that produces the most coherent and well-separated clusters.

We evaluate each dataset version using k-means clustering and three complementary quality metrics. For each version v, we compute a composite score:

$$Q(v) = \alpha \cdot \text{silhouette}(v) + \beta \cdot \frac{1}{1 + \text{davies\_bouldin}(v)} + \gamma \cdot \frac{1}{1 + \text{inertia}(v)} \tag{7}$$

where α, β, and γ are weighting factors ($\alpha + \beta + \gamma = 1$), and:

- silhouette(v) measures cluster cohesion and separation [5].
- davies_bouldin(v) evaluates cluster compactness and separation [17].
- inertia(v) represents within-cluster sum of squared distances [9].

The optimal dataset version v^* is selected as:

$$v^* = \arg\max_{v} Q(v) \tag{8}$$

In our experiments, we choose $\alpha = \beta = \gamma = 1/3$ to give equal weight to each metric, though these weights can be adjusted based on the specific characteristics of the dataset.

4 Experimental Study

The implementation architecture includes a list of elements designed to efficiently process missing values in spatial GIS datasets, such as:

1. A data handling component that loads the data, processes it, finds missing values, and organizes the spatial information to find neighboring points.
2. A domain handling component that determines valid values for each type of data:
   - For numbers: finds minimum and maximum values
   - For categories: collects all possible values
   - For dates: establishes valid date ranges
3. A strategies component that:
   - Generates different versions of the dataset
   - Applies different strategies to missing values
   - Uses clustering to evaluate each version
   - Picks the best version based on quality scores

Memory requirements grow with dataset size and feature count, while processing time increases significantly with the number of missing features. The method also depends on a sufficient amount of complete data to define reliable domains. These considerations underline the need to account for dataset characteristics during implementation.

5 Results

5.1 Dataset Description

Geographic Information Systems (GIS) are specialized systems designed to capture, store, manipulate, and analyze all types of geographical data. In GIS, spatial data is commonly stored in shapefiles, a popular geospatial vector data format that stores both the geometry and attribute information of geographic features.

The dataset used in this study is a real-world GIS dataset stored in shapefile format, specifically focusing on manhole data from a municipal infrastructure project. The shapefile contains various attributes related to manholes, including their physical characteristics, material types, installation dates, and precise geographical locations. The key features of the dataset are as follows:

- *DIAMETRE*: Numerical attribute representing the diameter of the manhole. This feature has approximately 75.5% missing values.
- *MATERIAU*: Categorical attribute indicating the material of the manhole.
- *TYPE*: Categorical attribute specifying the type of manhole.
- *TYPEREG*: Categorical attribute with 80.5% missing values, representing the type of regulation.
- *COMMUNE*: Categorical attribute indicating the commune where the manhole is located.

- *TYPE_ VOIE*: Categorical attribute specifying the type of road.
- *FOURNISSEU*: Categorical attribute representing the supplier.
- *ENTREPRISE*: Categorical attribute indicating the company responsible for the installation.
- *DATE_ POSE*: Date attribute representing the installation date of the manhole.

The dataset also includes several columns that are not used in the analysis due to 100% missing rates or irrelevance, such as *ID_ NODE*, *CODCOMM*, *NOTES*, *NUMERO*, *NUM_ VOIE*, *SOURCE*, and *NOM_ VOIE*, as shown in Figure 1.

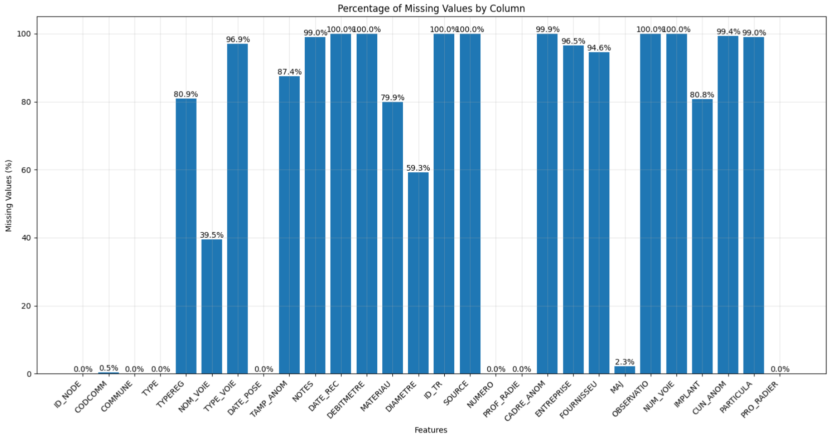

Fig. 1. Percentage of missing values in the manhole dataset attributes.

