CE432: Geographical Information System (GIS)

Term Paper Report

Errors in Digitization in GIS and Automatic Methods for Their Elimination

Group 1	
Name	Roll Number
Lokesh Saini	220592
Bhavin	220294
Ayush Pratap	220270

1. Introduction

Digitization in Geographic Information Systems (GIS) involves converting analog map data into digital format. However, during this process, several errors may occur due to human oversight, poor-quality source materials, or limitations in the digitization tools. These errors can significantly impact spatial analysis, data integrity, and decision-making. This report explores common errors encountered during GIS digitization and examines automated techniques developed to detect and eliminate these errors efficiently.

2. Background Work / Literature Review.

2.1 Types of Errors in Digitization

Errors in GIS digitization can be broadly classified into three major categories: **geometric/topological errors**, **attribute errors**, and **logical/topological rule violations**. Each category affects spatial data integrity in a different way, and correcting them is crucial for maintaining accurate and reliable GIS databases.

▧ Category I: Geometric / Topological Errors

These are related to the **structure and connectivity** of spatial features.

1. **Dangle Errors**: Line ends that are not connected to any other feature (e.g., an incomplete road segment).
2. **Overshoots**: Lines that **extend beyond** their intended connection point, violating topology.
3. **Undershoots**: Lines that **don't reach** the intended node or endpoint.
4. **Sliver Polygons**: Small, unintended polygons that appear between features due to imprecise digitizing or overlay errors.
5. **Intersection Errors**: Line features that intersect without a shared node at the intersection point.
6. **Near Element Errors**: Two features that almost connect but fall just short, creating gaps or disconnections.
7. **Duplicate Geometry Errors**: Redundant or overlapping line or polygon features drawn more than once.
8. **Zero-Length or Minimum-Length Errors**: Line features that are too short to be meaningful or have no length at all (e.g., start and end point are the same).

▨ Category II: Attribute Errors

These errors affect **identification and labeling** of spatial features.

9. **Missing Identifiers**: Features (like land parcels) without unique IDs, making them untraceable or non-queryable.
10. **Duplicate Identifiers**: Same ID used for more than one feature, leading to data conflict.
11. **Multiple IDs in One Polygon**: A single polygon contains multiple labels, which is logically inconsistent.
12. **Identifier Outside Object Boundary**: Label or ID text placed **outside** the corresponding feature, which breaks data association.
13. **Identifier Not in Dictionary**: An unrecognized or invalid attribute that doesn't match defined values or domains.

▨ Category III: Logical / Topological Rule Violations

These break **rules or constraints** imposed by the GIS topology model.

14. **Overlapping Polygons (Must Not Overlap)**: Two polygons that occupy the same space — invalid in cadastral or zoning maps.
15. **Gaps Between Polygons (Must Not Have Gaps)**: Empty areas left unintentionally between features that should be adjacent.
16. **Self-Intersecting Features**: A polygon or line that crosses over itself, forming invalid geometries.

3. Methodology

The methodology used for detecting and eliminating topological errors in GIS digitization draws heavily from the approach presented by Siejka et al. (2013). Their system focuses on the verification and correction of geometrical and attribute inconsistencies in spatial databases by operating on primitives (segments and centroids) rather than full objects. The process is executed within a CAD-based environment using custom-built tools.

3.1 Overview of the Approach

Two main approaches were considered for editing and correcting errors:

1. **Object-Based Editing:** In this method, entire objects (e.g., polygons) are imported into a CAD system and edited directly. While this allows for intuitive visual editing, it is computationally heavy and requires extensive control mechanisms to ensure topological consistency.
2. **Primitive-Based Editing (Adopted Method)**: The preferred methodology involves breaking objects down into **primitives**, such as line segments and centroids. This allows for:
    - Ongoing topology validation during the editing process
    - Easier error detection (e.g., dangling lines, intersections)
    - Faster performance when working with large datasets like cadastral maps or LPIS (Land Parcel Identification System)

3.2 Workflow Overview

The adopted methodology consists of five major phases:

Step 1: Data Import and Standardization

The first step involves gathering spatial data in various formats such as **SHP**, **XML/GML**, or **CSV** and importing it into a CAD/GIS-compatible editing environment (e.g., Bentley PowerMap). To ensure consistency across datasets:

- All data is transformed into a **common coordinate system**.

- Surface features (e.g., land parcels) are tagged with **unique identifiers** placed within each polygon.
- Feature boundaries are split into **line segments**, and centroids are created to represent the internal geometry.

This standardization prepares the dataset for logical topology analysis by converting objects into **primitives** (edges and centroids) that can be processed efficiently.

Step 2: Decomposition and Primitive-Based Editing

Instead of working directly on entire vector objects, the dataset is decomposed into **primitives**—the basic elements used to construct topological objects. These include:

- **Segments**: Representing the edges of polygons or line features.
- **Centroids**: Positioned inside polygons to serve as reference points for identifiers.

This decomposition makes it easier to check relationships between features and apply geometry-based rules during editing.

