Development of a Comprehensive Spatial Database for Bus Stops of a Public Transit Network

A geospatial approach to streamline KSRTC's transit infrastructure data

By

Dr. Shehna Basheer, Project Consultant

Indu P. Nair, Senior GIS Analyst

Anoja B.V., Senior GIS Analyst

Renjitha R. N., Junior Consultant

KIIFCON Pvt. Ltd.
Thiruvananthapuram
Kerala - 695001
June 2025

1. INTRODUCTION

Bus stops are fundamental components of any public transportation system. They serve as the primary points of access to public transportation system for passengers and form the backbone of transit operations. For transit agencies that use a stage-based fare collection system, bus stops fundamentally serve as fare stages as well, which are designated at specific intervals according to the fare structure prescribed by the government. Without reliable geospatial data of bus stops, transit agencies cannot effectively manage fare collection, optimize routes, assess demand patterns, or provide real-time passenger information, all of which are critical for modern transit operations and for the adoption of Intelligent Transportation Systems (ITS) and Advanced traveller information systems (ATIS). Moreover, for interstate services, where fare structures vary by state, accurate stop and fare stage data enable the transit agency to correctly apply the respective government's fare policies once crossing state borders. Comprehensive bus stop inventories also support demand forecasting, service reliability, and infrastructure planning, thereby enhancing overall transit system performance and passenger experience.

The Kerala State Road Transport Corporation (KSRTC) is one of India's oldest and largest state-run public transport providers, offering an extensive network of services across the state as well as interstate routes. KSRTC operates approximately 35,000 trips every day, connecting the nooks and corners of the state and acting as a major service provider for interstate travel at competitive prices compared to private operators in the domain. The corporation manages a diverse fleet, including electric and diesel buses, and serves more than 1.5 million passengers daily, making it a critical lifeline for mobility, economic activity, and social inclusion in the region. Given the scale and complexity of its operations, KSRTC's efficiency and service quality are highly dependent on accurate operational data, particularly regarding bus stops and fare stages, which supports its stage-based fare collection system and route management.

KSRTC had been trying to digitize the data related to its daily operations by adopting Vehicle Location Tracking Devices (VLTD) and working with various agencies to spatially represent the fundamental units of operation and for system development. However, these attempts were not as fruitful as desired due to the lack of accurate information on bus stops and the inability to link bus stops to corresponding trips. Knowledge of stop locations and fare stages existed informally among local depot staff and bus crews, with no standardized records or reliable geographic references. Often, the data was stored in multiple Excel sheets and different versions of the same data. The data frequently contained different spelling variations for a single place, and there were multiple locations with the same name that may appear as duplicates, but in reality, these places belonged to different districts in the state.

This lack of structured data posed significant challenges for route planning/rationalization, fare management, vehicle tracking and monitoring systems, and the adoption of advanced transit technologies. Without a comprehensive inventory, it was nearly impossible for any expert agency to develop a reliable spatial data framework for KSRTC or to implement essential functions such as route optimization, accurate fare stage assignment, and real-time passenger information—capabilities increasingly expected in modern transit systems. Consequently, the absence of reliable information on the location and infrastructure of bus stops and fare stages across the state became a critical obstacle in the organization's efforts to digitize its data. Additionally, fare stage names were not linked to route or trip-related data, making it impossible to verify their geographic positions or sequence within a trip.

Recognizing the fundamental importance of bus stop data, KIIFCON Pvt. Ltd., which was entrusted with the route rationalisation initiative of KSRTC, focused on building a comprehensive and reliable

dataset for KSRTC bus stops. Since GPS data was unavailable for the KSRTC fleet, an alternative approach was necessary to collect bus stop information. To ensure completeness and accuracy, KIIFCON utilized a smartphone-based survey application that garnered the GPS capabilities of smartphones and supported the visualization of data with GIS layers.

KSRTC supported the initiative by organizing the field data collection teams composed of bus conductors, who possess in-depth knowledge of the location of transit stops and the route network. By leveraging the local expertise of these conductors, KIIFCON's approach enabled the capture of detailed information on bus stop infrastructure, geographic coordinates, and photographic documentation of transit stops.

This white paper outlines the background, methodology, and outcomes of the bus stop data collection effort, highlighting its significance for KSRTC's modernization and offering insights for other transit agencies facing similar data challenges.

2. BUILDING A DATABASE FOR BUS STOPS

While the geographic locations of bus stops were an identified requirement from KSRTC and recognized as a fundamental attribute for a transit stop database, additional fields and attributes were determined with a view toward integration with Advanced Public Transportation Systems (APTS). Essential fields required for a GTFS (General Transit Feed Specification) database were also considered during the data collection process. The initial step in designing the database involved understanding the specific requirements of the transit organization and envisioning the fields necessary for successful integration with both the organization's existing databases and potential future integration with ATIS/ITS (Advanced Traveler Information Systems/Intelligent Transportation Systems).

The key attributes identified for the transit stop database included the name of the stop, other local names, the geo-coordinates of the stop and the photographs of the bus stop. Additionally, information on the services that stop at the particular stop and whether the bus stop is a fare stage or not were also collected. Details of additional attributes collected are detailed in the next section.

3. SMARTPHONE-BASED SURVEY APPLICATION FOR BUS STOP DATA COLLECTION

3.1. Survey Application

Given the widespread penetration of smartphones in Kerala—ranked second in India with 121.89 phones per 100 people, according to the latest data from the Telecom Regulatory Authority of India—a smartphone-based approach was identified as the most practical and scalable solution for the KSRTC bus stop data collection initiative. A smartphone-based application with integrated GIS capabilities was identified and then subsequently utilized to create, deploy, and analyze field surveys in a structured and efficient manner for the data collection. This application enabled the creation and deployment of flexible, user-friendly survey forms that allowed for the efficient capture of both spatial and non-spatial data directly from the field. After tailoring the application to suit the data collection needs, the enumerators were instructed to download it from the Play Store onto their personal smartphones. The

widespread availability of GPS-enabled smartphones among the survey staff facilitated accurate and seamless data collection, even in remote and network-challenged areas.

