

MUTra-CDMX: Multisource Urban Traffic Dataset for the Insurgentes Sur Corridor in Mexico City

Arturo Rodríguez-Roman ^{1,2}, Alicia Martínez-Rebollar ^{1,*}, Hugo Estrada Esquivel ¹, Ernesto de la Cruz-Nicolás ^{1,3} and Eddie Clemente ¹ 

¹ Tecnológico Nacional de México, CENIDET, Cuernavaca 62490, Morelos, Mexico; d23ce176@cenidet.tecnm.mx (A.R.-R.); hugo.ee@cenidet.tecnm.mx (H.E.E.); ernesto.delacruz@cuautila.tecnm.mx (E.d.l.C.-N.); eddie.ct@cenidet.tecnm.mx (E.C.)

² Tecnológico de Iguala, Tecnológico Nacional de México, Iguala 40030, Guerrero, Mexico

³ Tecnológico de Cuautla, Tecnológico Nacional de México, Cuautla 62745, Morelos, Mexico

* Correspondence: alicia.mr@cenidet.tecnm.mx

Abstract

The growing complexity of urban mobility requires datasets that integrate dynamic traffic observations with meteorological, geometric, and urban-context information. This study presents MUTra-CDMX, a multisource urban traffic dataset covering a 14.72 km section of the Insurgentes Sur corridor in Mexico City. Traffic data were obtained from TomTom at five-minute intervals for 20 consecutive road segments from 1 November 2024 to 28 February 2025. Hourly meteorological data were retrieved from Meteosource, while segment-level geometry, topology, signalized locations, and nearby points of interest were derived from TomTom metadata and OpenStreetMap. The primary analytical file contains 691,200 segment–timestamp records and 12 variables describing traffic and free-flow conditions, meteorological information, derived operational indicators, and reconstruction status. Of these records, 682,264 are original observations and 8936 are reconstructed segment–timestamp combinations, identified by the Boolean variable *is_imputed*. Technical validation confirmed complete temporal coverage, preservation of original traffic observations, consistent weather alignment, and reconstruction performance through artificial masking. Predictive utility was evaluated through chronological travel-time forecasting under a leakage-controlled protocol. At the 30 min horizon, XGBoost achieved a mean absolute error of 12.84 s, a root mean squared error of 37.91 s, and a coefficient of determination (R^2) of 0.771, outperforming a persistence baseline. MUTra-CDMX supports congestion analysis, imputation studies, spatiotemporal modeling, and travel-time forecasting.

Dataset: <https://doi.org/10.5281/zenodo.21617011>.

Dataset License: Creative Commons Attribution 4.0 International (CC BY 4.0)

Keywords: urban traffic dataset; travel-time forecasting; multisource data; traffic-data imputation; spatiotemporal analysis; urban corridor; intelligent transportation systems; Mexico City

1. Introduction

Urban traffic prediction is a key component of Intelligent Transportation Systems (ITS) because it supports congestion management, route optimization, and decision-making in



Academic Editor: Tianwen Zhang

Received: 9 June 2026

Revised: 2 August 2026

Accepted: 4 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

complex urban environments [1–6]. Vehicular congestion increases travel times, energy consumption, and pollutant emissions while reducing travel-time reliability [7].

Despite advances in predictive models, the availability of suitable datasets remains a major limitation [8–10]. Widely used datasets, such as METR-LA, PEMS-BAY, and other PeMS variants [11], have played an important role in the evaluation of spatiotemporal forecasting methods. However, many of these resources focus primarily on traffic speed or flow and provide limited information on meteorological conditions, road geometry, traffic-control infrastructure, and the surrounding urban environment [12–14]. This limitation restricts the analysis of how external conditions and segment-level characteristics influence traffic behavior.

Recent traffic-prediction research has demonstrated the importance of representing spatial relationships, network structure, and auxiliary information. For example, software-defined vehicular-network approaches have been used to estimate traffic parameters and predict urban congestion [2], while auxiliary-information-aware graph convolutional models have incorporated weather and calendar information to capture dynamic spatiotemporal relationships [3]. Furthermore, deep learning architectures—such as bidirectional LSTM models applied to capture sequential traffic dependencies—have been extensively evaluated to improve forecasting accuracy under varying operational conditions [15]. These developments reinforce the need for datasets that provide not only historical traffic observations but also contextual and spatial information.

Dataset development has also moved toward multimodal representations. The Beijing Text-Traffic dataset (BjTT), for example, combines more than 32,000 traffic time-series records from over 1200 roads with textual descriptions of time, location, and events [16]. BjTT demonstrates the value of associating traffic measurements with contextual information for unusual-event representation and long-term prediction. However, its text-traffic structure and city-scale Beijing coverage differ from the segment-level meteorological, geometric, topographic, infrastructure, and reconstruction-provenance information provided by MUTra-CDMX.

Recent studies have emphasized the value of combining traffic observations with contextual information, including weather, road structure, points of interest, and network topology, to improve the representation of urban mobility conditions [13,14,17,18]. Nevertheless, openly available multisource datasets remain limited, particularly for Latin American cities, where differences in infrastructure, congestion patterns, interactions with public transport, and urban form may limit the direct transferability of datasets collected in other regions.

To address this need, this study presents MUTra-CDMX, a multisource urban traffic dataset covering a 14.72 km section of the Insurgentes Sur corridor in Mexico City. The dataset contains traffic-speed and travel-time observations for 20 consecutive road segments, recorded at five-minute intervals over 120 days, from 1 November 2024 to 28 February 2025. Traffic information was obtained from TomTom [19], hourly meteorological observations from Meteosource [20], and segment-level geometry, traffic-control infrastructure, topology, and nearby points of interest from TomTom metadata and OpenStreetMap [19,21].

The primary analytical file contains traffic and free-flow variables, meteorological information, derived operational indicators, and a record-status flag. Static geometric and urban-context attributes are provided separately as segment-level metadata and can be joined to the temporal observations through `segment_id`. This organization preserves the distinction between time-varying observations and fixed segment characteristics while supporting integrated spatiotemporal analysis.

MUTra-CDMX supports the training and evaluation of predictive models for travel-time forecasting across different horizons [17]. The inclusion of contextual variables enables researchers to examine relationships among traffic conditions, meteorological factors, road geometry, traffic-control infrastructure, and the surrounding urban environment [18].

The dataset-construction workflow involved several key processes: source-specific data acquisition, temporal-grid generation, missing-record detection, traffic-series reconstruction, meteorological alignment, spatial aggregation of contextual variables, and the calculation of relative-speed and congestion indicators. The original traffic dataset contained 682,264 observations. Temporal completion added 8936 missing segment-timestamp combinations to obtain the expected 691,200-record grid. Original observations were preserved without modification, reconstructed records were identified through `is_imputed`, and record-level reconstruction methods were documented in the provenance files.

The dataset was evaluated through complementary validation procedures addressing structural integrity, temporal completeness, preservation of original observations, meteorological alignment, reconstruction performance, and predictive utility. Artificial-masking experiments quantified reconstruction error across traffic variables and gap durations, while a chronological travel-time forecasting experiment assessed whether the dataset supports supervised prediction under a leakage-controlled protocol. At the 30 min horizon, an optimized XGBoost model reduced the mean absolute error (MAE) and root mean squared error (RMSE) while increasing the coefficient of determination (R^2) relative to a persistence baseline. Together, these validation results support the use of MUTra-CDMX for congestion analysis, imputation studies, spatiotemporal modeling, and travel-time forecasting.

The principal contribution of MUTra-CDMX is the integration of high-frequency traffic observations with hourly meteorological data, segment-level geometry and topography, signalized locations, nearby educational and healthcare-related points of interest, spatial topology, reconstruction provenance, artificial-masking validation, and leakage-controlled predictive-utility evidence within a single documented data package. By releasing an integrated and documented data package, MUTra-CDMX builds on earlier initiatives aimed at promoting the reuse of enriched urban traffic data that combine traffic observations with contextual information [13,14]. The dataset provides a resource for congestion characterization, spatiotemporal analysis, algorithm evaluation, and reproducible travel-time forecasting research [9,12], particularly in a Latin American urban context.

2. Data Description

The final dataset, consolidated in the primary file `MUTra_CDMX_completed.csv`, comprises 691,200 tabular observations collected along a 14.72 km section of the Insurgentes Sur corridor in Mexico City. It forms a temporally complete five-minute grid suitable for urban mobility analysis, data mining, and time-series modeling.

The observations are distributed across 20 road segments and cover a continuous period of 120 days, from 1 November 2024 at 00:00:00 to 28 February 2025 at 23:55:00. Traffic measurements were collected at five-minute intervals, yielding 288 observations per day and 34,560 records per segment over the four-month study period.

The primary file consists of 12 columns, including two spatiotemporal identifiers, four traffic and free-flow variables, three hourly meteorological variables, two derived traffic indicators, and one Boolean flag indicating whether a traffic record was reconstructed through imputation. It is complemented by segment-level metadata describing road geometry, functional class, topography, signalized locations, and the surrounding urban environment. Additional repository files provide segment topology, data provenance, validation results, a data dictionary, source documentation, and a file manifest.

