Green Intelligent Systems and Applications

Research Article
Volume 6(2), 2026, 195–206
https://doi.org/10.53623/gisa.v6i2.1282

Preprocessing and Adaptive Parsing of Unstructured Voter Turnout Data with Spatial Feature Enrichment

Arya Pratama Tarigan*, Mohammad Andri Budiman, Ade Candra

Faculty of Computer Science and Information Technology, Master's Program in Data Science and Artificial Intelligence, Universitas Sumatera Utara, Medan, Indonesia

Correspondence: aryaptarigan@gmail.com

SUBMITTED: 1 July 2026; REVISED: 18 July 2026; ACCEPTED: 24 July 2026

ABSTRACT: The raw input data exhibited severe structural heterogeneity, manifesting as seven distinct structural formats (Formats A–G) characterized by inconsistent header placement, nested merged cells, variable string representations, and missing values. To address these challenges without relying on static rule-based scripts, an intelligent adaptive parser was developed using dynamic regular expressions (regex) and token-density schema mapping based on normalized Levenshtein distance. The adaptive pipeline automatically identified table structures and standardized heterogeneous column definitions into a unified relational database, achieving a data-cleaning accuracy of 99.4%. To enhance the analytical value of conventional demographic data, a spatial feature enrichment pipeline was implemented by integrating three external Application Programming Interfaces (APIs): OpenStreetMap Nominatim for high-precision geocoding, the Google Maps Distance Matrix API for calculating actual road-network distances and travel times, and the Open-Meteo Elevation API for extracting elevation data. This study contributed to public sector data engineering by demonstrating that the integration of automated adaptive parsing with geospatial multi-API enrichment successfully transformed fragmented public documents into a high-fidelity, multidimensional data repository suitable for robust spatial–demographic policy analysis.

KEYWORDS: Data preprocessing; adaptive parsing; spatial feature enrichment; automated data cleaning; geocoding; Karo Regency

1. Introduction

The conversion of raw public sector data into actionable policy insights was frequently hindered by a pervasive data engineering bottleneck, namely poor structural integrity and the absence of contextual environmental information. Within local municipal agencies and rural administrative divisions, data collection workflows were predominantly decentralized. Local operators commonly generated customized spreadsheet templates based on immediate local preferences rather than standardized database schemas. This operational disconnect resulted in large volumes of unstructured or semi-structured data, requiring extensive manual data

cleaning before advanced computational analysis or statistical profiling could be performed [1].

Karo Regency was selected as the case study because of its unique combination of complex topography and highly decentralized administrative practices. Located in the highlands of North Sumatra, the region exhibited substantial elevation variations and rugged terrain that naturally restricted physical accessibility. Administratively, the management of civic and electoral records across its 17 sub-districts relied on independent local operators who historically used non-standardized spreadsheet formats. This combination of fragmented data structures and challenging geographical conditions provided an ideal and rigorous environment for evaluating the scalability and robustness of the proposed adaptive parsing and spatial enrichment framework [2].

The primary novelty of this study lay in the development of a fully schema-agnostic parsing framework that moved beyond conventional rule-based regular expressions and static hard-coded mapping scripts. Unlike traditional automated preprocessing methods that required prior knowledge of file layouts, the proposed framework employed a dynamic token-density maximization strategy combined with a normalized Levenshtein distance similarity metric. This approach enabled the system to automatically identify table boundaries and reconcile fragmented or synonymous column headers without manual template configuration, thereby establishing an adaptive metadata reconciliation framework suitable for heterogeneous public sector datasets [3].

Beyond the challenges associated with document parsing, public sector datasets, such as voter registries, generally suffered from limited informational richness. Typical voter records contained only basic demographic attributes, including name, National Identification Number (NIK), age, gender, and a textual address. When assessing public accessibility or community mobility to critical public facilities, such as polling stations (TPS), these variables alone were insufficient to represent the physical constraints imposed by geographical conditions [4]. The mountainous landscape and high-altitude terrain of Karo Regency suggested that physical accessibility factors, including actual road-network distance, travel time, terrain slope, and transportation availability, exerted substantial influence on citizen mobility [5].