Recognizing that many locations across the state experience inconsistent network coverage and that enumerators may use different network providers with varying levels of connectivity, a critical requirement for offline data collection was identified during the planning phase and validated through a pilot survey. To address this, an offline data collection setup was incorporated into the application, allowing enumerators to download maps of their assigned areas when network connectivity was available and store them on their devices. While in the field, enumerators could reference these offline maps to accurately capture geospatial coordinates, even in areas without network access. Data collected in offline mode was temporarily stored on the device and could be transmitted to the central database once connectivity was restored, either automatically or by manual upload.

The application featured advanced form logic, including conditional questions, multiple choice inputs, dropdown menus, and options for multimedia attachments (ex: photographs). Geotagging capabilities were also integrated, automatically capturing the geographic coordinates of each survey entry, which was essential for accurate mapping and spatial analysis. Designed with a spreadsheet-compatible form standard, the tool allowed for easy configuration and data entry, making it accessible even to non-technical users. Its intuitive interface enabled a wide range of field personnel, regardless of their digital literacy, to participate effectively in the data collection process.

Serving as a digital bridge between field surveyors and the central data repository, this GIS-enabled application empowered depot-level staff to capture comprehensive information on each bus stop, including location coordinates, shelter and sidewalk availability, and photographic documentation, all in a consistent and standardized format. Real-time monitoring and validation were facilitated through a central dashboard, ensuring transparency and data integrity throughout the project.

3.2. Attributes of Stop Database

The survey form covered a wide array of data fields including volunteer and team details, local identifiers (stop names, landmarks), operational designations (route names, directions, fare stage information), physical infrastructure data (shelters, sidewalks), spatial coordinates, and visual validation through photographs along with the metadata (object IDs, creation dates). The various attributes were considered and included in survey based on requirements for analysis. The structure of the data collection form was carefully designed by the KIIFCON team, with each field purposefully included to fulfil a specific role in ensuring data quality and utility.

The key attributes collected through the form included and the rationale behind each field collected are detailed below:

- Bus Stop Name: The most commonly used/popular name of the bus stop
- Other Local Names of Bus Stop (If Any): Captures alternate names by which the stop is locally known. This is essential for standardization, as a stop might be referred by a different name across various service categories
- Landmarks nearby: Used to distinguish stops with similar names located nearby. For example, in Trivandrum, multiple stops are referred to as 'Statue'. Landmarks like 'Statue near SBI' or 'Statue near the junction' help to differentiate them.

- District: The name of the district in which the stop is located

Bus Stop Name, District in which it lies, Other Local Name, And Landmark: These fields are essential to unify and standardize the identity of each bus stop. Bus stops often have multiple names across different communities or documents. Collecting all variations ensures data completeness, improves matching with legacy systems, and aids in resolving ambiguities.

- Volunteer ID: Identifies the Permanent Employee Number (PEN) number of the staff member conducting the survey.
- Team member: Records the staff member who collected the stop information. This helps in allowance allocation and accountability of data.
- Unit: Indicates the depot name from which the staff belongs

Volunteer ID, Team member & Unit: These fields capture the identity and responsibility of the data enumerator. While the Team Member field allows recording the name of the enumerator, the Volunteer ID serves as a unique identifier, resolving cases where multiple enumerators might share the same name. This facilitates tracking the quantity and quality of data collected by each staff member.

- From and To: This attribute captures the relative position of a bus stop with respect to a line reference and helps resolve the directionality context of the stop. The enumerator enters the head node, previous bus stop, or fare stage as the ‘From’ point, and the subsequent stage or stop as the ‘To’ point. Once the geolocation is collected, the line segment defined by the ‘From’ and ‘To’ fields serves as a reference line to determine the absolute position of the stop. This approach is particularly useful because different smartphones have varying GPS accuracies, which can result in collected stop points not aligning precisely with their true ground locations. Additionally, some bus stops may have different names depending on the travel direction, or a junction may have multiple stops based on the outgoing direction. Capturing both ‘from’ and ‘to’ information helps contextualize the stop name accurately and ensures correct identification within the network.
- Direction: Indicates whether the stop lies in the ‘Up’ or ‘Down’ direction of the defined line segment. For example, if the line segment is defined by ‘From: Trivandrum’ and ‘To: Pattom’, the ‘Up’ direction might represent travel from Trivandrum to Pattom, while the ‘Down’ direction could signify the reverse (Pattom to Trivandrum). This attribute contextualizes the stop’s position relative to the route’s directional flow, resolving ambiguities caused by bidirectional naming conventions or overlapping stops at junctions.

Direction, From, and To: These fields validate the directional sense of the trip. They help in accurately capturing the segment that the bus stop services, thus helping understand the movement patterns and route flows.

- Stop of Services: Lists all service categories that halt at the stop (e.g., Ordinary, Superfast, Super Express, Super Deluxe, Low Floor, Fast Passenger).
- Whether Stop is a Fare Stage: This is a binary attribute with possible responses of "Yes" or "No", to indicate if the stop is a fare stage or is just a stop.
- Fare Stage Details: If the stop is a fare stage, this fields specifies which service categories recognize the stop as a fare stage.

Stop of Services, Whether Stop is a Fare Stage, Fare Stage Details: These fields are important for fare structuring and planning. Knowing which stops are fare stages helps validate and analyse fare collection policies and service provisions.