2.1. Data Format

MUTra-CDMX is distributed as a structured collection of comma-separated value (CSV) files organized into four primary directories: data, metadata, provenance, and validation. The data directory includes the main dataset file, while the additional folders contain attributes specific to each segment, documentation related to data processing and integration, and the corresponding quality assessment results. This structure distinguishes the spatiotemporal data utilized for analysis and forecasting from the auxiliary information necessary for interpretation, traceability, and validation.

The main data file (data/MUTra_CDMX_completed.csv) comprises the time-indexed observations used for analysis and predictive modeling. Static attributes describing road geometry, functional class, topography, signalized locations, and surrounding educational and healthcare-related points of interest are stored separately in metadata/segment_metadata.csv and can be linked to the main file through segment_id. The spatial ordering and adjacency of the monitored road segments are provided in metadata/segment_topology.csv, allowing the temporal observations to be interpreted in relation to the physical structure of the corridor.

The provenance directory documents the origin and processing history of selected dataset components. Specifically, imputed_records.csv identifies the segment–timestamp combinations reconstructed during temporal completion, while weather_alignment_provenance.csv records the association between hourly meteorological observations and the corresponding five-minute traffic records. Additional files document the aggregation of points of interest and signalized locations. The validation directory contains the quality-control outputs used to assess temporal completeness, preservation of original observations, imputation performance, physical consistency, and weather alignment.

The main files in the MUTra-CDMX package are related through a small set of shared identifiers. As summarized in Table 1, segment_id links the spatiotemporal observations to the static segment characteristics and contextual attributes, whereas the composite key formed by segment_id and datetime identifies reconstructed records. Hourly meteorological provenance is associated with the traffic data through the corresponding hourly interval.

Table 1. Main files and linking keys in the MUTra-CDMX package.

File	Main Content	Linking Key
data/MUTra_CDMX_completed.csv	Spatiotemporal traffic, weather, and derived variables	segment_id + datetime
metadata/segment_metadata.csv	Static segment characteristics	segment_id
metadata/segment_topology.csv	Segment order and connectivity	segment_id
provenance/imputed_records.csv	Reconstructed records and applied methods	segment_id + datetime
provenance/weather_alignment_provenance.csv	Hourly weather alignment	weather_hour
provenance/poi_provenance.csv and signalized_locations_provenance.csv	Contextual-variable provenance	segment_id

Detailed variable definitions and file-specific relationships are provided in the data dictionary distributed with the repository.

2.2. Data Dictionary and Variable Organization

A complete file-level data dictionary is provided in metadata/data_dictionary.csv. For each field, the dictionary identifies the file in which it appears and specifies its variable name, data type, unit of measurement, description, source or generation method, permitted

values or expected range, missing-value policy, and relevant usage notes. The dictionary covers the main dataset, segment-level metadata, provenance records, and validation outputs, thereby providing the definitions required to interpret the complete data package.

The variables intended for traffic analysis and predictive modeling originate from different data sources and processing stages. Table 2 summarizes their organization according to analytical function and source or generation method.

Table 2. Organization and source of the analytical variables in MUTra-CDMX.

Variable Group	Variables	Source
Spatiotemporal identifiers and record status	segment_id, datetime, is_imputed	Dataset construction and temporal-completion workflow
Traffic variables	current_speed, free_flow_speed, current_travel_time, free_flow_travel_time	TomTom Traffic Flow data; missing segment-timestamp records reconstructed when required
Meteorological variables	temperature_c, precipitation_hourly_total_mm, wind_speed_ms	Meteosource historical weather data
Derived traffic indicators	relative_speed_index, congestion_index_0_10	Calculated from TomTom speed variables
Segment and urban-context variables	segment_length_m, functional_road_class, endpoint_gradient_percent, sinuosity, signalized_locations_count, education_poi_count_100m, health_poi_count_100m	TomTom segment information, OpenStreetMap, and geometric calculations

Traffic observations were obtained primarily from TomTom Traffic Flow data, meteorological variables from Meteosource historical records, and road and urban-context attributes from TomTom segment information, OpenStreetMap, and subsequent geometric calculations. Variables not directly supplied by these sources were generated through the documented derivation or temporal-completion procedures.

The first four variable groups are associated with individual segment-timestamp observations, whereas the segment and urban-context variables describe characteristics that remained constant for each road segment during the study period. These static attributes can be incorporated into the main time series through segment_id. The following subsections describe each analytical group in greater detail. Fields used exclusively for provenance, quality control, and file integrity are not reproduced in Table 2 because they support traceability and validation rather than direct traffic analysis. Their complete definitions remain available in the repository data dictionary.

2.3. Traffic Variables

Traffic variables describe the operating conditions of each road segment at five-minute intervals. Each record in the main dataset represents one road segment at one five-minute timestamp. The variable segment_id identifies the monitored road segment, while datetime records the corresponding date and time. Together, these variables form the composite key used to uniquely identify each spatiotemporal observation.

Traffic conditions are represented by current_speed, free_flow_speed, current_travel_time, and free_flow_travel_time. These variables were directly provided by TomTom based on the originally observed records. The current_speed and current_travel_time indicate the traffic conditions for a specific segment at a given timestamp. In contrast, free_flow_speed and free_flow_travel_time reflect the reference conditions that occur under uncongested traffic.

The original TomTom travel-time values were preserved whenever an observation was available. Reconstruction was applied only when an expected segment_id–datetime combination was absent during the temporal-completion process. In these cases, speed and travel-time values were reconstructed using the procedures outlined in Section 3.3.2. The Boolean variable `is_imputed` is used to differentiate between originally observed records (False) and those generated through the temporal-completion workflow (True). Detailed information on the reconstruction methods can be found in `provenance/imputed_records.csv`.

2.4. Meteorological Variables

The meteorological variables describe the hourly atmospheric conditions associated with the monitored corridor. Meteorological conditions are represented by `temperature_c`, `precipitation_hourly_total_mm`, and `wind_speed_ms`, obtained from Meteosource historical weather records. These variables describe air temperature in degrees Celsius, total accumulated precipitation during the corresponding hour in millimeters, and wind speed in meters per second, respectively.

The meteorological source has an hourly temporal resolution, whereas the traffic observations are recorded at five-minute intervals. Therefore, each hourly weather record was assigned to the twelve traffic timestamps contained within the interval $[h, h + 1)$ for every road segment. Temperature and wind speed were treated as hourly contextual conditions and were repeated without subhourly interpolation.

The precipitation variable represents the total precipitation reported for the entire hour. Its value was repeated across the corresponding five-minute traffic records, but was neither divided by twelve nor interpreted as precipitation accumulated during each five-minute interval. Consequently, users should not sum the repeated values within the same hour. The temporal association between weather and traffic records is documented in `provenance/weather_alignment_provenance.csv`.

2.5. Derived Traffic Indicators

Derived traffic indicators summarize the operating condition of each segment relative to its free-flow reference. Let i denote the road-segment index and t the timestamp. The current speed and free-flow speed are represented by $v_{i,t}$ and $v_{i,t}^{ff}$, respectively. These symbols correspond to the dataset variables `current_speed` and `free_flow_speed`. The `relative_speed_index` expresses the observed speed as a proportion of the free-flow speed:

$$R_{i,t} = \frac{v_{i,t}}{v_{i,t}^{ff}}, \quad (1)$$

where $R_{i,t}$ denotes the relative speed index, stored in the dataset as `relative_speed_index`.

Values close to 1 indicate that the observed speed is similar to the free-flow reference, whereas lower values indicate progressively greater speed reduction. Values above 1 may occur when the reported current speed exceeds the reference free-flow speed.

The variable `congestion_index_0_10` transforms the relative speed reduction into a non-negative congestion scale:

$$C_{i,t} = \max(0, 10(1 - R_{i,t})), \quad (2)$$

A value of 0 represents conditions at or above the free-flow reference, while increasing values indicate greater speed deterioration relative to that reference. Although the variable name identifies a 0–10 interpretive scale, users should consult the documented valid range and quality-control rules before assuming a fixed upper bound when processing the data externally.

Both indicators were calculated independently for each segment–timestamp record. For reconstructed traffic records, the indicators were derived only after the corresponding

speed values had been completed. In forecasting applications, derived indicators should be calculated or selected exclusively from information available at or before the prediction origin to prevent temporal leakage.

2.6. Segment-Level and Urban-Context Variables

Segment-level and urban-context variables describe the static physical, functional, and surrounding characteristics of each monitored road segment. Static characteristics of the monitored road segments are provided in `metadata/segment_metadata.csv`. Unlike the variables in the main time-series file, these attributes describe properties that remained constant for each segment during the study period and can be joined to the spatiotemporal observations through `segment_id`.

Road structure is described by `segment_length_m`, `functional_road_class`, `end-point_gradient_percent`, and `sinuosity`. The `segment_length_m` represents the nominal segment length in meters, while `functional_road_class` records the road hierarchy assigned to the segment. The `endpoint_gradient_percent` approximates the longitudinal gradient based on the elevation difference between the segment endpoints and the corresponding nominal length. `sinuosity` represents the ratio between the route length of the segment and the straight-line distance between its endpoints, providing an indicator of geometric curvature.