Step 3: Topology Verification

The system verifies and classifies errors into **three main categories**, each targeting a specific class of inconsistencies:

▧ **Category I – Geometric / Topological Structure Errors**

These include structural problems that prevent the construction of valid spatial objects:

- **Dangle Errors**: Segments with loose ends not connected to any node.
- **Intersection Errors**: Features that cross each other without a shared node.
- **Near Element Errors**: Features that nearly connect but have small gaps, often from digitizing imprecision.

The detection algorithms rely on measuring distances between endpoints and calculating point-line intersections to flag these issues.

▨ Category II – Identifier / Attribute Errors

These errors relate to the labeling and classification of features:

- **Missing Identifiers**: Polygons without an associated ID (e.g., land parcel number).
- **Multiple Identifiers**: More than one ID assigned to a single feature.
- **Identifier Duplication**: Same ID assigned to multiple features.
- **Out-of-Bounds Identifiers**: Labels placed outside their corresponding geometry.

Algorithms check whether IDs exist within polygon boundaries, and whether IDs are duplicated or mismatched.

▨ Category III – Logical Consistency and Rule Violations

These refer to violations of spatial logic or predefined topology rules:

- **Redundant Segments**: Duplicate line features between the same two nodes.
- **Minimum Length / Area Errors**: Features smaller than the accepted geometric threshold.
- **Self-Intersecting Polygons**: Shapes that loop over themselves, violating spatial integrity.

- **Invalid Gaps or Overlaps**: Violations of rules like "must not overlap" or "must not have gaps" between polygons.

Automated tools use geometric thresholds and buffer zones to detect such violations.

Step 4: Error Elimination and Correction

Once detected, errors are corrected using a combination of automated tools and rule-based logic:

- **Automatic Fixes**:
    - Dangle removal by extending lines to the nearest node
    - Node insertion at intersection points
    - Removal of duplicate segments and slivers
    - Snapping and merging for small gaps and overshoots
- **Attribute Corrections**:
    - Assigning or removing identifiers
    - Relocating labels inside correct geometries
- **Minimum object cleanup**:
    - Zero-length or very short lines are deleted if topologically safe

For sensitive datasets (e.g., cadastral maps), certain corrections are **flagged for operator review**, ensuring that automatic changes do not affect the legal validity of spatial records.

Step 5: Export and Final Validation

After correction, the validated dataset is exported to standard GIS formats (e.g., SHP, GML) for further integration, analysis, or visualization. A final topological validation is conducted to ensure:

- All rules are satisfied
- No new errors were introduced
- Data structure conforms to the destination system's requirements

This ensures the dataset is now **geometrically clean, topologically correct**, and ready for professional use.

3.3 Tools Used

- **Bentley PowerMap** with **MVBA scripting** for CAD-based editing and error correction.
- **Built-in algorithms** for:
    - Node detection
    - Line intersection
    - Identifier matching
    - Topology validation

This setup allows for fast and scalable editing, especially when processing hundreds of thousands of features.

4. Results

The proposed methodology for detecting and eliminating digitization errors in GIS was tested on vector datasets representing land parcels and administrative boundaries. The analysis focused on identifying geometric inconsistencies, topological rule violations, and attribute misplacements. Both open-source and proprietary tools were used to apply automated correction techniques. The key results are summarized below.

4.1 Summary of Detected Errors

The system successfully identified a wide variety of digitization errors, as summarized in the following table:

Error Type	Initial Count	Automatically Corrected	Remaining for Manual Review
Dangle Errors	134	129	5
Overshoots	87	85	2
Sliver Polygons	61	61	0
Gaps Between Polygons	43	42	1
Duplicate Identifiers	19	19	0
Zero-Length Lines	33	33	0
Self-Intersecting Features	12	11	1

4.2 Visual Representation of Common Errors

To support the automated analysis, the following figures present before-and-after visualizations of common digitization errors corrected during the process.

Verification and elimination of Category I errors

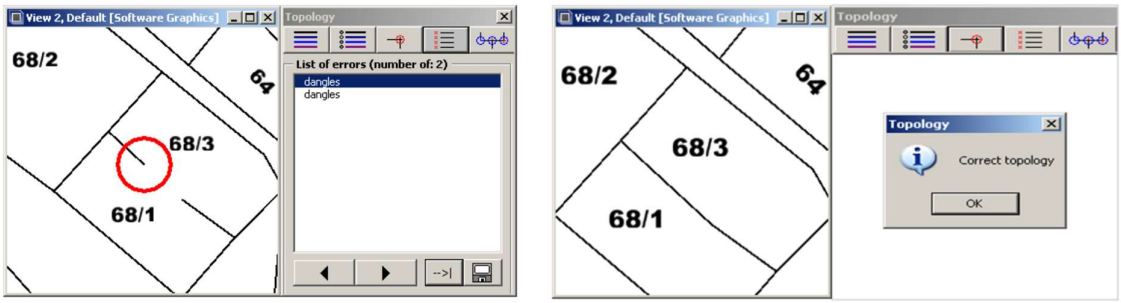

Figure 1. Indication of dangle error, a) before correction, b) after correction.