- Bus Stop Type: Categorizes the physical and functional nature of each bus stop. (e.g., Bus Stand, Sign Only, KSRTC Depot, Single Pole, Closed Bus Bay, Bus Bay, In-lane Bus Stop, Bypass Rider Hub, Mobility Hub, Others).
- Whether Bus Shelter Available: This is a binary attribute with possible responses of "Yes" or "No", allowing the survey to capture the presence or absence of protective infrastructure for passengers.
- Whether Sidewalk Available: This is a binary attribute with possible responses of "Yes" or "No", allowing the survey to capture the presence or absence of sidewalk for passengers.

Bus Stop Type, Whether Bus Shelter Available, Whether Sidewalk Available: These fields record the physical infrastructure at each bus stop. This information is valuable for city planners and transport departments. Availability of bus shelters influences commuter comfort, while sidewalks are critical for pedestrian and wheelchair accessibility. In particular, sidewalk design can affect persons with disabilities when heights or ramps are not friendly.

- X and Y: Geolocation data collected either manually via map pinning (in the absence of network) or automatically through the device's GPS.
- Bus Stop Change: To know if any road renovation is going on in the location and if there is a possibility to change the bus stop location in the future.
- Address: The address helps pinpoint the location of the stop

X, Y and Bus Stop Change, Address: This field identifies temporary or provisional bus stops, often a result of rerouting due to construction or road works. Tracking such changes helps maintain an up-to-date database.

- Photo: Capture the image of the bus stop for future maintenance and location identification.

Photographs serve dual purposes—ITS integration for advanced passenger information systems, and a visual reference for both planners and users to verify stops.

Metadata fields such as Object ID and Creation Date indicate the unique ID of the bus stop point data collected and the date on which the data was collected. These fields are auto generated by the application itself. As KSRTC operations are categorized zone-wise (South, North, and Central zones), the data collection form was also grouped in a similar manner. This created unique Object IDs within each zone; however, when data from various zones are combined, the Object ID field may not remain unique. The Object ID was used as the unique identifier for a stop during the initial data processing.

These attributes allowed for a detailed dataset covering both personnel and physical bus stop characteristics—ranging from staff identification to stop location, infrastructure, and operational roles. The integration of robust offline data collection capabilities proved to be a significant advancement, enabling reliable and uninterrupted data collection across diverse and challenging field conditions.

3.3. Strategy of Field Data Collection

The success of the bus stop data collection effort was highly dependent on effective field administration and strong coordination within the survey team. The implementation of the application for bus stop data collection followed a well-structured strategy, jointly led by KSRTC in collaboration with KIIFCON. It was recognized that bus conductors possess detailed knowledge of stop locations and related attributes, as they regularly traverse the same routes as part of their scheduled duties. This local expertise was leveraged for data collection both within Kerala and in neighbouring states.

Initially, data collection was proposed to be performed during regular bus service operations, with conductors collecting stop data as buses halted for passenger boarding and alighting. However, this approach would have caused delays and disrupted passenger schedules, as collecting all required attributes at each stop would significantly impact bus operation timings. To avoid such disruptions, conductors were instead asked to use their personal vehicles or departmental vehicles for data collection. When personal vehicles were used, staff were reimbursed for fuel expenses. This approach ensured that data collection was thorough and accurate, without interfering with regular bus services.

To oversee the process, a senior team of KSRTC staff was formed as the bus stop data collection supervisory team (Core Team), responsible for managing the deployment of bus conductors from various depots for field data collection. Bus conductors were strategically selected by the Core Team based on their familiarity with the route network and service duty patterns. Each staff member was assigned a specific geographic area—typically routes and stops they were already acquainted with—ensuring contextual accuracy in the data collected.

KIIFCON conducted training and knowledge transfer sessions to familiarize the Core Team with the application and the data input process. The Core Team subsequently visited each depot and trained the dedicated personnel on the smartphone application, including raising awareness about the offline data collection mode. Preliminary groundwork included listing all the bus stop names from local depot data and raising awareness among conductors. As each route was attempted for data collection, it was ensured that all the listed stops were covered. If any stops were missed during the initial listing, conductors were able to fill these data gaps in the field using their knowledge of bus stop availability.

The survey team comprised Senior Surveyors, Junior Surveyors, Depot Admins, and Field Enumerators. The field enumerators collected the data, which was initially checked by the Depot Admins for data coverage and fuel reimbursement approval. The next level of validation was carried out by the Junior Survey Team, which operated under the Core Team. The Core Team finally approved the validation of data and the completion of the data collection process within each depot's vicinity.

A total of 200 enumerators were deployed for the statewide bus stop survey. Each enumerator collected an average of 500 data points, traveling approximately 100 kilometres to map stops in both travel directions. In total, 115,695 bus stop data points were successfully collected and recorded through the mobile GIS application.

A custom dashboard was developed by KIIFCON team for monitoring and validation, with the collected data uploaded to a centralized platform in real-time or upon reconnection. This dashboard was shared with KSRTC officials and included administrative privileges for authorized personnel. Access levels were assigned based on user profiles: enumerators had view-only access to verify route coverage, while Senior and Junior Surveyors and Depot Admins were granted edit access for quality

checks and validation. The dashboard also facilitated real-time validation and correction of data, ensuring a collaborative and transparent data management process.

Figure 1 shows the dashboard elements. The classification of data collection points according to the PEN number helps to assess the areas/routes covered by each enumerator.