The surrounding infrastructure and urban context are represented by `signalized_locations_count`, `education_poi_count_100m`, and `health_poi_count_100m`. These variables report the number of signalized locations associated with each segment and the number of educational and healthcare-related points of interest identified within a 100 m surrounding area. They describe potential sources of recurring roadside activity and traffic friction rather than direct traffic measurements.

Segment-level characteristics were obtained from TomTom segment information, OpenStreetMap, and subsequent geometric or spatial calculations. Missing `sinuosity` values are retained as unavailable rather than being interpreted as a value of 1, because a missing geometric estimate does not necessarily indicate a perfectly straight segment. The applicable sources, calculation procedures, units, and missing-value policies are documented in `metadata/data_dictionary.csv` and the corresponding provenance files.

2.7. Dataset Coverage and Descriptive Statistics

Dataset coverage summarizes the spatial extent, temporal duration, observation frequency, and overall composition of MUTra-CDMX. The dataset covers 20 road segments with a combined nominal length of 14.72 km, and each segment contains 34,560 five-minute observations over the 120-day study period. Table 3 summarizes the length, number of records, average speed, average travel time, and average free-flow speed for each segment.

The descriptive statistics show substantial differences in operating conditions along the monitored corridor. Segments 18 and 19 are the longest, with nominal lengths of 2000 m and 1900 m, respectively, and have the highest average speeds, at 60.96 km/h and 57.23 km/h. In contrast, Segment 14 is the shortest segment, with a nominal length of 150 m, and has the lowest average speed, at 12.85 km/h. These differences demonstrate that the dataset captures distinct spatial and operational conditions across the monitored segments. The statistics were calculated from the completed dataset and therefore include both originally observed and reconstructed segment–timestamp records.

The operational distribution of the completed dataset was also examined using the relative speed index (RSI), stored in the dataset as `relative_speed_index`, which compares the current speed of each observation with its corresponding free-flow speed. Based on the thresholds summarized in Table 4, 60.80% of the records correspond to free-flow conditions,

22.27% to moderate-flow conditions, 15.34% to moderate congestion, and 1.59% to severe congestion. The presence of observations across all four operational states shows that the dataset represents both recurrent uncongested conditions and less frequent congestion episodes, supporting its use in traffic analysis and predictive modeling experiments.

Table 3. Spatial and operational profile of the monitored road segments.

Segment ID	Length (m)	Records	Mean Speed (km/h)	Mean Travel Time (s)	Mean Free-Flow Speed (km/h)
1	500	34,560	18.51	114.55	22.59
2	170	34,560	15.02	119.06	18.23
3	210	34,560	18.48	44.08	22.69
4	1100	34,560	22.75	190.12	26.31
5	270	34,560	15.19	77.44	20.09
6	500	34,560	21.43	98.78	32.39
7	800	34,560	21.41	139.81	25.61
8	750	34,560	20.67	141.21	25.15
9	200	34,560	34.75	22.84	40.28
10	170	34,560	13.96	44.61	17.29
11	1100	34,560	24.84	163.03	28.57
12	700	34,560	20.91	146.92	27.83
13	1900	34,560	31.16	172.28	34.49
14	150	34,560	12.85	39.59	16.69
15	350	34,560	19.05	54.16	25.69
16	1000	34,560	29.30	142.30	35.46
17	350	34,560	34.48	38.82	36.73
18	2000	34,560	60.96	120.37	61.56
19	1900	34,560	57.23	127.17	60.95
20	600	34,560	23.78	102.44	27.42
Total/ Average	14,720	691,200	25.84	104.98	30.30

Table 4. Distribution of completed records by operational state.

Operational State	RSI Criterion	Records	Records (%)	Interpretation
Free Flow	$RSI \geq 0.80$	420,256	60.80	The segment preserves at least 80% of its free-flow speed.
Moderate Flow	$0.60 \leq RSI < 0.80$	153,920	22.27	There is a relatively small drop in speed from free-flow circumstances.
Moderate Congestion	$0.30 \leq RSI < 0.60$	106,056	15.34	Speed is significantly reduced relative to free-flow conditions.
Severe Congestion	$RSI < 0.30$	10,968	1.59	The segment operates well below its reference speed.

3. Proposed Methods

The construction of the MUTra-CDMX dataset integrates information from multiple sources and organizes it systematically by road segment and timestamp. The main goal is to ensure consistent alignment between traffic observations, meteorological conditions, and complementary segment-level characteristics. The dataset production process comprises six distinct stages designed to support transparency and reproducibility, as described below.

3.1. Description of the Study Area

The Insurgentes Sur corridor was selected as the study area because it exhibits a wide range of traffic conditions influenced by multiple signalized intersections, dedicated public transport infrastructure (Metrobus, a bus rapid transit system), school and commercial zones, and variations in topography. Together, these factors contribute to heterogeneous

traffic behavior and make the corridor representative of complex urban mobility environments. Figure 1 shows the location and segmentation of the monitored section.

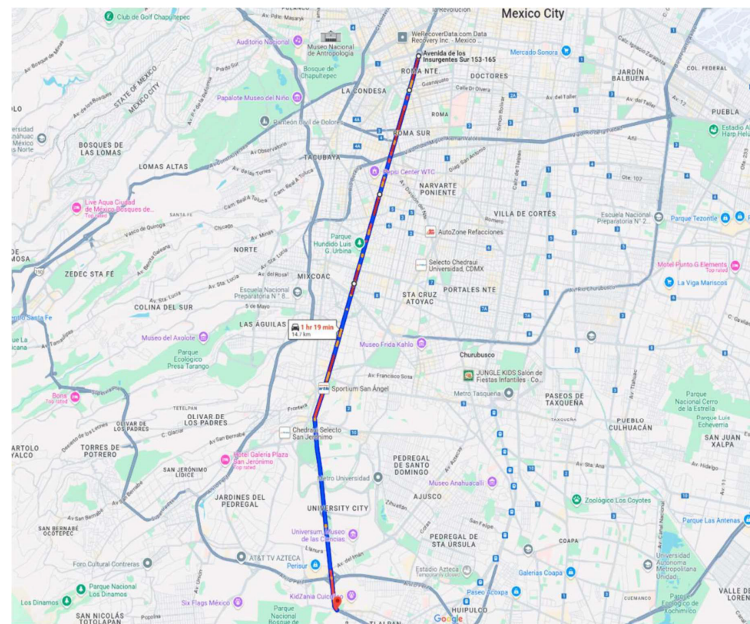


Figure 1. Location and segmentation of the monitored section of the Insurgentes Sur corridor.

The monitored area covers a total length of 14.72 km and is divided into 20 directional road segments, as defined by the TomTom Traffic Flow service architecture. This segmentation provides spatial granularity and allows traffic conditions to be characterized at the segment level rather than treating the corridor as a single homogeneous unit. Each segment therefore constitutes a spatial observation unit that associates dynamic traffic measurements with fixed road and urban-context attributes, including nominal length, functional road class, endpoint gradient, sinuosity, signalized locations, and nearby points of interest.

The order and adjacency of the 20 segments define the spatial structure of MUTra-CDMX. Each traffic record represents the operating condition of one segment at a specific timestamp, while metadata/segment_topology.csv documents the corresponding upstream and downstream relationships. This structure supports both segment-level temporal analysis and methods that incorporate spatial relationships along the monitored corridor.

3.2. Data Collection Procedure

The development of MUTra-CDMX began with the collection of data from independent sources, with the aim of capturing various aspects of urban traffic behavior. At this stage, the information had not yet been integrated into a unified structure; each source generated its dataset in separate tables, corresponding to different aspects of the phenomenon: dynamic traffic conditions, meteorological variables, and physical or urban characteristics of the corridor.

Data collection was organized into three main groups: traffic data obtained through TomTom's Traffic Flow service [19], meteorological data retrieved through MeteSource Weather's Time Machine service [20], and geospatial attributes and urban infrastructure derived from OpenStreetMap's Overpass service [21]. Maintaining these sources separately during collection preserved the provenance of each variable and allowed its spatial or temporal resolution to be documented before the subsequent cleaning and integration stages.

3.2.1. Traffic Data

To collect dynamic traffic data, geographic query points were defined for each of the 20 segments of the corridor. Each point served as a spatial reference for retrieving the operational status of the corresponding road segment. Queries were performed using the TomTom Traffic application programming interface (API), specifically the Flow Segment Data service.

The geographic query points used for traffic-data acquisition correspond to the segment coordinates documented in the file `metadata/segment_metadata.csv`. These points were used to request traffic conditions for the road fragment nearest to each query location. The request followed the structure required by the Flow Segment Data endpoint. A simplified example of a request used in the data collection process is shown below, with the access key anonymized for security reasons:

https://api.tomtom.com/traffic/services/4/flowSegmentData/absolute/10/json?point=19.40592172314550,-99.16861928048740&key=YOUR_API_KEY (accessed on 1 November 2024)

The endpoint is composed of the following elements: the TomTom API server (<https://api.tomtom.com>), the Traffic API service path and version (`/traffic/services/4/flowSegmentData`), the flow style (`absolute`), the zoom level (`10`), the response format (`json`), the access key (`key`), and the geographic query point (`point`). The point parameter contains the latitude and longitude used to retrieve the traffic-flow information for the nearest road fragment.

