A LiDAR-Based Digital Twinning Workflow for Traffic Monitoring and Simulation

Maulana Ikram Wibisana [1], Mila Koeva [1], Pirouz Nourian [1], Dessislava Petrova-Antonova [2], Kaloyan Karamitov [2]

[1] University of Twente, ITC Faculty, 7522 Enschede, Netherlands – maulanaikramwibisana@student.utwente.nl; m.n.koeva@utwente.nl; p.nourian@utwente.nl
[2] The Big Data for Smart Society Institute (GATE), 1604 Sofia, Bulgaria – dessislava.petrova@gate-ai.eu; kaloyan.karamitov@gate-ai.eu

Keywords: Urban Digital Twins, LiDAR Data, Traffic Data Enrichment, Traffic Monitoring, SUMO, Transportation Planning.

Abstract

The process of ensuring efficient and safe urban transportation is closely linked to urban planning, particularly through the aspects of transportation planning. Transportation planning is a pivotal concern for urban regions worldwide, reflecting the growing need to increase mobility while ensuring safety and sustainability in densely populated areas. This research focuses on developing a novel digital-twin-based approach for micro-traffic simulation to support data-driven decision-making for increasing traffic safety through scenario planning. Leveraging the traffic data obtained through monitoring one of the busiest intersections in Sofia city, this research workflow shows the effective integration of LiDAR data and the urban digital twin concept in intelligent transportation systems (ITS). The research addresses problems related to moving object classification, trajectory analysis, and reclassification of unrecognised objects by processing the LiDAR data, pre-processed in a .osef format, thereby transforming it to make it suitable for simulation. The proposed solution for the monitoring of urban traffic is demonstrated by the usage of SUMO (Simulation of Urban MObility) for performing simulations and a Random Forest model for unrecognized object reclassification to pre-existing vehicles and pedestrian classes. The architecture of the proposed workflow can possibly be applied in other similar urban settings, providing a scalable solution for both traffic management and urban planning. The study’s results support the wider use of urban digital twin principles in ITS by highlighting the value of advanced modelling tools and high-quality data in addressing today's urban transportation challenges.

1. Introduction

The development of effective and safe transportation systems within cities is significantly influenced by urban planning, particularly in the area of transportation planning. Increased mobility enhances certain types of social and economic activity in the cities. Despite the benefits, transportation congestion remains the most prevalent issue, with increased delays recorded in 58% of metropolitan locations (Pishue, 2023). A 2023 traffic report reveals the average driver lost 51 hours to congestion in 2022, a 15-hour increase from 2021 (Fernandez, 2023). Furthermore, traffic congestion not only causes delays but also has detrimental environmental impacts, contributing to urban pollution.

Another concern is traffic accidents. An estimated 35,766 fatal Traffic accidents were reported in the United States alone in 2020, while globally, road crashes cause 1.3 million deaths and 20-50 million non-fatal injuries annually (Zhang, 2020). This highlights a global epidemic of traffic accidents causing millions of deaths and injuries every year. Economically, traffic congestion in Europe costs the continent around 1% of its GDP (Gross Domestic Product) annually. This underscores the urgent need for sustainable urban mobility solutions (Rodrigues et al., 2021).

In response to these challenges, European communities have invested in bike and pedestrian infrastructures to lessen dependence on motorized vehicles, thereby reducing intraurban traffic and pollution. This movement towards sustainable urban transportation emphasizes data-driven policy and green infrastructure (European Commission, 2022). Bulgaria, reflecting these trends, had the second-highest road fatality rate in the EU in 2021, worsening in 2022 with 175 road deaths between January and May (The Sofia Globe, 2022). Sofia faces similar issues, with traffic and pollution persisting despite public transit advancements. The city is committed to promoting environmentally friendly transportation and digitalizing transportation systems, as outlined in the Sustainable Urban Mobility Plan (SUMP) for 2019–2035 (Ministry of Transport and Communications of Republic of Bulgaria, 2017)

This paper proposes an end-to-end approach for micro-traffic analysis and simulation based on urban digital twins. This approach utilizes key functionalities of digital twins, developing a robust traffic data workflow covering data transformation, moving object reclassification, data storage, and exportation for traffic simulations. The workflow supports informed decision-making and provides a foundation for advancements in intelligent transportation systems. The paper is structured as follows: Section 2 outlines related work, Section 3 presents the study area and datasets, Section 4 explains the research methodology, Section 5 describes the results, and Section 6 concludes with future work.

2. Related Work

Recent developments in LiDAR technology have had a big influence on transportation applications by making it possible to perceive the environment in detail and accuracy. Real-time traffic data collecting is made easier by this technology, especially when it is used on roadside platforms. LiDAR's capability to capture precise geometric data enables intelligent transportation systems to create comprehensive holographic scenes of traffic conditions, underscoring its significance in contemporary transportation research for optimizing traffic management methods and alleviating congestion (Williams et al., 2013).

Accurate coordinate transformations are crucial for aligning local Cartesian coordinates with global geographical coordinate systems (CRS), significantly impacting spatial analysis outcomes, particularly for LiDAR data (Fan et al., 2014).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-411-2024 | © Author(s) 2024. CC BY 4.0 License.