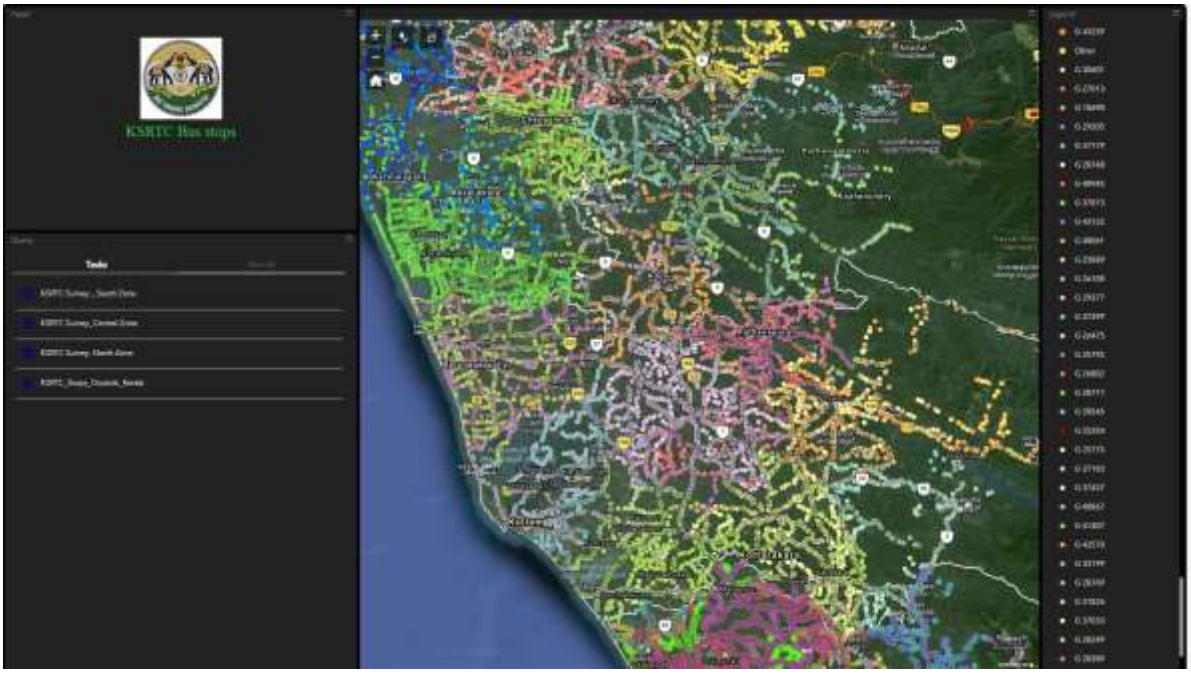

Figure 1 Dashboard depicting field data collected by different enumerators of KSRTC

Figure 2 displays the pop-up box of a selected bus stop point in the map. The attributes collected are displayed in the pop-up box in the dashboard.

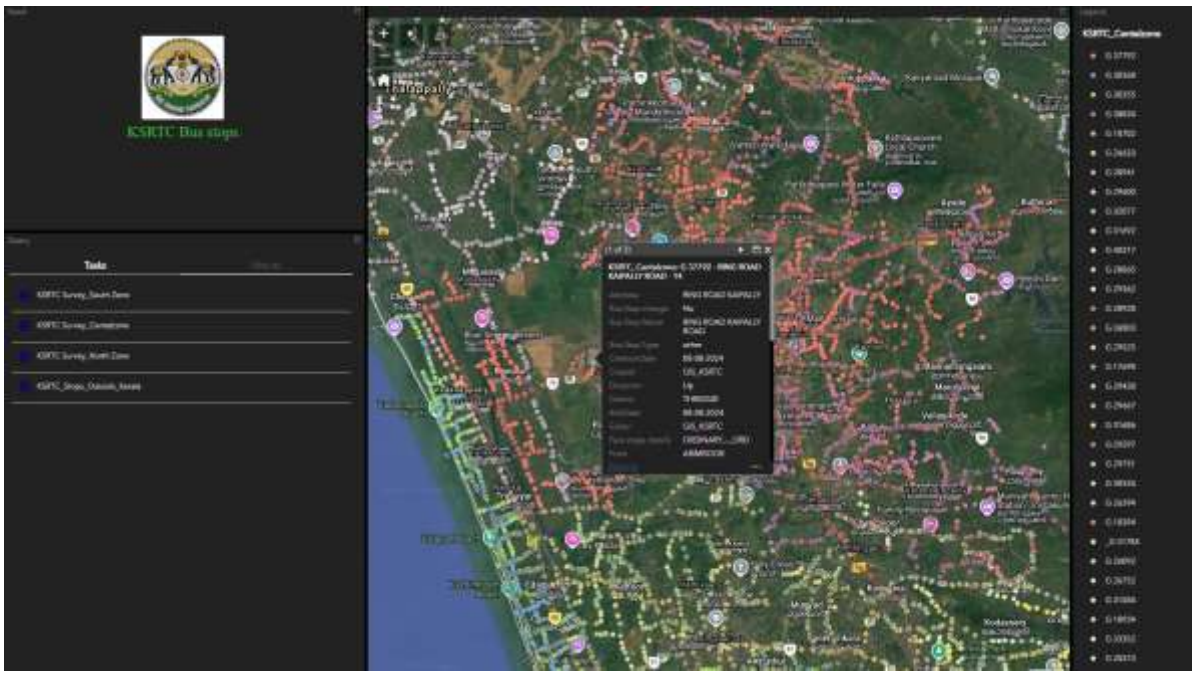

Figure 2 Dashboard displaying *the pop-up box with essential attributes of field collected bus stop data point*

3.4. Survey Deployment and Coverage Statistics

The entire process—from initial development to validation—spanned over 12 to 17 months:

- Application Identification and Development Phase: May–July 2023
- Testing of Application and Pilot Survey Phase: August–December 2023
- Field Data Collection of Bus Stops: January–June 2024
- Data Validation and Finalization: July–September 2024

This phased approach ensured a robust deployment and validation mechanism, allowing real-time feedback and continuous improvement throughout the project lifecycle. The strategic approach to implementation not only streamlined the data collection process but also ensured the quality, consistency, and usability of the data for downstream transit planning and system integration efforts.