Queries to the Traffic Flow service were executed periodically every five minutes during the period from 1 November 2024 to 28 February 2025. For each valid response, TomTom directly provided `current_speed`, `free_flow_speed`, `current_travel_time`, and `free_flow_travel_time`. The current variables describe the operating conditions reported for the corresponding segment and timestamp, whereas the free-flow variables represent the reference conditions supplied by the service. Consequently, the travel-time values in the original records were preserved as reported by TomTom and were not systematically recalculated from nominal segment length and speed.

During this stage, the data were retained with their capture timestamp and spatial segment identifier. Periodic real-time acquisition did not yield an observation for every expected segment–timestamp combination, resulting in gaps in the original traffic series. These missing combinations were subsequently identified against the complete five-minute temporal grid and reconstructed during the temporal-completion stage described in Section 3.3. All successfully retrieved TomTom observations were preserved without modification. In the completed dataset, `is_imputed` distinguishes the original observations (`False`) from the segment–timestamp records added through the reconstruction workflow (`True`).

3.2.2. Meteorological Data

Meteorological data describe the atmospheric conditions associated with the monitored corridor during the traffic-observation period. These data were retrieved on demand for the same period during which traffic observations were recorded, using the central coordinates of the study corridor as the reference location. The data were obtained through the Meteosource Weather API, specifically the Time Machine service, which provides historical weather information for a specified location and date. Unlike the traffic data, which were collected every five minutes, the meteorological information was obtained at hourly resolution and later aligned with the five-minute traffic time grid during the data integration process. A simplified example of a request to the Time Machine service is shown below, with the access key anonymized for security reasons:

https://www.meteosource.com/api/v1/flexi/time_machine?lat=19.422086&lon=-99.163494&date=2024-11-01&timezone=America/Mexico_City&units=metric&key=YOUR_API_KEY (accessed on 1 March 2024)

The endpoint is composed of the following elements: the Meteosource API server (<https://www.meteosource.com>), the API service path and version (/api/v1/flexi/time_machine), the geographic reference location defined by latitude and longitude (lat and lon), the historical date requested (date), the time zone used to interpret the timestamp (timezone), the unit configuration (units), and the access key (key). The lat and lon parameters contain the central coordinates of the study corridor used to retrieve the meteorological conditions for the selected date. The access key was anonymized and was not stored in the dataset or released in the public repository.

The retrieved variables were air temperature, hourly accumulated precipitation, and wind speed, subsequently stored as temperature_c, precipitation_hourly_total_mm, and wind_speed_ms, respectively. The request used the America/Mexico_City time zone and metric units so that the weather timestamps were directly comparable with the local traffic timestamps. During the study period, Mexico City remained on standard time, and no daylight-saving-time transition occurred. At this stage, the weather observations were retained as an independent hourly data source and were not interpolated to five-minute intervals. Their subsequent alignment with the traffic records is described in Section 3.4.

3.2.3. Road Geometry and Urban-Context Data

Road geometry and urban-context data were incorporated to describe the static physical and surrounding characteristics of each monitored road segment. Segment coordinates, nominal length, and functional road class were obtained from the TomTom segment information used during traffic-data acquisition. These attributes were complemented with topographic, geometric, traffic-control, and point-of-interest (POI) information derived from external geospatial sources.

OpenStreetMap data were retrieved through the Overpass API to identify signalized locations and nearby educational and healthcare-related points of interest along the monitored corridor. The queries covered the geographic extent of the 20 road segments and selected OpenStreetMap features associated with traffic signals, educational facilities, and healthcare facilities. The retrieved entities were then spatially associated with the corresponding road segments according to their location relative to the segment geometry.

For each road segment, a geographic search area was defined using its start and end coordinates, which are documented in metadata/segment_metadata.csv. A bounding box enclosing the segment was generated and expanded by 0.0006° in each direction to retrieve nearby OpenStreetMap entities. Separate queries were executed for traffic-control elements, educational facilities, and healthcare facilities.

The procedure involved running separate queries by feature type and associating the retrieved results with the corresponding road segment. Traffic signals were identified using OpenStreetMap tags related to traffic control, whereas schools and hospitals were retrieved as points of interest located near each road segment. Following previous geographic information system (GIS)-based urban mobility studies that use POI-derived built-environment indicators around road segments to analyze travel behavior and road-speed variability, a conservative search margin smaller than 100 m was adopted to identify educational and healthcare facilities in the immediate surroundings of each segment [22–24]. The implemented expansion of 0.0006° corresponds to approximately 63 m in the east–west direction and 67 m in the north–south direction on each side of the segment bounding box.

A simplified example of an Overpass query used to retrieve schools within a bounding box is shown below:

[https://overpass.kumi.systems/api/interpreter?data=\[out:json\]\[timeout:30\]](https://overpass.kumi.systems/api/interpreter?data=[out:json][timeout:30]) (accessed on 8 June 2026);

node["amenity"="school"]({south_lat},{west_lon},{north_lat},{east_lon});out;

The request includes the following elements: the Overpass API server (<https://overpass.kumi.systems/api/interpreter> (accessed on 8 June 2026)), the output format (out:json), the timeout parameter (timeout:30), the OpenStreetMap element type to be searched (node), the amenity filter used to identify the selected point of interest, and the geographic bounding box defined by {south_lat}, {west_lon}, {north_lat}, and {east_lon}. The complete retrieval procedure considered node, way, and relation elements for educational and healthcare facilities. The bounding box defines the initial retrieval area around each road segment. The retrieved entities were then spatially filtered and aggregated according to the criteria established for each contextual variable.

Educational and healthcare-related points of interest were counted and stored as education_poi_count_100m and health_poi_count_100m, respectively. Duplicate POI records were identified using their OpenStreetMap identifier and element type before segment-level aggregation. Traffic-control infrastructure was represented by signalized_locations_count, which records the number of consolidated signalized locations associated with each segment. Because OpenStreetMap may contain multiple records referring to the same physical intersection or signalized location, nearby signal records were consolidated before the segment-level counts were calculated. The original entities, consolidation results, and segment assignments are documented in the corresponding provenance files.

Additional segment-level characteristics, including endpoint_gradient_percent and sinuosity, were calculated from the available segment coordinates, endpoint elevations, nominal route length, and geodesic distance. These calculations are described in the subsequent metadata-construction stage. The resulting static attributes were stored in metadata/segment_metadata.csv and can be linked to the main spatiotemporal dataset through segment_id.

Once data collection was completed, the traffic, meteorological, and geospatial data remained stored as independent sources. Because the traffic observations constitute the temporal and operational core of MUTra-CDMX, data cleaning and temporal completion were applied to the traffic series before multisource integration.

3.3. Data Cleaning

Given that traffic observations are the temporal basis of MUTra-CDMX, the data cleaning process started with an assessment of the temporal completeness of this dataset. The purpose of this stage was to identify missing records, null values, duplicate segment_id-datetime keys, or discontinuities in the observation series for each road segment before integrating external meteorological and infrastructure data sources.

Considering that data acquisition was performed at a scheduled interval of five minutes, each segment should have recorded 288 observations per day. For the 20 segments and the entire study period, covering the period from 1 November 2024 to 28 February 2025, the expected complete temporal structure comprised 691,200 records. A complete five-minute grid was therefore generated for every segment and compared with the originally retrieved traffic observations to explicitly identify all missing segment-timestamp combinations.

3.3.1. Missing Data Detection

Initial detection of missing data was performed by comparing the expected time series structure with the records actually observed in the main traffic table. The original file contained 682,264 observations, so 8936 missing segment-timestamp combinations were

identified out of the expected total of 691,200 records. This difference corresponds to an initial coverage of 98.71%.

The presence of gaps in the traffic series was to be expected, since the dataset was generated through periodic real-time queries. Consequently, some scheduled acquisitions did not yield a traffic observation for every expected segment–timestamp combination. The specific cause of each missing interval could not be determined from the retained traffic records; however, real-time acquisition may be affected by connectivity interruptions, unavailable server responses, or temporary absence of valid information from the provider. In contrast, meteorological data and infrastructure attributes did not exhibit the same type of gaps. Weather conditions were obtained through historical requests for the entire study period, while urban infrastructure data consisted of geospatial information associated with each segment rather than continuous high-frequency data capture.

During this stage, it was found that the missing records were not distributed randomly. Rather than appearing as isolated omissions, the missing data were grouped into 23 clearly defined time blocks. To analyze this distribution, heat maps of missing hours were generated by segment and day. Figure 2 shows that the absences are concentrated in specific intervals, allowing continuous interruptions to be distinguished from shorter gaps and isolated missing records.