The primary challenge with this dataset is the high percentage of missing values in several key attributes, which necessitates robust imputation methods to ensure data integrity and reliability for subsequent spatial analysis and decision-making processes.

5.2 Experimental Setup

In this section, we describe the experimental setup used to evaluate our proposed methods. The experiments were conducted using a dataset of manhole inspections, which includes various features such as diameter, material, type, and installation date. The dataset was preprocessed to handle missing values and categorical variables, and different clustering strategies were applied to analyze the data.

Table 2. Clustering quality metrics for the 10 best imputed dataset versions.

Idx	Silhouette	Davies-Bouldin	Inertia	Total Score
1072	0.825118	0.586900	97072.5	0.485096
313	0.825104	0.587020	113406.5	0.485075
1129	0.825108	0.587038	97176.5	0.485074
2082	0.825108	0.587038	97176.5	0.485074
512	0.824978	0.587381	97186.5	0.484986
3007	0.825016	0.587552	96416	0.484976
322	0.824993	0.587699	96540.5	0.484949
985	0.824576	0.587059	98603	0.484894
215	0.824574	0.587053	98587	0.484894
86	0.824552	0.587195	98688.5	0.484868

Data Preprocessing The dataset was cleaned and preprocessed to ensure the quality of the data. Missing values were handled using domain-specific strategies, and categorical variables were encoded using one-hot encoding and target encoding. Date features were converted to numerical features by extracting components such as year, month, and day, and adding cyclical features for periodic components.

Clustering Method Determining the optimal number of clusters (k) is a crucial aspect of any clustering analysis. While methods like the elbow method or silhouette analysis can help identify the optimal k, we chose k=3 primarily for computational efficiency given the combinatorial nature of our strategy evaluation process. With larger k values, the computational complexity would increase significantly, as we need to evaluate clustering quality for each strategy combination. This choice of k=3 provides a reasonable balance between cluster granularity and computational feasibility for our experimental validation. The clustering analysis was performed using the k-means algorithm with k=3 clusters. The features were preprocessed appropriately to enable clustering of the mixed-type data (numerical, categorical, and temporal features).

5.3 Results

From the space of strategy combinations, we select $S^f/10$ random combinations and evaluate the resulting clusters using the silhouette score, the Davies-Bouldin index, and the inertia. The best combination is selected based on the highest score. The best combination is then used to impute the missing values in the dataset. The resulting dataset is then used for further analysis.

Table 2 shows the clustering quality metrics for the 10 best imputed dataset versions compared to the original dataset. The results demonstrate that our method effectively imputes missing values while preserving the clustering structure of the data. The best imputed dataset version achieved a silhouette score of 0.83 (on a scale from -1 to 1, where 1 indicates optimal clustering), a Davies-Bouldin index of 0.59 (where lower values indicate better clustering, with 0 being

Table 3. Selection strategies for the 10 best clustering metrics.

Idx	**Diam.**	**Mat.**	**Type**	**TypeR**	**Comm.**	**TVoie**	**Fourn.**	**Entr.**	**Date**
1072	Max	Max	NBR-Max	Max	NBR-Med	NBR-Min	Max	NBR-Max	NBR-Med
313	NBR-Max	NBR-Max	Max	NBR-Max	NBR-Max	NBR-Med	Max	NBR-Min	NBR-Med
1129	NBR-Max	NBR-Max	NBR-Med	Max	Max	Min	Max	Min	Max
2082	Max	NBR-Max	Max	NBR-Max	NBR-Max	NBR-Min	NBR-Max	Min	Max
512	Max	NBR-Max	Max	Max	NBR-Min	Min	NBR-Max	NBR-Med	NBR-Max
3007	NBR-Max	NBR-Max	NBR-Med	Max	Med	Max	Max	NBR-Max	NBR-Max
322	Max	Max	NBR-Max	Max	NBR-Med	Max	Max	Min	NBR-Med
985	NBR-Max	Max	NBR-Min	NBR-Max	Max	NBR-Min	NBR-Med	NBR-Max	Max
215	NBR-Max	NBR-Max	NBR-Med	Max	NBR-Max	Med	NBR-Min	Max	Max
86	Max	NBR-Max	Min	Max	Max	Med	NBR-Min	Min	NBR-Med

Legend: Max = Maximum from feature domain, Min = Minimum from feature domain, Med = Median from feature domain, NBR-Max/Min/Med = Maximum/Minimum/Median from neighbors domain

optimal), and a total normalized score of 0.485 (on a scale from 0 to 1, where higher values indicate better overall performance).