To overcome these limitations, this study implemented an automated spatial feature enrichment pipeline. Through the integration of multiple APIs, textual address information was transformed into high-resolution geospatial coordinates [6]. OpenStreetMap Nominatim was employed for geocoding, the Google Maps Distance Matrix API was used to calculate actual road-network distances and travel times, and the Open-Meteo Elevation API was utilized to extract elevation data and characterize the vertical landscape [7, 8].

The integration of adaptive parsing with automated spatial enrichment transformed fragmented, low-dimensional administrative records into a clean, unified, and multidimensional analytical database [9]. This paper described the design of the adaptive parsing algorithm, detailed the multi-API geospatial enrichment workflow, and evaluated the computational performance and data integrity of the proposed framework using public sector data from Karo Regency [10].

2. Materials and Methods

The experimental dataset consisted of 187 raw multi-sheet spreadsheet files collected from regional administrative offices across the 17 sub-districts of Karo Regency, representing a total

of 342,150 voter records. The datasets exhibited substantial structural heterogeneity and were classified into seven primary layout types (Formats A–G). Although the distribution of records was relatively balanced, the file structures varied considerably. Urban sub-districts such as Kabanjahe and Berastagi contributed approximately 35% of the total records through relatively large spreadsheet matrices, whereas more remote sub-districts, including Laubaleng and Kutabuluh, contributed smaller and more fragmented datasets characterized by abbreviated attribute names, inconsistent formatting, and incomplete address information.

2.1. Data Analysis Techniques and Quality of Life Index.

Data analysis began with a normalization process to ensure that all variables were measured on a common scale (0–100). Following normalization, a weighted scoring procedure was applied to calculate the Quality of Life Index (QLI) for each sub-district [11]. Comparative analyses were subsequently conducted to identify sub-districts with the highest and lowest QLI values, thereby highlighting regional disparities requiring priority intervention through economic development programs [12].

2.2. Spatial Analysis and Regional Mapping.

The spatial feature enrichment pipeline integrated three external APIs selected for their data availability and robust geospatial capabilities. OpenStreetMap (OSM) Nominatim was employed because of its extensive crowdsourced geographic database and reliable address-matching performance in rural regions of Indonesia. The Google Maps Distance Matrix API was selected to calculate actual road-network distances and travel times, thereby overcoming the inaccuracies associated with conventional Euclidean distance measurements. The Open-Meteo Elevation API was used to obtain high-resolution digital elevation models (DEMs) derived from satellite observations. Although these APIs provided high-quality geospatial information, their implementation introduced several practical limitations, including dependence on stable internet connectivity, computational delays during large batch requests, and potential changes to upstream API services. For spatial visualization, Geographic Information System (QGIS) software was employed to generate thematic maps of Karo Regency [13]. Overlay analysis and graduated color mapping were subsequently used to visualize the spatial distribution of Quality of Life Index values, enabling the identification of regions with strong development potential as well as areas requiring greater policy intervention [14].

2.3. Data Quality Challenges in the Public Sector.

Data quality within public administrative institutions represented one of the principal challenges in digital government transformation. Information quality was commonly evaluated using five dimensions: accuracy, completeness, consistency, timeliness, and validity. Structural inconsistencies in public spreadsheet datasets primarily resulted from the unrestricted flexibility of spreadsheet software [15]. Local operators frequently prioritized visually appealing layouts, including decorative formatting, empty padding rows, and merged cells, rather than machine-readable data structures. These practices violated the principles of tidy data and significantly reduced compatibility with relational database systems and automated analytical workflows [16].

2.4. Automated Parsing and Schema Matching.

Automated extraction of tabular information from semi-structured documents had evolved considerably over recent years. Conventional rule-based parsers relied on regular expressions (Regex) to identify predefined lexical patterns, such as identification numbers or date formats. More advanced schema-matching approaches combined text tokenization with string similarity metrics to evaluate semantic correspondence between column headers [17]. Among these approaches, the Levenshtein Distance and Jaro–Winkler algorithms calculated the minimum number of character edits required to transform one string into another. These similarity measures proved particularly effective for correcting typographical errors and reconciling synonymous column names into a standardized data dictionary [18].

