An Automated Data Validation Approach for Power Distribution Networks using Grid Partitioning and Multi-faceted Quality Scoring

Oğuz Deniz[1], Süha Nur Arslan[1]

[1]SI GSW GIS, Siemens A.Ş., Ankara/Türkiye – oguzdeniz179@gmail.com, suha.arslan@siemens.com

Keywords: GIS Data Validation, Power Networks, Spatial Data Quality, AI-Assisted Data Integrity, Topological Error Detection

Abstract

Accurate Geographic Information System (GIS) data is fundamental to the reliable management, planning, and operation of modern power distribution networks. Conventional validation methods, however, often rely on network-wide rule-based checks or manual inspections, which are inefficient at identifying and localizing errors within vast, heterogeneous infrastructures. These approaches frequently fail to detect complex spatial or topological inconsistencies, leading to significant operational challenges and costly data remediation efforts. To address these limitations, a novel, automated validation pipeline has been developed with a modular, two-stage approach. The first stage, Smart Grid Partitioning, spatially divides the network into manageable cells using either fixed-size grids or a density-aware dynamic partitioning. This dynamic mode employs a bottom-up, clustering-inspired algorithm that adapts grid sizes to the local intensity of network equipment, effectively resolving issues of data sparsity and overload. The second stage, AI-Assisted Grid Validation, calculates a comprehensive Correctness Score for each resulting grid. This score provides a quantitative measure of data quality by synthesizing four weighted factors: (1) configurable rule-based attribute checks, (2) connectivity file conformance, (3) topological integrity assessed via advanced network trace functions, and (4) a series of representative graph-theoretic metrics. By generating an intuitive, color-coded map of data health, our framework allows utility providers to precisely localize data quality issues and prioritize remediation efforts. This targeted approach significantly enhances the efficiency of data maintenance, improves the integrity of foundational GIS data for critical power infrastructure, and streamlines integration with essential platforms like SCADA, OMS, and DMS.

1. Introduction

The transition towards smarter, more resilient, and efficient energy systems is critically dependent on the quality of the underlying data infrastructure. For power distribution networks, accurate and reliable Geographic Information System (GIS) data is the bedrock upon which essential functions—such as operational management, long-term planning, outage response, and the integration of distributed energy resources (DERs)—are built. The integrity of this data directly impacts the performance of critical systems like Supervisory Control and Data Acquisition (SCADA), Outage Management Systems (OMS), and Distribution Management Systems (DMS). However, ensuring this integrity across large, complex, and constantly evolving networks presents a formidable challenge.

Conventional data validation approaches, which predominantly rely on network-wide rule-based checks or laborious manual inspections, are increasingly proving inadequate. These methods struggle to scale with the growing complexity of modern grids and are often inefficient at pinpointing localized errors within sprawling, heterogeneous infrastructures. More importantly, they frequently overlook subtle yet critical spatial or topological anomalies that can compromise network analysis and lead to flawed operational decisions. The need for more sophisticated, automated, and targeted validation methodologies is therefore critical, a sentiment echoed across recent studies focused on data analysis and mining within smart grids (Li et al., 2024). The academic and industrial communities have made significant strides in addressing aspects of this challenge. Foundational work by (Wan et al. 2015) established a general framework for automated spatial data inspection, laying the theoretical groundwork for systematic quality assessment. Concurrently, the conceptualization of the power grid as a complex network has opened avenues for advanced analysis using graph theory (Pagani & Aiello, 2013), providing a robust theoretical basis for evaluating topological properties like density and connectivity. Building on these foundations, researchers have developed specific validation techniques, such as benchmark-driven feeder validation using statistical and operational metrics (Krishnan et al., 2020) and connectivity verification using data from Advanced Metering Infrastructure (AMI) (Luan et al., 2015). These methods have advanced the state-of-the-art but often focus on specific error types or data sources in isolation.

More recently, the advent of Artificial Intelligence (AI) and Machine Learning (ML) has introduced powerful new tools for geospatial analysis. For instance, Wang et al. (2023) demonstrated the use of ML with publicly accessible, multi-modal data to achieve high-precision geospatial mapping of distribution grids. Furthermore, the application of graph-based AI, including Knowledge Graphs and Graph Convolutional Networks (GCNs), has shown great promise for identifying complex topological errors that are invisible to standard rule-based systems (Chang et al., 2020; Fei et al., 2024b). Parallel advancements in spatial analysis have underscored the benefits of dividing extensive datasets into smaller, more manageable units.