The respective selection strategies for the 10 best clustering metrics are shown in Table 3. The results indicate that different strategies are optimal, with respect to the score, for different features, highlighting the importance of flexible imputation methods that can adapt to the unique characteristics of each attribute. Analysis of these selection strategies reveals important patterns about optimal imputation approaches for different attribute types in spatial datasets. Notably, for numerical attributes like *diameter*, maximum-based strategies (either global or neighbor-based) consistently perform best, appearing in 10 out of 10 top combinations. For categorical attributes such as *material* type and commune, neighbor-based strategies are predominant, suggesting that spatial context provides more valuable information for categorical imputation. For the temporal attribute (installation date), there's a clear preference for max/median values from the spatial neighbors, appearing in 6 of the top 10 combinations.

Table 4. Consistent imputation strategy performance across all features.

Strategy	**Silhouette**	**Davies-Bouldin**	**Inertia**	**Total Score**
NBR-Max	0.825028	0.587545	112668.5	0.484980
Max	0.825028	0.587545	145378.5	0.484979
Med	0.814290	0.600429	111722.5	0.479710
NBR-Med	0.814290	0.600429	111722.5	0.479710
Min	0.750307	0.693055	153803.	0.446987
NBR-Min	0.750307	0.693055	153803.	0.446987

When comparing these mixed strategy combinations to applying a consistent strategy across all features (Table 4), we observe that the best mixed strategy approach achieves a higher total score (0.485) than the best consistent strategy (0.484 for both NBR-Max and Max). This performance gap demonstrates the advantage of our flexible approach that tailors imputation strategies to individual

Table 5. Pareto-optimal imputation strategy combinations

Idx	Feature Imputation Strategies	Silh.	D-B	Inertia
42	Diam: Min, Mat: NBR-Max, Type: NBR-Max, TypeR: NBR-Max, Comm: NBR-Min, TVoie: NBR-Max, Fourn: NBR-Max, Entr: NBR-Max, Date: NBR-Max	0.83	0.66	158381
1072	Diam: Max, Mat: Max, Type: NBR-Max, TypeR: Max, Comm: NBR-Med, TVoie: NBR-Min, Fourn: Max, Entr: NBR-Max, Date: NBR-Med	0.82	0.58	97072
2964	Diam: Max, Mat: NBR-Max, Type: Max, TypeR: NBR-Min, Comm: Min, TVoie: Max, Fourn: Max, Entr: NBR-Med, Date: Med	0.82	0.59	88377

Legend: Max = Maximum from feature domain, Min = Minimum from feature domain, Med = Median from feature domain, NBR-Max/Min/Med = Maximum/Minimum/Median from neighbors domain

features. Interestingly, minimum-based strategies (Min and NBR-Min) perform significantly worse when applied consistently across all features, achieving scores of only 0.446, further supporting the need for feature-specific imputation strategies.

These patterns demonstrate that different attribute types benefit from different imputation approaches, validating our multi-strategy framework. The consistency of neighbor-based strategies across multiple attributes highlights the importance of incorporating spatial relationships in the imputation process, particularly for GIS datasets where geographic proximity often correlates with attribute similarity.

Table 2 shows the clustering quality metrics for the 10 best imputed dataset versions compared to the original dataset. The results demonstrate that our method effectively imputes missing values while preserving the clustering structure of the data. The best imputed dataset version achieved a silhouette score of 0.82 (on a scale from -1 to 1, where 1 indicates optimal clustering), a Davies-Bouldin index of 0.58 (where lower values indicate better clustering, with 0 being optimal), and a total normalized score of 0.485 (on a scale from 0 to 1, where higher values indicate better overall performance).

5.4 Pareto Optimization Analysis

While the previous analysis focused on selecting dataset versions based on a combined score, and since the choice of α, β, and γ directly affects the final selection, we also performed a Pareto optimization analysis [13] to identify solutions that represent optimal trade-offs between competing objectives. In multi-objective optimization problems, a solution belongs to the Pareto front if no other solution can improve one objective without degrading at least one other objective.