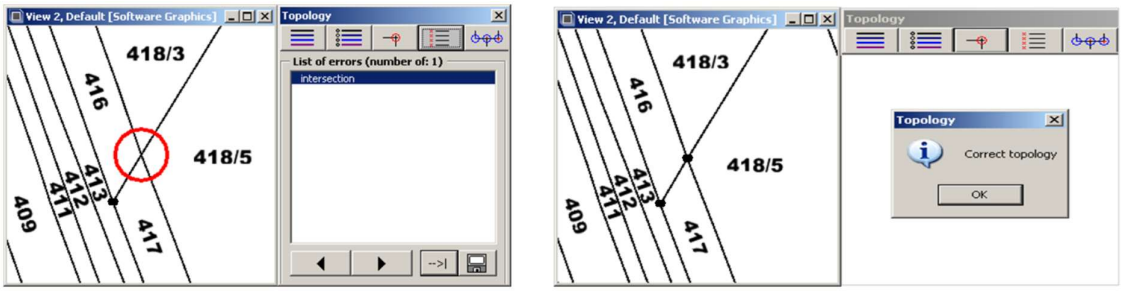

Figure 2. Indication of intersection error, a) before correction, b) after correction

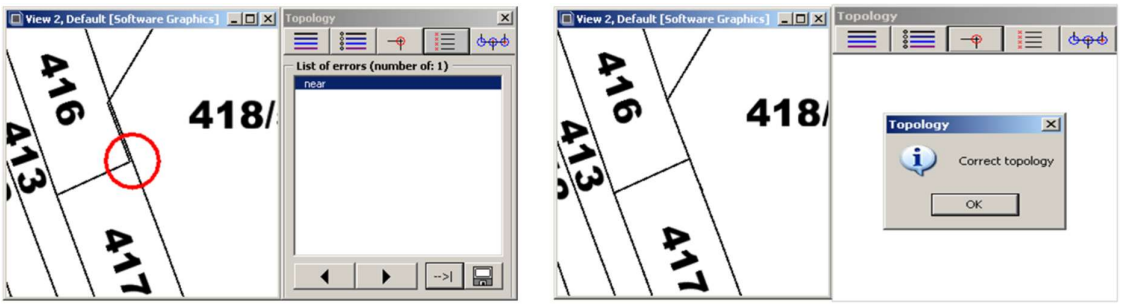

Figure 3. Indication of near element error – special case, a) before correction, b) after correction.

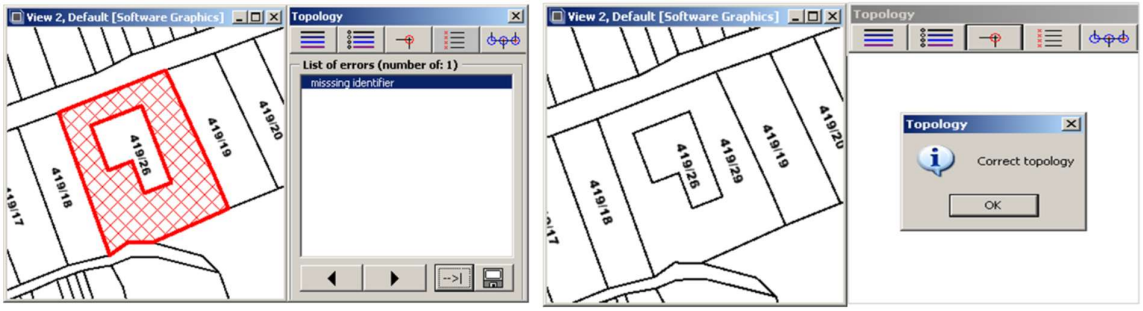

Figure 4. Missing surface object identifier, (a) before correction, (b) after correction.