Despite this progress, a critical gap remains: the absence of a holistic, density-aware, and grid-centric validation framework that integrates these disparate advancements into a unified, end-to-end pipeline. While individual tools for rule-based checks, connectivity verification, topological analysis, and spatial partitioning exist, they have not been systematically combined to offer a comprehensive solution tailored to the unique challenges of power distribution networks. This paper addresses this gap by introducing a two-stage, modular pipeline designed to automate and enhance GIS data validation. The first stage, Smart

Grid Partitioning, divides the network into either fixed-size or adaptively sized grids. The adaptive mode, inspired by density-based clustering, adjusts grid dimensions based on local equipment intensity, ensuring that both dense urban centers and sparse rural areas are partitioned effectively. The second stage, AI-Assisted Grid Validation, synthesizes multiple validation techniques into a single, unified Correctness Score for each grid. The composite score integrates weighted contributions from multiple assessment dimensions, combining rule-based attribute validations with connectivity-file conformance checks, higher-order topological integrity evaluations using network trace functions, and a comprehensive suite of graph-theoretic metrics.

Our framework empowers data engineers with more manageable workloads and precise error-correction strategies by localizing faults and supporting user-defined validation priorities. This approach not only minimizes the need for costly and time-consuming field investigations but also provides decision-makers with actionable, spatially-explicit insights for resource allocation and maintenance planning. Furthermore, the modular architecture enables seamless integration with existing utility management systems. High-quality grid outputs can be directly ingested into real-time SCADA, Outage Management System (OMS), and Distribution Management System (DMS) platforms, while lower-quality segments undergo offline refinement. This dual-track approach accelerates data readiness across large, multi-company projects while maintaining operational continuity, driving significant time and cost efficiencies in power distribution network management. Ultimately, by facilitating the rapid identification and correction of data quality issues, our pipeline accelerates the readiness of high-integrity data for critical real-time platforms, driving significant time and cost efficiencies in the management of modern power distribution networks.

2. Methodology

The proposed GIS data validation pipeline employs a sequential, multi-stage methodology specifically designed to systematically assess and localize data quality issues within power distribution networks. The framework architecture is built on a modular design philosophy that enables scalable processing, flexible configuration, and seamless integration with existing utility management systems. The methodology encompasses four primary phases: data ingestion and preprocessing, smart grid partitioning, AI-assisted grid validation, and results synthesis with temporal tracking. The system leverages spatial decomposition to account for variations in equipment density and complexity, replacing one-size-fits-all validation with adaptive, grid-based checks. By tailoring validation criteria to local network characteristics, we boost computational efficiency and ensure that inspections align with actual asset distributions.

The pipeline processes multiple data streams concurrently, including spatial geometries, attribute tables and connectivity data. The modular architecture ensures that individual components can be independently updated, configured, or replaced without affecting the overall system integrity. This design philosophy enables adaptation to diverse utility environments, regulatory requirements, and technical specifications while maintaining consistent validation quality.

2.1 Data Ingestion and Preprocessing

The validation pipeline begins by ingesting detailed **GIS asset data** encompassing spatial geometries and attributes for diverse equipment types including transformers, distribution lines, switches, and busbars. The system processes this spatial information alongside a **connectivity data file** that defines the topological relationships between infrastructure components, containing terminal-to-terminal connections and network structural information.

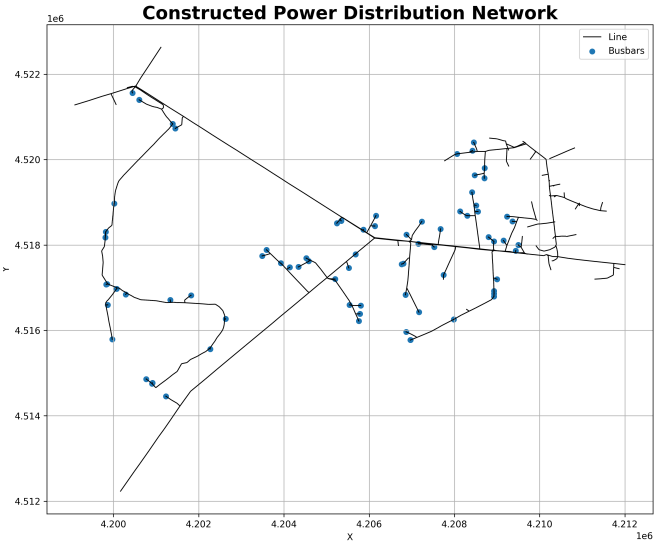

Figure 1. Power distribution network topology constructed after GIS data ingestion