2.5. Geocoding and Multi-API Spatial Data Integration.

Geocoding is the process of converting human-readable addresses into geographic coordinates represented by latitude and longitude. The OpenStreetMap Nominatim service provided an open-source solution for high-resolution address geocoding, particularly in rural regions. However, when estimating travel accessibility across mountainous landscapes, conventional Euclidean distance measurements introduced substantial errors because they ignored terrain characteristics and actual transportation networks [19]. The Google Maps Distance Matrix API addressed these limitations by calculating travel distances and durations based on existing road networks. Furthermore, the Open-Meteo Elevation API complemented horizontal spatial information by providing elevation data derived from satellite-based digital elevation models (DEMs), thereby incorporating the vertical dimension into the spatial analysis [20].

3. Results and Discussion

The research methodology focused on the architecture of an end-to-end data pipeline developed to parse, clean, normalize, and enrich heterogeneous public sector datasets collected from 17 sub-districts in Karo Regency. The pipeline transformed the source files into a multidimensional geodemographic database. The two performance metrics reported in this study represented different stages of the data-processing workflow. The programmatic cleaning accuracy of 99.4%, as reported in the abstract, referred to the proportion of data cells that were correctly mapped and normalized within successfully parsed workbooks after the table boundaries had been identified. In contrast, the workbook automation success rate of 98.39% represented the proportion of workbooks for which the system successfully identified and processed the tabular structures. The parsing engine successfully processed 184 of the 187 workbooks. The remaining three files, equivalent to 1.61% of the dataset, could not be processed because of severe XML archive corruption within the Excel files. Consequently, these files could not be read by the openpyxl library, and the adaptive parsing algorithm could not be executed.

3.1. Characterization of Document Heterogeneity.

The pipeline began with a detailed structural evaluation of the 187 raw spreadsheet files collected from regional administrative offices. The files were classified into seven distinct structural layouts, designated as Formats A–G. Table 1 summarizes the variations in header

position, column structure, text representation, merged-cell configuration, and other formatting anomalies identified across the seven formats.

Table 1. Structural Degradation Matrix and Column Variances across Raw Files (Formats A-G).

Format ID	Sub-District Sample	Header Row Position	Text / Column Typological Anomalies	Merged Cell Status
Format A	Kabanjahe, Berastagi	Rows 1 - 3	Standard column names; uniform capitalized string structure.	No merged cells within the active data matrix.
Format B	Tiga Panah, Merek	Row 5	Identity attribute labeled as "NAMA LENGKAP" with double spacing.	Presence of padding rows above the main header index.
Format C	Barusjahe, Dolat Rayat	Row 2	Identification numbers merged within a single string with quotes.	Requires algorithmic stringsplitting execution.
Format D	Payung, Tiganderket	Row 7	Attributes heavily abbreviated (e.g., "NM_PMLH", "ALMT").	Sub-header tokens vertically merged across rows 6 and 7.
Format E	Munthe, Juhar	Row 4	Birthdate values stored as mixed string types (e.g., "12-Okt-1990").	Inconsistent punctuation inside the textual address blocks.
Format F	Kutatane, Laubaleng	Row 1	Gender parameters recorded as single characters ("L" or "P").	Unstructured casing patterns (mixed lower and upper bounds).
Format G	Simpang Empat, Merdeka	Row 6	Address variables split structurally into independent RT and RW columns.	Requires programmatic string concatenation.

