IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 14 December 2025, accepted 23 January 2026, date of publication 3 February 2026, date of current version 9 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3660926

RESEARCH ARTICLE

Automatic Generation of Open Data-Based Traffic Simulation Model

HYUNYOUNG RYU[1], JITAEK LIM[2], AND MINSEOK SONG[1,2], (Member, IEEE)
[1]Future Open City Innovation Center, POSTECH, Pohang, Gyeongbuk 37673, Republic of Korea
[2]Department of Industrial and Management Engineering, POSTECH, Pohang, Gyeongbuk 37673, Republic of Korea

Corresponding author: Minseok Song (mssong@postech.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korean Government (Ministry of Science and ICT) under Grant 2022R1A4A101907113 and Grant RS-2024-00357330.

ABSTRACT Open transportation data is becoming more accessible, creating new opportunities for data-driven traffic simulation. However, transforming this data into a simulation-ready model is complex and time-consuming. This study introduces an automated framework to generate city-scale traffic simulation models using open data, including road networks, traffic signals, and vehicle volumes. Built on the Simulation of Urban Mobility (SUMO), the pipeline streamlines data processing and integrates real-world inputs into simulation components. The framework combines multiple open datasets through hybrid data synthesis, ensuring consistent road–signal coordination and enhancing traffic-flow realism and spatial resolution. Applied to a real urban center, the resulting model showed strong alignment with observed traffic patterns, supporting validity. As data collection, real-time monitoring, and spatial resolution advance, the framework's applicability and precision will improve. By enabling efficient and reproducible model generation, this approach contributes to scalable traffic simulation and lays the groundwork for urban digital twin applications.

INDEX TERMS Automation, data processing, open data, SUMO, traffic simulation.

I. INTRODUCTION

Traffic simulation plays a central role in improving traffic efficiency in transportation planning and management [1]. It is used to optimize traffic through signal control and to forecast traffic demand, and it enables the evaluation of alternative designs, the testing of policy scenarios, and the ex ante analysis of potential infrastructure impacts. A simulation model consists of core elements such as roads, signals, and vehicles, and its performance depends on the accuracy of the input data as well as the suitability of the model structure and configuration. Against this backdrop, the efficient design and rapid generation of accurate simulation models are essential for providing reliable decision support.

At the same time, the volume and complexity of traffic data used in simulation are steadily increasing [2]. As data-driven simulations that rely on observed data become more prevalent, the integration between simulation models and data-processing workflows is becoming increasingly important. Using empirically collected data in simulations enhances the credibility of the results, but constructing large-scale models for complex and heterogeneous urban road networks requires extensive deployment of advanced sensors and large-scale data collection. Moreover, substantial time and effort are still required for manual data processing, underscoring the need for technical tools that can process such data efficiently and generate simulation models rapidly.

In response to these needs, a variety of technical tools and systems have been proposed to support data-driven traffic simulation in practice, including workflows for integrating heterogeneous traffic datasets into simulation platforms and web applications for managing and visualizing large-scale simulation outputs. For example, Bautista et al. [3] proposed a SUMO-based traffic mobility generation tool that allows users to specify road networks, traffic control, vehicle types, and routes, and to collect performance indicators for evaluating vehicular networks. Xu et al. [4] presented a web-based interactive application that manages, stores, and visualizes

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyan Zhang.

results from multiple runs of large-scale traffic simulations, enabling users to compare scenarios using time-series graphs, spatial distributions, and summary indicators. In parallel, automated traffic-operation frameworks have been developed in the context of digital twins, which automatically generate various intersection geometries from geometric data and use them for real-time traffic management. For instance, Qin et al. [5] developed AUTOSIM, an online urban traffic-operation simulation system that automatically configures models for heterogeneous intersection layouts and calibrates model parameters using meta-learning with real-time traffic information. However, these approaches typically depend on the sensor infrastructure and data conditions of specific regions, which limits their scalability and reproducibility in other contexts.

One promising direction for alleviating these limitations is the use of open-source platforms and open data for generating traffic simulation models. By leveraging open government data and open-source tools, it is possible to reduce reliance on proprietary software and exclusive data resources, while building simulations more quickly for a wider range of cities and networks and enhancing reproducibility. This not only enables faster and more flexible use of simulation, but also allows scenarios to be tested across diverse networks and conditions without being constrained by specific vendor systems [6]. In particular, many studies have employed the open-source traffic simulation platform Simulation of Urban Mobility (SUMO) together with road networks derived from OpenStreetMap (OSM) [7]. There are also documented cases in which public data have been used to construct real-world, city-scale scenarios [8], [9], [10]. While these studies have substantially expanded the potential of open-source-based traffic simulation, systematic approaches remain scarce for efficiently combining diverse datasets and transforming them into complete, ready-to-run models. To fully exploit the accessibility and generality of open data, there is thus a need for frameworks that can efficiently process and integrate heterogeneous open datasets and flexibly and rapidly generate models applicable to diverse urban environments.

This study addresses this gap by proposing a data-processing workflow that automatically generates the key components required for city-scale traffic simulation using open data. Multiple urban datasets—including OSM-based road geometry, standardized signal data, national-level node and link records, external traffic counts, and building-level energy-use statistics as proxy indicators of internal building activity—are integrated within a single workflow to construct multi-intersection city-scale SUMO models, and the resulting traffic flows are quantitatively compared against observed data for validation. To this end, we develop a framework that automatically transforms these diverse open datasets into simulation input formats for SUMO, enabling flexible city-scale model generation that includes multiple intersections and supports data-driven traffic analysis and emerging urban digital-twin applications. The primary contribution of this work lies in providing a reproducible, end-to-end data-to-SUMO workflow that systematically integrates heterogeneous open datasets into executable city-scale traffic simulation models.

The remainder of this paper is organized as follows. We first review related work and present the overall methodology, including data collection, processing, and validation. We then apply the proposed automated procedure to real urban areas to generate and validate simulation models and discuss the results and their main implications. Finally, we summarize the limitations of this study and outline directions for future research.

II. RELATED WORK

Automatic simulation model generation has been widely studied as a means of maintaining consistency between complex real-world systems and their models [11]. In operations research and manufacturing, for example, approaches have been proposed that derive process structures and routings from layout data, equipment configurations, and other structured production data, and then automatically construct or update simulation models based on these inputs [12], [13], [14]. In this context, automated data-processing pipelines function as a key mechanism for transforming raw data into up-to-date simulation models, reducing the burden of manual model construction and enabling continuous monitoring and data-driven operational decision making.