Upon ingestion, a critical preprocessing phase ensures data consistency by harmonizing all spatial data into a single Coordinate Reference System (CRS), mapping disparate attribute schemas to a unified standard, and enforcing unique identifiers for every asset to eliminate ambiguity downstream. This module also validates basic data integrity—performing format checks, referential integrity verification, and initial topological consistency analyses—before network partitioning. The result is a cohesive spatial and structural foundation that meets minimum quality requirements, enabling effective partitioning and comprehensive validation in subsequent stages.

2.1.1 Pre-Partitioning Quality Metrics: The preprocessing phase generates key quality metrics—spatial coverage, attribute completeness, and topological connectivity statistics—that establish baseline indicators for tracking data quality improvements. It also performs statistical profiling of equipment distributions, analyzing density patterns, spatial clustering, and network topology (including spatial autocorrelation and density gradients) to inform and optimize adaptive partitioning parameters.

2.2 Smart Grid Partitioning

To manage the inherent complexity of large distribution networks and effectively localize data errors, the entire network is spatially divided into smaller, manageable grids using the **Smart Grid Partitioning** module. The partitioning strategy recognizes that power distribution networks exhibit natural spatial hierarchies, from high-density urban cores to sparse rural areas, and provides two distinct operational modes to accommodate different network characteristics and analysis requirements.

2.2.1 Fixed-Size Partitioning Methodology: In this mode, a fixed-size partitioning approach overlays the network with a uniform grid of rectangular cells, offering consistent spatial

resolution and predictable computational demands; users can either define cell dimensions manually or rely on an integrated optimal grid size calculator. The optimal grid size calculator implements an automated routine that computes grid dimensions by balancing a target number of cells N against a user-specified minimum equipment count per cell E_{min}. The calculator determines grid dimensions that satisfy three essential criteria:

- **Complete Coverage:** Grid cells completely cover the network's bounding box
- **Target Grid Count:** Produces approximately N cells across the network
- **Minimum Density:** Ensures expected equipment density per cell $\geq E_{min}$

Given a network bounding box with dimensions $W \times H$ and total equipment count E_{total}, the optimization process determines grid cell dimensions (w, h) that minimize the objective function:

$$f(w,h) = \alpha|N_{actual} - N_{target}| + \beta \max(0, E_{min} - \rho_{expected}) \quad (1)$$

The actual number of grid cells, denoted as $N_{\text{actual}} = \lceil \frac{W}{w} \rceil \times \lceil \frac{H}{h} \rceil$, is computed based on the width and height of the network extent divided by the chosen cell dimensions.

The expected equipment density per cell is given by $\rho_{\text{expected}} = \frac{E_{\text{total}}}{N_{\text{actual}}}$, where E_{total} is the total number of equipment points. Additionally, the optimization process incorporates two weighting parameters, α and β, to balance multiple partitioning objectives. The algorithm iteratively evaluates candidate grid dimensions within feasible bounds and selects the configuration that best satisfies the competing objectives of achieving target grid counts while maintaining adequate equipment density for reliable validation.

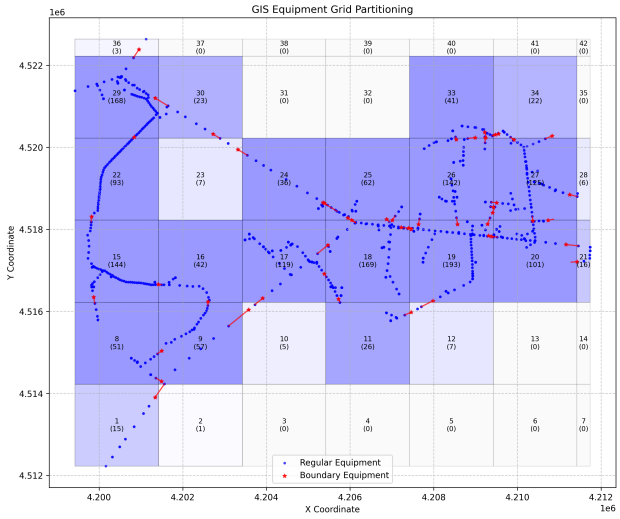

Figure 2. An example partitioned grid with fixed-size mode

2.2.2 Dynamic-Size Partitioning Methodology: Power distribution networks are often spatially heterogeneous, with high equipment density in urban centers and sparse distribution in rural or suburban areas. To address this, the dynamic-size partitioning mode adapts grid cell sizes to the local equipment intensity. This approach avoids the pitfalls of fixed-size grids, which can result in either empty, uninformative cells in sparse areas or overly congested cells in dense areas.