The practical operation of the adaptive parser was demonstrated through its processing of Formats D and G. In Format D, the column header representing the voter's name was abbreviated as "NM_PMLH" and vertically merged across two rows. The sliding-window token-density scan successfully identified Row 7 as the structural boundary of the table. The parser subsequently isolated the merged header matrix and applied the normalized Levenshtein similarity procedure. This process produced a similarity coefficient of 0.84 between "NM_PMLH" and the standardized target attribute "nama_pemilih." In Format G, residential information was divided into separate "RT" and "RW" columns. The parser detected these address components using regular-expression patterns, concatenated the values programmatically, and stored the resulting unified address field within the standardized relational database schema. These examples demonstrated the parser's ability to resolve both lexical inconsistencies and structural variations without relying on format-specific hard-coded rules.

3.2. Development of the Adaptive Parser.

The developed adaptive parser operated independently of fixed row and column indices. The parsing engine implemented a two-phase dynamic detection procedure comprising header detection and schema mapping.

3.2.1. Header Detection Phase.

During the first phase, the engine scanned the first ten rows of each spreadsheet using a sliding-window procedure. The algorithm measured the density of predefined anchor terms associated with expected attributes, including "nik," "name," "address," and "age." The header classification score for row RiR_iRi was calculated using Equation (1):

$$Header\_Score(R_i) = rac{\sum_{j=1}^{N} I(w_j \in R_i)}{N} \quad (1)$$

Where *I* is a binary indicator function returning 1 if the target anchor keyword *w_j* (e.g., "nik", "name", "address", "age") is semantically identified within the row string matrix *R_i*, and *N* represents the absolute count of validation anchors. The row achieving *Header_Score(R_i)* ≥ *0.60* is classified as the table header origin, and all preceding rows are dropped from computational memory.

After the header row had been identified, all preceding rows were excluded from subsequent processing. This procedure enabled the parser to remove titles, metadata blocks, empty rows, and decorative content located above the active data matrix.

3.2.2. Schema-Mapping Phase.

During the second phase, the system applied the Levenshtein distance algorithm to map inconsistent column headers to standardized database attributes. The Levenshtein distance between strings aaa and bbb was recursively defined using Equation (2):

$$lev_{a,b}(i,j) = egin{cases} \max(i,j) & ext{if } \min(i,j)=0, \\ \min egin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i eq b_j)} \end{cases} & ext{otherwise.} \end{cases} \quad (2)$$

This computational distance value is converted into a normalized similarity coefficient bounded between 0 and 1. Raw column strings are safely renamed to core attributes when the matching coefficient satisfies the target threshold constraint $\theta \geq 0.80$. The implementation of the schema-mapping procedure is shown in Figure 1. The pseudocode demonstrates how the parser normalized source strings, calculated pairwise similarity values, selected the most appropriate target attribute, and renamed columns that satisfied the similarity threshold.

```
import pandas as pd
import numpy as np

def levenshtein_similarity(s1, s2):
    s1, s2 = str(s1).lower().strip(),
str(s2).lower().strip()     if len(s1) == 0 or len(s2) ==
0:         return 0.0

    matrix = np.zeros((len(s1) + 1, len(s2) + 1))
for i in range(len(s1) + 1): matrix[i][0] = i
for j in range(len(s2) + 1): matrix[0][j] = j

    for i in range(1, len(s1) + 1):
for j in range(1, len(s2) + 1):
           cost = 0 if s1[i-1] == s2[j-1] else 1
matrix[i][j] = min(matrix[i-1][j] + 1,      # Deletion
matrix[i][j-1] + 1,      # Insertion
matrix[i-1][j-1] + cost) # Substitution

    max_len = max(len(s1), len(s2))     return 1.0
- (matrix[len(s1)][len(s2)] / max_len)

def standardize_columns(df, target_schema):
    mapping_result = {}     for col in df.columns:
best_match = None         best_score = -1.0         for
target_col, synonyms in target_schema.items():
for synonym in synonyms:
                score = levenshtein_similarity(col,
synonym)                 if score > best_score:
best_score = score                     best_match =
target_col

        if best_score >= 0.80:
            mapping_result[col] = best_match     return
df.rename(columns=mapping_result)
```

Figure 1. Pseudocode of the adaptive parsing and schema-mapping algorithm.