Table 1 Summary of Bus Stop Data Collected

Sl. No.	District	Zone	No. of Bus stops
1	Alappuzha	Central	6827
2	Ernakulam	Central	8455
3	Idukki	Central	5007
4	Kannur	North	10121
5	Kasaragod	North	5960
6	Kollam	South	8904
7	Kottayam	Central	8007
8	Kozhikode	North	10414
9	Malappuram	North	7931
10	Palakkad	North	6060
11	Pathanamthitta	South	5505
12	Thiruvananthapuram	South	13773
13	Thrissur	Central	12762
14	Wayanad	North	5000
15	Outside State	Outside State	969
Total			**115695**

In addition to the points collected through field surveys, KSRTC contributed 688 geocoded locations which was manually geocoded by the KSRTC team for identifying the boarding and alighting stops for reservation ticket-enabled services. This dataset was integrated with the field-collected bus stop database to create a unified and comprehensive inventory. When points from both datasets overlapped or were located in close proximity, geospatial clustering methods were employed to identify and merge them as single stops, eliminating redundancy. Furthermore, fare stage information extracted from Electronic Ticket Machine (ETM) data during the route rationalization project was also incorporated into the database. By combining these sources and applying rigorous post-processing, the resulting dataset represents an exhaustive collection of all KSRTC fare stages and operational bus stops, supporting accurate mapping and analysis across the state.

Figure 3 below depicts the bus stop data collected for KSRTC bus services. The state has extensive network of bus stop data points which is visible as dense concentration of points in each district. The stop point data for services that operate outside the state are sparse when compared to stops within the state, as the data collection primarily focussed on capturing of information for fare stages in these areas. Very few data points for bus stops that are not fare stages were captured by enumerators outside the state.

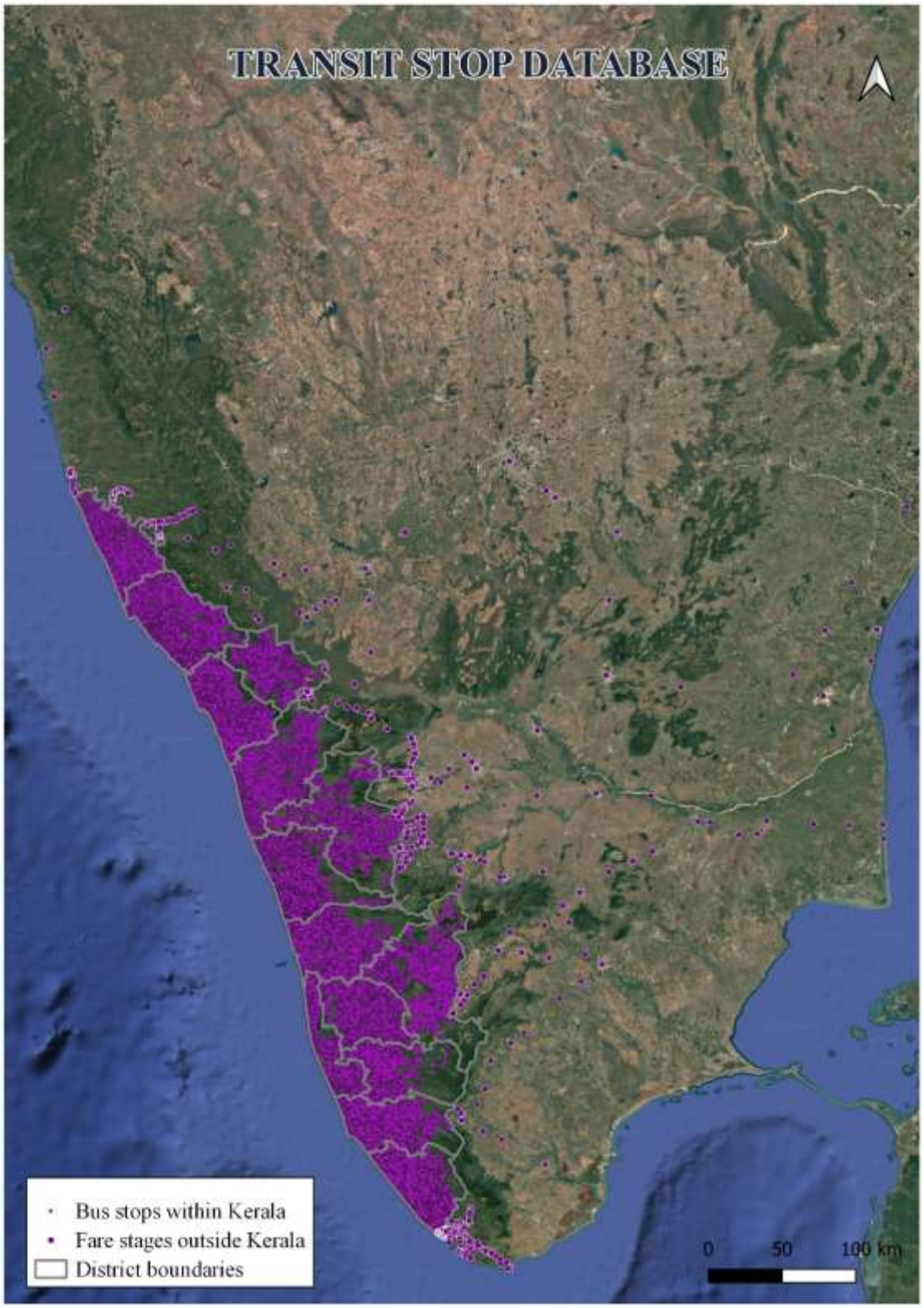

Figure 3. GIS visualization of Transit Stop Database of KSRTC

As clear from the data, Thiruvananthapuram has highest density in terms of bus stop and Wayanad has the lowest density, attributed to predominantly rural and hilly nature of the terrain with lower population density and fewer roads compared to coastal and urban districts.

For better visualization and to depict the density of bus stop points, the map layout of bus stop data of Thiruvananthapuram district is shown in *Figure 4*. The areas with lesser bus stops are forest areas with lesser road density and hilly terrain. The district has good coverage of bus stop points indicative of better coverage of transit services.