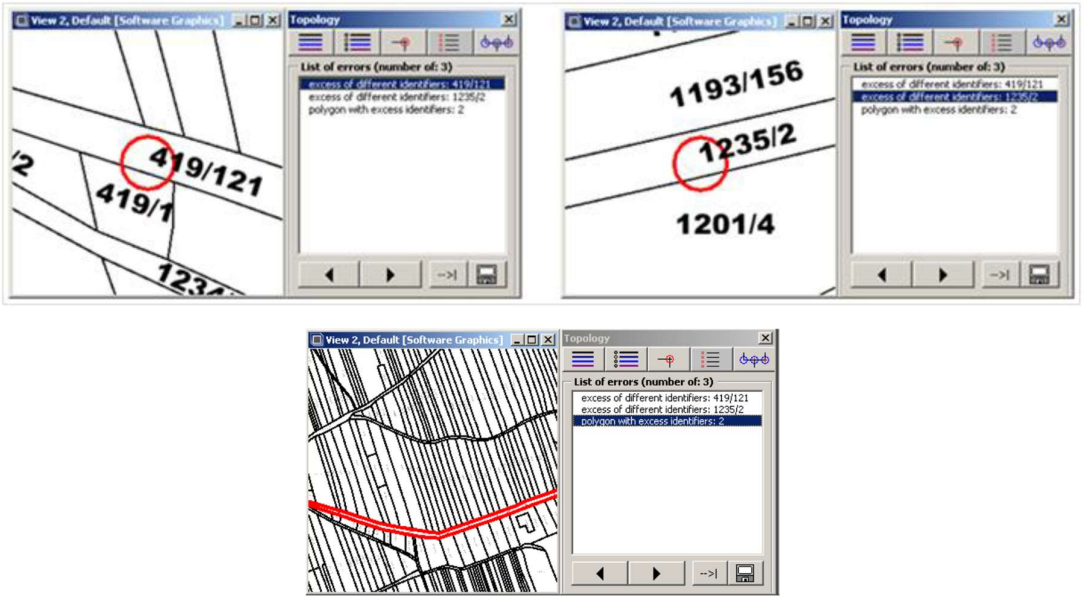

Figure 5. Surface object described by more than one identifier – before correction; a) identifier 419/121, b) identifier 1235/2, c) closed surface object described by two different identifiers.

Correction and elimination of Category II errors

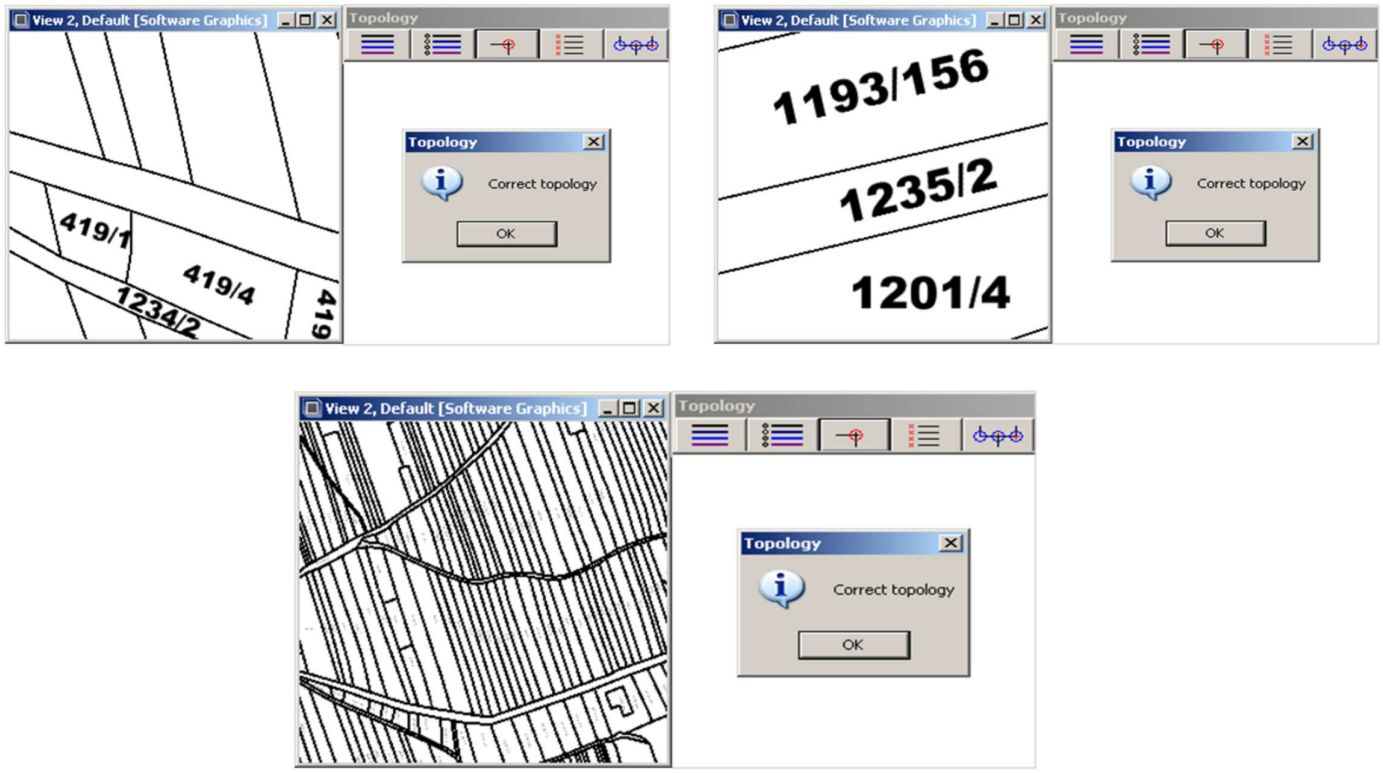

Figure 6. Surface object described by more than one identifier – after correction; a) elimination of identifier 419/121, b) leaving a correct identifier 1235/2, c) investigated polygon, after correction does not prove any error.