Improved spatial data alignment, essential for traffic simulation, relies on high-resolution satellite imagery for precise road extraction, often supplemented by manual digitization to enhance segmentation accuracy at the lane level. Understanding traffic dynamics involves analysing GPS data to interpret vehicle paths, with spatial analysis playing a key role in identifying travel behaviours (Zheng, 2015). Handling missing labels in these datasets is effectively managed by artificial intelligence, particularly Random Forest (RF), which excels in classifying complex datasets and improving traffic simulation accuracy (Breiman, 2001).

Parallel to the development of LiDAR applications, the concept of digital twins (DTs) has gained popularity in the Intelligent Transportation System (ITS) field. A digital twin is a dynamic digital model of a physical object or system that combines sensor data and analytics to reflect its real-world status, enabling real-time monitoring, simulation, and decision-making (Digital Twin Geohub, 2023). The research by Kušić et al. (2023) on the digital twin model of the Geneva Motorway (DT-GM) exemplifies the use of the microscopic traffic simulator SUMO (Simulation of Urban MObility) to model and simulate synchronized virtual representations of transportation dynamics.

This research aims to improve the operational efficiency of dynamic traffic management by integrating real-time LiDAR data into traffic simulation models, focusing on the Open Serialization Format (OSEF) dataset. Developed by Outsight, this dataset employs Type-Length-Value (TLV) encoding, tailored for LiDAR sensor data, streamlining processing and delivering relevant information efficiently (Vincent, 2023). While visualizing real-time data provides an immediate snapshot, traffic simulation allows for the analysis of potential scenarios and the evaluation of traffic management strategies, leading to a more comprehensive understanding.

3. Study Area and Datasets

3.1 Study Area

The research area is located in one of the busiest intersections in Sofia, Bulgaria, adjacent to a big shopping center, in the heavily populated Lozenets district. The development of the shopping center in a neighborhood featuring a mix of high-speed secondary and tertiary roads, smaller residential streets, bi-directional roads, and various public transportation options has led to significant congestion issues. A high density that adds to traffic challenges is evident in the Lozenets district, whereas, of June 15, 2023, there were 63,214 residents living at their current address and 67,093 at their permanent residence (Ministry of Regional Development and Public Work of Bulgaria, 2023).

The intersection is composed of a single lane (Blvd. "Cherni vrah") that runs from south to north and a dual-lane road that splits into two streets: "Srebarna" that runs northeast and "Henrik Ibsen" that leads southwest. The numerous types of vehicles, such as cars, two-wheelers, high-load trucks, and buses, impact the traffic dynamics in this area. The traffic is monitored by a LiDAR system, comprising 6 sensors, which are the primary data source for this study. They are positioned along various borders surrounding the main intersection, record the movement of vehicles and pedestrians over time. The study area and sensor distribution are shown in Figure 1

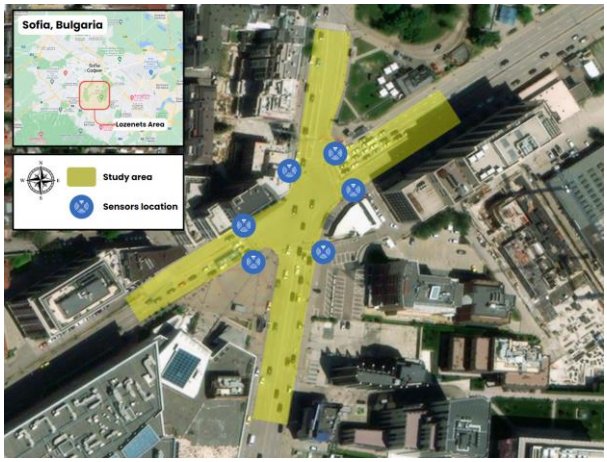

Figure 1. Study area and sensors distribution.

The LiDAR sensor, Ouster OS1-128, features a 45° field of view across 64 layers and uses an ALB processing module for classifying cars, trucks, pedestrians, and two-wheelers. It measures 85 mm in diameter and 58.45–73.5 mm in height, consuming 14–20W of power at 22–26V. Operating temperatures range from -53°C to 60°C. It is IP68 and IP69K rated, shock-resistant (IEC 60068-2-27), and vibration-resistant (IEC 60068-2-64). The sensor offers 0.3 cm resolution, detects as close as 0.3 m, and provides ±3 cm Lambertian and ±10 cm retroreflector accuracy.

3.2 Research Datasets

The LiDAR data collected from the sensors is processed in real-time and made available for further analysis in .osef format, which is the main input dataset for this research. The .osef data format was developed to make LiDAR point cloud technology more manageable and accessible, addressing integration problems into ITS applications. The .osef datasets offer precise and comprehensive spatial measurements, making them invaluable across diverse applications. Its key advantages include adaptability, which simplifies processing; efficiency, by reducing processing overhead; simplicity, through straightforward parsing; robustness, offering versatile data management; compatibility, facilitating integration with new features; and scalability, accommodating varying data volumes.

The real-time download of the .osef dataset was managed using a TCP (Transmission Control Protocol) stream. The time-stamped data, specifically in GMT+3 to match the local time zone of each recorded frame, is organized systematically by the preprocessor of the sensors. The datasets used in the research vary, ranging from a 2-minute period in the morning to a 4-minute period in the afternoon; both were utilized as foundational data for specific research steps. As shown in Figure 2, the data can be parsed to distinguish between objects that are considered dynamic and those that are static, belonging to the urban infrastructure.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-411-2024 | © Author(s) 2024. CC BY 4.0 License.