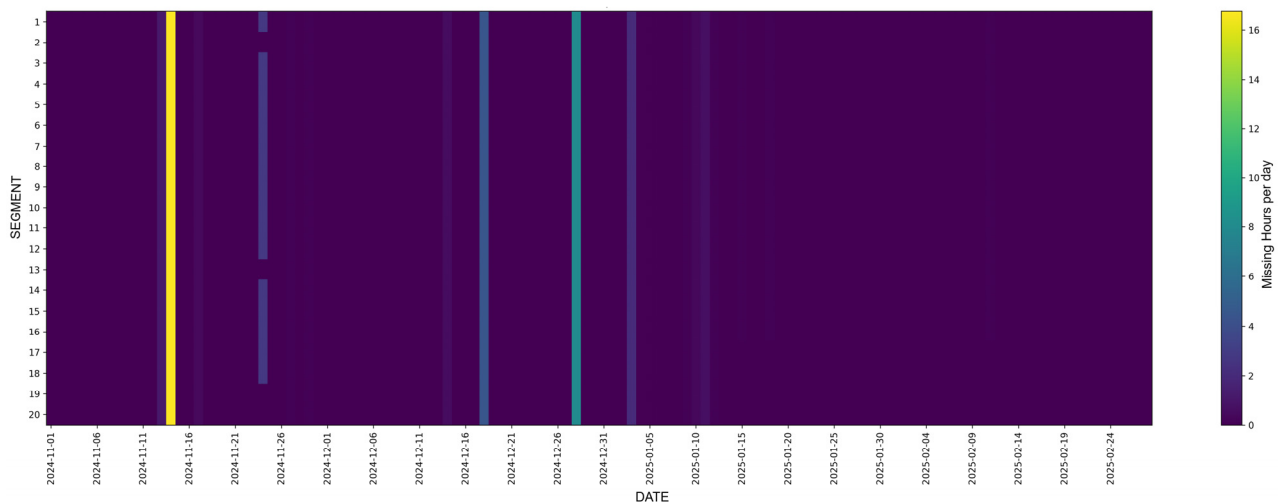


Figure 2. Heat map of missing hours by segment and day.

Analysis of the time blocks showed that only four critical blocks accounted for most of the missing records, as summarized in Table 5. This concentration demonstrated that not all gaps had the same impact on the temporal integrity of the dataset.

Table 5. Critical blocks of missing data in the traffic series.

Date	Start Time	End Time	Duration (h)	Affected Segments	Missing Records	Cumulative Missing Segment-Hours
14 November 2024	07:40	17:40	10.00	20	2400	200.00
28 December 2024	08:20	16:25	8.08	20	1940	161.67
13 November 2024	22:50	06:40	7.83	20	1880	156.67
18 December 2024	11:25	15:50	4.42	20	1060	88.33

Note: The period beginning on 13 November 2024 extends until 6:40 a.m. on 14 November, spanning two consecutive days.

Combined, these four blocks accounted for 7280 missing records, equivalent to 606.67 cumulative hours of missing data across the 20 affected segments. This represents approximately 81.47% of all missing records identified in the original traffic table.

This evidence confirmed that most of the missing data did not correspond to scattered failures, but rather to prolonged interruptions that simultaneously affected all segments of the corridor.

The identification of these blocks allowed the gaps to be classified according to their duration and temporal impact. Short gaps could be addressed using nearby information within the same time series, while prolonged blocks required a more conservative approach to avoid generating artificial values not supported by immediate neighboring observations. This classification served as the basis for defining the temporal-completion strategy described in the following section.

3.3.2. Missing Data Imputation

Based on the above analysis, the missing data were classified according to their duration into three intervals associated with the reconstruction policy. Short gaps corresponded to sequences lasting up to 60 min. These cases retained sufficiently close valid observations to support interpolation within the same segment. Intermediate gaps were defined as intervals lasting from 65 to 120 min, for which a persistence strategy was considered. Long gaps exceeded 120 min and required historical temporal information because immediate neighboring observations were insufficient to represent the missing traffic conditions.

This classification enabled the application of a hybrid imputation strategy. For gaps of up to 60 min, linear interpolation by segment was applied, provided that valid values existed before and after the missing interval. For gaps from 65 to 120 min, the defined strategy consisted of persistence using the nearest valid observation; however, no actual gaps within this duration range were identified, and therefore this method was not applied to the final missing records. For gaps longer than 120 min, an approach based on the historical temporal profile was employed, calculated solely from valid original observations. The procedure followed a hierarchical support strategy that prioritized, first, comparable records from the same segment, day of the week, hour, and minute; next, patterns from the same segment by hour and minute, from the same segment by hour, from the entire segment, global patterns by hour and minute, global patterns by hour, and finally, the global median of the variable when there were insufficient matches.

Linear interpolation was applied to the speed variable because it directly represents the traffic conditions of each road segment. More formally, for a missing value at time t , located between two valid observations t_0 and t_1 , the imputed value was estimated as:

$$\hat{v}_{i,t} = v_{i,t_0} + \frac{t - t_0}{t_1 - t_0} (v_{i,t_1} - v_{i,t_0}), \quad (3)$$

where $\hat{v}_{i,t}$ denotes the reconstructed speed of segment i at time t , while v_{i,t_0} and v_{i,t_1} denote the valid speed observations immediately before and after the missing interval, respectively.

Once the speed series had been completed, missing current travel-time values were reconstructed according to the physical-consistency characteristics of each segment. For segments whose original TomTom observations were consistent with the nominal segment length, speed expressed in kilometers per hour was converted to meters per second. The reconstructed travel time was calculated as:

$$\hat{T}_{i,t} = \frac{3.6L_i}{\hat{v}_{i,t}}, \quad (4)$$

where $\hat{T}_{i,t}$ denotes the reconstructed travel time, in seconds, for segment i at time t ; L_i is the nominal segment length in meters; and $\hat{v}_{i,t}$ is the reconstructed speed in kilometers per hour. The factor 3.6 accounts for the conversion between kilometers per hour and meters per second. The result is expressed in seconds. This nominal-length derivation was used for 7640 reconstructed records in regular segments. This reconstruction ensured consistency

among the main physical variables in the dataset, preventing the reconstructed travel time from becoming disconnected from the actual length of the segment and the imputed speed.

For Segments 2, 13, and 15, the original TomTom observations exhibited persistent effective-coverage regimes that were not consistently represented by a single nominal-distance relationship. For 1276 nonambiguous missing records in these segments, travel time was reconstructed using the corresponding regime-conditioned physical relationship. The remaining 20 regime-ambiguous records were reconstructed through local interpolation of the travel-time series.

In addition to `current_speed` and `current_travel_time`, the reference variables associated with free-flow conditions, `free_flow_speed` and `free_flow_travel_time`, were completed to maintain the consistency of the traffic records and the derived indicators that depend on them. Because these variables represent reference conditions for the segment and exhibit relatively stable behavior, missing values were imputed using the observed median per segment, calculated solely from valid original records. When no value was available for a given segment, the overall median of the corresponding variable was used as a fallback.

At the end of this stage, the dynamic traffic data were organized into uniform five-minute intervals per segment, with no missing values in the main variables, and were ready for the subsequent integration of weather, geometry, and urban infrastructure data. The reconstruction process added 8936 missing segment-timestamp records, while all 682,264 original TomTom observations and their traffic values were preserved without modification. Each added record was identified through `is_imputed`, and its specific reconstruction methods were documented in `provenance/imputed_records.csv`.

3.4. Climate, Geometry, and Infrastructure Data Integration

Once the traffic time series had been cleaned, external sources were incorporated into the main table or linked through complementary metadata so that the completed traffic observations could be analyzed together with dynamic, environmental, and urban information. The traffic dataset, organized into uniform five-minute intervals per segment, was used as the reference time frame.

Meteorological variables, such as temperature, precipitation, and wind speed, were incorporated by associating them with the corresponding hourly interval $[h, h + 1)$, preserving their original resolution without applying additional interpolation. Each hourly temperature and wind-speed value was assigned to the twelve five-minute traffic observations within the corresponding hour for every segment. Hourly precipitation was also repeated across these records as the total reported for the complete hour; it was not divided among the twelve intervals.

In parallel, the static attributes of each segment, such as nominal length, functional road class, signalized locations, and nearby educational and healthcare-related points of interest, were organized at the segment level and linked to the traffic observations through `segment_id`. These attributes were retained in `metadata/segment_metadata.csv` rather than replicated across all timestamps in the main dataset. The main file therefore combines traffic and meteorological variables at the segment-timestamp level, while the complementary metadata file provides the corresponding road-geometry, infrastructure, and urban-context characteristics.

As a result, MUTra-CDMX was organized as a multisource data package in which each record in the main table was defined by a segment and a timestamp and could be complemented with static segment attributes using the shared identifier. This structure provided the basis for the subsequent stage of feature engineering, in which derived variables were constructed to represent relative speed and congestion conditions, while geometric characteristics were calculated and retained as segment-level metadata.

3.5. Feature Engineering

Following the integration of dynamic and contextual data sources, a feature engineering process was applied to represent traffic operation and segment-level physical conditions. This stage enabled the development of derived indicators that facilitate comparisons between segments with varying free-flow conditions, geometry, infrastructure, and surrounding urban characteristics.