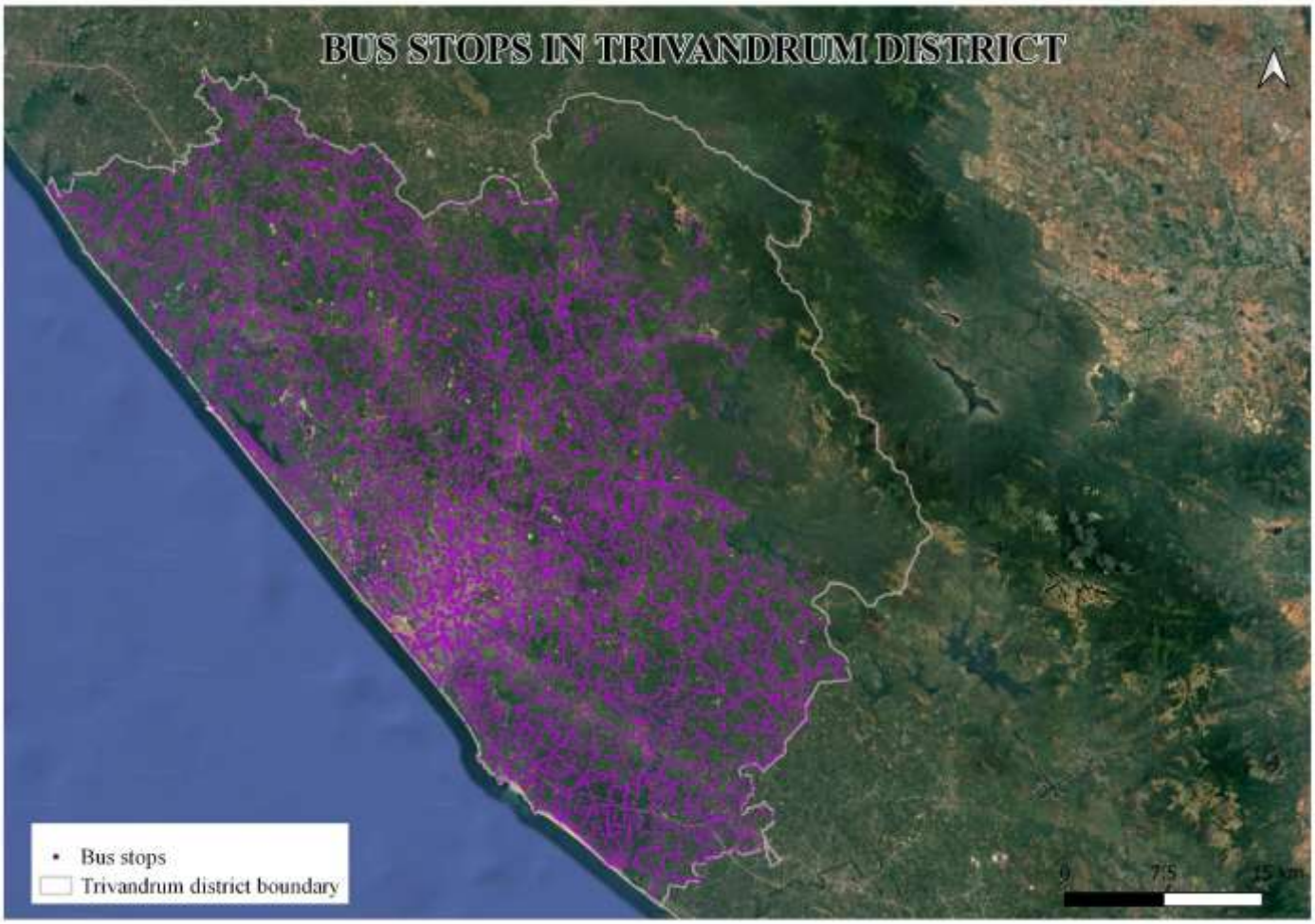

Figure 4 GIS Visualization for Transit Stops in Trivandrum District

4. POST DATA PROCESSING

4.1. Quality Check of Data

The field data collected for bus stops underwent an initial data quality assessment. Enumerators used a variety of smartphone models and network service providers, resulting in significant variation in GPS accuracy. This variation often caused discrepancies in the geolocation of bus stop points, with many points failing to align precisely with the road network. As a result, clusters of stop points were observed, particularly for stops located on opposite sides of the road. Directionality and positional

accuracy were sometimes ambiguous, despite the inclusion of "from" and "to" segment information in the survey form. Inconsistent and incorrect spelling of location names by enumerators introduced further variability into the dataset. Additionally, errors were noted in district information entered by enumerators. There were also instances where multiple bus stop points at different legs of intersections were assigned the same name, leading to naming ambiguities. Photographic documentation was also incomplete: out of 115,695 points collected, only 109,903 had associated images. The absence of photographs was primarily due to limitations in phone memory of enumerators, concerns about camera quality, or apprehension about data charges when uploading images. For bus stops outside the state, only limited field data was collected due to the logistical challenges of long-distance travel, resulting in a focus on fare stage information rather than comprehensive stop data. During the data quality assessment, certain data points were also found to be erroneously located in the sea due to incorrect capturing of geo-coordinates.

To address these data quality issues, the following steps were adopted:

- The "from" and "to" segment information was used to clarify directionality and correct the positioning of bus stop points relative to the road network.
- Spelling inconsistencies and errors in location names were systematically corrected using automated tools such as EZGeocode and fuzzy string-matching techniques (detailed in section below), resulting in standardized naming.
- For points lacking photographs, available images were reviewed and retained, while missing images were documented as data limitations.
- Incorrect district details were corrected through geospatial analysis, ensuring accurate assignment of each bus stop to its proper administrative region.
- Naming ambiguities at intersections were resolved by leveraging "from" and "to" information to differentiate stops at different legs.
- For bus stops outside the state, fare stage information was supplemented with data from existing databases to ensure adequate coverage.
- Points with geo-coordinates falling outside the defined terrestrial boundaries were flagged and reported back to the respective enumerators for correction of their geographic locations.