In the transportation domain, similar concepts of automatic simulation-model generation have been applied to build models from real traffic data and network information. Several research strands have focused on automatically generating individual simulation components such as road networks and travel demand. On the network side, methods based on OpenStreetMap and other road data have been proposed to automatically construct large-scale, simulation-ready networks. Meng et al. [15] developed a topology-preserving simplification approach that cleans and reduces OSM-derived networks while maintaining routing-relevant structure for large-scale SUMO simulations. Zhang et al. [16] propose a data-based framework that fuses OSM, GIS, and operational traffic data via nearest-neighbor matching and dynamic network editing to automatically generate and update multi-modal microscopic road networks.

On the demand and traffic-flow side, several data-driven approaches have been proposed to generate or calibrate large-scale SUMO demand using real-world traffic measurements. Ma et al. [17] adopt a strategy based on population distribution, using publicly available population data with SUMO's ActivityGen to generate demand for an urban case area and comparing simulated flows with observed conditions, showing that population-based demand generation can reasonably approximate real-world traffic. In contrast, Qiu et al. [19] introduce a machine-learning framework that trains models on detector data to predict link-level flow rates and uses these predictions to generate

SUMO demand, demonstrating that ML-based demand reproduces observed flows more accurately than random-trip or simple replay baselines. Codeca et al. [18] construct the Luxembourg SUMO Traffic (LuST) scenario by deriving 24-hour, city-scale traffic demand from detector and signal data and building a corresponding SUMO network, generating time-dependent OD matrices and link flows from real-world measurements and validating the resulting demand by comparing simulated and observed volumes and daily traffic patterns. Although these approaches are not fully automated across the entire workflow, they collectively show how traffic scenarios can be systematically constructed and calibrated from empirical data.

Furthermore, research on transportation digital twins (DT) points toward increasingly automated, data-driven pipelines that can continuously update simulation models and support iterative experimentation. In particular, recent studies [20], [21] propose layered system architectures in which sensor, vehicle, and infrastructure data are collected and standardized, then passed through automated pipelines for storage, model updating, and decision support, with a dedicated data-processing layer for data quality control and heterogeneous data fusion highlighted as critical. In line with this, Kušić et al. [24] implement a pipeline that automatically collects and aggregates continuous traffic-count streams from roadside detectors and injects them into a SUMO-based digital twin network to keep the simulated traffic state synchronized with the field. While these studies highlight the operational value of digital twin systems, the challenges associated with initial data harmonization, preprocessing, and model construction from heterogeneous open datasets remain relatively underexplored.

In summary, while data-driven automation improves the scalability and reproducibility of traffic simulation, prior work has largely focused on component-level automation and digital twin frameworks, with comparatively limited attention to data-integration challenges posed by heterogeneous open datasets. Accordingly, this study shows how such datasets can be cleaned, aligned, and structured within a unified pipeline and transformed into executable simulation models, thereby bridging the gap between data availability and runnable simulations.

III. METHOD AND MATERIALS

A. OVERALL FRAMEWORK

The overall framework for automatic generation of an open-data–based traffic simulation model is shown in Fig. 1, comprising three steps: open-data acquisition, data processing, and simulation model generation. We collect relevant datasets and assess their usability by format, coverage, and quality. Each dataset is processed through a reproducible pipeline into simulation-ready components. These components are integrated to build the model and run simulations. Then, the effectiveness of the generated model is evaluated by validating simulation outputs.

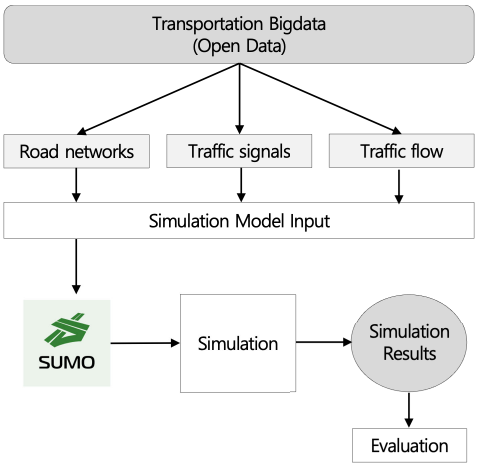


FIGURE 1. Framework for automated open data processing in traffic simulation modeling.

Simulations are implemented in SUMO, an open-source microscopic traffic simulator developed by the German Aerospace Center and continuously extended with diverse tools [25]. Assuming city-level multi-intersection signal optimization as the primary objective, the simulation model consists of three components—road network, traffic signal, and traffic flow.

For the road network, we primarily define edges (or links) and junctions (or nodes), with multiple lanes inside each edge that generate connections. Each connection is equipped with dedicated traffic signals, and each direction of signal involves a specific time phase. Additionally, traffic is represented by routes as ordered sequences of edges and by origin–destination volumes that generate vehicles; each vehicle is characterized by a departure lane, arrival lane, departure time, and route. The simulator then updates its time-varying state (position and speed). We derive these components from open data and integrate them into a unified simulation model.

B. DATA COLLECTION

The data used in this study include road network, signal control, and traffic volume; detailed sources are summarized in Table 1.

Road geometry data is sourced from OpenStreetMap (OSM) and augmented with Standard Node-Link (STNL) data. OSM provides detailed geometry and attributes and serves as the primary source, supplied as static XML covering roads, lanes, and intersections. STNL represents the network as simplified nodes and links but records detailed intersection-level attributes for roadway asset management. Provided by the Ministry of Land, Infrastructure and Transport, it covers nearly all roadway assets within the administrative boundaries of the Republic of Korea.

Traffic signal data are administered by the National Police Agency and are available upon request through

TABLE 1. Simulation input data sources.

Data type	Device	Spatial	Tempora	Format	Provider
Road geometry	Map	Road, lanes, intersections	Static	XML	OpenStreetMap [26]
Node, Link	Road management system	Node, Link	Static	SHP	National Transport Information Center [27]
Signals, Phases	Signal system	Intersection	Static	CSV	National Police Agency [28]
Traffic count	Traffic monitoring CCTV	Camera location	1 hour	CSV	Municipal Government (Pohang UTIS) [29]
Building energy	Statistical data	Building location	Daily	CSV	National Construction Data Open System [30]

the Korean government open-data portal, where they are provided in CSV format. The signal dataset comprises per-intersection controller records specifying intersection location and approach directions, signal-group IDs, phase sequence, phase splits (green, yellow, all-red), cycle length, offset, and time-of-day schedules, along with controller metadata on type and applicability. We implemented the main program at each intersection in the SUMO model, and stored all other programs in a time-of-day database.