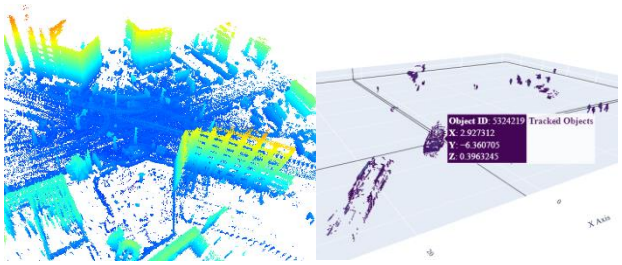

Figure 2. Intersection point cloud visualisation; static objects (Left), and dynamic objects (Right).

The nested TLV tree structure is traversed while processing binary data into an array of data points or other formats, such as .csv. The .osef data contains information on tracked objects and the augmented cloud, in addition to the previously described information pertaining to the base data. Unique object identifiers (IDs), classes consisting of CAR, PERSON, TRUCK, TWO_WHEELER, and UNKNOWN, speed in km/h (convertible to other units of measurement), volume computed using bounding boxes, coordinates (local pose x, y, and z, or Cartesian coordinates) and zones are among the information that can be extracted from tracked objects.

However, it is important to note a few issues with the dataset. Firstly, the dataset currently records coordinates only in a local pose, conversion into geo-coordinates is needed. Second, the data contain an unknown class that needs to be identified. Therefore, addressing these issues as practically as possible should be the initial step in this research.

4. Research Methodology

The purpose of this research is to develop a digital twin-based workflow that can integrate real-time LiDAR data into traffic simulations for effective traffic monitoring and assessment. The research methodology illustrated in Figure 3 explores the potential of OSEF data integration into traffic simulation while addressing data issues such as unknown classes and spatial alignment problems for the traffic simulation input.

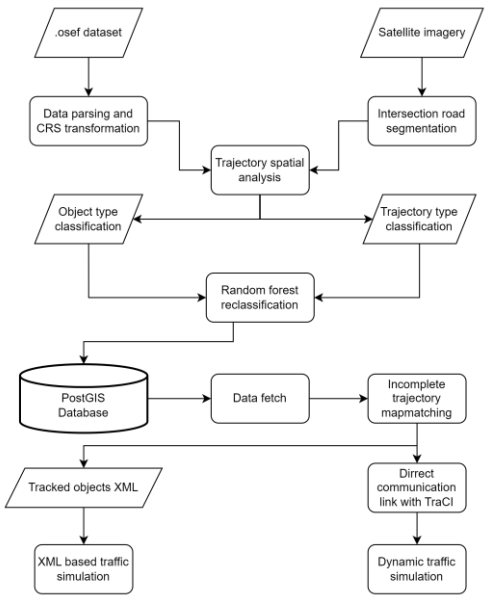

Figure 3. Research methodology workflow.

The methodology involves several steps: collecting data from the .osef dataset and satellite imagery, parsing and transforming it into an appropriate CRS, and classifying object types within each frame. Trajectory spatial analysis classifies each moving object’s trajectory, providing features for random forest reclassification to address unknown classes and enhance data enrichment. Properly classified trajectories are essential for accurate traffic simulations and must align with the simulation network. Processed data is stored in a PostGIS database, with incomplete trajectories corrected through map-matching. Data is fetched into the simulation using XML-based traffic simulations and the dynamic TraCI method for real-time monitoring and assessment. The .osef dataset is parsed using a Python script and OSEF 3.0 libraries, efficiently processing large data volumes.

The output of geopackage (.gpkg) data from road segmentation informs trajectory spatial analysis, determining object trajectories and classifying vehicles based on intersection locations. RF reclassification addresses unknown classes, enhancing model performance. The reclassified data is uploaded to the PostGIS database, which includes trajectory and object information. SUMO traffic simulation retrieves this data, transforming it into XML format for input. The simulation replicates real-world traffic conditions, using XML and TraCI to dynamically adjust the simulation, creating a comprehensive pipeline for processing, storing, and simulating traffic data.

4.1 Local Coordinates Transformation to Geo-coordinates

Cartesian coordinates or a local pose array are the two types of coordinates that are available from the .osef dataset. This dataset was first transformed into global CRS (Coordinate Reference System) geo-coordinates in order to be used for spatial analysis and to ensure that it is aligned with the SUMO road network. To accurately convert local coordinates to the desired geo-coordinates, Well-Known Text (WKT) is used. This is a text format that defines geographical properties and geometries.

With given WKT information, there are multiple processes involved in converting local coordinates (x, y, z) to geographical coordinates (longitude, latitude) using PyProj. A Python interface to PROJ, a general coordinate transformation software, is used in this process using the PyProj package. The position of a point in three dimensions, where $x, y, and\ z$ represent distances from an origin in a localized coordinate system, are converted to a position on the surface of the Earth, which is represented by latitude (north-south position from the equator) and longitude (east-west position from the Greenwich meridian).

