Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Enriched traffic datasets for the city of Madrid: Integrating data from traffic sensors, the road infrastructure, calendar data and weather data

Iván Gómez[*], Sergio Ilarri

Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, I3A, Zaragoza, Aragón 50018, Spain

ARTICLE INFO

Article history:
Received 3 June 2024
Revised 19 August 2024
Accepted 20 August 2024
Available online 29 August 2024

Dataset link: Enriched Traffic Datasets for Madrid (Original data)

Keywords:
Traffic data analysis
Traffic sensor data
Weather data
Calendar data
Road infrastructure data
Localization data
Madrid

ABSTRACT

The proliferation of urban areas and the concurrent increase in vehicular mobility have escalated the urgency for advanced traffic management solutions. This data article introduces two traffic datasets from Madrid, collected between June 2022 and February 2024, to address the challenges of traffic management in urban areas. The first dataset provides detailed traffic flow measurements (vehicles per hour) from urban sensors and road networks, enriched with weather data, calendar data and road infrastructure details from OpenStreetMap. This combination allows for an in-depth analysis of urban mobility. Through preprocessing, data quality is ensured by eliminating inconsistent sensor readings. The second dataset is enhanced for advanced predictive modelling. It includes time-based transformations and a tailored preprocessing pipeline that standardizes numeric data, applies one-hot encoding to categorical features, and uses ordinal encoding for specific features. In constructing the datasets, we initially employed the k-means algorithm to cluster data from multiple sensors, thereby highlighting the most representative ones. This clustering can be adapted or modified according to the user's needs, ensuring flexibility for various analyses and applications.

[*] Corresponding author.
E-mail addresses: ivan.gomez@unizar.es (I. Gómez), silarri@unizar.es (S. Ilarri).
Social media: @ivangomezz21 (I. Gómez), @silarri (S. Ilarri)

https://doi.org/10.1016/j.dib.2024.110878

This work underscores the importance of advanced datasets in urban planning and highlights the versatility of these resources for multiple practical applications. We highlight the relevance of the collected data for a variety of essential purposes, including traffic prediction, infrastructure planning, studies on the environmental impact of traffic, event planning, and conducting simulations. These datasets not only provide a solid foundation for academic research but also for designing and implementing more effective and sustainable traffic policies. Furthermore, all related datasets, source code, and documentation have been made publicly available, encouraging further research and practical applications in traffic management and urban planning.

Specifications Table

Subject	Transportation Management
Specific subject area	Datasets for urban traffic analysis in Madrid, created using traffic sensor data, data about weather conditions, calendar data and road data.
Type of data	Analyzed Filtered Processed Multi-source
Data collection	We exploited data obtained through urban sensors deployed across the road network in the city of Madrid, complemented with meteorological data from a local weather station and the city's work calendar, all accessible through Madrid's Open Data Portal [1]. The information on the road infrastructure was extracted from OpenStreetMap [2] and processed using OSMnx [3]. To streamline the integration of these diverse data sources, custom Python scripts were developed. To ensure the integrity and quality of the data, a preprocessing process was implemented. This included the removal of anomalies, discarding incomplete data (NaNs), the cross-validation of sensor information, and the normalization of categorical variables using advanced encoding techniques. To facilitate the analysis and understanding of traffic patterns, the k-means [4] algorithm was employed, allowing for the effective grouping of data from various sensors. These datasets, product of extensive integration of publicly available data sources and processing work applied to prepare and enhance the resulting dataset, are novel and provide a unique perspective on urban mobility.
Data source location	Country: Spain. Region: Community of Madrid. City: Madrid Coordinates: 40°24′59" latitude and 3°42′09" longitude.
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/697ht4f65b.1 Direct URL to data: https://data.mendeley.com/datasets/697ht4f65b/1

1. Value of the Data

- The datasets built and provided are valuable because they offer a comprehensive view of urban traffic patterns, combining real-time sensor data with weather conditions, the city's work schedule, and urban road data. This unique combination allows for a nuanced understanding of traffic dynamics in relation to various external factors, essential for urban planning and environmental studies.

- Providing two distinct datasets significantly enriches the value of this collection for the scientific community. It gives researchers the flexibility to tackle a wide range of research questions, from descriptive analysis to the development and validation of complex predictive models.
- The datasets can be useful to develop and test predictive models of traffic flow, enabling the anticipation of congestion and the optimization of traffic management strategies. This has implications not only for urban mobility but also for reducing environmental impacts through more efficient vehicle flow.
- Environmental scientists and urban planners could analyse the correlation between traffic patterns and pollution levels, leveraging the data to propose interventions that mitigate urban air pollution. This is crucial for cities looking to improve air quality and public health.
- The openness of the code for dataset generation, combined with its high parameterization through its variables, grants researchers the capability to adjust or extend the code used for the processing and integration of data according to other potential existing needs. This unlocks a broad spectrum of possibilities for revolutionizing how data-driven approaches can influence urban design and public policy formulation. The source code, documentation, required data, and datasets are located in a GitHub repository [5], which will be accessible to interested parties.