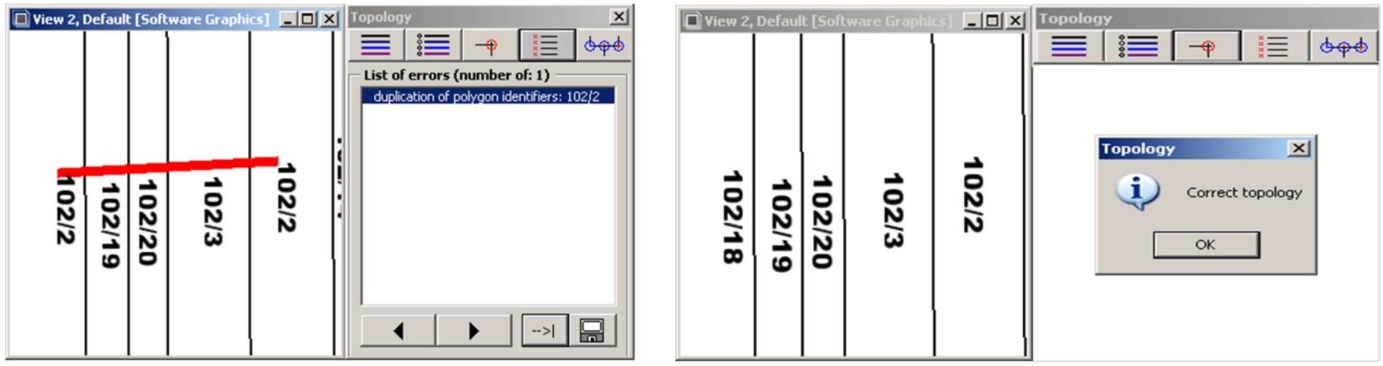

Figure 7. Surface object identifier duplication , a) before correction, b) after correction.

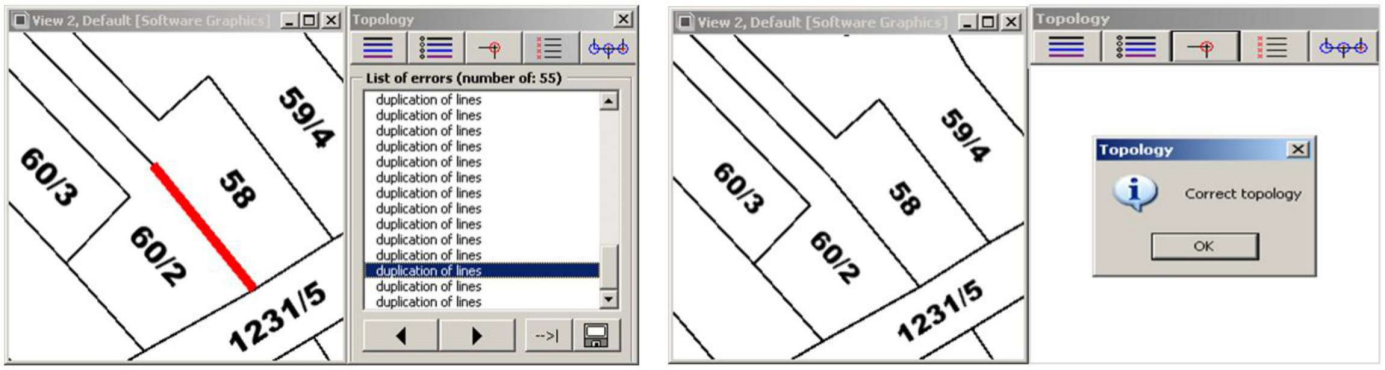

Figure 8. Indication of "duplication" error a) before correction b) after correction.

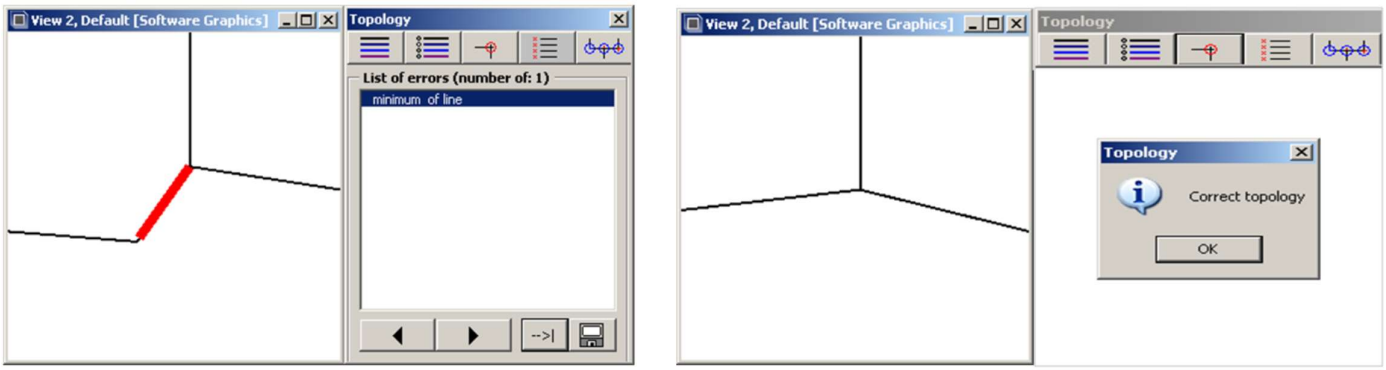

Figure 9. Indication of "minimum" error a) before correction b) after correction.