The local pose, which refers to point or object position and orientation within local coordinate systems, can be transformed directly into WGS84 coordinate systems using WKT information. To ensure the alignment with SUMO's road network projection, which follows an ellipsoid rather than a 2D plane projection, EPSG:4326 was determined to be the appropriate coordinate system for this research. EPSG:4326 uses an ellipsoid to model the surface of the Earth. The transformation procedure requires a datum shift to adjust the coordinates from the local datum relative to WGS84.

4.2 Semantic Road Segmentation

Road segmentation plays an important part in initiating the process of classifying objects and their trajectories. It allows the understanding of semantic information about the road edges at the intersection, which aids in the spatial identification of areas classified as roads from the ones that are not. In order to

correctly classify object trajectories, the key objective of semantic segmentation is to identify the main lanes and junctions. During this process, roads are segmented into polygons, following the real-world shape of the intersection, and traced from satellite imagery —USGS Landsat 7 ETM+ C2 L1 with a resolution of 15-meter panchromatic blend— and ESRI’s world imagery service (up to 1-meter resolution) using ArcGIS software. The main focus is to identify the main road segments, the turning point, the lane distribution details, and the junctions that existed in the study area. The output is stored as a .gpkg. The segment name and junction type constitute the semantic information structure; segment names are modified in accordance with the SUMO road network naming format.

"In/out_name_lanenumber" is the format used for segmentation names. "In/out" indicates the direction of the road flow, i.e., whether the segment is traveling inward or outward from the main intersection junction. "Name" indicates the road edges' name based on their location (for example, if they are in the north, their name will have a "n" affixed to it), and "lane number" indicates the exact lane number of those road edges. Lane numbers (e.g., 1, 2, 3,...) are assigned in ascending sequence from the outside to the inner portion of the road boundaries. When it comes to junction types, they are referred to according to the type of junction, such as intersection or fork (a junction where several roads merge into one or diverge into more than one).

4.3 Trajectory Spatial Analysis

Spatial analysis of objects is carried out using the outcomes of road segmentation and local coordinate transformations. This analysis includes object type reclassification (classifying possible vehicles or non-vehicles) and trajectory type classification (categorizing points into complete, short-complete, violation, and incomplete trajectories). It determines where road segmentation intersects with the frame sequence for each object ID, generating unique trajectories. Geopandas is the main library used in this analysis.

The logic behind object classification is straightforward: an object ID is classified as a possible vehicle if the majority of its geo-coordinate sequence $(Lon_1, Lat_1, Lon_2, Lat_2, \dots)$ intersects within a road segment (such as a lane or junction). This suggests that although it is presumably a vehicle, it might also be a person. On the other hand, it is categorized as a non-vehicle or a person if the majority of the data-point sequence of geo-location does not traverse the road segment.

Trajectory-type classification is more complex and refined than object-type classification, requiring semantic road segment information and generating multiple unique instances for detailed classification. When object data-point sequences are primarily "non-vehicle," their trajectories are classified as "none" due to lack of alignment with road segments. For "possible vehicle" classifications, objects are filtered and assumed as vehicles based on their frame sequence within road segments. Specific segment types (inward, outward, intersection, special turns) are analysed to classify trajectory types. Complete trajectories follow the pattern (in, junction, out), with an 8-meter threshold from the junction distinguishing complete from short trajectories, essential for traffic simulation XML adjustments. Additional assessments identify unexpected direction changes indicating rule violations, such as illegal turns or same-side violations, where trajectories start and end on the same roadside. This meticulous approach ensures that trajectories are appropriately identified and saved as a separate column in the main array of data points.

4.4 Random Forest Reclassification

Given the issues associated with many objects that are labelled as "unknown," reclassification using the Random Forest (RF) algorithm as a supervised machine learning approach is used. It is considered to be appropriate for handling complicated information with a lot of uncertainties, notably volume, speed, and the trajectory and object types that come from spatial trajectory analysis. RF is able to deal with mixed data types in a dataset, especially when the dataset contains both categorical (e.g., object type and trajectory type) and numerical (e.g. speed and volume) data. Additionally, RF reduces overfitting by using an ensemble approach, where multiple decision trees influence the results, producing broadly applicable and reliable predictions across various datasets.

Furthermore, RF provides insights into feature significance, identifying key features like speed, volume patterns, object type, and trajectory type for reclassification. Feature engineering calculates average speed and volume for each object ID and changes in volume to improve prediction precision. The dataset is split into 70% training and 30% test sets, with the test set further divided into 15% with unknown labels and 15% without, to assess model predictions on unseen data. Refined class adjustments resolve labels inconsistencies by prioritizing precise labels over ambiguous ones, creating a uniform training set. This approach enhances the model's ability to recognize patterns and improves accuracy for previously unknown cases.

RF processes data internally as arrays, with each row frame denoting a data point and each column representing a feature. A random subset of features is taken into consideration at each split in a tree during the training phase when feature arrays are randomly selected (with replacement) to form subsets for training individual trees. The process of classifying new data point (x) for the reclassification issues, where there are N trees and classes as C, can be defined as follows:

$$\text{Predicted Class} = \underset{c \in C}{\operatorname{argmax}} \sum_{i=1}^{N} I(\text{prediction of tree}_i(x) = c), \qquad (1)$$

where I is an indicator function that evaluates whether the prediction of ith decision tree (tree_i) for the data point x matches the spesific class c. If it is true, the indicator functions return 1, otherwise 0. Based on all the trees in the forest, this equation indicates that for data point x, the projected class is the one that has received the most votes.