2. Background

The datasets were compiled in the context of increasing urbanization and enhanced mobility in metropolitan areas, with a specific focus on the city of Madrid in Spain [6]. The main motivation to build this dataset was to have an integrated and rich collection of data that can be used to study the challenges of traffic congestion, a perennial problem in large cities that affects economic and environmental sustainability [7], as well as the quality of life. Traditional traffic management strategies often fall short due to their isolated handling of data sources, without capturing the dynamic interaction between various factors influencing traffic flow. These datasets integrate diverse data sources, including urban sensor data (see Fig. 1), weather conditions, and road infrastructure information, along with calendar information to provide a comprehensive view of traffic dynamics. This type of dataset can contribute to the development of sustainable urban mobility solutions by facilitating detailed analysis and modelling of traffic behaviours, thus supporting more informed decisions in policy and urban planning.

3. Data Description

In this study, two main datasets have been created to analyse urban traffic dynamics in Madrid. These data cover the period from June 1, 2022, to February 29, 2024, and are published in CSV format on Mendeley Data [8], facilitating their accessibility and manipulation for research purposes. Besides, we also provide the code used to generate them [5], so it could be revised, fine-tuned and extended (if needed) according to other potential existing needs. To illustrate the depth and application of these datasets, Fig. 2 presents an interactive map displaying the locations of all traffic sensors used, with features enabling viewers to see real-time geographic coordinates and meteorological conditions for each sensor. As another example, Fig. 3 provides a temporal graph showing weekly traffic patterns, highlighting daily and weekly traffic fluctuations. These visual aids foster a comprehensive understanding of the data collection and the analytical possibilities they offer.

On the one hand, the first dataset that we provide, that we call ‘DADAS’ (Descriptive Analysis DAtaSet), constitutes an exhaustive compilation of traffic intensity in Madrid, gathering 18,012,128 rows of data and covering 300 unique sensor identifiers, collected from June 2022 to February 2024. This rich source of information comes from an advanced network of urban

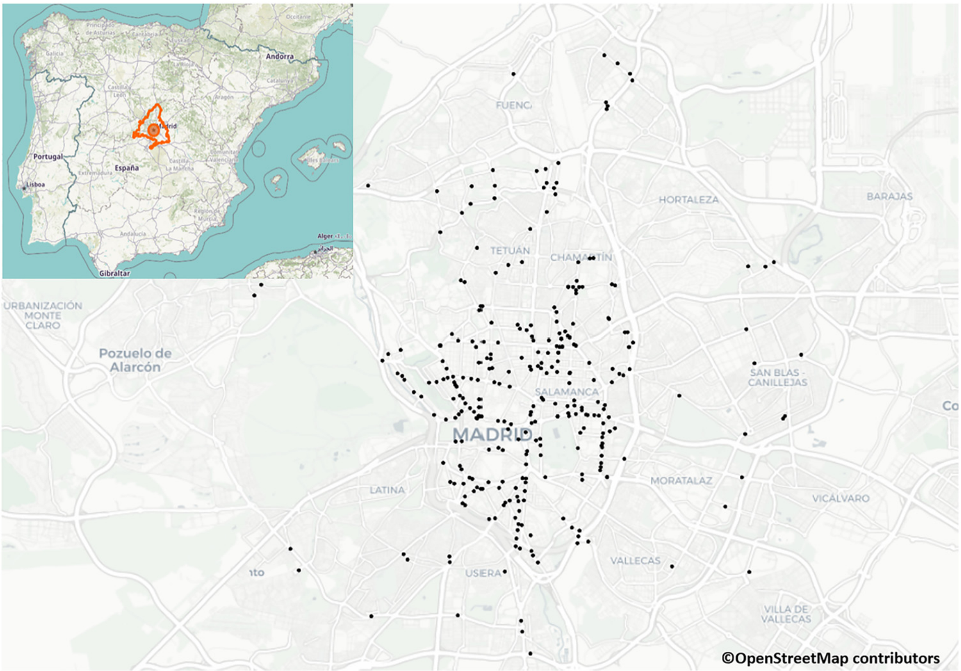
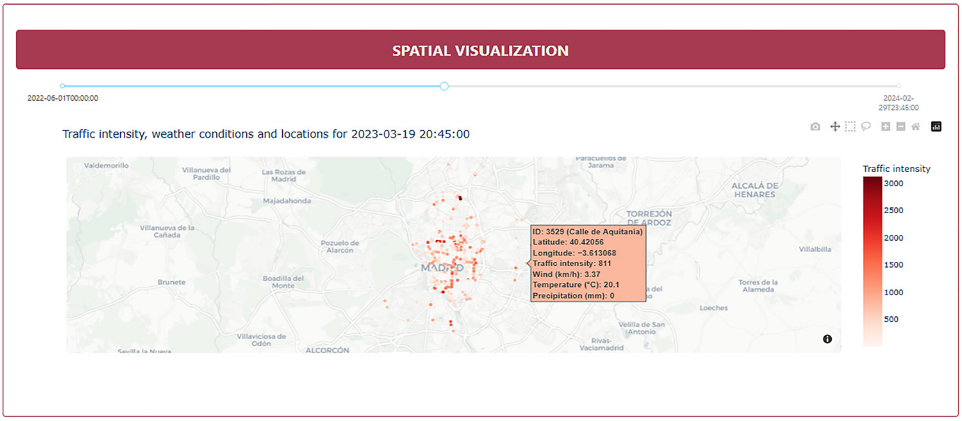


Fig. 1. Localization of traffic sensors (black dots) from the datasets — Madrid, Community of Madrid, Spain.