Traffic volumes for OD matrix construction were derived from two sources: observed traffic counts and trip volumes estimated from building energy use. Traffic count data are provided by municipal governments as hourly, daily, or monthly aggregates. Although collection relies on real-time aggregation of CCTV images and can support live reporting, the publicly released open data are provided at hourly resolution. In addition, nationwide building-energy data including building coordinates and daily electricity consumption are obtained from the National Construction Data Open System. We assume that trip generation is proportional to building energy use, treating higher electricity consumption as a basic indicator reflecting relative levels of activity intensity when spatially allocating trips across the study area [31], [32]. Additional OD nodes are defined at building locations to fill gaps where observed counts are unavailable and to improve spatial resolution and completeness.

C. DATA STRUCTURE

Collected data are structured into a relational database and modeled using an entity relationship diagram (ERD) in Fig. 2. The ERD formalizes entities, their attributes, and the relationships between them. The schema is organized into three blocks: network topology, signal control, and travel demand. Primary keys include `edge_id`, `lane_id`, `node_id`, `conn_id`, `tls_id`, `phase_id`, `route_id`, and `veh_id`, while foreign keys enforce consistency across blocks and enable reproducible SUMO input generation.

The network topology centers on Edge, Lane, Junction, and Connection. An Edge entity represents the road segments

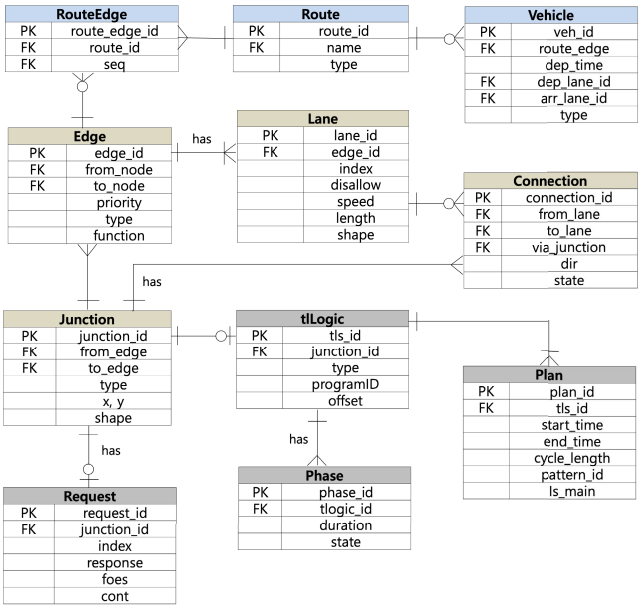


FIGURE 2. ERD for SUMO.

within the network and is connected to multiple Lane entities, which indicate the lanes that compose each road segment. A Junction entity represents intersections within the network and is connected to incoming and outgoing Edge entities. A Connection entity represents the links between different Edges and Lanes, ensuring continuity within the network. It is linked to Edge and Lane entities, detailing the connections between road segments and lanes.

Signal control is modeled with `tlLogic` and `Phase` entities attached to a `Junction`. A `tlLogic` entity is composed of multiple `Phase` entities, which specify the sequence and timing of traffic light phases. The `Request` entity captures right-of-way relations at a `Junction` (yield and priority among conflicting `Connection` entities). Time-of-day operation is stored as a plan linked to `tls_id` with start and end times, cycle length, and pattern identifiers, and the main plan is flagged for simulation runs.

Travel demand is modeled with `Route` and `Vehicle` entities. A `Route` is an ordered list of edges recorded in a child table (e.g., `route_edge(route_id, seq, edge_id)`), and many `Vehicle` records can reference the same route. Each `Vehicle` stores type, departure time, departure and arrival lanes, and its `route_id`.

This ERD binds topology (`Lane–Edge–Junction`), control (`Connection–tlLogic–Phase–Request`), and demand (`Vehicle–Route`) at the lane level. It preserves referential integrity from data ingestion through the creation of SUMO input files and supports efficient validation and regeneration of inputs.

D. DATA PROCESSING

The overall workflow is illustrated in Fig. 3 and comprises two stages: data preprocessing and simulation creation.

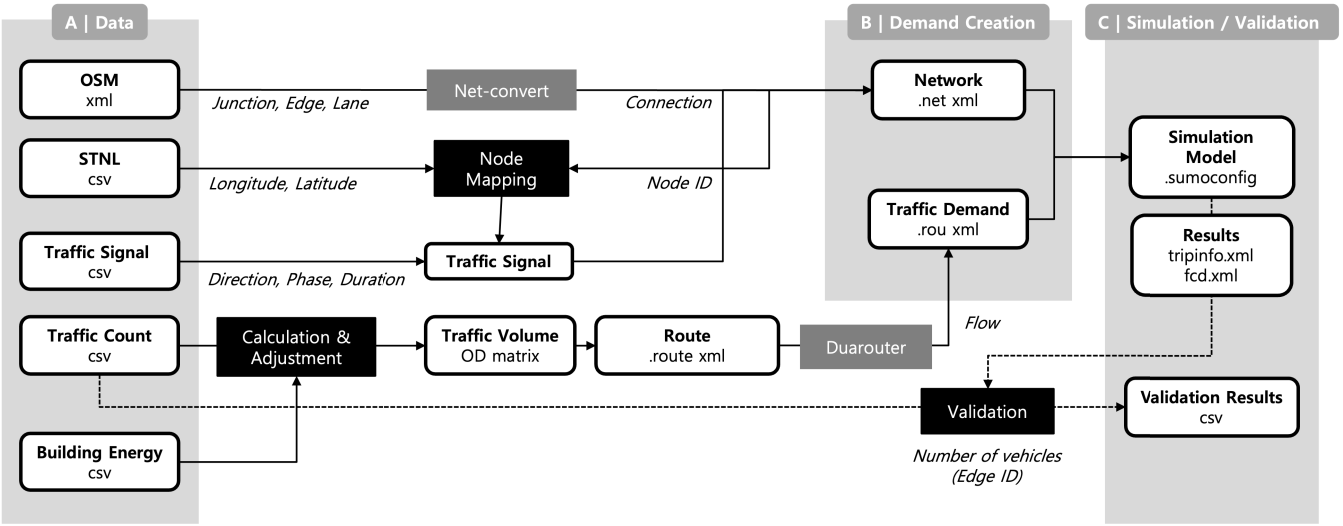


FIGURE 3. Data processing workflow.

In preprocessing (A to B), the road network is built from OSM and aligned with STNL by nearest-neighbor matching of nodes to create integrated node IDs; signal records are then linked to the corresponding STNL node IDs to complete the network. Then, the OD matrix is constructed by combining observed traffic counts with trip generation estimated from building energy use: external in/out and internal OD points are defined at data-collection locations, observed volumes are distributed to in/out nodes, and internal demand is allocated using building energy weights, followed by calibration against observed counts. OD pairs are assigned shortest-path routes on the finalized network, and SUMO's Duarouter generates vehicle routes from the OD volumes.