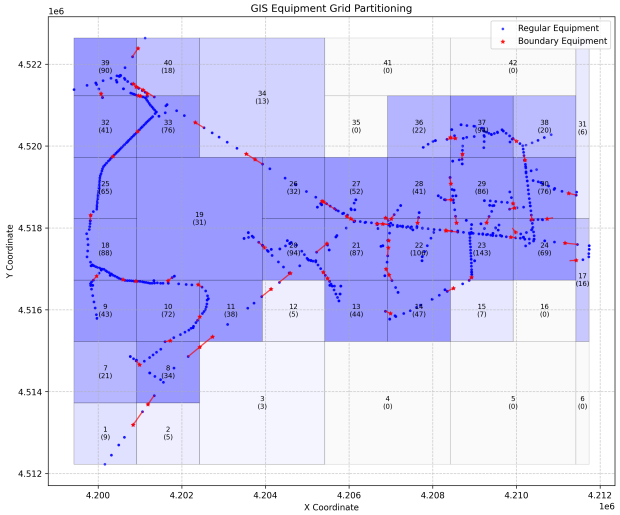

Figure 3. An example partitioned grid with dynamic-size mode

The dynamic partitioning algorithm uses a bottom-up, clustering-inspired approach based on DBSCAN (Ester et al., 1996) to adapt grid sizes to local equipment intensity. It begins with a fine-resolution grid and iteratively merges adjacent cells whose combined equipment density falls below a defined threshold. The full procedure is detailed in Algorithm 1.

Algorithm 1 Dynamic-Size Partitioning

Require: Fine-resolution grid G, equipment locations $\mathcal{E}$, intensity threshold T, maximum cell size $S_{\max}$
{Initialize counts}
for all cell $c \in G$ **do**
  $\text{count}[c] \leftarrow |\{e \in \mathcal{E} : e \text{ in } c\}|$
end for
while $\exists c \in G$: $\text{count}[c] < T$ **and** $\text{size}(c) < S_{\max}$ **do**
  {Select merge candidates}
  $(c_i, c_j) \leftarrow \arg\min_{(x,y) \text{ adjacent in } G} \big(\text{count}[x] + \text{count}[y]\big)$
  {Merge cells}
  Remove c_i, c_j from G; add merged cell $c' = c_i \cup c_j$
  $\text{count}[c'] \leftarrow \text{count}[c_i] + \text{count}[c_j]$
end while
return G

The process, detailed in Algorithm 1, operates as follows:

1. **Initialization:** The network's bounding box is first overlaid with a fine-resolution base grid. The number of equipment assets within each initial cell is counted.
2. **Iterative Merging:** The algorithm iteratively identifies adjacent grid cells where the combined equipment count falls below a user-defined intensity threshold (T). The pair of adjacent cells with the lowest combined count is merged into a single, larger cell.
3. **Termination:** This merging process continues until no adjacent cells can be merged without violating the intensity threshold or exceeding a maximum allowable cell size ($S_{\max}$).

2.2.3 Boundary Equipment Handling: Equipment spanning multiple grid cells requires handling strategies to maintain data integrity and prevent validation errors. Our system provides three configurable approaches for boundary equipment management, each optimized for different validation scenarios and accuracy requirements.

1. **Duplication Strategy:** The default approach includes boundary equipment in all intersecting grid cells, ensuring comprehensive coverage while accepting potential redundancy. This strategy proves optimal for critical equipment validation where missing components would create significant operational risks. The system maintains cross-reference tables that track duplicated equipment to prevent double-counting in aggregate statistics.

2. **Primary Grid Assignment:** This approach allocates boundary equipment exclusively to the grid containing the largest portion of the equipment geometry (e.g., greatest length for a line, greatest area for a polygon).

   By calculating geometric intersections to assign each piece of equipment to a primary grid cell, the algorithm minimizes redundancy and preserves single-source accountability for validation—although this can sacrifice important contextual information in neighboring cells.

3. **Geometric Splitting:** The equipment's geometry is physically split at the grid boundaries, with each resulting segment assigned to its corresponding cell. This is the most geometrically precise method but is computationally more intensive and requires careful handling of attribute data for the newly created features.