SPATIAL VISUALIZATION

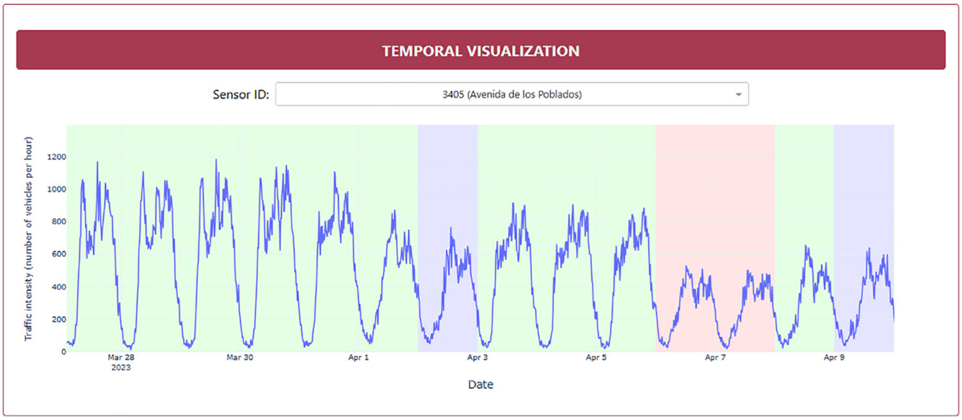
2022-06-01T00:00:00 2024-02-29T23:45:00

Traffic intensity, weather conditions and locations for 2023-03-19 20:45:00

ID: 3529 (Calle de Aquitania)
Latitude: 40.42056
Longitude: −3.613068
Traffic intensity: 811
Wind (km/h): 3.37
Temperature (°C): 20.1
Precipitation (mm): 0

Fig. 2. Interactive traffic sensor map of Madrid.

and highway sensors, capturing traffic intensity in 15 min intervals, expressed in vehicles per hour. The capture devices, predominantly electromagnetic loops installed under the pavement [9], offer precise measurements of the vehicles passing by. The raw base traffic sensor data are published in the Open Data Portal of Madrid [10], but we have performed significant processing over these data and integrated them with other data sources. Through a meticulous preprocessing process, traffic data were filtered to eliminate inconsistencies. The preprocessing involved several steps crucial for ensuring data quality and usability. Outliers, duplicates, and incomplete data were excluded, along with data from sensors recording fewer than 1,000 samples per month—well below the expectation of 2,880 samples per month (according to the sensor sam-



Sensor ID: 3405 (Avenida de los Poblados)

Fig. 3. Temporal visualization of a selected sensor (Orange: Holidays, Blue: Weekends, Green: Weekdays).

pling rate of one measurement every 15 min). The threshold for excluding sensors that record fewer than 1,000 samples per month is configurable, enabling easy adjustments to meet different analysis needs or data availability constraints. After applying this threshold, 976 sensors remain in the dataset. Additionally, the sensor data are collected in 15 min intervals and measured as the number of vehicles per hour; however, the code is structured to allow quick adaptation to other coarser time granularities that might be required. The data were further enriched with climatic information and road infrastructure data through OSMnx [3], based on OpenStreetMap [2], facilitating a multidimensional analysis of traffic flow.

Initially, simple clustering is applied to select the most representative sensor data, making the dataset more focused and relevant for analysis. The number of sensors included in the clustering is set to 300. The clusters are based on scaled characteristics (mean, median, and standard deviation of traffic intensity). Each sensor is assigned to one of the 300 clusters. In the end, each sensor will have an associated cluster number. For each of the 300 clusters, a representative sensor is selected. This sensor is the one with the smallest distance to the group's mean (centroid). This selection captures the variability and geographic diversity of the data, ensuring a statistically significant sample while maintaining manageable computational requirements. Based on statistical calculation, 273 sensors are sufficient to ensure a statistically significant sample for our population size, assuming a maximum variability (proportion of 0.5, which is a frequently-used defaut value, as it provides a worst-case sample size), a 95% confidence level, and a 5% margin of error; this calculation is for, determining a suitable sample size in a finite population, using Cochran's formula adapted for finite populations [11]. Moreover, the computational cost analysis showed that clustering with 300 sensors requires approximately 30 % of the computation time needed when using all 947 sensors, making it significantly more manageable. The choice of 300 sensors serves as an example, with this parameter being configurable within the code. Users can adjust the number of sensors or clusters as needed, or even disable clustering altogether. The implementation of the k-means algorithm [12] for optimizing cluster analysis and the use of efficient search techniques such as KD trees [13] for the geographical assignment of measurement points underscore the rigorous approach adopted to maximize data accuracy and usefulness. The removal of outliers, duplicates, and incomplete data, along with the adaptation of the measurements to hourly values, ensure a coherent and analytically viable database. This holistic approach guarantees that the datasets are not only representative of traffic flow in Madrid but also optimal for future research in the field of urban mobility.