In the simulation creation stage (B to C), the network and demand produced above are exported as SUMO inputs, network (`.net.xml`) with signal information and routes (`.rou.xml`) containing vehicle information, and compiled in a configuration file (`.sumocfg`) to run the model. The simulation outputs two logs: TripInfo (`tripinfo.xml`), which records per-vehicle summaries, and Floating Car Data (`fcd.xml`), which provides time-stamped trajectories of position and speed at fixed intervals. We aggregate these outputs to edge-level measures for validation against observations and for visualization.

1) ROAD NETWORK

- Obtain the base road map from OSM and convert it to SUMO format using `netconvert`, with default setting, generating edges, junctions, and lanes.
- Filter OSM edges using a custom SUMO edge type to retain only road types used for signalized vehicular traffic (highway, motorway, trunk, primary, secondary, tertiary) and exclude residential, pedestrian, and other non-motorized links.
- Extract node IDs and coordinates from the OSM network $\mathcal{N}^{\text{OSM}} = \{(o_i, \mathrm{x}_{o_i})\}$ and the STNL nodes $\mathcal{N}^{\text{STNL}} = \{(s_k, \mathrm{x}_{s_k})\}$.
- For each STNL node s_k, find the closest OSM node within a search radius $d_{\max}$ using geodesic distance. The distance threshold $d_{\max}$ is set based on roadway width ($3 - 3.5$ m) and lane configuration at signalized intersections.
- Treat nearby OSM nodes as a multi-node representation of the same STNL node when their distances are sufficiently similar.
- The output is a mapping $\mathcal{M} = \{(s_k, \{o^*_{k,1}, \ldots, o^*_{k,m_k}\})\}$ that links each STNL node ID to one or more matched OSM node IDs.

2) TRAFFIC SIGNAL

- From the road network, derive all movements (connections) between incoming and outgoing edges at each intersection.
- Raw signal records are given by tuples (ϕ, d, τ), where phase index ϕ, direction index d (clockwise from true north), and duration τ (s) are specified:

$$\mathcal{S}_{\text{raw}} = \{(\phi, d, \tau)\}.$$

- For each phase–direction pair, assign a signal state $\sigma(\phi, d) \in \{G, g, y, r\}$, where G: priority green, g: permitted green, y: yellow, r: red.
- Movements are classified as $m \in \{t, r, s, L\}$ (U-turn, right, straight, left), and U-turn/right movements are mapped to non-priority green by default.
- For each signalized junction j, construct a time-ordered sequence of state strings over all connections $c \in \mathcal{C}_j$:

$$6_j = \left(\sigma_{\phi_1}(c_1) \ldots \sigma_{\phi_1}(c_{|\mathcal{C}_j|}), \ldots, \sigma_{\phi_P}(c_1) \ldots \sigma_{\phi_P}(c_{|\mathcal{C}_j|})\right).$$

and write this as `tlLogic` in the SUMO network.

- Add `tl` and `linkIndex` attributes to each `connection` so that every movement is linked to its corresponding signal logic.

3) TRAFFIC FLOW

Traffic flow is constructed in two stages: OD matrix generation from observed traffic count data, followed by traffic flow generation via network-based path assignment.

a: OD GENERATION

- Let V_i^{IN} and V_j^{OUT} denote observed vehicle counts at IN and OUT gateway edges, respectively. The total inbound and outbound volumes are

$$V_{\mathrm{tot}}^{\mathrm{IN}} = \sum_i V_i^{\mathrm{IN}}, \qquad V_{\mathrm{tot}}^{\mathrm{OUT}} = \sum_j V_j^{\mathrm{OUT}}.$$

- Define internal edges $\mathcal{E}^{\mathrm{NO}}$ from building energy data by selecting non-industrial buildings with $E_z \geq 0.001\,E_{\mathrm{tot}}$ and located more than 500 m from any IN/OUT node, so that internal destinations represent typical daily activity locations, keep the simulation computationally efficient, and avoid excessively short OD paths.
- For each IN edge i, split flows into through and internal trips using fixed proportions $\alpha_{\mathrm{through}}$ and α_{int} (e.g., 0.6 and 0.4 for commuting; heuristically set based on local knowledge):

$$V_i^{\mathrm{through}} = \alpha_{\mathrm{through}} V_i^{\mathrm{IN}}, \qquad V_i^{\mathrm{intIN}} = \alpha_{\mathrm{int}} V_i^{\mathrm{IN}}.$$

- Distribute V_i^{through} from IN edge i to OUT edges j (case a) according to weights w_j^{OUT} (e.g., proportional to V_j^{OUT}), and distribute V_i^{intIN} to internal destination edges $k \in \mathcal{E}^{\mathrm{NO}}$ (case b) using internal weights w_k^{NO} (e.g., proportional to building energy).
- If $V_{\mathrm{tot}}^{\mathrm{OUT}} > V_{\mathrm{tot}}^{\mathrm{IN}}$, allocate the deficit $\Delta V = V_{\mathrm{tot}}^{\mathrm{OUT}} - V_{\mathrm{tot}}^{\mathrm{IN}} > 0$ as additional internal-origin–OUT-destination trips (case c), using the same weight structure. Otherwise, set $\Delta V = 0$.
- Assign internal–internal trips (case d) as a fixed share of the total gateway volume, e.g.,

$$V_{\mathrm{tot}}^{\mathrm{int}} = 0.1\big(V_{\mathrm{tot}}^{\mathrm{IN}} + V_{\mathrm{tot}}^{\mathrm{OUT}}\big),$$

  and distribute $V_{\mathrm{tot}}^{\mathrm{int}}$ over all pairs of distinct internal edges $k \neq \ell$ in $\mathcal{E}^{\mathrm{NO}}$ using weights proportional to $w_k^{\mathrm{NO}} w_\ell^{\mathrm{NO}}$.
- The final OD matrix element between origin o and destination d is

$$Q_{od} = Q_{od}^{(a)} + Q_{od}^{(b)} + Q_{od}^{(c)} + Q_{od}^{(d)},$$

  excluding all $o = d$ pairs.
- A final calibration step rescales the OD matrix Q_{od} so that the simulated inbound and outbound volumes at IN/OUT edges match V_i^{IN} and V_j^{OUT} as closely as possible.