Hyperparameter tuning is also considered in the reclassification, depending on the performance of the model and how well can they effectively adapt the model to accurately reclassify the unknown classes. Metrics such as the number of tree (n_estimators), maximum depth of the tree (max_depth), minimum sample split (min_sample_split), and the number of features (max_features) are considered. Parameter grid search techniques are used to determine the initial hyperparameter tuning. It involves systematically testing a predefined set of hyperparameters to find the optimal combination by evaluating the performance of the model across different parameter settings.

4.5 Traffic Simulation

After local coordinate transformation, trajectory spatial analysis, and reclassification of unknown objects, the dataset is uploaded into a PostGIS database. The traffic simulation platform SUMO, which relies on XML, retrieves this data using the SQLAlchemy

library and xml.tree libraries to export it into an XML format compatible with SUMO. The traffic simulation workflow includes two processes: one using XML input to initiate the simulation and another using TraCI libraries to dynamically add vehicles without XML. The data fetching process incorporates map-matching to handle incomplete trajectories, using Sumolib to transform them into complete routes by aligning geo-coordinates with the nearest road network.

For XML based approach, data is directly fetched, transformed, and saved in the required XML format. Vehicles use the {\trip>} attribute, and pedestrians use the {<person>} attribute to define the XML configuration for vehicle input. In separate trips.xml files, every vehicle class is stored. The characteristics selected for the XML configuration are intended to match real-world behaviour as much as feasible while also considering the data that is contained in the database. The list below shows the details about each trips.xml configuration structure.

- Vehicle (car, truck, two-wheeler):

<**vtype** id vClass accel decel sigma color/>
<**trip** id type depart from to via departLane departPos departSpeed arrivalLane arrivalPos arrivalSpeed/>

- Pedestrian (person):

<**vType** id vClass color/>
<**person** id type depart departPos/>
<**walk** edges speed arrivalPos/>

Where acceleration (acceleration) and deceleration (deceleration) are derived from speed changes over time, vehicle class (vClass) classifies vehicles. Vehicle behaviour is represented by sigma, while its appearance is assigned by colour. A type (vType for cars) defines every vehicle or pedestrian (object id), containing details about the vehicle's exact position, lane usage, departure time, and initial and destination coordinates (from/to). Trajectory continuity is guaranteed via speed settings, where "last" denotes adherence to the speed of the vehicle ahead to avoid simulation errors.

The dynamic approach, on the other hand, does not store data locally. During the simulation time step, TraCI utilize the base XML input and continuously adding vehicles based on timestamps. A road network created in SUMO, imported from OpenStreetMap and modified to match the research area and road segmentation, is necessary for initiating the traffic simulation and completing trajectories identified as incomplete by the spatial analysis.

5. Results

In this section, the results of the research are presented. The sections are ordered according to the research workflow. Starting from the coordinate transformation to the final DT traffic simulation workflow.

5.1 Local Coordinates Transformation to Geo-coordinates

All of the arrays of data-points had their local pose translation (x, y, z) are converted to WGS84 EPSG:4326 geo-coordinates. This process is done as part of the .osef data parsing algorithm. Where each frame's local coordinates are transformed into geo-coordinates, following the methodology section. To further confirm that the tracked object points have been transformed successfully, the Plotly Express library's function called Mapbox was used. The array of data points must have precise geographic coordinates (longitude, latitude) for Mapbox to work properly. The georeferenced plotting of tracked objects in Mapbox can be seen in Figure 4 below and footnote (1).

Figure 4. Mapbox plotting of the geo-referenced data points.

As can be observed, the tracked items were successfully reprojected onto the intersection's true geo-coordinates, displaying a pattern that matches the form of the intersection. Following this coordinates transformation, trajectory spatial analysis can be performed on this dataset.

5.2 Semantic Road Segmentation

Road segmentation focused on lane and junction construction, with segment polygons manually digitized and aligned to the satellite images. The lanes were named based on their positions relative to the road borders, encompassing four road directions: northern, northeastern, southern, and southwestern. The northern segment has two outward lanes and three inward lanes toward the main intersection. The northeastern segment comprises two outward lanes and 3-4-3 inward lanes, with a single turning lane connecting to the northern outbound lane, as shown in Figure 6. The southern edges feature three inward and three outward lanes, while the southwestern segment has 2-1 outward lanes and 2-3 inward lanes, including a turning segment connecting the southern outward lanes to the bidirectional lanes.

In total, there are five junctions that constitute the road segmentation: four fork junctions and one main intersection junction. Fork junctions, for instance, facilitate division and merging in the case of a single turning route in the northeast that links two separate road segments. The semantic segmentation details can be seen in Table 1 and Figure 5.

Type	Segments	Count
Road	In_North	3 lanes
	In_North-East	4 lanes, 1 turning lane
	In_South-West	3 lanes
	In_South	3 lanes
	Out_North	3 lanes
	Out_North-East	2 lanes
	Out_South-West	2 lanes, 1 turning lane
	Out_South	3 lanes
Junctions	Main Junction	1
	Forks	4

Table 1. Road segmentation details.