The chosen strategy is applied consistently across the network, and the resulting grid-specific connectivity information is updated accordingly.

2.2.4 Grid Metadata and Temporal Tracking: For auditability and temporal analysis, each grid cell maintains a comprehensive timestamped metadata record. This record includes the cell's unique ID and precise spatial extent, complete inventories of its assets and their assignment modes, and connectivity snapshots reflecting the network structure within the grid.

The metadata is stored hierarchically by execution time, enabling efficient temporal queries and change tracking. Each partition captures metrics like total and empty grids, equipment counts, and distribution percentiles, summarized in Table **??**. This structure facilitates monitoring data corrections and quality trends, identifying problem areas, and maintaining multi-resolution data for efficient retrieval and comprehensive audit trails.

Metric	Value
Total grids	79
Empty grids	22 (27.8%)
Min equipment per grid	0
Max equipment per grid	86
Average equipment per grid	21.8
Median equipment per grid	16

Table 1. An Example Grid Equipment Distribution Analysis

2.3 AI-Assisted Grid Validation

Once the network is partitioned, each grid undergoes an independent validation process. The AI-Assisted Grid Validation module evaluates each grid against a series of criteria and computes a composite *Correctness Score* S, a normalized value between 0 and 100 that quantifies the overall data quality. The score is calculated as:

$$S \;=\; 100 - \big(w_1 f_1 + w_2 f_2 + w_3 f_3 + w_4 f_4\big), \quad \sum_{i=1}^{4} w_i = 1 \quad (2)$$

Where f_i are normalized factor scores and w_i are user-adjustable weights that enable customization for different validation priorities, operational requirements, and regulatory contexts. The scoring formulation ensures that perfect validation yields the maximum score of 100, while increasing errors progressively reduce the score toward zero.

2.3.1 Rule-Based Attribute Checks (f_1): This factor assesses the syntactic and semantic quality of the attribute data for all equipment within a grid. The system employs a configurable suite of rule-based checks that can be tailored to any network data schema. Standard checks include detecting null or duplicate IDs, invalid attribute values (e.g., incorrect voltage levels), missing or null coordinate information, and incomplete boundary information for assets connected to adjacent grids. The score (f_1) is calculated based on the error rate within the grid:

For a grid containing equipment set $\mathcal{E}$, let $\text{errors}_k \subseteq \mathcal{E}$ represent the subset of features violating rule category k. The total error count and error rate are calculated as:

$$\text{total\_errors} = \sum_k |\text{errors}_k| \quad (3)$$

$$\text{error\_rate} = \min\Big(1,\ \frac{\text{total\_errors}}{|\mathcal{E}|}\Big) \quad (4)$$

The final rule-based score is normalized to ensure that $f_1 = 1$ indicates an error-free grid, with values decreasing toward zero as errors accumulate:

$$f_1 = 1 - \text{error\_rate} \quad (5)$$

2.3.2 Connectivity File Conformance (f_2): Connectivity file conformance assessment employs systematic analysis to validate topological relationships and identify structural inconsistencies. The system processes connectivity matrices and terminal definitions to ensure electrical network integrity through comprehensive error classification.

The validation process classifies connectivity errors into five primary categories: duplicate terminals, high-order connectivity violations, invalid terminals, missing terminal pairs, and missing connectivity records. Each category represents a different type of topological inconsistency that can impact network analysis and operational reliability.

Given a connectivity table of size N_c and a dictionary of error lists vr, the total error count and conformance rate are calculated as:

$$\text{total\_errors} = \sum_k |\text{vr}_k| \tag{6}$$

$$\text{error\_rate} = \begin{cases} \dfrac{\text{total\_errors}}{N_c}, & N_c > 0, \\ 1, & N_c = 0\,. \end{cases} \tag{7}$$

$$f_2 = 1 - \text{error\_rate} \tag{8}$$

This formulation ensures that grids with complete, accurate connectivity information receive high scores, while those with missing or inconsistent connectivity data receive proportionally lower scores.