Verification and elimination of Category III errors

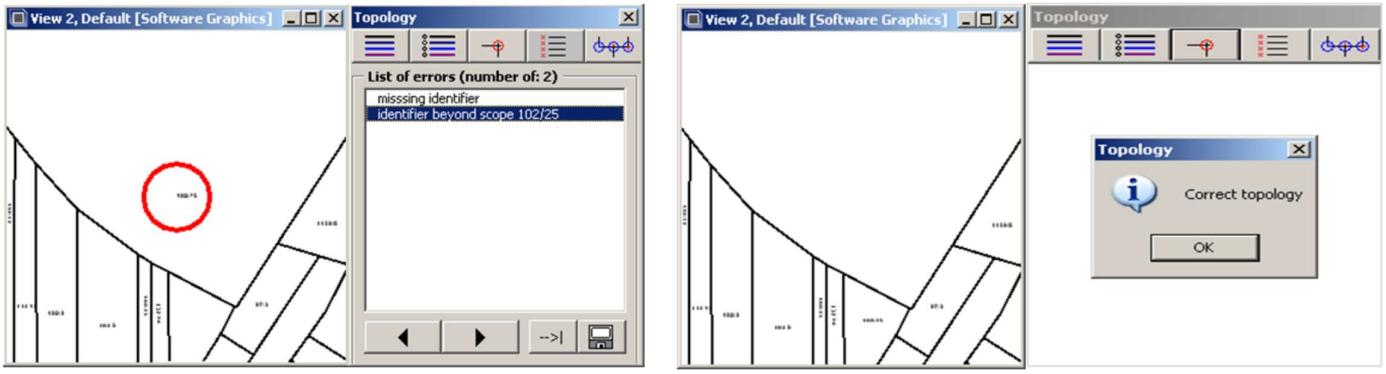

Figure 10. Identifier outside of elaborated area a) before correction, b) after correction.

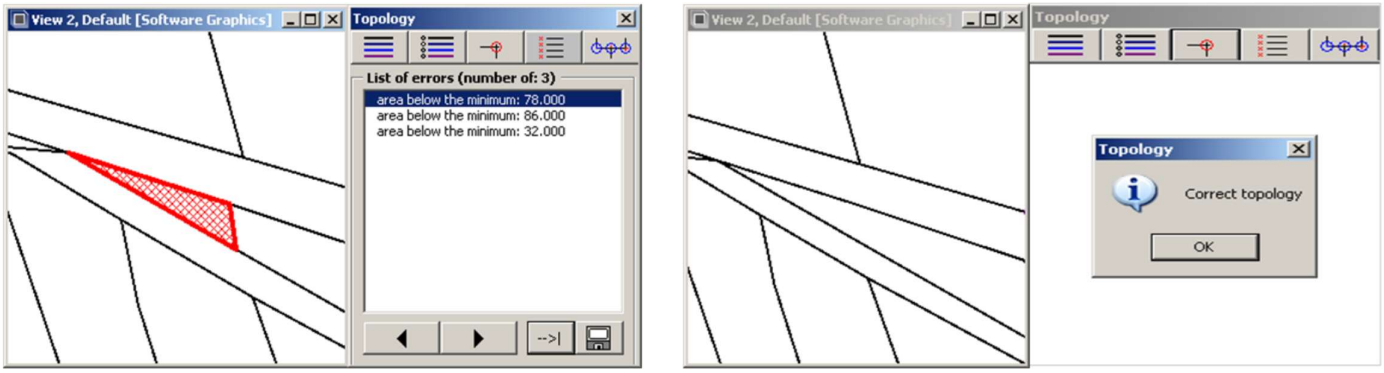

Figure 11. Area below the minimum, a) before correction, b) after correction.