First, two geometric attributes of the segment were introduced: sinuosity and gradient. Sinuosity represents the degree to which a road segment deviates from a straight line between its endpoints, serving as a contextual attribute for analyzing variations in speed and travel time. The literature recognizes the fact that the geometric properties of the road are relevant for estimating travel time and explaining differences in operating speed [14,17]. To calculate sinuosity, the mathematical notation was mapped directly to the variables in the dataset: the segment identifier i corresponds to `segment_id`; S_i denotes the sinuosity value obtained from this operation and subsequently stored in `sinuosity`; and the nominal segment length L_i corresponds to `segment_length_m`. The straight-line distance D_i was derived from the geometric metadata of each segment, specifically from the start and end coordinates of its geometry. Sinuosity was computed as:

$$S_i = \frac{L_i}{D_i}, \quad (5)$$

A value of $S_i = 1$ indicates that the nominal route length is equal to the straight-line distance between the segment endpoints. Higher values mean that the segment deviates more from the straight-line path. If the available coordinates and nominal segment length did not allow for a valid sinuosity estimate, the value was kept as missing instead of being assigned a value of 1.

Gradient represents the longitudinal inclination of a road segment and is defined as the ratio between the change in elevation and the length of the segment. This variable incorporates a topographic condition of the corridor into the dataset. Operationally, segments with higher gradients can affect traction effort, energy consumption, and operating speed, particularly for heavy vehicles or under low-speed conditions. For this reason, gradient was included as a geometric attribute representing a potential source of physical resistance along the segment.

To compute this variable, the same base notation defined for the geometric attributes of the segment was retained: the index i identifies the road segment, and L_i represents its nominal length in meters. In addition, the initial and final elevations of the segment, $z_{start,i}$ and $z_{end,i}$, were incorporated as topographic metadata stored in meters. Using these values, the signed gradient of the segment was calculated as a percentage according to:

$$G_i = 100 \left(\frac{z_i^{end} - z_i^{start}}{L_i} \right), \quad (6)$$

where G_i denotes the gradient of segment i , expressed as a percentage; z_i^{start} and z_i^{end} are the elevations, in meters, at the start and end of the segment; and L_i is the nominal segment length in meters. Positive values indicate an increase in elevation from the start to the end of the segment, whereas negative values indicate a decrease. Because the calculation uses only the endpoint elevations, G_i represents an approximation of the segment-level longitudinal gradient rather than the complete road-elevation profile.

The relative speed index and congestion indicator were calculated according to Equations (1) and (2), respectively. For each segment i and timestamp t , $R_{i,t}$ represents the ratio between the current speed $v_{i,t}$ and the corresponding free-flow speed $v_{i,t}^{ff}$, while $C_{i,t}$ represents the non-negative congestion indicator derived from $R_{i,t}$. The resulting values were stored as `relative_speed_index` and `congestion_index_0_10`, respectively.

Values of $R_{i,t}$ close to 1 indicate operation near free-flow conditions, whereas lower values reflect a relative reduction in speed and potential congestion. Values above 1 may occur when the reported current speed exceeds the corresponding free-flow reference. A value of $C_{i,t} = 0$ indicates operation at or above the free-flow reference, whereas increasing values represent progressively greater speed deterioration. The factor of 10 scales the relative speed reduction into an interpretable operational range.

Once the derived variables were computed, a final verification step was performed to ensure that the derived dynamic indicators did not introduce unexpected missing values into the main spatiotemporal file. The two dynamic indicators, `relative_speed_index` and `congestion_index_0_10`, were stored in the main spatiotemporal file, whereas `sinuosity` and `endpoint_gradient_percent` were retained in `metadata/segment_metadata.csv`. The resulting multisource data package preserves a continuous five-minute traffic and weather series that can be complemented with segment-level geometry, infrastructure, and urban-context attributes through `segment_id`.

4. Data Records

The MUTra-CDMX dataset is publicly available through the Zenodo repository at <https://doi.org/10.5281/zenodo.21617011>. The Zenodo record provides the complete data package as `MUTra_CDMX_v4.0.rar`. The compressed package contains the primary analytical file, `data/MUTra_CDMX_completed.csv`, which includes 691,200 records organized in a spatiotemporal tabular structure. Each row represents the traffic and associated meteorological conditions of a specific road segment at a given timestamp, with observations collected every five minutes from 1 November 2024, to 28 February 2025. The dataset covers 20 road segments and the primary file contains 12 variables comprising spatiotemporal identifiers, traffic and free-flow conditions, hourly meteorological information, derived traffic indicators, and the `is_imputed` flag used to distinguish original and reconstructed records.

The repository includes supporting documentation to help users interpret and reuse the dataset. Static segment-level attributes can be found in `metadata/segment_metadata.csv`. These attributes encompass start and end coordinates, nominal segment length, functional road class, endpoint elevations, endpoint gradient, sinuosity, consolidated signalized-location counts, and points of interest related to education and healthcare in the vicinity. The spatial order and adjacency of the monitored road segments are detailed in `metadata/segment_topology.csv`. These complementary files can be linked to the primary analytical file through `segment_id`.

Additional documentation and provenance files describe the origin, processing, and validation of the published data. The data dictionary, provided in `metadata/data_dictionary.csv`, specifies variable names, descriptions, data types, units, source or generation method, expected ranges, and missing-value policies. `provenance/imputed_records.csv` identifies each reconstructed segment–timestamp record and the method applied, while the `weather`, `point-of-interest`, and `signalized-location` provenance files document the corresponding alignment and aggregation procedures. The repository also contains quality-control outputs, predictive-utility results, source documentation, and a file manifest with integrity information.

Together, the repository files provide the complete information required to reproduce the dataset structure, understand the origin and meaning of each variable, and use MUTra-CDMX for exploratory analysis, congestion characterization, imputation studies, feature engineering, and travel-time prediction tasks.

5. Technical Validation

Technical validation was conducted to assess the structural integrity, temporal completeness, reconstruction quality, multisource alignment, and predictive utility of MUTra-CDMX. The validation workflow included verification of the segment–timestamp grid, preservation of the original TomTom observations, assessment of the weather-integration procedure, artificial-masking experiments for the reconstructed traffic variables, and a chronological travel-time forecasting experiment. Detailed validation outputs are included in the validation directory of the repository.

5.1. Dataset Integrity and Temporal Completeness

Dataset integrity and temporal completeness were evaluated against the expected five-minute grid for the 20 monitored segments and the complete 120-day study period. The final dataset contains the expected 691,200 unique segment–timestamp combinations, corresponding to 34,560 records per segment. No duplicate composite keys or missing timestamps remained after temporal completion.

The preservation audit confirmed that all 682,264 originally retrieved TomTom observations were retained without modification. The completed dataset added only the 8936 previously absent segment–timestamp combinations, each identified through *is_imputed*. Comparisons between the original and completed files verified that the current-speed, free-flow-speed, current-travel-time, and free-flow-travel-time values associated with the original keys remained unchanged.

Additional checks verified chronological ordering, five-minute continuity within each segment, valid ranges for the traffic variables, and consistency of the derived indicators, as summarized in Table 6. Detailed results are provided in the corresponding completeness, preservation, and quality-control files distributed in the validation directory.

Table 6. Structural and temporal integrity of the completed dataset.

Validation Item	Result
Expected segment–timestamp records	691,200
Final records	691,200
Records per segment	34,560
Original observations preserved	682,264
Reconstructed records added	8936
Duplicate composite keys	0
Missing segment–timestamp combinations	0
Final temporal coverage	100%

5.2. Weather-Integration Validation

Weather integration was validated to confirm temporal consistency between the hourly Meteosource observations and the five-minute traffic records. All meteorological timestamps were expressed in the America/Mexico_City time zone, matching the local-time reference used for the traffic dataset. Each hourly weather observation at time h was associated with the traffic records whose timestamps fell within the interval $[h, h + 1)$.

Because the dataset contains 20 road segments and twelve five-minute intervals per hour, each hourly meteorological observation was expected to be associated with 240 traffic records. The validation confirmed this relationship throughout the study period. Temperature and wind speed were repeated as hourly contextual values without subhourly interpolation, while precipitation was preserved as the total reported for the entire hour and was not divided among the twelve five-minute intervals.

The audit also verified that every traffic record had an associated meteorological observation, that no timestamp was assigned to more than one hourly interval, and

that the hourly precipitation values remained unchanged during integration, as detailed in Table 7. Detailed alignment records and validation summaries are provided in provenance/weather_alignment_provenance.csv and the corresponding files in the validation directory.

Table 7. Summary of weather-integration validation.

Validation Item	Result
Weather time zone	America/Mexico_City
Weather resolution	1 h
Traffic resolution	5 min
Five-minute intervals per hour	12
Road segments	20
Traffic records per weather hour	240
Subhourly interpolation	Not applied
Missing weather assignments	0
Weather-alignment status	PASS

5.3. Imputation Validation

The reconstruction strategy was evaluated through artificial-masking experiments applied to originally observed traffic records. In each experiment, known observations were temporarily removed, reconstructed using the same rules as those applied during dataset completion, and then compared with their original values. The validation considered five random seeds, seven gap durations, and the four traffic variables included in the completed dataset: `current_speed`, `current_travel_time`, `free_flow_speed`, and `free_flow_travel_time`.