(1) https://vimeo.com/948598952?share=copy

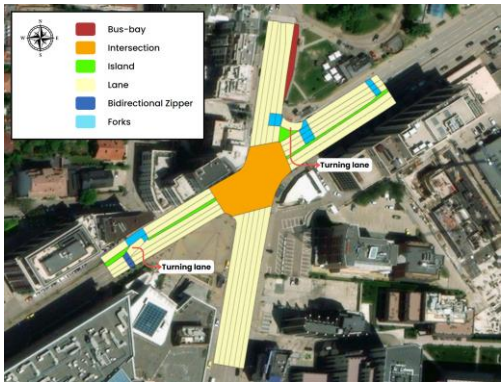

Figure 5. Road semantic information map of study area.

5.3 Trajectory Spatial Analysis

Geospatial analysis is performed to determine object and trajectory categories using the results of CRS transformation and semantic road segmentation. The first step in the analysis is to find the object type, or whether the objects are possibly vehicles or non-vehicles, by intersecting the road segment with the geographical coordinates of each object. Here, the rational presumption is that an object is possibly a vehicle if its frame sequence begins and ends on a road segment. Otherwise, they are non-vehicles.

Figure 6 shows the distribution of object categories over a two-minute period, including 516 tracked objects. It indicates that about 309 unique objects are classified as possible vehicles, though this category also includes persons and unknown objects. Additionally, 194 objects are classified as non-vehicles, likely pedestrians. However, anomalies exist, such as four cars and four two-wheelers being misclassified as non-vehicles, likely due to inconsistent labelling (associated with more than two classes).

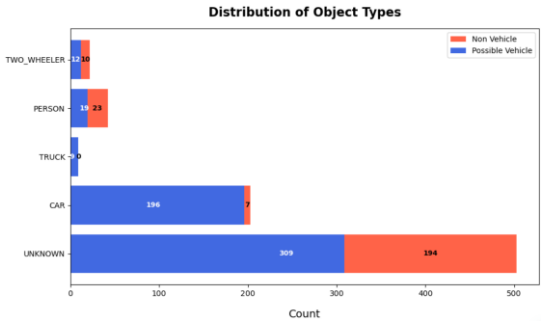

Figure 6. Object type classification distribution graph.

The next step involves identifying the trajectories and analysing their patterns to understand movement behaviours or detect anomalies. According to the research goal of determining appropriate vehicle trajectories, a thorough categorization of trajectories was only done for possible vehicles; non-vehicles were not taken into consideration further. The methods section outlines the process for determining trajectories, and the spatial analysis result showed that 500 of the 516 recorded objects had incomplete trajectories. The analysis revealed that only 10 object had complete trajectories, 4 cars had (short) complete trajectories, and 2 objects had violation trajectories due to illegal lane changes or U-turns. This is observed in the trajectory plot shown in Figure 7.

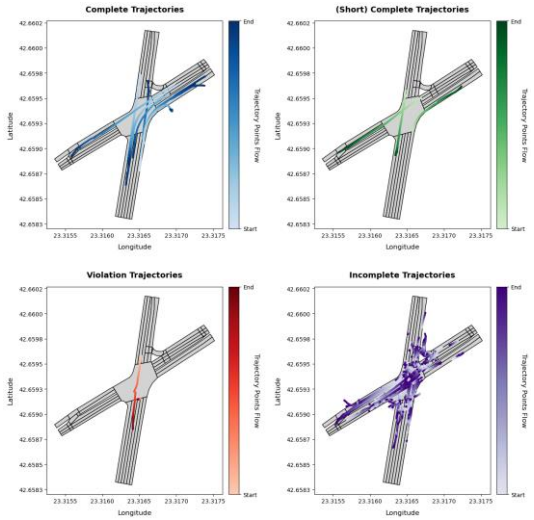

Figure 7. Trajectory type plot of each object id, consist of complete, (short) complete, violation, and incomplete track.

Upon closer inspection of Figure 7, especially the incomplete trajectories, we find that the majority of them end at the centre of the primary intersection junction. This could be due to either the object moving outside of the LiDAR sensors' range for some reason or the time frame was cut short during the downloading of the .osef data.

5.4 Random Forest Reclassification

Understanding the data structure is crucial before beginning to train the RF model. Analysis of a 2-minute morning .osef dataset sample, comprising 58,939 rows and 516 tracked objects, revealed complex class classifications, adding noise and reducing prediction accuracy. The dataset includes three categories: recognized objects (one unknown among two classes, classified as the known class), consistent objects (single class), and unidentified objects (changing classes more than twice). There are 11 unidentified, 241 recognized, and 264 consistent objects. To reduce noise, the training process prioritized identified classes over unknowns, focusing predictions on unknown or unidentified multi-class objects.

The model was trained using two distinct .osef datasets, one from an afternoon session lasting four minutes and the other from a morning session lasting two minutes, comprising 108,156 frames. The first training round is done with the default settings of reclassification model training. It shows a high accuracy of 0.998 when tested with the test set. This indicates close-to-perfect performance, which is unlikely. Since the aim of this process is to reclassify the unknown labelled objects and generalize the classes as much as possible throughout the frames, the number of objects with multiple classes needs to be minimized. In this first attempt, there are still 88 objects that have multiple classes after predictions, and 11 of them consist of more than two classes.