On the other hand, our second dataset, that we call 'MLDAS' (ML-oriented DAtaSet), is a dataset derived from the previous dataset DADAS. It is primarily intended for predictive analysis and pattern detection in sensor behaviour. The preprocessing of the data transforms the 'date'

Table 1
Feature comparison of the two datasets provided: DADAS vs. MLDAS.

Feature	DADAS	MLDAS
Data Source	Urban and highway sensors in Madrid enriched with weather conditions, calendar data and road data.	Derived from DADAS, further processed to tailor for machine learning applications.
Purpose	Useful for detailed and descriptive analysis.	Useful to facilitate predictive modelling and model validation.

Table 2
DADAS statistics of the traffic flow intensity attribute.

Statistic	Value
Count	18,012,128
Mean	566.73
Standard Deviation	548.87
Minimum	0.00
25th Percentile	151.00
Median (50%)	397.00
75th Percentile	819.00
90th Percentile	1,329.00
Maximum	47,784.00

column into a datetime object (Facilitates time-based analysis), extracts the month, hour, and day of the week, and applies trigonometric encoding to the hour to effectively capture cyclical patterns. A preprocessing pipeline standardizes numeric features, applies one-hot encoding to categorical features, and uses ordinal encoding for specific features, thus preparing the data for advanced machine learning models (e.g., neural networks, support vector machines, and complex ensemble methods). Table 1 succinctly captures the primary differences and applications of each dataset, illustrating how they serve distinct but complementary purposes in urban mobility research.

Traffic intensity, measured as the number of vehicles per hour passing through a checkpoint, is the backbone of the analysis of vehicular flow in Madrid. A detailed examination of this variable, as summarized in Table 2, reveals a complex and multifaceted panorama of traffic behaviour in the city. The dataset, covering from June 2022 to February 2024, offers valuable insights into urban dynamics and allows the identification of congestion patterns, peak times or weather conditions on traffic. The observed variability, from quiet streets to congested arteries, underscores the importance of an adaptable and well-informed approach to mobility planning and traffic management interventions.

Tables 3 and 4 provide detailed descriptions of each column in the DADAS and MLDAS datasets, including the data type, units of measurement (where applicable), and a brief description of the information they provide. Both tables summarize the structure and type of information contained in each dataset, providing a clear foundation for analyzing and interpreting traffic patterns in Madrid. By detailing the units and describing each column, these tables enable understanding of how each attribute contributes to the study of traffic and its interaction with external factors such as the weather and road infrastructure.

4. Experimental Design, Materials and Methods

Our study utilizes a comprehensive methodological framework to process and assemble traffic data in Madrid. The sequence of data processing steps from initial data retrieval through to DADAS preparation is outlined in Fig. 4. This includes integrating multiple data sources, such as traffic intensity data, sensor locations, work calendar, weather conditions, and road information. Fig. 5 illustrates the conversion of the Descriptive Analysis Dataset (DADAS) into the ML-

Table 3
Description of attributes in the DADAS dataset.

Attribute	Data Type	Units	Description
id	Integer		Unique identifier of the sensor.
date	Date and Time		Date and time of the measurement, indicating the moment of data collection.
longitude	Float	Degrees	Geographical longitude of the sensor.
latitude	Float	Degrees	Geographical latitude of the sensor.
traffic_intensity	Integer	Vehicles/hour	Traffic intensity, measured as the number of vehicles passing by the sensor per hour.
day_type	Category		Type of day (Working day, Holiday, Sunday or Saturday) to characterize traffic according to the calendar.
wind	Float	m/s	Wind speed at the time of the measurement.
temperature	Float	°C	Ambient temperature at the time of the measurement.
precipitation	Float	mm	Accumulated precipitation during the hour prior to the measurement.
original_point	WKT (Well-Known Text)		Geographic location of the sensor in geographic coordinates format.
closest_point	WKT (Well-Known Text)		Closest geographic point on the road network, used to map the sensor onto the street network.
oneway	Boolean		Indication of whether the street where the sensor is located is one-way or not (true/false).
lanes	Integer		Number of lanes of the street where the sensor is located.
name	Text		Name of the street or road where the sensor is located.
highway	Category		Type of road according to OpenStreetMap's classification (e. g., secondary, primary, motorway, residential).
maxspeed	Float	km/h	Maximum speed limit on the road corresponding to the sensor.
length	Float	Meters	Length of the street segment associated with the sensor.

oriented Dataset (MLDAS). The transformation involves loading DADAS, refining the date column into more specific features, and encoding these features to prepare the dataset for future machine learning applications.

More in detail, we performed the following steps, to integrate data from the different data sources considered:

Table 4
Description of attributes in the MLDAS dataset.