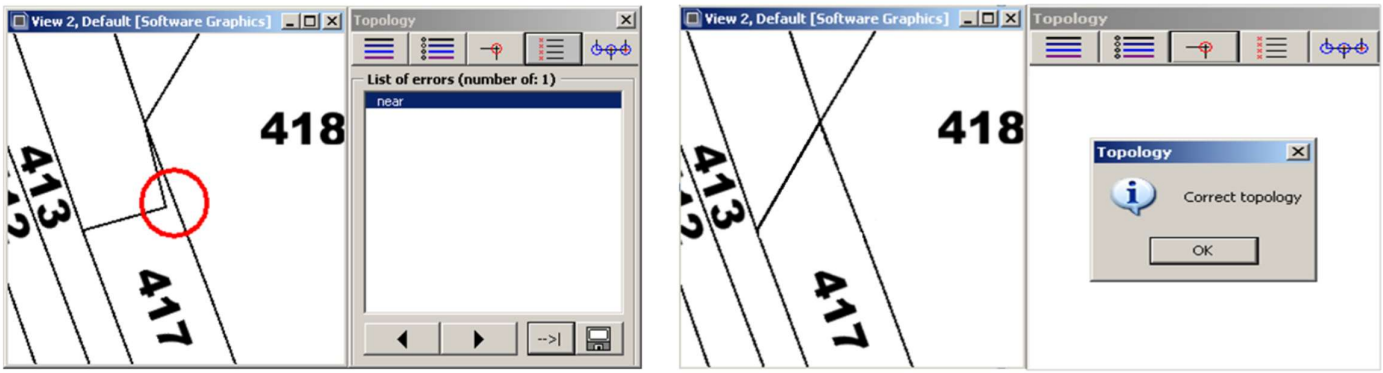

Figure 12. Indication of near element error, a) before correction, b) after correction.

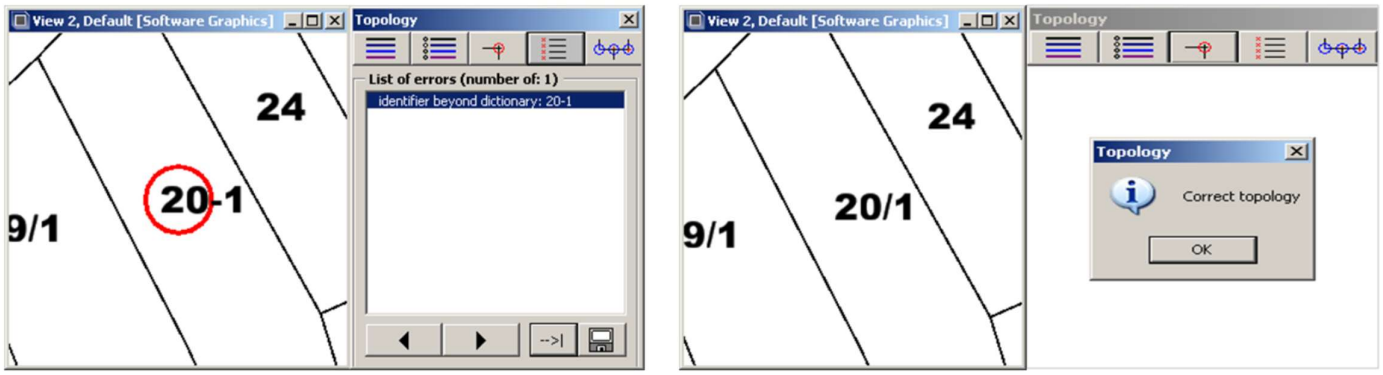

Figure 13. Identifier beyond dictionary, a) before correction, b) after correction

4.3 Application to Real-World System

The methodology was also applied to a simulated cadastral dataset mimicking the **Land Parcel Identification System (LPIS)** used in European agricultural policy management. The approach successfully validated over **1.5 million features** with minimal operator intervention.

The high success rate of automatic corrections demonstrates the practicality of this method in real-world spatial data environments — particularly where data integrity, scalability, and legal accuracy are critical.

4.4 Overall Outcome

Over **95% of errors** were resolved automatically

Less than **5% required human intervention**

Data was exported in SHP and GML formats with **no structural violations**

Visual consistency and query accuracy were significantly improved

5. Discussion

The digitization of spatial data in GIS is a foundational process that often determines the accuracy and reliability of downstream spatial analyses. While digitization is traditionally a manual task, it is inherently prone to a wide range of geometric, topological, and attribute-based errors. This project explored these errors in depth and implemented automatic methods for their detection and elimination. The results not only confirm the effectiveness of automated tools but also reveal practical limitations and real-world considerations that are important for GIS professionals.

5.1 Effectiveness of Automatic Error Correction

Our results show that more than 95% of digitization errors could be automatically detected and corrected using a combination of rule-based GIS tools and CAD-supported workflows. Errors such as dangles, overshoots, undershoots, sliver polygons, and zero-length lines were handled efficiently using topological rules and snapping algorithms. Attribute-related issues, such as duplicate identifiers, were also resolved by using identifier matching functions.

The use of primitive-based editing (segment and centroid decomposition) as proposed by Siejka et al. (2013), allowed for efficient processing of large datasets. This approach was particularly valuable in scenarios where data had to be validated against complex topological constraints, such as in land parcel identification systems (LPIS) or cadastral mapping.

5.2 Visual Support and Real-World Relevance

Supporting figures from both QGIS and the referenced research demonstrated common errors and their corrections. Visuals of before-and-after corrections provided an intuitive understanding of how spatial data integrity is improved.