2.3.3 Trace-Based Topological Integrity (f_3): To move beyond simple checks and assess the functional integrity of the grid, this factor leverages network tracing algorithms. The trace algorithms implement advanced logic for handling complex network configurations:

- $trace\_up$: Traverses upstream from loads to sources/injection points
- $trace\_down$: Traverses downstream to map load-distribution paths
- $trace\_all$: Conducts bidirectional traversal to ensure full connectivity coverage

To evaluate network integrity and issue severity, let N be the total node count, T the number of traversed nodes, and D_1, D_2, D_3 the counts of disconnections, dead ends, and switch issues respectively, each weighted by severity factors $w_1 > w_2 > w_3$. The weighted issues and coverage metrics are calculated as:

$$\text{weighted\_issues} = w_1 D_1 + w_2 D_2 + w_3 D_3 \tag{9}$$

$$\text{coverage} = \frac{T}{N} \tag{10}$$

The final trace score integrates issue detection with coverage assessment:

$$f_3 = \begin{cases} \left[1 - \dfrac{\text{weighted\_issues}}{N}\right] \times \text{coverage}, & N > 0, \\ 0\,, & N = 0\,. \end{cases} \tag{11}$$

2.3.4 Graph-Theoretic Metrics Assessment (f_4): Graph-theoretic validation constructs undirected graph representations of the network topology within each grid and computes comprehensive structural metrics specifically designed for power distribution network analysis. The implementation recognizes the unique characteristics of electrical networks, including their sparse topology, hierarchical structure, and specific branch-to-node relationships.

After constructing the undirected graph for each grid, the system computes key structural metrics including node count (n), edge count (m), density (δ), connected component count (c), isolated nodes (i), clustering coefficient (γ), average path length, and electrical network-specific metrics such as branch-to-node ratios.

The graph-theoretic score employs a penalty-based system that assesses multiple structural characteristics specific to power distribution networks. The composite score is calculated as:

$$f_4 = \max(0, 1 - \sum_j P_j) \tag{12}$$

where P_j represents individual penalty components:

Penalty Components:

1. **Component Fragmentation Penalty:**

$$P_{\text{comp}} = \min\big(0.3,\ 0.1\,(c-1)\big) \tag{13}$$

Penalizes networks with multiple disconnected components beyond the expected single component.

2. **Network Density Penalty:**

$$P_{\text{density}} = \min\big(0.2,\ \delta\big) \tag{14}$$

Penalizes networks with density above 0.1, as electrical networks are typically sparse.

3. **High-Degree Nodes Penalty:**

$$P_{\text{degree}} = \min\Big(0.2,\ 0.02\,\big|\{v\colon d(v) > 8\}\big|\Big) \tag{15}$$

Penalizes nodes with degree greater than 8, which may indicate modeling errors in power networks.

4. **Isolation Penalty:**

$$P_{\text{isolation}} = \min\Big(0.3,\ \frac{i}{n}\Big) \tag{16}$$

Penalizes isolated nodes that indicate connectivity problems.

5. **Branch-to-Node Ratio Penalty:**

$$P_{\text{ratio}} = \begin{cases} \min\big(0.2,\ 0.4\,(0.5-r)\big), & r < 0.5, \\ \min\big(0.2,\ 0.1\,(r-3)\big), & r > 3, \\ 0, & 0.5 \le r \le 3\,. \end{cases} \tag{17}$$

where r represents the branch-to-node ratio. Electrical networks typically maintain ratios between 0.5 and 3.0.

6. **Clustering Anomaly Penalty:**

$$P_{\text{clustering}} = \min\big(0.1,\ \gamma\big) \tag{18}$$

Penalizes clustering coefficients above 0.3, which are unusual in power distribution networks.

2.3.5 Composite Scoring & Visualization: The four factor scores (f_1 to f_4) are combined using user-defined weights (w_1 to w_4) to yield a final Correctness Score for each grid. That score—together with all individual factor scores and the associated metadata—is recorded for audit and analysis. For intuitive interpretation, the results are rendered as a thematic map in which each cell is color-coded along a green-to-red gradient. This immediate, at-a-glance view of network data health enables managers to pinpoint problematic areas and efficiently prioritize remediation efforts.

3. Results

To evaluate the efficacy and scalability of the proposed data validation pipeline, we conducted experiments on two real-world, medium-voltage power distribution networks of significantly different scales. The experimental validation demonstrates the framework's capability to handle networks ranging from compact urban deployments to extensive regional infrastructure.

3.1 Experimental Setup & Datasets

The validation pipeline was tested on two distinct networks:

- **Small Network:** A network covering a 132.6 km² area, comprising 1,588 equipment assets.
- **Large Network:** A significantly larger network spanning a 53,732 km² area, containing 97,967 equipment assets.