Attribute	Data Type	Units	Description
id	Float		Unique identifier for the sensor, converted to a float.
hour_sin	Float		Sine transformation of time (minutes from midnight), capturing the cyclical nature of time within the day. A standard scaler was used for normalizing values.
hour_cos	Float		Cosine transformation of time (minutes from midnight), capturing the cyclical nature of time within the day. A standard scaler was used for normalizing values.
week_day	Float		Numeric representation of the weekday. A standard scaler was used for normalizing values.
latitude	Float	Degrees	Geographical latitude of the sensor. A standard scaler was used for normalizing values.
longitude	Float	Degrees	Geographical longitude of the sensor. A standard scaler was used for normalizing values.
wind	Float	m/s	Wind speed at the time of measurement. A standard scaler was used for normalizing values.
precipitation	Float	mm	Accumulated precipitation during the hour prior to the measurement. A standard scaler was used for normalizing values.
temperature	Float	°C	Ambient temperature at the time of measurement. A standard scaler was used for normalizing values.
lanes	Float		Number of lanes on the street where the sensor is located. A standard scaler was used for normalizing values.
maxspeed	Float	km/h	Maximum speed limit on the road where the sensor is located. A standard scaler was used for normalizing values.
length	Float	Meters	Length of the road segment associated with the sensor. A standard scaler was used for normalizing values.
day_type_Holiday	Float		Indication of whether the day is a holiday, derived from the one-hot encoding of the day type.
day_type_Saturday	Float		Indication of whether the day is a Saturday, derived from the one-hot encoding of the day type.
day_type_Sunday	Float		Indication of whether the day is a Sunday, derived from the one-hot encoding of the day type.
day_type_Working day	Float		Indication of whether the day is a working day, derived from the one-hot encoding of the day type.
month_1 to month_12	Float		One-hot encoded representation of the month, where each column corresponds to a month of the year.
highway	Float		Type of road according to the OpenStreetMap classification, processed via ordinal encoding.
oneway	Float		Indication of whether the street where the sensor is located is one-way, processed via ordinal encoding.
traffic_intensity	Float	Vehicles/ hour	Number of vehicles passing per hour. Directly from DADAS.

4.1. Traffic Intensity Data from Sensors

The traffic intensity data analyzed in this study were obtained from Madrid's Open Data Portal [10]. A monthly collection of the information was carried out, which was initially stored in CSV files. While these files encompass multiple columns, for the study's objective only those pertaining to sensor ID, date and time of the record, and traffic intensity were chosen. Other columns, such as the average speed 'vmed', which is solely measured at interurban points of the M30, were excluded to maintain the focus on traffic intensity analysis.

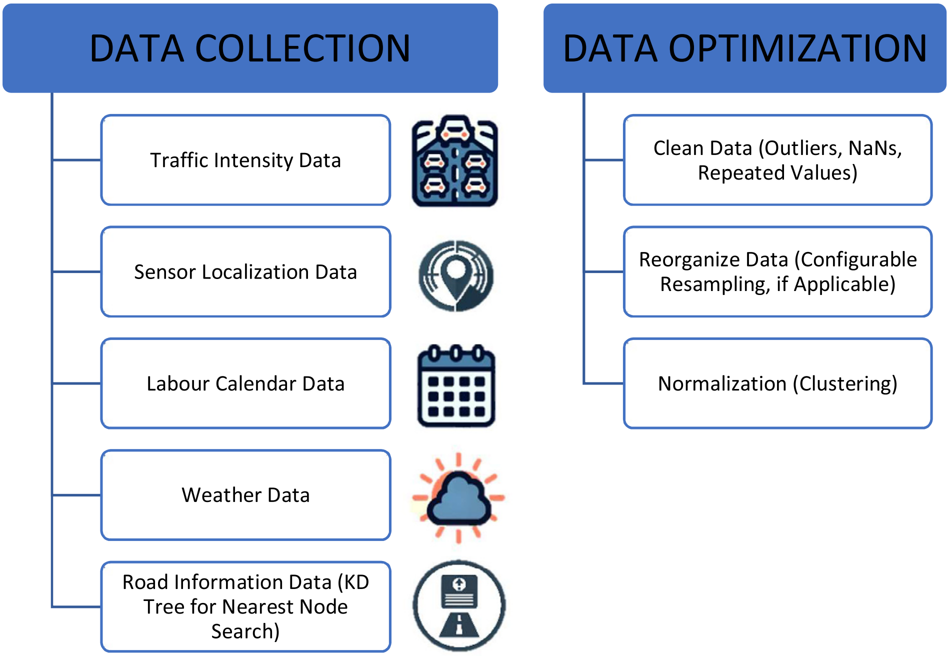

DATA COLLECTION
Traffic Intensity Data
Sensor Localization Data
Labour Calendar Data
Weather Data
Road Information Data (KD Tree for Nearest Node Search)
DATA OPTIMIZATION
Clean Data (Outliers, NaNs, Repeated Values)
Reorganize Data (Configurable Resampling, if Applicable)
Normalization (Clustering)

Fig. 4. Methodology for collecting and optimizing data from various sources.