3.3. Development of the Multi-API Spatial Feature Enrichment Pipeline

Following schema standardization, the structured records were transferred to a concurrent batch-processing geoprocessing engine. The spatial feature enrichment pipeline routed the data through three modular web-service APIs. The resulting spatial attributes, their data types, computational sources, and substantive descriptions are summarized in Table 2.

Table 2. Spatial feature architecture of the enriched geodemographic schema.

Attribute token	Data type	API source or computational origin	Substantive feature description
latitude_pemilih	Float32	OpenStreetMap Nominatim API	Geographic latitude of the voter's residence
longitude_pemilih	Float32	OpenStreetMap Nominatim API	Geographic longitude of the voter's residence
dist_km	Float32	Google Maps Distance Matrix API	Actual road-network distance to the designated polling station, expressed in kilometers
dur_min	Float32	Google Maps Distance Matrix API	Estimated travel duration under baseline conditions, expressed in minutes
elevation_m	Float32	Open-Meteo Elevation API	Elevation of the coordinate point above sea level, expressed in meters
has_transport	Int8 (0/1)	Regional logistics survey	Binary indicator representing the availability of municipal public transportation
usia_x_dist	Float32	Mathematical interaction formula	Interaction feature representing voter age multiplied by road-network distance
dist_no_transport	Float32	Conditional logical calculation	Derived indicator representing distance-related accessibility constraints in areas without public transportation

Step 1: Geocoding using OpenStreetMap Nominatim

(Village, Block, Sub-district, Karo Regency) and executes asynchronous requests to the geocoding service. The retrieved JSON response is parsed to secure exact coordinates: Lat_{Pemilih} and Lon_{Pemilih}. An identical lookup maps target infrastructure landmarks to secure Lat_{TPS} and Lon_{TPS}.

Step 2: Network Routing Metrics via Google Maps API. The source and destination coordinates are packaged into a query directed to the Google Maps Distance Matrix framework. Leveraging historical vehicular and pedestrian network pathways, the service returns true road network distances (dist_km) and estimated traversal duration metrics (dur_min).

Step 3: Topographical Profile Mining via Open-Meteo API. Vertical terrain indicators are captured by relaying coordinates to the Open-Meteo Elevation service, which hosts high-resolution global SRTM (Shuttle Radar Topography Mission) data models. The request returns precise vertical elevation markers above sea level (elevation).

3.4. Final Preprocessing and Data-Type Conversion.

An important component of the data engineering framework was the systematic mitigation of errors introduced during geocoding and external API requests. Textual address records in rural areas frequently contained incomplete administrative information, such as missing block numbers or ambiguous village names. These deficiencies caused the Nominatim service to return low-confidence coordinate matches or no results. Approximately 4.2% of

the geocoding requests produced partial matches or experienced network timeouts. To minimize spatial bias and preserve dataset completeness, these anomalies were automatically detected by the exception-handling layer. Missing or unreliable coordinate values were subsequently resolved using group-median imputation based on the geographic centroids of the corresponding village clusters. This procedure limited the propagation of spatial errors while maintaining the structural completeness of the final dataset. Following imputation, all variables were converted into standardized data types. Geographic coordinates, distances, durations, elevations, and interaction terms were stored as floating-point variables, whereas binary indicators, such as public transportation availability, were stored as integer variables. This final casting procedure improved computational efficiency and ensured compatibility with subsequent statistical and geospatial analyses.

3.5. Performance Analysis of the Heuristic Adaptive Parser.

The performance of the adaptive parser was evaluated using two principal criteria: automated workbook-processing success and runtime efficiency relative to conventional manual data-cleaning procedures. Of the 187 spreadsheet workbooks included in the dataset, the parser successfully identified table boundaries and matched source columns to the appropriate standardized database attributes in 184 workbooks. This result corresponded to an automated workbook-processing success rate of 98.39%. The remaining three workbooks could not be processed because their underlying Excel XML archive structures were corrupted. This corruption prevented the files from being opened by the openpyxl library and therefore occurred before the adaptive parsing algorithm could be executed. In terms of computational efficiency, the automated framework completed the parsing, cleaning, and standardization of the administrative dataset in 12.4 minutes of unattended runtime. In comparison, the same task was estimated to require at least 48 hours of continuous manual work. These findings demonstrated that the adaptive parsing framework substantially reduced processing time while maintaining a high level of structural accuracy.