A second training attempt involved hyperparameter tuning, setting n_estimators to 400 and using 'Adjusted Class' labels to minimize multiple predictions. This improved the accuracy to 0.999 and reduced objects with multiple class predictions to 36. To further enhance robustness, a final iteration used a parameter grid search to optimize hyperparameters: 400 n_estimators, min_sample_split of 10, max_depth of 9, and max_features set to 'none.' Figure 8 shows the learning curve of this tuned model, demonstrating a steady accuracy increase and good

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-411-2024 | © Author(s) 2024. CC BY 4.0 License.

generalization, with the curves converging at the end, indicating good generalization.

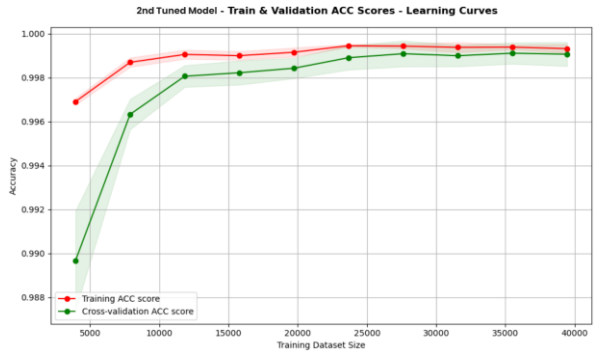

Figure 8. 2nd Tuned model learning curves.

This pattern suggests that the model can predict new, unseen data more effectively and generalize better than the previous model. Although this model has a similar accuracy of 0.998 compared to the first model, its generalization is far better, reducing multiple class predictions for an object ID by 76%. Table 2 below shows the differences in each model's performance in minimizing multiple class predictions.

RF Model	Accuracy	Multiple Class Prediction	
		2 Classes	>2 Classes
Default model	0.998	88	11
1st Tuned model	0.999	36	4
2nd Tuned model	0.998	21	1

Table 2. RF model generalization performance.

The 2nd tuned model was determined to be the ideal model for the purpose of this research. It is directly incorporated into the final pipeline, where new, unseen data with an unknown label will be reclassified. To handle the remaining multiple class object IDs, a logic to take the majority class per frame is implemented. Consequently, the entire dataset will have only one true class for each object ID. The data overview of before and after prediction is displayed in Figure 9. After reclassification where all the object ID have 1 true class, the dataset is then uploaded to the database for the next processing phase.

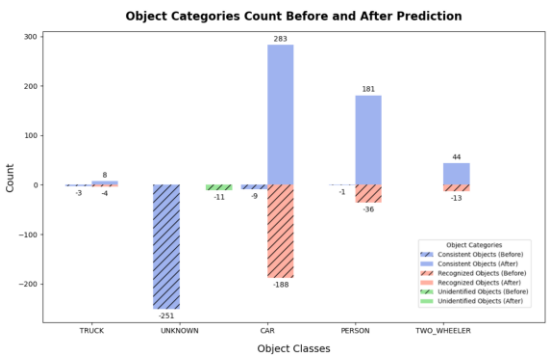

Figure 9. Object categories overview from reclassified dataset.

5.5 Traffic Simulation

To simulate traffic in real-time, the traffic simulation program SUMO fetches the parsed dataset from the database. As the methodology section explains, there are two types of simulation experiments: xml based and dynamic using TraCI. To map-match incomplete trajectories to the SUMO road network, the data fetch function additionally includes the map-match logic utilizing sumolib. By comparing the incomplete trajectories with the missing sequences, this map-matching algorithm finds the edges and lanes that are closest to the end or starting point.

XML dataset is then generated considering the map-match logic. There are five XML files produced in the workflow: *"passenger (CAR).trips.xml"*, *"truck(TRUCK).trips.xml"*, *"motorcycle(TWO WHEELER).trips.xml"*, and *"pedestrian(PERSON).trips.xml"*. In contrast to the other XML formats, the pedestrian XML format applies map-matching only to pedestrian walkways, denoted by lane 0 in the road network. Sumo-gui —a visualization tool for SUMO— then runs the simulation and opens the configuration automatically. Successful execution of the XML-based approach, as can be seen in footnote (2), results in an accurate simulation representation of the tracked objects from the .osef data, as shown in Figure 10.

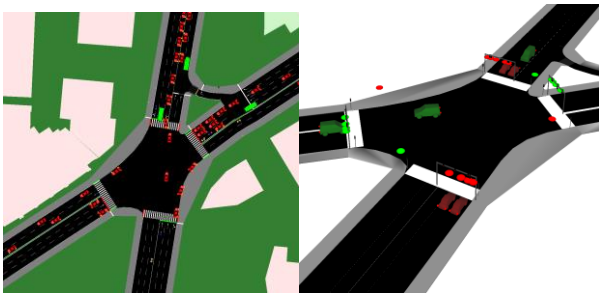

Figure 10. (Left) XML-based approach and (Right) Dynamic approach with TraCI for SUMO traffic simulation.

On the other hand, in a dynamic approach with TraCI shown in footnote (3), the database is fetched directly without the need to store data locally. This workflow employs TraCI as middleware. TraCI adds and removes vehicle IDs when they arrive at their destinations, updating them in real-time based on the timestep. The dynamic changes with TraCI (Figure 13 right image) have yet to be able to designate lanes like the XML-based approach does. Route distribution is necessary to identify certain lanes; nevertheless, considering the size of the dataset and the range of route distributions, this is a challenging process. However, the dynamic flow functions as intended, automatically determining the optimal lane for a vehicle.