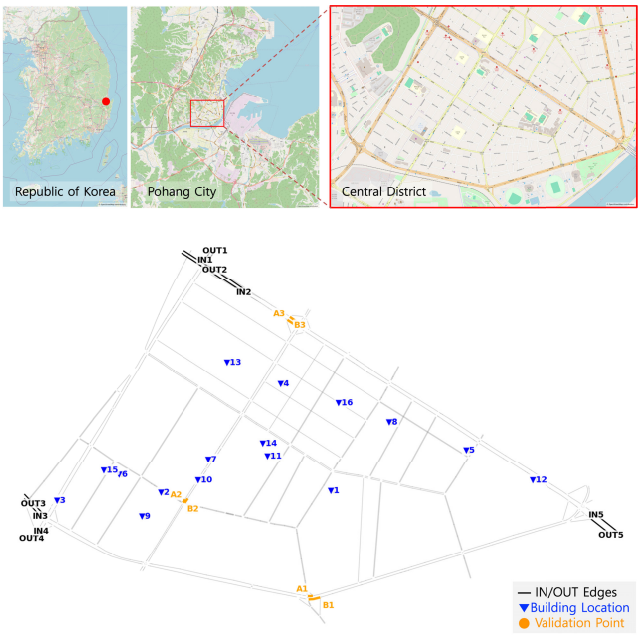


FIGURE 4. Test site: Central district of Pohang, Korea.

b: ROUTE GENERATION

- The calibrated origin–destination (OD) matrix Q_{od} is combined with the SUMO network, and each OD pair is converted into a shortest (or fastest) path using `duarouter` with default settings.
- For each vehicle, a record containing the vehicle identifier (ID), route identifier, departure time, vehicle type, departure lane, and departure speed is generated and written to the traffic demand file `.rou.xml`.
- This procedure distributes vehicles over the network according to the OD matrix while respecting the network topology and link travel times.

4) SIMULATION

The completed road network with traffic signals (`.net.xml`) and the traffic demand (`.rou.xml`) are assembled into a SUMO configuration file (`.sumocfg`) to run the simulation. The file (`.sumocfg`) lists input files, sets the simulation window and step length, and specifies outputs (e.g., `tripinfo.xml`, `fcd.xml`). For the experimental setup, simulation settings are kept at their default values, with a simulation horizon of 3600 s (1 h). Consequently, the simulation model generates traffic flows in a fully automated, end-to-end process driven by the specified hourly traffic volumes. Although input datasets differ across sites, model generation is consistently handled within the same automated framework.

E. VALIDATION

To evaluate the automated process, we validate simulation outputs against observed traffic counts. The simulation output used for validation is the edge-level vehicle volume,

defined as the total number of simulated vehicles passing each validation edge during the target hour, extracted from `tripinfo.xml`. The reference data for validation are edge-level traffic counts collected at independent count locations that are not part of the OD matrix construction. The simulation results are first visualized to examine spatial patterns of traffic flow and major congestion corridors, providing an intuitive check against observed data. Subsequently, for each validation edge, we compute an empirical range (minimum, mean, maximum) from selected days and verify whether simulated volumes fall within these bounds. Lastly, model performance is evaluated using the coefficient of determination (R^2) and mean absolute percentage error (MAPE). R^2 measures the variance explained, and MAPE reports the average percentage error. Together, these measures provide a balanced assessment of both statistical performance and practical accuracy, supporting the reliability of the proposed modeling framework.

The simulation model is tested for the city center of Pohang, Republic of Korea, which comprises a mix of industrial, commercial, and business districts. Major roads also serve as commuting corridors from residential areas to the industrial complex across the river. The data points used for the simulation in the Pohang area are illustrated in Fig. 4. For validation, traffic count data from six additional points—three directional pairs—outside the IN/OUT edges are used. A total of 140 weekday samples from September to December 2024 during the morning peak hour (7:00–8:00 AM) are analyzed.

IV. RESULTS AND DISCUSSION

Using the proposed framework, we documented how raw datasets from the study site were transformed into final simulation inputs, identifying which data elements were incorporated through the automated pipeline and where residual manual checks or interventions remained. We also evaluated how well the automatically generated simulation model reproduced observed edge-level traffic volumes at independent count locations, characterizing overall goodness-of-fit as well as remaining local discrepancies.

A. SIMULATION GENERATION RESULTS

Table 2 summarizes the process of generating the simulation model, including the key components such as the number of edges, nodes, traffic signals, and OD matrix elements incorporated from the raw data to the final model.

For road network construction, a total of 3,275 OSM edges and 11,923 OSM nodes were filtered to 718 edges and 478 nodes representing signalized arterial and collector roads, integrated as junctions and intersections within the model. Among 57 STNL nodes, 18 matched nodes were located within 20 m of their respective OSM counterparts. Of these, 17 exhibited a direct one-to-one correspondence with OSM nodes, while one showed a one-to-many (1:N) relationship, requiring manual geospatial validation. Traffic signals were automatically assigned to each matched node.

TABLE 2. Data processing summary.

Process	Simulation components	Data collected (n)	Processed (n)	Final model (n)
Edge filtering	OSM-based edges	3,275	718	718
Node matching	OSM-based nodes	11,923	478	478
	STNL-based nodes	57	18	17
Signal integration	Traffic signal	54	17	17
Traffic flow generation	OD matrix IN/OUT nodes (traffic volume)	10	10	10
	OD matrix inner nodes (building energy)	1,048	86	16

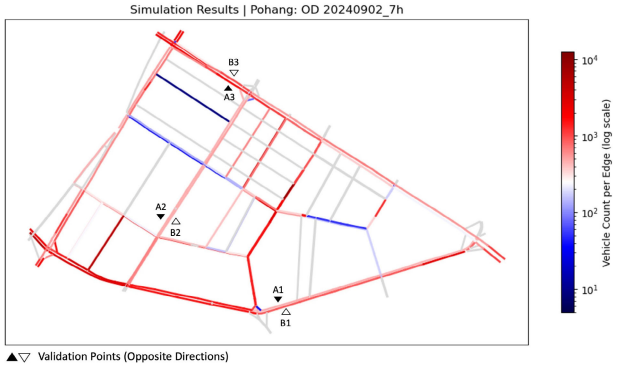


FIGURE 5. Simulation results.

Traffic flow generation was based on traffic count data from ten major edges designated as `IN/OUT` points, along with building energy consumption data for internal nodes. Among the 1,048 buildings in the study area, only those accounting for at least 0.1% of total energy consumption were selected. Facilities with high industrial-type energy consumption, such as water treatment facilities, were excluded because they are not typical daily trip destinations. In addition, buildings within 500 m of `IN/OUT` nodes were excluded to avoid routing distortions caused by excessively short OD paths. The processed set of 86 buildings was consolidated into 16 internal nodes through edge-level aggregation of buildings mapped to the same network edges. Together with five inflow–outflow node pairs, this configuration defined a final 21×21 OD matrix.