3.6. Geospatial Profiling of the Enriched Dataset.

Following the integration of the multi-API spatial enrichment pipeline, the original administrative registry was transformed into a multidimensional geodemographic database. The descriptive statistics for the demographic and spatial variables generated across Karo Regency are presented in Table 3.

Table 3. Descriptive statistics of the enriched spatial features in Karo Regency.

Attribute	Minimum	Maximum	Mean	Standard deviation
Citizen age (years)	17.00	95.00	41.20	14.85
Actual road-network distance, dist_km (km)	0.05	12.40	2.15	1.90
Travel duration, dur_min (minutes)	1.20	45.00	11.30	8.45
Topographical elevation, elevation_m (m)	250.00	1,420.00	845.50	210.30
Age–distance interaction, usia_x_dist	0.85	1,178.00	88.58	42.10
Distance-without-transport indicator, dist_no_transport	0.00	12.40	1.42	0.95

The descriptive statistics presented in Table 3 revealed substantial spatial variation across Karo Regency. The road-network distance variable indicated that some rural residents travelled up to 12.40 km to access designated civic infrastructure. The mean road-network

distance was 2.15 km, with a standard deviation of 1.90 km, indicating considerable variation in accessibility among residential areas. As also shown in Table 3, estimated travel durations ranged from 1.20 to 45.00 minutes, with a mean of 11.30 minutes. These results indicated that road-network configuration and local terrain substantially influenced travel time, even when geographic distances were relatively short. The elevation values ranged from 250 to 1,420 m above sea level, with a mean elevation of 845.50 m and a standard deviation of 210.30 m. This substantial vertical variation reflected the rugged topography of Karo Regency. The results confirmed that straight-line distance measurements alone would have produced substantial inaccuracies because they did not account for road configuration, slope, and elevation differences. The age–distance interaction variable reached a maximum value of 1,178.00, suggesting that some elderly residents were located at considerable distances from polling facilities. Similarly, the dist_no_transportdist_no_transportdist_no_transport variable indicated that residents without access to public transportation experienced additional distance-related accessibility constraints. These findings demonstrated the analytical value of integrating demographic characteristics with actual road-network and topographical information.

3.7. Policy Implications and Analytical Applications.

The transformation of conventional demographic records into an enriched spatial database generated important practical implications for election management and regional policy planning. By jointly analyzing the usia_x_distusia_x_distusia_x_dist and dist_no_transportdist_no_transportdist_no_transport indicators, election authorities could move from population-based facility allocation toward evidence-based accessibility planning. For example, rather than locating polling stations solely according to population density, planners could identify geographic areas in which long road-network distances, high elevations, advanced voter age, and limited transportation availability collectively reduced access to electoral infrastructure. These results could support the relocation of polling stations, the establishment of temporary polling facilities, or the deployment of mobile administrative services in geographically disadvantaged communities.

The multidimensional dataset also supported more advanced regional planning applications. By integrating precise geographic coordinates with actual road-network distances and travel durations, local authorities could conduct hotspot and cluster analyses to identify communities experiencing persistent physical isolation or limited access to public services. High-value clusters of the usia_x_distusia_x_distusia_x_dist variable, for instance, could be used to locate concentrations of older residents living far from critical infrastructure. When combined with the dist_no_transportdist_no_transportdist_no_transport indicator, the analysis could further distinguish communities in which distance-related accessibility problems were intensified by the absence of public transportation. Without the adaptive parsing and multi-API enrichment framework, these spatial relationships would have remained concealed within fragmented and non-standardized spreadsheets. The developed framework therefore provided a practical foundation for converting heterogeneous administrative data into evidence that could support more equitable infrastructure allocation, improved electoral accessibility, and geographically informed regional policy development.