For both experiments, the **Dynamic-Size Partitioning** mode was employed to effectively handle the heterogeneous equipment densities characteristic of real-world networks. The Duplication strategy was used for handling boundary equipment. The composite Correctness Score was calculated using a consistent set of weights across both networks, prioritizing rule-based and topological checks: (w_1=0.35) (Rule-based Checks), (w_2=0.25) (Connectivity Checks), (w_3=0.25) (Trace Checks), and (w_4=0.15) (Graph Metrics).

3.2 Performance & Network-Wide Analysis

The pipeline demonstrated efficient performance on both datasets. For the small network, the entire process from ingestion to final reporting completed in 7.92 seconds. For the large network, the execution time was 245.05 seconds, showcasing the tool's scalability. Dynamic partitioning module divided the small network into 40 grids (35 non-empty) and the large network into 457 grids (368 non-empty). A summary of the validation results for both networks is presented in Table **??**.

Metric	Small Network	Large Network
Total Grids Analyzed	35	368
Average Score	57.36	50.36
Minimum Score	35.11	0.00
Maximum Score	68.86	74.00
Total Execution Time (s)	7.92	245.05
Time per Grid (s)	0.226	0.667

Table 2. Network Validation Summary Statistics

The results indicate that both networks contain data quality issues, with the larger network exhibiting a lower average score and a wider range of quality. The primary output of this analysis is a color-coded thematic map (results are shown in Figures 4 and 5), which provides an immediate visual guide to the spatial distribution of these data quality problems.

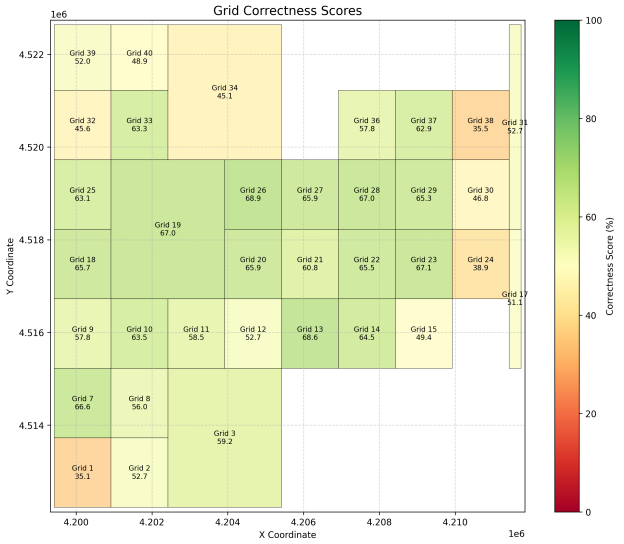

Figure 4. Color-coded thematic map for the small network.

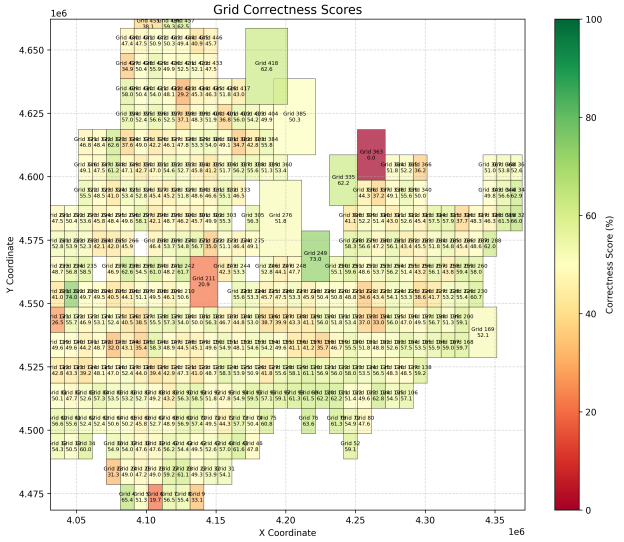

Figure 5. Color-coded thematic map for the large network.

The dynamic partitioning algorithm successfully created larger grid cells in sparse rural areas while maintaining compact cells in dense urban zones, optimizing the balance between computational efficiency and validation granularity. The high utilization rates of 87.5% for the small network and 80.5% for the large network demonstrate effective adaptation to equipment density patterns.

Small network achieved an average correctness score of 57.36 out of 100, with scores ranging from 35.11 to 68.86, indicating moderate quality with consistent maintenance needs across all grid regions. The large network demonstrated an average correctness score of 50.36 out of 100, with a wider score range from 0.00 to 74.00, indicating greater variability and heterogeneous data quality across different network regions.