B. VALIDATION RESULTS

Fig. 5 presents the simulated vehicle counts for a single weekday peak hour. The spatial distribution highlights major commuting corridors, with key arterial roads experiencing the highest congestion during peak periods.

In addition, Fig. 6 shows edge-wise comparisons with observed traffic ranges at each validation point. At the designated validation points, most simulated volumes were within the observed traffic ranges across key segments. Statistical validation results were within acceptable thresholds ($R^2 = 0.757$, MAPE $= 15.1\%$). These results indicate overall model

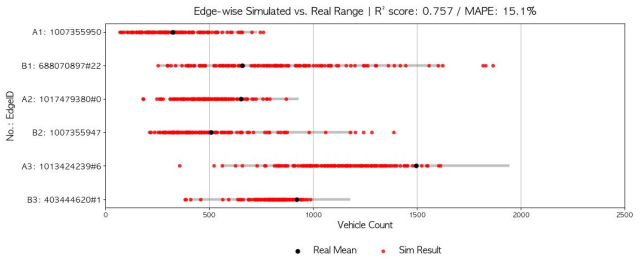


FIGURE 6. Generated and observed traffic validation.

reliability, although minor discrepancies were noted in some segments.

C. IMPLICATIONS, LIMITATIONS, AND FUTURE IMPROVEMENTS

The automated framework for generating traffic simulation models from open data on the SUMO platform systematically converts road network, traffic signal, and traffic volume datasets into simulation-ready inputs. Applied to the central district of Pohang, it successfully transformed raw OSM edges and nodes into a filtered network of 718 edges and 478 nodes, producing a 21×21 OD matrix and 17 operational signal controllers that were directly usable in SUMO.

In particular, the systematic integration of OSM with the national STNL road network, and the linkage of standardized node IDs with signal records, allowed intersection geometry and signal control to be automatically encoded in the model rather than manually scripted. By combining these two datasets into a unified node-link structure and embedding rule-based mapping logic from standardized node identifiers to signal phases, the framework automatically converted raw signal entries into simulation-ready signal plans, demonstrating a reusable procedure for nationwide standard-based signal integration. Similarly, the OD matrix, constructed by integrating observed gateway traffic counts with building-level energy consumption as a proxy for internal trip generation, provides a consistent and data-grounded representation of both external and internal traffic flows; validation shows that simulated traffic patterns and observed counts exhibit generally high agreement across locations. This indicates that the combined use of measured and inferred data can approximate actual traffic conditions with reasonable accuracy, while preserving consistency with the underlying input data used to generate the demand structure.

At the same time, several limitations emerged that clarify the current bounds of applicability. From a data perspective, many open datasets were partial, aggregated, or available only in non-standard or non-digital formats, such that only a subset could be directly exploited for detailed traffic simulation. This indicates that, relative to the overall volume of collected data, only a limited fraction can currently be used directly in simulation. Notably, the absence of intersection-level traffic observations limited the application of the framework to a broader range of sites and prevented more precise node-level calibration and validation.

Signal integration was also not fully automated for a small subset of intersections. When applying the automated phase–direction–movement mapping logic and inspecting the signal input data, we found that the logic was effective primarily for regular three- and four-leg intersections, whereas mismatches between geometric structure and recorded signal logic occurred frequently at irregular or complex intersections, accounting for approximately 4% of candidate signalized nodes. These issues typically arose when directional information did not clearly represent approach legs, when the ordering of signal phases did not align with the ordering of road connections in the network, or when a single physical intersection was represented by multiple nodes. In such cases, geospatial validation was performed by visually cross-checking OSM geometry, STNL node definitions, and signal controller records, after which the observed signal phases were manually encoded into the SUMO network.

In addition, the building energy data used for OD estimation exhibited structural limitations: small buildings with very low energy consumption were effectively excluded, and OD flows were generated using fixed-percentage rules and ratio-based distributions that function as heuristics rather than empirically calibrated relationships. Specifically, the initial distribution proportions were set based on limited local contextual information, including approximate estimates of flows toward representative destinations rather than zone based data, and some degree of arbitrariness was unavoidable. These proportions therefore required subsequent iterative adjustment through simulation based scaling to ensure consistency with observed gateway traffic volumes.

Within this context, the proposed approach is valuable in that it offers generality and reproducibility using nationwide open datasets and a consistent procedure for demand model construction; however, it inevitably falls short of the flexibility and accuracy of highly refined models that are built on rich observational datasets and repeatedly calibrated.

Future improvements should therefore focus on integrating richer and more standardized datasets—such as intersection-level traffic volumes, vehicle trajectories, high-resolution signal logs, and detailed land-use and activity information—and on introducing explicit calibration stages that iteratively adjust OD patterns and signal settings. A hybrid strategy that combines automated structural generation with selective manual refinement and machine-learning-based calibration [33], [34] is a promising option for simultaneously enhancing model robustness and transferability across diverse urban contexts. In parallel, recent studies have begun to exploit large language models to generate simulation scenarios, configurations, and code directly from natural language prompts, with applications to SUMO-based traffic scenario generation, traffic data analysis, and simulation input design [35], [36]. The automated data-processing framework proposed in this study can serve as an upstream process for such LLM-based interfaces,

providing standardized, simulation-ready inputs that these systems can build upon.

V. CONCLUSION

This study proposed and evaluated an automated framework for generating city-scale traffic simulation models from open data on the SUMO platform. By systematically converting road network, traffic signal, and traffic volume datasets into simulation-ready inputs and integrating OSM-based geometry with a nationwide road network and standardized signal records, the framework enabled realistic representation of multi-intersection urban traffic conditions using only public datasets. The OD matrix construction approach, which combines gateway traffic counts with building-level energy consumption, produced plausible internal and external flow patterns, and validation against observed counts in real urban areas indicated generally high agreement, demonstrating that credible traffic models can be derived from nationwide open data with minimal manual coding.

The contribution of this study lies in presenting an end-to-end, reproducible data-to-SUMO processing pipeline that systematically transforms heterogeneous open datasets into city-scale, multi-intersection SUMO models. Although the current implementation is constrained by data quality, spatial and temporal coverage, and heuristic OD estimation, it establishes a robust and extensible foundation onto which more precise calibration procedures and richer data sources can be layered. The framework can incorporate improved traffic count coverage, higher-resolution land use or activity data, and dynamic information such as probe vehicle or detector data, enabling tighter calibration of route choice and flow distributions. Further extensions may also include systematic uncertainty analysis, sensitivity testing of OD generation parameters, and integration of behavioral or multimodal components to expand the range of scenarios that can be assessed.