The Pareto analysis reveals interesting trade-offs that a single-objective approach would miss. For example, solution 42 achieves the highest silhouette score (0.83) but has higher inertia, while solution 1072 has the lowest Davies-Bouldin index (0.58) and low inertia but a slightly lower silhouette score. Table 5 shows the imputation strategies for selected Pareto-optimal solutions.

This multi-objective perspective offers decision-makers flexibility in choosing imputation strategies based on their specific priorities. For applications where cluster cohesion is paramount, solutions on the Pareto front with higher silhouette scores would be preferred, while applications requiring well-separated clusters might prioritize solutions with lower Davies-Bouldin indices.

6 Discussion and Conclusion

In this paper, we proposed a set-based domain optimization framework to address missing values in GIS datasets. The approach integrates domain knowledge with clustering-based optimization to support imputation while maintaining spatial coherence and respecting domain constraints. The experimental results yield several observations. The set-based domain representation contributes to maintaining data integrity by ensuring imputed values remain within valid ranges. The clustering-based optimization shows potential in selecting suitable imputation strategies for different feature types. The method also handles mixed-type data, including numerical, categorical, and temporal variables, with reasonable effectiveness.

Despite encouraging results, several challenges remain. Computational complexity grows rapidly with the number of features, highlighting the need for more scalable optimization techniques. Additionally, dependence on clustering quality metrics may not fully capture all types of spatial relationships. Future work could explore more efficient optimization methods beyond basic aggregation operators, incorporate richer spatial relationship metrics, and evaluate performance using alternative criteria beyond traditional clustering scores.

Acknowledgements

This research has received support from the European Union's Horizon research and innovation programme under the MSCA-SE (Marie Sklodowska-Curie Actions Staff Exchange) grant agreement 101086252; Call: HORIZON-MSCA-2021-SE-01; Project title: STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management). This work has also been supported by the CHIST-ERA project ATLAS: GeoAI-based Augmentation of Multi-Source Urban GIS (grant number CHIST-ERA-23-MultiGIS-02) and also by the French National Research Agency (ANR) under grant number ANR-24-CHR4-0005. This research has also received support from the French national project ANR CROQUIS (Collecte, représentation, complétion, fusion et interrogation de données hétérogènes et incertaines de réseaux d'eaux urbains) [grant number ANR-21-CE23-0004].

References

1. Guofeng Cao, Shaowen Wang, Myunghwa Hwang, Anand Padmanabhan, Zhenhua Zhang, and Kiumars Soltani. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, 51:70–82, 2015.
2. James R Carpenter, Jonathan W Bartlett, Tim P Morris, Angela M Wood, Matteo Quartagno, and Michael G Kenward. *Multiple imputation and its application*. John Wiley & Sons, 2023.
3. Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.
4. Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
5. Andrzej Dudek. Silhouette index as clustering evaluation tool. In *Classification and Data Analysis: Theory and Applications 28*, pages 19–33. Springer, 2020.
6. Chung-Yi Li, Wei-Lun Su, Todd G McKenzie, Fu-Chun Hsu, Shou-De Lin, Jane Yung-jen Hsu, and Phillip B Gibbons. Recommending missing sensor values. In *2015 IEEE international conference on big data (big data)*, pages 381–390. IEEE, 2015.
7. Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
8. David J Maguire, Michael F Goodchild, and David W Rhind. Geographical information systems. 1991.
9. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011.
10. Christos Platias and Georgios Petasis. A comparison of machine learning methods for data imputation. In *11th Hellenic Conference on Artificial Intelligence*, pages 150–159, 2020.
11. Quinten AW Raaijmakers. Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, 59(5):725–748, 1999.
12. Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
13. Thomas A Runkler. Pareto optimality of cluster objective and validity functions. In *2007 IEEE International Fuzzy Systems Conference*, pages 1–6. IEEE, 2007.
14. Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
15. Xiang Su, Pingjiang Li, Jukka Riekki, Xiaoli Liu, Jussi Kiljander, Juha-Pekka Soininen, Christian Prehofer, Huber Flores, and Yuhong Li. Adaptive recovery of incomplete datasets for edge analytics. In *Pervasive Computing and Communications (PerCom), 2018 IEEE International Conference on*, pages 1–9, 2018.
16. Nzar A Ali Wafaa Mustafa Hameed. Comparison of seventeen missing value imputation techniques. *Journal of Hunan University Natural Sciences*, 49(7), 2022.
17. Junwei Xiao, Jianfeng Lu, and Xiangyu Li. Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, 21(6):1327–1338, 2017.
18. Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.