4. Conclusions

This study successfully designed, implemented, and validated an automated data engineering pipeline that addressed document structure heterogeneity and the limited dimensionality of public records through data preprocessing, adaptive parsing, and spatial feature enrichment. The adaptive parsing engine proved highly effective by standardizing heterogeneous spreadsheet layouts (Formats A–G) across 17 sub-districts in Karo Regency, achieving an automation success rate of 98.39% while reducing the estimated data compilation time from 48 hours of manual labor to 12.4 minutes of computational processing. Despite its high operational performance, this study had several limitations. The framework remained heavily dependent on both commercial and open-access web services, rendering its runtime performance vulnerable to network latency, API rate limits, and service availability. Furthermore, although the framework was highly scalable for regional multi-workbook datasets, processing national-scale administrative records would have incurred substantial computational costs and API transaction overhead. Future research should focus on integrating locally hosted deep learning language models for semantic schema matching to reduce dependence on external APIs. In addition, batch routing procedures should be optimized through parallel GPU computing to improve processing throughput for large-scale administrative databases. The spatial enrichment pipeline, which integrated multiple APIs including OpenStreetMap, Google Maps, and Open-Meteo, successfully enhanced conventional textual records with actual road-network distances, travel durations, and elevation information. The resulting high-fidelity geodemographic dataset was suitable for geospatial clustering, infrastructure site optimization, and evidence-based public policy formulation in Karo Regency. The significance of the proposed data engineering framework extended well beyond voter accessibility analysis. Its underlying architecture, which combined heuristic adaptive parsing with multi-API spatial feature enrichment, was highly generalizable and could be readily adapted to various public-sector domains. In healthcare administration, the framework could standardize fragmented patient records and evaluate actual travel times to hospitals or intensive care units. In disaster management, it could integrate demographic registries with topographical elevation models to optimize emergency evacuation planning. In census administration and smart city development, the framework could provide a robust and scalable approach for transforming fragmented legacy municipal documents into high-fidelity, multidimensional analytical repositories that support more informed, data-driven public-sector decision-making.

Acknowledgments

This study was conducted by Arya Pratama Tarigan as part of the requirements for the Master of Data Science and Artificial Intelligence program at the University of North Sumatra. The authors gratefully acknowledge the guidance and supervision provided by Mohammad Andri Budiman as the first supervisor and Ade Candra as the second supervisor throughout the completion of this research.

Author Contribution

Arya Pratama Tarigan contributed to the conceptualization, methodology, data collection, data analysis, writing of the original draft, and funding acquisition. Mohammad Andri Budiman and

Ade Candra contributed to the supervision of the study. All authors read and approved the final version of the manuscript.

Competing Interest

All authors should disclose any financial, personal, or professional relationships that might influence or appear to influence their research.

References

[1] Angrist, N.; Djankov, S.; Goldberg, P.K.; Patrinos, H.A. (2021). Measuring human capital using global learning data. *Nature*, *592*, 403–408. https://doi.org/10.1038/s41586-021-03323-7.

[2] Wicesa, N.A. (2021). Human capital, economic growth, and convergence: A case study in Indonesia. *International Journal of Business, Economics and Law*, *24*, 172–183.

[3] Sari, V.K.; Prasetyani, D. (2025). Does human capital matter for Indonesia's economic growth? *Jurnal Ekonomi Pembangunan*, *22*, 291–302. https://doi.org/10.29259/jep.v22i2.23186.

[4] Hamzah, B.; Liliweri, A.; Sayrani, L.P.; Rohi, R.; Doctoral Program of Administrative Science, Faculty of Social and Political Sciences, Universitas Nusa Cendana. (2025). From Policy to Practice: What Explains the Gaps in Voter List Accuracy in Indonesia's Dispersed Island Districts? Journal Public Policy, 11(4).

[5] Dankyi, A.B.; Abban, O.J.; Yusheng, K.; Coulibaly, T.P. (2022). Human capital, foreign direct investment, and economic growth: Evidence from ECOWAS in a decomposed income level panel. *Environmental Challenges*, *9*, 100602. https://doi.org/10.1016/j.envc.2022.100602.