Beyond the specific application in this study, the proposed approach is designed so that comparable open datasets from other regions and periods can be ingested with minimal reconfiguration to regenerate updated models and perform temporal analyses of network or demand changes. Its modular, database-backed structure also makes it compatible with emerging traffic analytics and control pipelines, including machine-learning–based OD calibration, data-driven signal optimization, and LLM-assisted scenario generation that require standardized, simulation-ready inputs. By serving as a generic data-to-simulation infrastructure that can be repeatedly executed as new data streams become available, the framework can support more routine use of microscopic traffic simulation in operational monitoring, policy scenario evaluation, and experimentation with advanced AI-based traffic management tools and urban digital twins.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) under Grant Nos. 2022R1A4A101907113 and RS-2024-00357330. The authors would like to thank Hyun Joon Jeong for his valuable contributions to data processing and simulation validation. They used OpenAI ChatGPT for language polishing and limited coding assistance. All outputs were reviewed and approved by the authors.

REFERENCES

[1] J. Barceló, "Models, traffic models, simulation, and traffic simulation," in *Fundamentals of Traffic Simulation*. New York, NY, USA: Springer, 2010, pp. 1–62, doi: 10.1007/978-1-4419-6142-6_1.

[2] W. Jiang and J. Luo, "Big data for traffic estimation and prediction: A survey of data and tools," *Appl. Syst. Innov.*, vol. 5, no. 1, p. 23, Feb. 2022, doi: 10.3390/asi5010023.

[3] P. B. Bautista, L. F. Urquiza-Aguiar, and M. A. Igartua, "STGT: SUMO-based traffic mobility generation tool for evaluation of vehicular networks," in *Proc. 18th ACM Symp. Perform. Eval. Wireless Ad Hoc, Sensor, Ubiquitous Netw.*, Nov. 2021, pp. 17–24. [Online]. Available: https://dl.acm.org/doi/10.1145/3479240.3488523

[4] H. Xu, C. Wang, A. Berres, T. LaClair, and J. Sanyal, "Interactive web application for traffic simulation data management and visualization," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2676, no. 1, pp. 274–292, Jan. 2022, doi: 10.1177/03611981211035760.

[5] Y. Qin, W. Hua, J. Jin, J. Ge, X. Dai, L. Li, X. Wang, and F.-Y. Wang, "AUTOSIM: Automated urban traffic operation simulation via meta-learning," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 9, pp. 1871–1881, Sep. 2023, doi: 10.1109/JAS.2023.123264.

[6] G. Tamminga, P. Knoppers, and J. W. C. van Lint, "Open traffic: A toolbox for traffic research," *Procedia Comput. Sci.*, vol. 32, pp. 788–795, 2014, doi: 10.1016/j.procs.2014.05.492.

[7] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic traffic simulation using SUMO," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2575–2582, doi: 10.1109/ITSC.2018.8569938.

[8] L. Bieker, D. Krajzewicz, A. Morra, C. Michelacci, and F. Cartolano, "Traffic simulation for all: A real world traffic scenario from the city of Bologna," in *Modeling Mobility With Open Data* (Lecture Notes in Mobility). Cham, Switzerland: Springer, 2015, pp. 47–60, doi: 10.1007/978-3-319-15024-6_4.

[9] E. R. Maiorov, I. R. Ludan, J. D. Motta, and O. N. Saprykin, "Developing a microscopic city model in SUMO simulation system," *J. Phys., Conf. Ser.*, vol. 1368, no. 4, Nov. 2019, Art. no. 042081, doi: 10.1088/1742-6596/1368/4/042081.

[10] M. L. Clemente, "Building a real-world traffic micro-simulation scenario from scratch with SUMO," in *SUMO Conf. Proc.*, vol. 3, 2022, pp. 215–230, doi: 10.52825/scp.v3i.109.

[11] S. Behrendt, T. Altenmüller, M. C. May, A. Kuhnle, and G. Lanza, "Real-to-sim: Automatic simulation model generation for a digital twin in semiconductor manufacturing," *J. Intell. Manuf.*, pp. 1209–1224, Jan. 2025, doi: 10.1007/s10845-025-02572-x.

[12] M. Lütjen, U. Clausen, and J. Kallrath, "Automatic simulation model generation in the context of micro manufacturing (WIP)," in *Proc. Spring Simul. Multiconference*, 2015, pp. 1–8. [Online]. Available: https://dl.acm.org/doi/10.5555/2872965.2873000

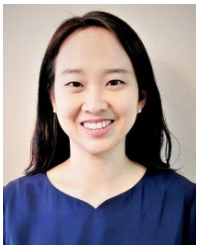
[13] G. S. Martinez, S. Sierla, T. Karhela, and V. Vyatkin, "Automatic generation of a simulation-based digital twin of an industrial process plant," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2018, pp. 3084–3089, doi: 10.1109/IECON.2018.8591464.

[14] K. Kluska, "Automatic simulation modelling of warehouses," *Logforum*, vol. 17, no. 1, pp. 59–69, Mar. 2021, doi: 10.17270/j.log.2021.547.

[15] Z. Meng, X. Du, P. Sottovia, D. Foroni, C. Axenie, A. Wieder, D. Eckhoff, S. Bortoli, A. Knoll, and C. Sommer, "Topology-preserving simplification of OpenStreetMap network data for large-scale simulation in SUMO," in *Proc. SUMO Conf.*, vol. 3, Sep. 2022, pp. 181–197, doi: 10.52825/scp.v3i.111.

[16] Q. Zhang, Y. Wang, R. Yin, W. Cheng, J. Wan, and L. Wu, "A data-based framework for automatic road network generation of multi-modal transport micro-simulation," *Electron. Res. Arch.*, vol. 31, no. 1, pp. 190–206, 2023, doi: 10.3934/era.2023010.