4.2. Sensor Location Data

The location data for each sensor used in the traffic data collection also comes from Madrid's Open Data Portal [9]. This dataset includes detailed information on the geographic position of the sensors, necessary for traffic analysis focused on specific areas of the city. Each location record includes fields such as the type of element (for example, "URB" for urban locations or "M30" for locations along the M30 ring road in Madrid), the district, sensor ID, center code, location name (e.g., "DR. ESQUERDO N-S(MONTANO-GAVINET)"), UTM coordinates (X and Y), and most crucially, the longitude and latitude coordinates. When integrating the location data into the traffic dataset, these coordinates are added in the Well-Known Text (WKT) format [14]. The use of the WKT format is crucial, as it is a standardized text markup language for representing vector geometry objects on a map, simplifying the integration and manipulation of geographic data within various spatial databases and GIS (Geographic Information Systems) software. The process of integrating traffic data with location data was carried out by removing unnecessary columns like the type of element from the traffic dataset, and then combining both datasets using the sensor ID as a common key. This process results in a new dataset that includes the columns of sensor ID, date and time of the record, traffic intensity, and the longitude and latitude coordinates. This combination provides a detailed perspective of traffic activity in a specific geographic context.

4.3. Labour Calendar Data

The data on Madrid's labour calendar, which includes information on workdays, holidays, and festive Sundays, were also obtained from Madrid's Open Data Portal [15]. This information allows for the analysis of how traffic patterns vary depending on the different types of days.

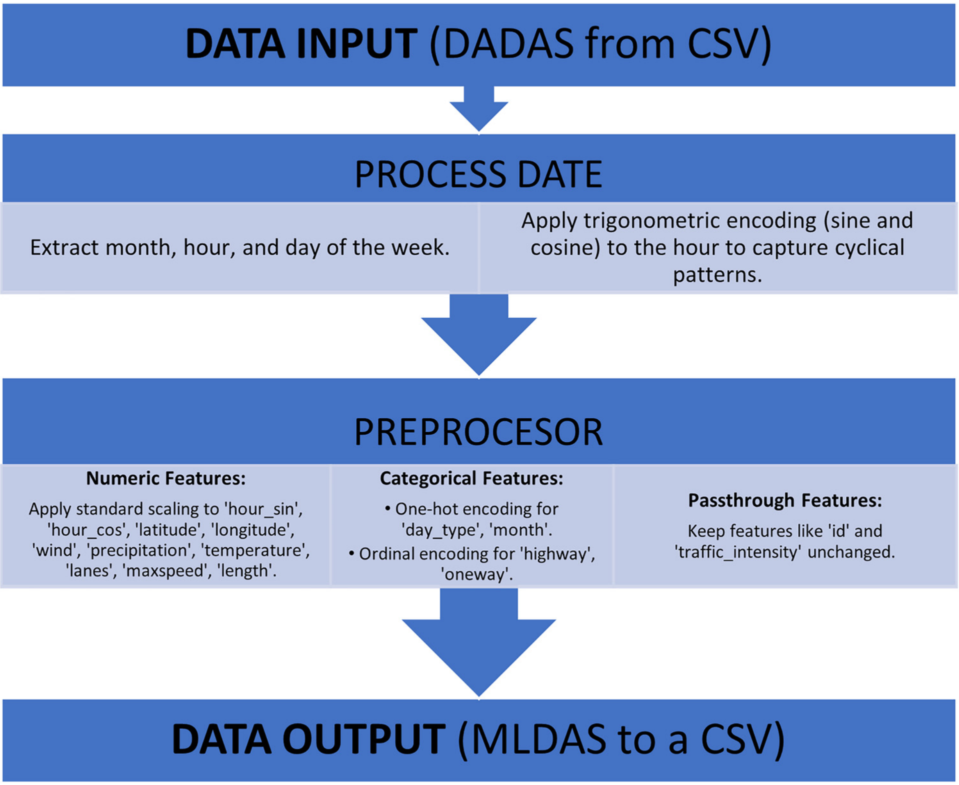

DATA INPUT (DADAS from CSV)
PROCESS DATE
Extract month, hour, and day of the week.
Apply trigonometric encoding (sine and cosine) to the hour to capture cyclical patterns.
PREPROCESOR
Numeric Features:
Apply standard scaling to 'hour_sin', 'hour_cos', 'latitude', 'longitude', 'wind', 'precipitation', 'temperature', 'lanes', 'maxspeed', 'length'.
Categorical Features:
• One-hot encoding for 'day_type', 'month'.
• Ordinal encoding for 'highway', 'oneway'.
Passthrough Features:
Keep features like 'id' and 'traffic_intensity' unchanged.
DATA OUTPUT (MLDAS to a CSV)

Fig. 5. Transformation process of DADAS to MLDAS.

The process of integrating labour calendar data with traffic data was carried out by combining both datasets based on the dates. Rows in the calendar where the column indicating whether the day is a workday, holiday, or festive Sunday is empty are identified and filled in based on the day of the week. Saturdays and Sundays are marked accordingly, and weekdays are classified as workdays. After completing this information, unnecessary columns from the calendar dataset were removed. For example, the ‘Festividad’ column, which represents the name of the holiday as designated by our country’s legislation, was excluded. The date column in the traffic data was converted to the datetime format to ensure proper merging with other datasets. The result is a dataset that includes the original traffic data and its geographic location along with a classification of each day in terms of its working or festive character.