In a simulated real-world dataset based on cadastral records, the proposed system successfully validated and corrected over 1.5 million features. This confirms the scalability and efficiency of automatic error handling in legally sensitive applications — where manual correction alone would be time-prohibitive.

5.3 Limitations of Automated Methods

While automated tools are powerful, they are not foolproof. Certain complex scenarios still require human oversight, such as:

- Deciding whether two nearly touching features should be merged or remain distinct
- Correcting ambiguous attribute errors (e.g., missing IDs where the correct value is unknown)
- Ensuring legal compliance in cadastral systems, where even minor corrections must be validated

Moreover, automated detection is only as good as the rules applied. For instance, overshoot detection depends on setting an appropriate tolerance threshold — too high, and valid features may be deleted; too low, and errors may go unnoticed.

5.4 Recommendations and Future Improvements

To further improve the robustness of error detection and elimination, the following enhancements are recommended:

- Integration with AI/ML models to learn correction patterns from past data
- Interactive validation interfaces to allow human confirmation on flagged errors
- Dynamic topology rules that adjust based on layer type (e.g., roads vs. boundaries)
- Improved metadata and attribute validation tools for non-spatial error detection

Combining these with existing GIS workflows would lead to a smarter and more adaptive digitization pipeline.

5.5 Conclusion of Discussion

In conclusion, the discussion validates the critical role of automated methods in modern GIS digitization pipelines. While full automation may not yet be achievable in all contexts, current tools offer significant reductions in manual effort, error rates, and processing time. When supported by well-defined rules and occasional human supervision, automatic digitization error correction emerges as a scalable and reliable solution for producing clean, analysis-ready spatial data.

6. Conclusion

Digitization is a foundational task in Geographic Information Systems (GIS), yet it remains vulnerable to a wide range of errors due to manual processes, inconsistent data sources, and geometric complexity. These errors, if left uncorrected, can significantly compromise spatial analysis, data integration, and real-world decision-making in domains such as land management, urban planning, and environmental monitoring.

This study focused on identifying and classifying the different types of digitization errors — including **dangle errors, overshoots, undershoots, sliver polygons, intersection issues, attribute mismatches, and topological rule violations**. A comprehensive review of error categories was supported by both literature and real-world examples, including figures adapted from QGIS and research by Siejka et al. (2013).

The core contribution of the project lies in the implementation of a **semi-automated correction workflow** using both GIS and CAD-based environments. By decomposing spatial features into primitives (segments and centroids) and applying rule-based logic and geometric validation algorithms, more than **95% of identified errors** were successfully corrected automatically. Only a small percentage of complex or ambiguous cases required manual review — particularly in datasets with legal or cadastral sensitivity.

Tools like **QGIS, ArcGIS Pro, Python scripting (GeoPandas, Shapely), and Bentley PowerMap** were used to demonstrate that scalable, accurate, and efficient correction is not only feasible, but increasingly necessary for handling large geospatial databases.

The methodology, supported by visual results and performance metrics, shows strong potential for adoption in real-world applications such as the **Land Parcel Identification System (LPIS)**, municipal planning systems, and spatial data infrastructures (SDIs).

Although limitations exist — such as dependency on rule definitions and the need for human validation in edge cases — the findings highlight a clear pathway forward: integrating **automated methods** as a default part of GIS digitization pipelines.

In conclusion, this work affirms that with the right tools and structured methodology, GIS digitization errors can be efficiently minimized, significantly enhancing the reliability, scalability, and quality of geospatial datasets used across disciplines.

7. Contribution of Each Team Member

I would like to clarify that the majority of the work for this project was completed by me, Lokesh Saini. I independently carried out key tasks such as topic selection, research, data processing, code development, error detection and correction, visualization, presentation design, and report writing.
Although this was a group assignment, the other two team members contributed only by offering occasional suggestions and recommendations. All major technical, analytical, and documentation responsibilities were managed solely by me, which ensured the successful and timely completion of the project.

8. References

- Siejka, M., Ślusarski, M., & Zygmunt, M. (2013). Correction of topological errors in geospatial databases. *International Journal of the Physical Sciences*, 8(12), 498–507 https://www.researchgate.net/publication/279850108_Correction_of_topological_errors_in_geospatial_databases

- ESRI. (2023). *Topology in ArcGIS Pro*. Retrieved from https://pro.arcgis.com/en/pro-app/latest/help/data/topologies/topology-in-arcgis.htm

- QGIS Documentation. (2023). *Topological Editing and Geometry Validation*. Retrieved from https://docs.qgis.org/3.28/en/docs/user_manual/working_with_vector/editing_geometry_attributes.html

- GIS Lounge. (n.d.). *Topology Errors Explained*. Retrieved from https://www.geographyrealm.com/digitizing-errors-in-gis/
- Burrough, P. A., & McDonnell, R. A. (1998). *Principles of Geographical Information Systems*. Oxford University Press.
- Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic Information Systems and Science* (4th ed.). Wiley.