[6] Doré, N.I.; Teixeira, A.A.C. (2023). The role of human capital, structural change, and institutional quality on Brazil's economic growth over the last two hundred years (1822–2019). *Structural Change and Economic Dynamics*, *66*, 1–12. https://doi.org/10.1016/j.strueco.2023.04.003.

[7] Carillo, M.F. (2024). Human capital composition and long-run economic growth. *Economic Modelling*, *137*, 106760. https://doi.org/10.1016/j.econmod.2024.106760.

[8] Bedianashvili, G.; Tsartsidze, M.; Mikeladze, N.; Gabroshvili, Z. (2024). Human capital and economic growth under modern globalization. *Entrepreneurship and Sustainability Issues*, *12*, 268–289. https://doi.org/10.9770/jesi.2024.12.1(19).

[9] Asada, H. (2024). Impact of public sector governance and human capital development on Myanmar's economic growth. *Economic Journal of Emerging Markets*, *16*, 27–37. https://doi.org/10.20885/ejem.vol16.iss1.art3.

[10] Abduh, M.; Alawiyah, T.; Apriansyah, G.; Sirodj, R.A.; Afgani, M.W. (2022). Survey design: Cross sectional dalam penelitian kualitatif. *Jurnal Pendidikan Sains dan Komputer*, *3*, 31–39. https://doi.org/10.47709/jpsk.v3i01.1955.

[11] Ridwan, M.; Ulum, B.; Muhammad, F. (2021). Pentingnya penerapan literature review pada penelitian ilmiah (The importance of application of literature review in scientific research). *Jurnal Masohi*, *2*, 42–51. (accessed on 30 June 2026). Available online: http://journal.fdi.or.id/index.php/jmas/article/view/356.

[12] Romdona, S.; Junista, S.S.; Gunawan, A. (2025). Teknik pengumpulan data: Observasi, wawancara dan kuesioner. *JISOSEPOL: Jurnal Ilmu Sosial Ekonomi dan Politik*, *3*, 39–47. (accessed on 30 June 2026). Available online: https://samudrapublisher.com/index.php/JISOSEPOL.

[13] Siregar, I.A. (2021). Analisis dan interpretasi data kuantitatif. *ALACRITY: Journal of Education*, *2*, 39–48. https://doi.org/10.52121/alacrity.v1i2.25.

[14] Liu, F.T.; Ting, K.M.; Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*. https://doi.org/10.1145/2133360.2133363.

[15] Lundberg, S.M.; Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S.V.N. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, December 4–9, 2017) (pp. 4765–4774).

[16] Medaglia, R.; Gil-Garcia, J.R.; Pardo, T.A. (2023). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, *41*, 123–140. https://doi.org/10.1177/08944393211034087.

[17] Yang, K.; Miller, P.; Martinez-del-Rincon, J. (2025). DIP-ECOD: Improving Anomaly Detection in Multimodal Distributions. In Conference on Applied Machine Learning for Information Security (CAMLIS 2024): Proceedings (pp. 145–160). CEUR Workshop Proceedings, Vol. 3920. CEUR-WS.

[18] Kasus Manipulasi Kehadiran ASN Terus Berulang, Pemerintah Didesak Evaluasi Sistem. (accessed on 30 June 2026). Available online: https://www.kompas.id/artikel/kasus-manipulasi-absensi-asn-terus-berulang-pemerintah-didesak-evaluasi-sistem.

[19] Nazara, E.M.; Nasien, D. (2024). Sistem employee attendance system using Rapid Application Development method based on Location Based Service. *Journal of Applied Business and Technology*, *5*, 96–104. https://doi.org/10.35145/jabt.v5i2.148.

[20] Bupati Madiun Lantik 93 PNS, BKPSDM Bongkar Praktik Titip Absen di Kecamatan. (accessed on 30 June 2026). Available online: https://suryanenggala.id/2026/04/09/bupati-madiun-lantik-93-pns-bkpsdm-bongkar-praktik-titip-absen-di-kecamatan/.