4.2. Creation of Fare Stage Database

Following the field data collection, the consolidated bus stop database underwent a comprehensive post-processing phase, with a particular focus on creating a comprehensive database for Fare stages by integrating fare stage information derived from Electronic Ticket Machine (ETM) data. KIIFCON leveraged ETM dataset analysis carried out as part of the route rationalization project to enhance the accuracy and utility of the bus stop data for KSRTC’s operational needs.

The post-processing of bus stop data began with the extraction of fare stage points from the field-collected dataset. This was accomplished by filtering survey responses using the 'fare stage or not' attribute in the data collection application, thereby isolating those stops explicitly marked as fare stages by enumerators. Out of 115,695 points collected from the field, 24,713 were marked as fare stages for at least one category of service by the enumerators.

It was common for a single fare stage location to be recorded as multiple points—typically at least two, one for each direction of travel. At intersections, this number could increase to four, reflecting stops at each leg of the intersection. To address this multiplicity and to ensure spatial consistency, a clustering method was applied to group fare stage points that were geographically co-located. The centroid of each cluster was designated as the representative fare stage point and assigned the corresponding bus stop name.

Clustering was performed exclusively for fare stage points to accurately consolidate the locations where fare changes occur. The underlying bus stop points themselves were not aggregated; instead, the cluster membership information for each bus stop is retained in the database. This means that for every bus stop, exactly which fare stage cluster it belongs to is stored.

In cases where the fare stage points were visibly different in different direction of travel was kept as such as per the data collected from the field. This structure allows for flexible integration with the trip database: when mapping trip data, each stop can be traced back to its fare stage cluster, facilitating accurate fare calculation and route analysis. By storing the relationship between bus stops and their respective clusters, the system ensures that future analyses or integrations—such as linking with trip databases—can utilize this cluster information to identify all stops associated with a fare stage, or to map fare stage transitions along a route.

In the post-processing stage, the next step was to assign a unique identification code to each extracted fare stage centroid. Assigning unique IDs was essential for ensuring consistency, system-wide uniqueness, and easy tracking of fare stages—especially since the same fare stage could be used in multiple trips or appear in different localities. This step was also necessary to comply with GTFS (General Transit Feed Specification) standards and to facilitate future integration with ITS-based systems.

Initially, a simple format was considered, where each fare stage ID would start with the first four letters of the district name, followed by a serial number (for example, THIR_1 for Thiruvananthapuram). This approach aimed to make it easy to identify the district from the ID. However, as the dataset expanded to include services spanning multiple districts and even inter-state routes, it became clear that this method could lead to confusion and duplication. For instance, multiple districts such as Thiruvananthapuram (Kerala) and Thiruvarur (Tamil Nadu) could share the same initial four letters ("THIR").

Additionally, continuous serial numbering was not practical as the number of fare stages exceeded 5,000, making logical grouping and scalability difficult. To address these issues, the ID format was redefined by appending the Regional Transport Office (RTO) code associated with each district. This ensured that the prefix was both district-specific and unique. RTO numbers were assigned only to fare stages within Kerala. For fare stages outside Kerala, the district names were retained, and a separate sequence of serial numbers was used, starting from the number where Kerala’s RTO-based codes ended.

For example, a fare stage in Thiruvananthapuram (RTO code 01) would have an ID like THIR01_001, while a fare stage in Thiruvarur, Tamil Nadu (assigned code 43 for differentiation), would be THIR43_001. In both cases, "THIR" indicates the district, and the appended code distinguishes between districts with similar abbreviations.

The final structure of the fare stage code is:

<district_code><RTO_number>_<serial_number>

This system provided several key benefits:

- District identification directly from the fare stage code
- Avoidance of code duplication across districts and states
- Scalable numbering for future additions

By combining district abbreviations with RTO codes and numeric suffixes, a robust and scalable ID system was implemented, ensuring that every fare stage across Kerala and beyond had a globally unique and easily traceable identifier.

Ensuring the completeness of fare stage database involved extracting stage names and their sequence from the ETM database, which records the progression and direction of trips through various fare stages. However, the information was only textual and not geocoded information. To standardize the fare stage names in the ETM data, KIIFCON employed a combination of automated and manual methods. String matching based on the Levenshtein distance was used to identify minor spelling variations and phonetically similar names. Geocoding APIs, such as Nominatim, were queried using all available name variants. These services, especially when provided with contextual information like "Kerala" or the relevant district, often returned the correctly spelled place name along with associated geographic metadata. The ezGeocode tool further supported this process by efficiently handling batch queries and matching input strings to known place names, enhancing the reliability of automated corrections.

Despite these automated approaches, certain localities and minor bus stops—particularly those not well-represented in global databases like OpenStreetMap—did not yield satisfactory matches. To address these gaps, KIIFCON supplemented the automated results with a manually curated correction dictionary and applied additional fuzzy string-matching techniques against a verified list of place names. After identifying all spelling variations, the correct spelling was established for each fare stage in the ETM database, which then served as the reference dataset for further integration.

The fare stage information identified from the ETM was subsequently overlayed against the field collected bus stop data. However, inconsistencies in naming across bus stop database and fare stage database —stemming from manual entry errors, abbreviations, or regional language variations—posed significant challenges. The field collected dataset comprised 115,695 bus stop entries, many of which contained inconsistently spelled names. Accurate identification and consolidation of stop names were critical, as fare stage information is fundamental for mapping, fare calculation, and further analysis. A custom Python program was developed to perform both exact and approximate string matching, with the Levenshtein distance metric at its core. This solution combined two approaches:

- Exact dictionary matching: Each input name was compared against a predefined dictionary of known fare stage names identified from Electronic Ticket Machine data. If a perfect match was found, the corresponding standardized name was assigned.
- Fuzzy string matching: For entries without an exact match, the fuzzywuzzy library was used to calculate similarity scores between the input name and reference list values. If the score exceeded a defined threshold, the closest standardized name was selected.