The lower average score and wider score range in the large network reflect the challenges of maintaining consistent data quality across extensive geographical areas, highlighting the value of localized validation approaches for large-scale infrastructure management.

Validation framework demonstrated exceptional performance across all dimensions. The trace function accuracy reached a perfect 100% detection rate for interrupted and disconnected equipment. Meanwhile, rule-based validation consistently flagged attribute errors exactly as specified by the configured rules, connectivity problem localization reliably pinpointed equipment experiencing connectivity issues, and graph-based metrics provided crucial insights into topological anomalies and structural inconsistencies.

3.3 Operational System Integration Benefits

By delivering cleaner, more accurate network data, the framework streamlines SCADA monitoring, enhances state estimation, and reduces false alarms—ultimately supporting smarter control decisions. In OMS applications, precise topology accelerates outage localization, improves restoration planning, and sharpens customer communications. And within DMS, validated connectivity and attributes power more reliable power-flow analyses, better capacity and volt-var planning, and enable advanced distribution management functions.

Partitioning the network into validated grid regions **allows multiple teams to operate in parallel**, each applying their specialized expertise while adhering to uniform quality standards. Seamless integration with GIS update workflows automates ongoing quality checks, maintains a comprehensive audit trail for compliance, and offers real-time change-impact scoring to guide data modification decisions.

4. Conclusion

We developed an automated, grid-based validation pipeline that systematically ensures the quality and accuracy of GIS data in power distribution networks. By integrating a density-aware spatial partitioning algorithm with AI-assisted validation across four complementary dimensions —rule-based checks to enforce data standards, connectivity conformance to ensure proper linkage between components, trace-based topology integrity to verify network structure, and graph-theoretic metrics to assess structural properties- the framework delivers both computational efficiency and comprehensive error detection.This integrated approach not only enhances computational efficiency by localizing analysis but also ensures comprehensive detection of data quality issues throughout the power distribution network. Dynamic partitioning adapts to network density, producing larger cells in rural areas and finer granularity in urban zones, while the unified correctness score simplifies quality assessment.

Experimental evaluations on two real-world medium-voltage networks demonstrate the pipeline's readiness for operational deployment. Individual grid segments are validated in just 0.226 seconds for smaller grids and 0.666 seconds for larger ones, while end-to-end assessments of entire networks complete in under 8 seconds and 245 seconds, respectively. By pinpointing errors to specific spatial partitions, the system enables precise, grid-level remediation and facilitates parallel workflows across multiple engineering teams. Furthermore, seamless interoperability with SCADA, OMS, and DMS platforms enhances real-time network visibility, accelerates outage response, and drives more effective distribution optimization.

Our contributions fill a notable gap in infrastructure informatics, offering the first holistic, density-aware validation framework tailored to power distribution. Beyond immediate utility applications, the methods and mathematical foundations presented here lay the groundwork for continuous real-time monitoring, predictive quality assessment, and transfer to other infrastructure domains.

5. References

1. Wang, Changgang, Jun An, and Gang Mu. "Power system network topology identification based on knowledge graph and graph neural network." *Frontiers in Energy Research*, 8: 613331.

2. Li, X., Zhu, Z., Zhang, C. et al. "Power data analysis and mining technology in smart grid." *Energy Inform 7, 93 (2024).*

3. Krishnan, Venkat, et al. "Validation of synthetic US electric power distribution system data sets." *IEEE Transactions on Smart Grid 11.5* (2020): 4477-4489.

4. Wan, Yiliang, et al. "A general framework for spatial data inspection and assessment." *Earth Science Informatics 8.4* (2015): 919-935.

5. Luan, Wenpeng, et al. "Smart meter data analytics for distribution network connectivity verification." *IEEE transactions on smart grid 6.4 (2015)*: 1964-1971.

6. Pagani, Giuliano Andrea, and Marco Aiello. "The power grid as a complex network: a survey." *Physica A: Statistical Mechanics and its Applications 392.11* (2013): 2688-2700.

7. Wang, Zhecheng, Arun Majumdar, and Ram Rajagopal. "Geospatial mapping of distribution grid with machine learning and publicly-accessible multi-modal data." *Nature Communications 14.1* (2023): 5006.

8. Fei, Shuyu, et al. "A Power Grid Topological Error Identification Method Based on Knowledge Graphs and Graph Convolutional Networks." *Electronics (2079-9292) 13.19* (2024).

9. Ester, Martin, et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 226-231.