In total, 28 global variable–duration combinations were evaluated. Across these combinations, the mean absolute percentage error (MAPE) ranged from 7.00% to 17.54% for `current_speed`, from 9.27% to 20.61% for `current_travel_time`, from 1.98% to 3.09% for `free_flow_speed`, and from 4.44% to 5.45% for `free_flow_travel_time`. The results show that reconstruction error generally increased with gap duration and was higher for the dynamic traffic variables than for the more stable free-flow variables.

Segment-level validation revealed that reconstruction difficulty was not uniform across the corridor. Some segments and longer artificial gaps produced higher errors, particularly where traffic behavior showed greater temporal variability. These cases were retained in the validation outputs to provide a transparent description of the limitations of the reconstruction procedure rather than reporting only global averages.

A separate preservation audit confirmed that the 682,264 originally observed records were not overwritten or modified. The completion process added only the 8936 missing segment–timestamp combinations, as summarized in Table 8. Each added record is identified by `is_imputed = True`, while the corresponding reconstruction method and confidence level are documented in provenance/imputed_records.csv. Detailed results by variable, gap duration, random seed, and road segment are provided in the imputation-validation files distributed in the validation directory.

Table 8. Summary of artificial-masking validation results.

Variable	Mean MAPE Range
<code>current_speed</code>	7.00–17.54%
<code>current_travel_time</code>	9.27–20.61%
<code>free_flow_speed</code>	1.98–3.09%
<code>free_flow_travel_time</code>	4.44–5.45%

5.4. Predictive Utility Assessment

The predictive utility of MUTra-CDMX was evaluated through a chronological travel-time forecasting experiment. The objective of this assessment was not to establish a comprehensive ranking of forecasting algorithms, but to verify that the dataset contains sufficient temporal and contextual information to support supervised prediction under a leakage-controlled protocol.

Input sequences were constructed independently for each road segment using twelve consecutive five-minute observations, equivalent to 60 min of historical context. Forecasting horizons of 5, 15, 30, and 60 min were considered. The data were partitioned chronologically according to the target timestamp: November and December 2024 were used for training, January 2025 for validation, and February 2025 for testing. No random shuffling was applied across the temporal partitions.

To prevent information leakage, only variables available up to the prediction origin were included in each input sequence. Future observed weather values and dynamic variables associated with timestamps after the prediction origin were excluded. The sequence audit confirmed that all input windows were consecutive, that each target occurred strictly after the final input timestamp, and that the primary evaluation used observed target records.

A persistence baseline was first evaluated at the four forecasting horizons. Test MAE increased from 7.83 s at the 5 min horizon to 15.90 s at the 60 min horizon, reflecting the expected reduction in predictability as the forecasting interval increased. At the 30 min horizon, persistence achieved an MAE of 13.84 s, an RMSE of 48.48 s, a MAPE of 10.67%, and an R^2 of 0.628.

A gradient-boosted tree model optimized with Optuna was subsequently trained for the 30 min forecasting horizon. The final XGBoost model achieved an MAE of 12.84 s, an RMSE of 37.91 s, a median absolute error (MedAE) of 3.95 s, a MAPE of 10.94%, and a coefficient of determination (R^2) of 0.771 on the February 2025 test partition. Compared with persistence, the trained model reduced MAE and RMSE and increased R^2 , indicating that the dataset contains exploitable temporal and contextual structure beyond the most recent observed travel time.

The benchmark results should be interpreted as evidence of dataset usability rather than as an exhaustive comparison of forecasting architectures, as presented in Table 9. The common sample index, chronological partitions, selected features, model parameters, performance metrics, and leakage-audit outputs are distributed with the predictive-validation materials in the repository.

Table 9. Predictive-utility results for the 30 min forecasting horizon.

Model	MAE (s)	RMSE (s)	MedAE (s)	MAPE (%)	R^2
Persistence	13.84	48.48	—	10.67	0.628
XGBoost	12.84	37.91	3.95	10.94	0.771

6. User Notes

MUTra-CDMX is a multisource urban traffic dataset organized at the road-segment level. Although travel time can be used as the primary target variable for forecasting applications, the dataset also supports congestion analysis, mobility characterization, imputation studies, feature construction, and the development of intelligent transportation models. Operational, meteorological, geometric, and urban-context variables provide complementary information for analyzing traffic behavior along the monitored corridor.

To ensure proper use of the dataset, the following technical guidelines are recommended:

- Data ordering. Records in `data/MUTra_CDMX_completed.csv` are structured in a segment-timestamp tabular format. For time-series analysis, the data should be ordered by `segment_id` and `datetime`. Temporal operations should be performed independently for each road segment to avoid combining observations from different segments within the same sequence.
- Combination with segment metadata. Static physical and urban-context attributes are provided separately in `metadata/segment_metadata.csv` and can be joined to the main file through `segment_id`. Variables such as nominal segment length, endpoint gradient, sinuosity, functional road class, signalized-location counts, and nearby educational and healthcare-related points of interest should be interpreted as segment-level contextual characteristics rather than time-varying observations.
- Interpretation of point-of-interest variables. The fields `education_poi_count_100m` and `health_poi_count_100m` represent deduplicated OpenStreetMap records retrieved from the expanded bounding box associated with each segment. The `_100m` suffix is retained as the published contextual label, but the implemented procedure used an expansion of 0.0006° , corresponding to approximately 63–67 m per side at the latitude of the study area. Therefore, these counts should not be interpreted as measurements obtained from an exact 100 m line-based buffer.
- Use of reconstructed records. The `is_imputed` field distinguishes originally observed records from reconstructed segment–timestamp combinations. Users should retain this flag during analysis and consult `provenance/imputed_records.csv` when the reconstruction method or confidence level is relevant. For model evaluation, it is advisable to report results using observed target records and, when appropriate, provide a separate analysis including reconstructed observations.
- Meteorological variables. Weather data are reported at hourly resolution and aligned with five-minute traffic records according to the interval $[h, h + 1)$. Temperature and wind speed are repeated across the twelve five-minute timestamps belonging to the same hour and should not be interpreted as independently measured subhourly observations. `precipitation_hourly_total_mm` represents the total precipitation reported for the complete hour; repeated values must not be summed across the twelve traffic rows.
- Sequence generation. Predictive models should construct sequences separately for each `segment_id` using consecutive timestamps. For example, a sliding window of 12 five-minute observations represents 60 min of historical context. Before generating a sequence, continuity should be verified so that adjacent records remain separated by exactly five minutes.
- Prevention of information leakage. Training, validation, and test partitions should preserve chronological order and should be assigned according to the target timestamp. Random partitions that mix past and future observations should be avoided. Scaling, transformation, feature selection, and hyperparameter optimization should be fitted using only the training partition.
- Forecast-origin restrictions. Only variables available at or before the prediction origin should be included as model inputs. Future values of `current_speed`, `current_travel_time`, `relative_speed_index`, and `congestion_index_0_10` must not be used. Observed future weather should also be excluded unless the experiment explicitly evaluates the use of externally available weather forecasts.
- Interpretation of derived indicators. `relative_speed_index` and `congestion_index_0_10` are calculated from traffic variables at the same timestamp. They are suitable as historical input features, but they must not be calculated from information occurring after the prediction origin. Values of `relative_speed_index` greater than 1 indicate

that the observed speed exceeded the corresponding free-flow reference, whereas `congestion_index_0_10` is bounded below at zero.

- Geometric limitations. `endpoint_gradient_percent` represents the elevation difference between the endpoints divided by the nominal segment length and should not be interpreted as a complete longitudinal road profile. Missing sinuosity values should be retained as missing and must not be automatically replaced with 1, because that value would imply a straight segment without supporting geometric evidence.

Finally, the dataset contains no personal or directly identifying information, individual trajectories, license-plate data, images, or user identifiers. Its release is intended for aggregated traffic analysis, congestion characterization, reconstruction and imputation studies, contextual feature analysis, and the development and evaluation of travel-time forecasting models for Intelligent Transportation Systems.

7. Conclusions

MUTra-CDMX provides a documented multisource urban traffic dataset covering a 14.72 km section of the Insurgentes Sur corridor in Mexico City. The dataset contains 691,200 segment–timestamp records for 20 consecutive road segments at five-minute intervals over a 120-day period. Its organization separates the main spatiotemporal observations from complementary segment-level metadata, topology, provenance, and validation files, allowing traffic, meteorological, geometric, infrastructure, and urban-context information to be analyzed in an integrated and traceable manner.

The temporal-completion workflow preserved the 682,264 originally retrieved Tom-Tom observations and added 8936 previously missing segment–timestamp combinations. Reconstructed records are explicitly identified through `is_imputed`, while their corresponding methods and confidence levels are documented in the provenance files. This distinction enables users to analyze the completed series while maintaining transparency regarding which records were directly observed and which were reconstructed.