[17] X. Ma, X. Hu, T. Weber, and D. Schramm, "Evaluation of accuracy of traffic flow generation in SUMO," *Appl. Sci.*, vol. 11, no. 6, p. 2584, Mar. 2021, doi: 10.3390/app11062584.
[18] L. Codeca, R. Frank, S. Faye, and T. Engel, "Luxembourg SUMO traffic (LuST) scenario: Traffic demand evaluation," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 2, pp. 52–63, Summer. 2017, doi: 10.1109/MITS.2017.2666585.
[19] A. Qiu, P. A. Sathish, D. Wang, and H. D. Schotten, "Advanced traffic demand generation in SUMO: ML-based prediction of flow rate based on real-world measured datasets," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2024, pp. 1–7, doi: 10.1109/vtc2024-spring62846.2024.10683300.
[20] M. S. Irfan, S. Dasgupta, and M. Rahman, "Toward transportation digital twin systems for traffic safety and mobility: A review," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 24581–24603, Jul. 2024, doi: 10.1109/JIOT.2024.3395186.
[21] C. Ge and S. Qin, "Digital twin intelligent transportation system (DT-ITS)—A systematic review," *IET Intell. Transp. Syst.*, vol. 18, no. 12, pp. 2325–2358, Dec. 2024, doi: 10.1049/itr2.12539.
[22] Y. Gao, S. Qian, Z. Li, P. Wang, F. Wang, and Q. He, "Digital twin and its application in transportation infrastructure," in *Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Beijing, China, Jul. 2021, pp. 298–301, doi: 10.1109/DTPI52967.2021.9540108.
[23] Y. Chen, X. Ma, L. Zheng, H. Yan, C. Wang, S. Li, H. Zhang, and X. Zhao, "Design and implementation of a digital twin simulation platform for autonomous transportation systems," in *Proc. Int. Conf. Artif. Intell. Auto. Transp. (AIAT)*. Singapore: Springer, 2025, pp. 166–176, doi: 10.1007/978-981-96-3965-6_18.
[24] K. Kušić, R. Schumann, and E. Ivanjko, "A digital twin in transportation: Real-time synergy of traffic data streams and simulation for virtualizing motorway dynamics," *Adv. Eng. Informat.*, vol. 55, Jan. 2023, Art. no. 101858, doi: 10.1016/j.aei.2022.101858.
[25] D. Krajzewicz, "Traffic simulation with SUMO—Simulation of urban mobility," *Int. Ser. Oper. Res. Manag. Sci.*, vol. 145, pp. 269–293, 2010, doi: 10.1007/978-1-4419-6142-6_7.
[26] OpenStreetMap Contributors. *OpenStreetMap*. Accessed: Jul. 25, 2024. [Online]. Available: https://www.openstreetmap.org
[27] National Transport Information Center. *Node-Link Data Service*. Accessed: Jul. 22, 2024. [Online]. Available: https://www.its.go.kr/nodelink
[28] Ministry of the Interior and Safety. *Public Data Portal (data.go.kr)*. Accessed: Aug. 29, 2024. [Online]. Available: https://www.data.go.kr
[29] Pohang City. *UTIS: Urban Traffic Information System*. Accessed: Oct. 21, 2024. [Online]. Available: https://utis.pohang.go.kr
[30] Ministry of Land, Infrastructure and Transport. *National Construction Data Open System*. Accessed: Oct. 4, 2024. [Online]. Available: https://open.eais.go.kr
[31] P. Zhang and Z. Qian, "User-centric interdependent urban systems: Using time-of-day electricity usage data to predict morning roadway congestion," *Transp. Res. Part C:, Emerg. Technol.*, vol. 92, pp. 392–411, Jul. 2018, doi: 10.1016/j.trc.2018.05.008.
[32] A. Movahedi, A. B. Parsa, A. Rozhkov, D. Lee, A. K. Mohammadian, and S. Derrible, "Interrelationships between urban travel demand and electricity consumption: A deep learning approach," *Sci. Rep.*, vol. 13, no. 1, Apr. 2023, Art. no. 7083, doi: 10.1038/s41598-023-33133-y.
[33] T. Pamuła and R. Żochowska, "Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105550, doi: 10.1016/j.engappai.2022.105550.
[34] C. Xiang, P. Yang, F. Xiao, and X. Fan, "Urban traffic application: Traffic volume prediction," in *Multi-Dimensional Urban Sensing Using Crowdsensing Data*. Singapore: Springer, 2023, pp. 113–150.
[35] S. Li, T. Azfar, and R. Ke, "ChatSUMO: Large language model for automating traffic scenario generation in simulation of urban mobility," *IEEE Trans. Intell. Veh.*, vol. 10, no. 11, pp. 4962–4973, Nov. 2024, doi: 10.1109/TIV.2024.3508471.
[36] S. Zhang, D. Fu, W. Liang, Z. Zhang, B. Yu, P. Cai, and B. Yao, "TrafficGPT: Viewing, processing and interacting with traffic foundation models," *Transp. Policy*, vol. 150, pp. 95–105, May 2024, doi: 10.1016/j.tranpol.2024.03.006.



HYUNYOUNG RYU received the B.S. degree in landscape architecture and the M.S. degree in urban and regional planning from Seoul National University, Seoul, Republic of Korea, in 2009 and 2011, respectively, and the Ph.D. degree in sustainability science from the University of Tokyo, Tokyo, Japan, in 2016.

She is currently a Research Assistant Professor with the Future Open City Innovation Center, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea. Her research interests include urban data and process analytics, smart and sustainable cities, shrinking cities, and data-driven decision support for urban planning and mobility.

Dr. Ryu was a recipient of Sejong Science Fellowship from the National Research Foundation of Korea from 2022 to 2027.

JITAEK LIM received the B.S. degree in information technology management from Seoul National University of Science and Technology, Seoul, South Korea, in 2017, and the Ph.D. degree in industrial and management engineering from POSTECH, Pohang, Republic of Korea, in 2023.

He is currently a Senior Researcher with Korea Institute of Science and Technology Information (KISTI), Seoul, Republic of Korea. His research interests include business process management, information systems, graph neural networks, and large language models.

MINSEOK SONG (Member, IEEE) received the Ph.D. degree in industrial and management engineering from POSTECH, Pohang, Republic of Korea, in 2006. Prior to joining POSTECH, he was a Postdoctoral Researcher with Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands, and later held faculty positions with Ulsan National Institute of Science and Technology (UNIST).

He is currently a Professor with the Department of Industrial and Management Engineering, POSTECH, where he the Director of the Open Innovation Big Data Center and leads the Analytics and Information Management Laboratory. He is an Associated Partner Member of European Research Center for Information Systems. He has published more than 130 papers and has conducted over 30 research projects, including large-scale grants, such as the EU Horizon 2020 RISE-BPM project and grants from the National Research Foundation of Korea. His data-driven methods and software systems have been deployed in major organizations (e.g., Samsung Electronics, Samsung C&T, POSCO, Samsung Heavy Industries). His research interests include process mining, recommendation systems, business analytics, process simulation, healthcare information systems, manufacturing process analysis, and industrial AI. He has received several awards, including the Minister of Science and ICT Award.