The matching process was critical for ensuring that all operational fare stages were accurately represented and geolocated in the consolidated database. As a result of this process, a total of 6,847 centroid points of bus stop clusters were identified as fare stages for KSRTC. These fare stages are applicable to various service categories—including ordinary, fast passenger, superfast, super express, and higher classes—that operated at least 15 days in a month.

The validation approach adopted ensured that fare stages present in either dataset—whether field-collected or ETM-derived—were reconciled, with any unmatched stages flagged for further review. The final step involved a comprehensive review to confirm that no fare stage was omitted from either source, resulting in a unified, validated, and geospatially accurate fare stage database. The consistency and reliability of fare stage data across the system is thus greatly enhanced, supporting accurate mapping, fare management, and future transit planning initiatives.

5. CONCLUSION

In summary, the comprehensive mapping and validation of KSRTC's bus stop and fare stage database represent a significant advancement in Kerala's public transport data infrastructure. By combining robust field data collection with advanced post-processing techniques—including geospatial clustering, unique identifier assignment, and sophisticated string-matching algorithms—this project successfully overcame longstanding challenges of data inconsistency, duplication, and spatial inaccuracy. The integration of Electronic Ticket Machine (ETM) data, supported by both automated and manual correction methods, ensured that the final database is not only exhaustive but also highly reliable for operational, planning, and analytical purposes. This unified, geocoded, and standardized dataset now forms a critical foundation for modernizing transit operations, enabling data-driven decision-making, and supporting future integration with intelligent transport systems and digital platforms.

This project stands as a testament to KIIFCON Pvt. Ltd.'s commitment to delivering innovative, accurate, and actionable solutions in transport data management. Through meticulous fieldwork, advanced geospatial analytics, and intelligent data integration, KIIFCON team addressed the complex challenges of standardizing and validating KSRTC's bus stop and fare stage data. By leveraging cutting-edge tools, custom algorithms, and a deep understanding of both local context and global best practices, KIIFCON ensured that the final dataset is robust, scalable, and future-ready.

KIIFCON's approach—marked by technical rigor, transparency, and collaborative problem-solving—demonstrates the organization's capacity to handle large-scale, mission-critical projects with clarity and efficiency. The proven experience in integrating multiple data sources, resolving inconsistencies, and delivering GIS-powered solutions positions KIIFCON Pvt. Ltd. as a trusted partner for organizations seeking to modernize their transport infrastructure and embrace digital transformation.

ACKNOWLEDGEMENTS

The journey of this project has been marked by vision, collaboration, and unwavering commitment. We extend our heartfelt gratitude to Dr. K.M. Abraham, Chairman and CEO of KIIFCON Pvt. Ltd., whose visionary leadership and unwavering commitment enabled this project to move forward despite several constraints. His dedication to empowering KSRTC and advancing the state's public transport

infrastructure exemplifies the spirit of service and technical excellence that made this initiative possible, offered as a gesture of partnership and support to KSRTC's mission.

We sincerely thank Dr. Vijayadas S. J., Director, Projects and Engineering, for his unwavering support throughout the duration of the project. We express our deepest gratitude to Mr. Murali L.S., the Deputy Chief Consultant whose extensive experience and profound technical expertise were the cornerstone of this project. His exceptional technical acumen and steadfast guidance were pivotal in advancing the project and ensuring the utmost integrity of the work undertaken. Our thanks extend to Ms. Smitha R. Prasad, Senior Consultant, for her invaluable support. The authors further acknowledge the invaluable contributions of the KIIFCON Route Rationalization Team—Mr. Siddhartha V. (Project Consultant), Ms. Varsha Vijay (Technical Assistant Trainee), Ms. Grace Sophia S. (Technical Assistant), and Ms. Reshma Babu (Technical Assistant-GIS)—whose expertise and tireless efforts during the data processing stage were pivotal to the project's success.

We now express our deepest gratitude to KSRTC for their enthusiastic collaboration and support. KIIFCON Pvt. Ltd. is especially thankful to Mr. Biju Prabhakar, IAS, the then Managing Director of KSRTC, whose scientific temper and dedication to organizational excellence gave life to the bus stop data collection initiative. His belief in the transformative power of data and decisive leadership in constituting the KSRTC field data collection team laid the foundation for the project's success.

We are equally thankful to Mr. Pramoj Shankar, IOFS, who, as the succeeding Managing Director, continued to champion this initiative. His steadfast support ensured the project's momentum and completion, affirming the enduring value of this endeavour even amidst leadership transitions. Our thanks also go to Mr. Pradeep Kumar, Executive Director (Operations) of KSRTC, whose facilitation of resources was instrumental in turning vision into action. We further thank Mr. Joshua Bennett John, K.A.S., the then General Manager (IT) of KSRTC, and Mr. Nishanth Sudhakaran, Deputy General Manager (IT), KSRTC, for their valuable support during the data collection phase.

The seamless orchestration of field data collection, guided by Mr. Prashanth V., then KSRTC State RTA Cell Coordinator, ensured that the efforts of the KSRTC team were harmonious and effective. Mr. Prashanth V. also led the data validation team, which included Mr. Rejeesh P.R. and Mr. Prabhu V.R., bus conductors of KSRTC, who meticulously oversaw the coverage of routes and transit stops across the state. Finally, we sincerely thank the field surveyors (bus conductors of KSRTC) who diligently collected the data in the field, ensuring the success of this project.

To all who contributed their time, knowledge, and passion—their collaboration transformed a vision into a robust and enduring foundation for Kerala's public transport future. Without their support, this achievement would not have been possible.