Technical validation confirmed the structural integrity and temporal completeness of the final dataset, the preservation of original traffic values, and the consistent alignment of hourly meteorological observations with five-minute traffic records. Artificial-masking experiments also quantified reconstruction performance across traffic variables, gap durations, random seeds, and road segments, showing that uncertainty varies according to both the variable and the temporal extent of the missing interval. The mean absolute percentage error ranged from 7.00% to 17.54% for `current_speed`, from 9.27% to 20.61% for `current_travel_time`, from 1.98% to 3.09% for `free_flow_speed`, and from 4.44% to 5.45% for `free_flow_travel_time`.

The predictive-utility experiment further demonstrated that MUTra-CDMX contains exploitable temporal and contextual information for travel-time forecasting. Under a chronological and leakage-controlled protocol, the optimized XGBoost model achieved an MAE of 12.84 s, an RMSE of 37.91 s, and an R^2 of 0.771 at the 30 min horizon, reducing MAE and RMSE and increasing R^2 relative to the persistence baseline. These results should be interpreted as evidence of dataset usability rather than as an exhaustive comparison of forecasting architectures. Together, the structural audits, artificial-masking experiments, and chronological predictive benchmark provide complementary evidence of the dataset's integrity, reconstruction performance, and practical usability.

MUTra-CDMX therefore offers a reproducible resource for congestion characterization, imputation research, spatiotemporal analysis, contextual traffic modeling, and travel-time forecasting in an urban corridor within a Latin American setting. Its main limitations are the four-month observation period, the focus on a single corridor, the hourly resolution of the meteorological variables, and the presence of reconstructed records in the completed

temporal grid. In addition, the segment-level gradient is estimated from endpoint elevations rather than from a complete longitudinal road profile, and the point-of-interest counts were obtained from expanded segment bounding boxes rather than exact 100 m line-based buffers. Future extensions could include additional corridors, longer monitoring periods, incident and event information, higher-resolution environmental observations, more detailed road-geometry profiles, and externally generated weather forecasts.

Author Contributions: Conceptualization, A.R.-R., A.M.-R. and H.E.E.; methodology, A.R.-R., A.M.-R., H.E.E. and E.C.; data curation, E.d.l.C.-N. and E.C.; formal analysis, A.R.-R., A.M.-R. and H.E.E.; writing—original draft preparation, A.R.-R. and A.M.-R.; writing—review and editing, A.R.-R., A.M.-R., H.E.E., E.d.l.C.-N. and E.C.; supervision, A.M.-R., H.E.E. and E.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through a doctoral scholarship awarded to the first author (CVU: 1317887).

Data Availability Statement: The MUTra-CDMX dataset and its supporting materials are publicly available on Zenodo at <https://doi.org/10.5281/zenodo.21617011>. This record corresponds to the second Zenodo release and contains version 4.0 of the MUTra-CDMX dataset, distributed as MUTra_CDMX_v4.0.rar. The package contains the main five-minute spatiotemporal dataset, segment-level metadata and topology, provenance records, technical-validation outputs, predictive-benchmark materials, a data dictionary, and supporting documentation. The primary analytical file is located at data/MUTra_CDMX_completed.csv within the package. Access and reuse are subject to the licensing and third-party data conditions specified in the repository.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miao, Z.; Liao, Q. IoT-Based Traffic Prediction for Smart Cities. *IEEE Access* **2025**, *13*, 52369–52384. [[CrossRef](#)]
2. Yang, L.; Cheng, J.; Zhao, Y.; Ni, Z.; Mao, Q.; Gao, S. A Unified Software-Defined Autonomous Vehicle Network and Urban Congestion Prediction Method. *IEEE Trans. Netw. Sci. Eng.* **2025**, *12*, 2708–2721. [[CrossRef](#)]
3. Chen, L.; Zhao, Q.; Li, G.; Zhou, M.; Dai, C.; Feng, Y.; Liu, X.; Li, J. A Sparse Cross Attention-Based Graph Convolution Network with Auxiliary Information Awareness for Traffic Flow Prediction. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 3210–3222. [[CrossRef](#)]
4. Ma, H.; Qin, X.; Jia, Y. Dynamic Spatio-Temporal Graph Fusion Convolutional Network for Urban Traffic Prediction. *Appl. Sci.* **2023**, *13*, 9304. [[CrossRef](#)]
5. Buroni, G.; Lebichot, B.; Bontempi, G. AST-MTL: An Attention-Based Multi-Task Learning Strategy for Traffic Forecasting. *IEEE Access* **2021**, *9*, 77359–77370. [[CrossRef](#)]
6. Fan, X.; Xiang, C.; Gong, L.; He, X.; Qu, Y.; Amirgholipour, S.; Xi, Y.; Nanda, P.; He, X. Deep Learning for Intelligent Traffic Sensing and Prediction: Recent Advances and Future Challenges. *CCF Trans. Pervasive Comput. Interact.* **2020**, *2*, 240–260. [[CrossRef](#)]
7. Ulvi, H.; Yerlikaya, M.A.; Yildiz, K. Urban Traffic Mobility Optimization Model: A Novel Mathematical Approach for Predictive Urban Traffic Analysis. *Appl. Sci.* **2024**, *14*, 5873. [[CrossRef](#)]
8. Manibardo, E.L.; Laña, I.; Del Ser, J. Transfer Learning and Online Learning for Traffic Forecasting under Different Data Availability Conditions: Alternatives and Pitfalls. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–6. [[CrossRef](#)]
9. Wei, L.; Yu, Z.; Jin, Z.; Xie, L.; Huang, J.; Cai, D.; He, X.; Hua, X.-S. Dual Graph for Traffic Forecasting. *IEEE Access* **2025**, *13*, 122285–122293. [[CrossRef](#)]
10. Wang, S.; Lv, Y.; Peng, Y.; Piao, X.; Zhang, Y. Metro Traffic Flow Prediction via Knowledge Graph and Spatiotemporal Graph Neural Network. *J. Adv. Transp.* **2022**, *2022*, 2348375. [[CrossRef](#)]
11. He, H.H. *California Traffic Network Datasets: METR-LA, PEMS-BAY, PEMS04 and PEMS08 for Traffic Speed and Flow Analysis*; IEEE DataPort: New York, NY, USA, 2025. [[CrossRef](#)]
12. Zhang, Z.; Wang, X.; Tan, X.; Pi, J. DSSA-TCN: Exploiting Adaptive Sparse Attention and Diffusion Graph Convolutions in Temporal Convolutional Networks for Traffic Flow Forecasting. *PLoS ONE* **2025**, *20*, e0336787. [[CrossRef](#)] [[PubMed](#)]
13. Cerqueira, S.; Arsenio, E.; Henriques, R. On How to Incorporate Public Sources of Situational Context in Descriptive and Predictive Models of Traffic Data. *Eur. Transp. Res. Rev.* **2021**, *13*, 60. [[CrossRef](#)] [[PubMed](#)]

14. Gómez, I.; Ilarri, S. Enriched Traffic Datasets for the City of Madrid: Integrating Data from Traffic Sensors, the Road Infrastructure, Calendar Data and Weather Data. *Data Brief* **2024**, *57*, 110878. [[CrossRef](#)] [[PubMed](#)]
15. Abduljabbar, R.; Dia, H.; Tsai, P.-W. Development and Evaluation of Bidirectional LSTM Freeway Traffic Forecasting Models Using Simulation Data. *Sci. Rep.* **2021**, *11*, 23899. [[CrossRef](#)] [[PubMed](#)]
16. Zhang, C.; Zhang, Y.; Shao, Q.; Feng, J.; Li, B.; Lv, Y.; Piao, X.; Yin, B. BJT: A Large-Scale Multimodal Dataset for Traffic Prediction. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 18992–19003. [[CrossRef](#)]
17. Zhang, Y.; Zhao, T.; Gao, S.; Raubal, M. Incorporating Multimodal Context Information into Traffic Speed Forecasting through Graph Deep Learning. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 1909–1935. [[CrossRef](#)]
18. Zhang, Y.; Wang, Y.; Gao, S.; Raubal, M. Context-Aware Knowledge Graph Framework for Traffic Speed Forecasting Using Graph Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 3885–3902. [[CrossRef](#)]
19. TomTom International B.V. Flow Segment Data. Available online: <https://developer.tomtom.com/traffic-api/documentation/tomtom-maps/v1/traffic-flow/flow-segment-data> (accessed on 5 June 2026).
20. Meteosource. Historical Weather Data API—Time Machine. Available online: <https://www.meteosource.com/api-weather-historical> (accessed on 8 June 2026).
21. Raifer, M. Overpass Turbo. Available online: <https://overpass-turbo.eu/> (accessed on 8 June 2026).
22. Zhu, H.; Zhang, K.; Wang, C.; Jia, L.; Song, S. The Impact of Road Functions on Road Congestions Based on POI Clustering: An Empirical Analysis in Xi'an, China. *J. Adv. Transp.* **2023**, *2023*, 6144048. [[CrossRef](#)]
23. Nian, G.; Sun, J.; Huang, J. Exploring the Effects of Urban Built Environment on Road Travel Speed Variability with a Spatial Panel Data Model. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 829. [[CrossRef](#)]
24. Li, T.; Jiang, H.; Jing, P.; Zhang, M. Analyzing the Coordination Relationship between Urban Built Environment and Traffic Level. *J. Adv. Transp.* **2021**, *2021*, 9952306. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.