4.4. Weather Data

The integration of weather data into traffic analysis was done by acquiring climatic variables from Madrid’s Open Data Portal [16], which provides information on temperature, precipitation, and wind. The dates in the traffic dataset were adjusted to facilitate merging with the weather data, ensuring that the weather observations matched the days of the traffic records.

The weather data were filtered to obtain the relevant months and years, extracting the day of the month to accurately align with the daily traffic records. Weather magnitude codes were mapped to descriptive variables, and the columns of temperature, precipitation, and wind were added to the traffic dataset, filling them with daily averages based on the corresponding observations. This transforms the traffic dataset into a matrix that includes both traffic patterns and environmental conditions.

4.5. Road Information Data

The integration of road information started by transforming the traffic data into a GeoDataFrame [17], where each prediction point is specified by its longitude and latitude coordinates. Duplicates based on coordinates were eliminated to ensure that each point is unique. Subsequently, the coordinates of street geometries were collected from another GeoDataFrame that houses the edges of the road network, with these coordinates recorded as nodes in the data structure.

With the street coordinates in hand, a KD Tree [13] was applied to index the spatial data. This efficient data structure was used for nearest point searching in geospatial applications, enabling the quick identification of the nearest road network point to each traffic measurement point in the traffic dataset. For each traffic measurement point, a query determined the shortest distance to the nearest road node, returning the exact coordinates of that node and the distance to it. These results were then incorporated into the original traffic GeoDataFrame, thus embedding detailed road infrastructure information into each traffic record and associating each traffic point with a specific location on the street network. This approach highlights the importance of leveraging road network information to enhance urban transport management, as also noted by Myrovali et al. [18].

4.6. Optimization of the Descriptive Analysis Dataset (DADAS)

In the analysis of Madrid's traffic data, advanced techniques were applied to optimize the dataset. Essential steps were undertaken to clean the data before analysis, including the removal of records with outlier values in the 'intensity' column. This was done using the 99th percentile within rolling time windows and the 99.999th percentile for each ID as the upper limits. Only records that exceeded both thresholds were removed, ensuring that only representative values were retained. This method distinguishes between valid traffic peaks and non-representative outliers, as extreme traffic peaks rarely occur without context. Therefore, the removed outliers do not correspond to legitimate traffic peaks but to invalid data. Additionally, repeated 'intensity' values that persist for more than a number of consecutive records (12 by default) were identified and removed (it is strange that the sensor repeatedly measures exactly the same number of cars, so it can be assumed that the sensor is malfunctioning), to prevent distortions in the analysis of temporal trends, and records where the 'intensity' column contains NaN values were also eliminated.

The dataset has been configured to collect traffic data in 15 min intervals, with the results expressed in terms of vehicles per hour. Additionally, this time period can be adjusted to different time granularities as needed, enhancing the versatility of the data for various analytical purposes. The k-means algorithm was used to segment the data into a predefined number of clusters (300 by default), based on characteristics such as the mean, median, and standard deviation of 'intensity'. Each cluster was analyzed to select a representative sensor, chosen for its proximity to the cluster's centroid, focusing the analysis on data points that represent the characteristics of their group.

The process setup allows for the adjustment of parameters such as the number of clusters in the clustering analysis and the minimum threshold of records required to retain a sensor in the analysis. This facilitates the adaptation of the process to the quantity and quality of data available, allowing for customized analyses according to the specific needs of each project. These procedures establish a solid foundation for detailed analyses, providing useful information for urban planning and traffic management.

4.7. Generation of the ML-oriented Dataset (MLDAS)

MLDAS was generated from a process that begins with loading data from DADAS. The ‘datetime’ column was converted into a datetime object, from which the month, hour, and day of the week were extracted. Trigonometric encoding (sine and cosine) was then applied to the hour to capture the cyclical nature of time, which is crucial for analyzing patterns in time series data.

Numeric features like ‘hour_sin’, ‘hour_cos’, and others are standardized by removing the mean and scaling to unit variance. This method is preferred over others, such as min-max scaling, because it preserves the original distribution of the data without binding it to a specific range. This is particularly important for machine learning algorithms like SVM and k-nearest neighbours, which rely on distance calculations and perform better when the data maintains its natural variance and outliers. This approach ensures no single feature with a larger range can dominate the model’s predictions, maintaining a balanced input scale..

Categorical features such as ‘day_type’ and ‘month’ underwent one-hot encoding, transforming these variables into a format that can be effectively used in machine learning predictions (many machine learning algorithms require numerical attributes). Ordinal features like ‘highway’ and ‘oneway’ were encoded based on their inherent order, which is important for models where the order impacts performance. Passthrough features such as ‘id’ and ‘traffic_intensity’ were left unchanged because they serve as unique identifiers and the primary prediction target, respectively, within the dataset.