6. Conclusion

The digital twin workflow in this research, utilizing the .osef dataset for traffic simulation, has proven to be successful. The .osef raw dataset was parsed by the algorithm, which then pre-processed it to identify object types, trajectories, and unknown object reclassifications before storing it in the database. The data fetching procedure, which involved direct input into the traffic simulation and database retrieval, facilitated a smooth real-time data flow.

By automating the processing, enrichment through reclassification and usage for simulation, informed decisions are

(2) https://vimeo.com/929027122?share=copy

(3) https://vimeo.com/929027117?share=copy

enabled to resolve issues with the traffic and facilitate transportation planning. Various traffic objects and their movements are captured, allowing precise tracking and management of traffic flow, which is crucial for efficient urban traffic management. The continuous object and trajectory reclassifications ensure that the data remains accurate and relevant for real-time applications. Moreover, the ability to handle and reclassify unknown objects enhances the reliability of the traffic monitoring system. Thus, the proposed pipeline can support traffic control systems, improve intersection safety, and reduce congestion by providing timely and actionable insights.

With minor modifications in road segmentation, this adaptable digital twin-based methodology can be applied to other similar areas. Once trained, the RF classification model can be saved and applied to similar data structures. The hybrid approach for trajectory classification, using map-matching and spatial analysis, effectively identifies incomplete trajectories. Although dynamic flow offers more flexibility, lane assignment issues can be resolved with clear route assignments. However, larger datasets may complicate the classification of incomplete trajectories, impacting real-time adjustments and processing times. Future research should explore the timing constraints imposed by .osef datasets.

In the future, direct data parsing from TCP streams instead of using downloading .osef datasets may improve the traffic simulation workflow's performance. Thus, real-time data can be directly handled in simulation. Additionally, further refinement can be made to improve the classification performance of the RF model, with an emphasis on finding features that can enhance the model further, since it has been noted that increasing the frames training dataset would not increase the model performance.

Acknowledgements

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under agreement no. 857155, the enRichMyData project, funded by Horizon Europe research and innovation programme under agreement no. 101070284, and the INTEND project, funded the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101135576.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Digital Twin Geohub. (2023). *Digital Twinning for Urban and Rural Environmental Modelling*. https://www.utwente.nl/en/digital-society/research/digitalisation/digital-twin-geohub/#vision

European Commission. (2022). *Bulgaria’s recovery and resilience plan*. https://commission.europa.eu/business-economy-euro/economic-recovery/recovery-and-resilience-facility/country-pages/bulgarias-recovery-and-resilience-plan_en

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

Fernandez, C. (2023). *This city has the worst traffic in the U.S.—and it’s actually a good thing: “Congestion shows the economy is moving.”* https://www.cnbc.com/2023/08/17/north-america-cities-highest-traffic-delays-inrix-report.html

Kušić, K., Schumann, R., & Ivanjko, E. (2023). A digital twin in transportation: Real-time synergy of traffic data streams and simulation for virtualizing motorway dynamics. *Advanced Engineering Informatics*, *55*, 101858. https://doi.org/10.1016/j.aei.2022.101858

Ministry of Regional Development and Public Work of Bulgaria. (2023). *Tables of persons registered by permanent and current address in municipality of Sofia*. General Directorate of Civil Registration and Administrative Services. https://www.grao.bg/tna/isnt41nm-15-06-2023-2.txt

Ministry of Transport and Communications of Republic of Bulgaria. (2017). *Integrated Transport Strategy for the period until 2030*. https://www.mtc.government.bg/sites/default/files/integrated_transport_strategy_2030_eng.pdf

Pishue, B. (2023). *2022 Global Traffic Scorecard: Congestion is Up Despite High Oil Prices - INRIX*. https://inrix.com/blog/2022-traffic-scorecard/

Rodrigues, M., Teoh, T., Ramos, C., De Winter, T., Knezevic, L., Marcucci, E., Lozzi, G., Gatta, V., Antonucci, B., Cutrufo, N., Marongiu, L., & Cré, I. (2021). *Research for TRAN Committee - Relaunching Transport and tourism in the EU after COVID-19*.

The Sofia Globe. (2022, June 2). *Bulgaria’s road death toll in first five months of 2022 is 175*. https://sofiaglobe.com/2022/06/01/bulgarias-road-death-toll-in-first-five-months-of-2022-is-175/

Vincent, K. (2023, February 23). *What is a 3D LiDAR Preprocessor?* https://www2.outsight.ai/insights/whats-a-3d-lidar-preprocessor?ref=lidar-insighter.com

Williams, K., Olsen, M. J., Roe, G., & Glennie, C. L. (2013). Synthesis of Transportation applications of Mobile LIDAR. *Remote Sensing*, *5*(9), 4652–4692. https://doi.org/10.3390/rs5094652

Zhang, K. (2020, May). *How Urban Transport is Changing in the Age of COVID-19*. https://news.climate.columbia.edu/2020/07/10/urban-transport-changing-covid-19/

Zheng, Y. (2015). Trajectory data mining. *ACM Transactions on Intelligent Systems and Technology*, *6*(3), 1–41. https://doi.org/10.1145/2743025