All configurations and transformations are adjustable in the scripts, allowing for flexibility in how the data are processed depending on the specific needs of the analysis or simulations. This modularity ensures that the dataset can be tailored to meet various analytical requirements.

Limitations

We have only collected data for the period June 1, 2022, to February 29, 2024, but the dataset generation code provided could be easily used to build a dataset for a longer period, if required. Besides, as explained in this data article, some assumptions are made during the data processing and integration (e.g., concerning the interest in hourly measurements or the conditions used to detect a measurement as an outlier); all these assumptions are very reasonable but, if required for a specific use case, could be easily modified or dropped through minor modifications in the code provided (Python scripts). Similarly, other parameters that guide some processing steps can be easily adjusted in the code, if necessary.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Enriched Traffic Datasets for Madrid (Original data) (Mendeley Data).

CRediT Author Statement

Iván Gómez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft; **Sergio Ilarri:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This publication belongs to the project PID2020-113037RB-I00, funded by MICIU/AEI/10.13039/501100011033. The Next-gEnerATion dAta Management to foster suitable Behaviors and the resilience of cItizens against modErN ChallEnges (NEAT-AMBIENCE) project, led by Sergio Ilarri, aims to develop data management techniques to improve citizens' daily lives in the face of modern societal challenges, by improving decision-making. We also thank the support of the Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón (Government of Aragon: Group Reference T64_23R, COSMOS research group).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this data article.

References

[1] A. de Madrid, "Open data portal of the Madrid city council.2024 Accessed: August 17, 2024. [Online]. Available: https://datos.madrid.es/portal/site/egob.

[2] OpenStreetMap Contributors, "OpenStreetMap.2024 Accessed: August 17, 2024. [Online]. Available: https://www.openstreetmap.org/.

[3] G. Boeing, OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Comput. Environ. Urban Syst. 65 (2017) 126–139, doi:10.1016/j.compenvurbsys.2017.05.004.

[4] Scikit-learn developers, "KMeans clustering - Scikit-learn documentation.2024 [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[5] I. Gómez, S. Ilarri, Code for trafficdator-datasets, GitHub Repos. (2024). https://github.com/TrafficDator/TrafficDator-Datasets.

[6] A. de Madrid, Public space and mobility: the transformation of the Centro District of Madrid, Study Urban Transform. Mobil. (2020) [Online]. Available: https://www.madrid.es/UnidadWeb/Contenidos/Ficheros2015/centerpdf.pdf.

[7] R. Louf, M. Barthelemy, How congestion shapes cities: from mobility patterns to scaling, Sci. Rep. 10 (1) (2020), doi:10.1038/s41598-020-63614-3.

[8] I. Gómez, S. Ilarri, Enriched traffic datasets for Madrid, Mendeley Data (2024), doi:10.17632/697ht4f65b.1.

[9] A. de Madrid, "Information on the location of traffic measurement points.2024 Accessed: August 17, 2024. [Online]. Available: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=ee941ce6ba6d3410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

[10] A. de Madrid, "Traffic. Historical traffic data since 2013.2024Accessed: August 17, 2024. [Online]. Available: https://datos.madrid.es/sites/v/index.jsp?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

[11] W.G. Cochran, Sampling Techniques, 3rd ed., Wiley, New York, 1991.

[12] S. Wang, Y. Sun, Z. Bao, On the efficiency of K-means clustering: evaluation, optimization, and algorithm selection, in: Proceedings of the VLDB Endowment, 2020, pp. 163–175, doi:10.14778/3407790.3407856.

[13] The SciPy Consortium, "SciPy v1.12.0 Manual: KDTree.2024 Accessed: August 17, 2024. [Online]. Available: https://docs.scipy.org/doc/scipy-1.12.0/reference/generated/scipy.spatial.KDTree.html.

[14] Open Geospatial Consortium, "Geographic information — Well-known text representation of coordinate reference systems," *ISO/IEC 19162:2019*, 2019, Accessed: August 17, 2024. [Online]. Available: https://docs.ogc.org/is/18-010r7/18-010r7.html.

[15] A. de Madrid, "Madrid labor calendar.2024 Accessed: August 17, 2024. [Online]. Available: https://datos.madrid.es/sites/v/index.jsp?vgnextoid=9f710c96da3f9510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

[16] A. de Madrid, "Madrid daily weather data since 2019.2024 Accessed: August 17, 2024. [Online]. Available: https://datos.madrid.es/sites/v/index.jsp?vgnextoid=8d7357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

[17] GeoPandas contributors, "GeoPandas: python tools for geographic data.2024 PyPI, 2024. Accessed: August 17, 2024. [Online]. Available: https://geopandas.org/en/stable/.

[18] G. Myrovali, T. Karakasidis, A. Charakopoulos, P. Tzenos, M. Morfoulaki, G. Aifadopoulou, Exploiting the knowledge of dynamics, correlations and causalities in the performance of different road paths for enhancing urban transport management, in: Proceedings of the International Conference on Decision Support System Technology (ICDSST 2019), Lecture Notes in Business Information Processing, 313, 2019, pp. 88–99, doi:10.1007/978-3-030-18819-1_3. Springer, Cham.