Data in Brief 47 (2023) 108957

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Reconstructing secondary data based on air quality, meteorological and traffic data considering spatiotemporal components

Ditsuhi Iskandaryan, Francisco Ramos, Sergio Trilles*

Institute of New Imaging Technologies (INIT), Universitat Jaume I, Av. Vicente Sos Baynat s/n, Castelló de la Plana 12071, Spain

A R T I C L E I N F O

Article history:
Received 2 January 2023
Revised 22 January 2023
Accepted 1 February 2023
Available online 8 February 2023

Dataset link: Spatiotemporal Prediction of Air Quality Using Machine Learning Techniques (Reference data)

Keywords:
Spatiotemporal prediction
Nitrogen dioxide prediction
Geospatial analysis
Secondary data

A B S T R A C T

This paper introduces the reconstructed dataset along with procedures to implement air quality prediction, which consists of air quality, meteorological and traffic data over time, and their monitoring stations and measurement points. Given the fact that those monitoring stations and measurement points are located in different places, it is important to incorporate their time series data into a spatiotemporal dimension. The output can be used as input for various predictive analyses, in particular, we used the reconstructed dataset as input for grid-based (Convolutional Long Short-Term Memory and Bidirectional Convolutional Long Short-Term Memory) and graph-based (Attention Temporal Graph Convolutional Network) machine learning algorithms. The raw dataset is obtained from the Open Data portal of the Madrid City Council.

* Corresponding author.
E-mail address: strilles@uji.es (S. Trilles).
Social media: @Ditsuhiisk (D. Iskandaryan), @SergiTrilles (S. Trilles)

https://doi.org/10.1016/j.dib.2023.108957

Specifications Table

Subject	Computer Networks and Communications, Engineering.
Specific subject area	Study and Prediction of Air Quality in Smart Cities through Machine Learning Techniques Considering Spatiotemporal Components
Type of data	Text files (Comma Separated Values).
How data were acquired	Extracted from the Open Data portal of the Madrid City Council
Data format	Raw data, Reconstructed database
Data source location (of original dataset)	https://bit.ly/3FFRiQM
Description of data collection	The dataset presented consists of air quality, meteorological and traffic data from the period of January-June 2019 and January-June 2020, and the location of air quality and meteorological monitoring stations and traffic measurement points of the city of Madrid
Data source location	Madrid (Spain)
Related research article	Iskandaryan, D., Ramos, F. and Trilles, S., 2022. Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid. PloS one, 17(6), p.e0269295 [2]
Data accessibility	Zenodo [1] 10.5281/zenodo.7351424

Value of the Data

- The reconstructed dataset can serve as input to various machine learning and deep learning methods in order to predict air quality.
- The other researchers can use the reconstructed data for their analyses by saving resources and time in reconstruction of the raw data.
- The output of predictive analyses can help decision-makers to control air quality within a range of acceptable thresholds, and as a result, prevent negative consequences caused by poor air quality.

1. Objective

This data article adds value to published papers that have used the dataset by providing data files and detailed explanations of the reconstructed procedures. The secondary dataset is a combination of air quality, meteorological and traffic data in a grid format within a defined extent with the purpose to perform a nitrogen dioxide (NO_2) forecast.

The following work [2] is the most relevant research article for the reconstructed data already mentioned in the specification table. The aim of the work is to predict NO_2 in the spatiotemporal dimension by implementing Bidirectional Convolutional Long Short-Term Memory. The Bidirectional Convolutional Long Short-Term Memory used the reconstructed dataset as an input. This data article with a thorough explanation and refinement of the reconstruction stages provides external value to the published research article, enhances transparency to the implemented workflow, and can serve as a supplementary document describing the tedious work of preparing and reconstructing the input data.

2. Data Description

The essence of this work is to reconstruct the raw data into a format that can be used for geospatial analysis. The reason for focusing on geospatial analysis is that one of the ultimate goals of these data is to use them in further studies to predict air quality, and since air quality is controlled by many factors and components in spatiotemporal dimensions, the reconstruction

procedure must consider all these details capable of capturing geospatial and temporal dependencies. The dataset presented consists of air quality, meteorological and traffic data from the period of January-June 2019 and January-June 2020, and the location of air quality and meteorological monitoring stations and traffic measurement points of the city of Madrid. The raw data was acquired from the Open Data portal of the Madrid City Council [3] which later underwent the reconstructed procedures. There are twenty-four air quality and twenty-six meteorological control stations, and more than 4,000 traffic measurement points. Air quality stations are classified in three types: *urban background* (providing information of the exposure of the urban population), *traffic* (located in areas where their level of contamination is mainly influenced by emissions from a nearby street or highway) and *suburban* (located in areas with the highest ozone concentrations on the outskirts of the city). Regarding traffic measurement points, their location changes monthly and the Open Data portal provides location files for each month. The following variables are included in the dataset:

- Air Quality Data: NO_2 ($\mu g/m^3$).
- Meteorological Data: ultraviolet radiation (UV) (Mw/m^2), wind speed (m/s), wind direction, temperature (°C), relative humidity (%), barometric pressure (mb), solar irradiance (W/m^2), precipitation (l/m^2).
- Traffic Data: since the attributes of the traffic data can be specific to a certain area, below are the selected traffic attributes with their definition for the city of Madrid.
    - Intensity - intensity of the measurement point in a period of 15 min (vehicles/ h). A negative value implies the absence of data.
    - Occupancy time - measurement point occupancy time in a period of 15 min (%). For example, a 50% occupancy in a 15 min period means that vehicles have been positioned over the detector for 7 min and 30 s. A negative value implies the absence of data.
    - Load - vehicle loading in a 15 min period. This parameter represents an estimate of the degree of congestion, calculated from an algorithm that uses intensity and occupancy as variables, with certain correction factors. It establishes the degree of road use in a range from 0 (empty) to 100 (collapse). A negative value implies the absence of data.
    - Average traffic speed - an average speed of the vehicles in a period of 15 min (km/h). Only for M30 intercity measuring points. A negative value implies the absence of data.

It is worth to mention that among the other pollutants recorded in the city of Madrid only NO_2 was selected [recorded pollutants[1]: sulfur dioxide (SO_2) ($\mu g/m^3$), carbon monoxide (CO) (mg/m^3), nitric oxide (NO) ($\mu g/m^3$), nitrogen dioxide (NO_2) ($\mu g/m^3$), particles less than 2.5 micrometers in diameter ($PM_{2.5}$) ($\mu g/m^3$), particles less than 10 micrometers in diameter (PM_{10}) ($\mu g/m^3$), nitrogen oxides (NO_x) ($\mu g/m^3$), ozone (O_3) ($\mu g/m^3$), toluene (TOL) ($\mu g/m^3$), benzene (BEN) ($\mu g/m^3$), ethylbenzene (EBE) ($\mu g/m^3$), m-Xylene (MXY) ($\mu g/m^3$), p-Xylene (PXY) ($\mu g/m^3$), o-Xylene (OXY) ($\mu g/m^3$), total petroleum hydrocarbons (TCH) (mg/m^3), methane (CH_4) (mg/m^3), non-methane hydrocarbons (NMHC) (mg/m^3)]. The reason for choosing NO_2 is because of the work [4] related to premature mortality due to air pollution in European cities, which showed that Madrid has the highest NO_2 mortality burden.

Regarding temporal resolution, although the traffic data is captured every 15 min, however, since NO_2 and meteorological data are at hourly rates, the traffic data was filtered. Only hourly records were selected (for example, with entries at 13:00, 13:15, 13:30, 13:45 and 14:00, we selected the entries at 13:00 and 14:00 and the same logic was applied for the entire period).

After accessing the raw data, the subsequent important step is creating the secondary dataset through data integration and incorporation in spatial, as well as in temporal dimensions in order to capture and compute spatiotemporal dependencies. Since the location of the air quality

[1] Interpreter of air quality data files: https://bit.ly/3Utz9g5. Accessed February 15, 2023.

Table 1
Summary statistics of the periods January-June 2019 and January-June 2020 for each data type.

Phenomena	Descriptors	January-June 2019	January-June 2020
NO_2 ($\mu g/m^3$)	Mean (SD)	36.69 (30.85)	26.03 (25.35)
	Median [Min, Max]	27.0 [0.0, 328]	17.0 [0.0, 326]
UV (Mw/m^2)	Mean (SD)	15.83 (30.27)	-
	Median [Min, Max]	1.0 [0.0, 199]	-
Wind speed (m/s)	Mean (SD)	1.41 (1.11)	1.31 (1.05)
	Median [Min, Max]	1.14 [0.0, 8.75]	1.05 [0.0, 8.97]
Wind direction	Mean (SD)	167.80 (105.72)	140.82 (98.35)
	Median [Min, Max]	182.0 [0.0, 359]	135.0 [0.0, 359]
Temperature (°C)	Mean (SD)	13.38 (8.09)	13.63 (7.6)
	Median [Min, Max]	12.5 [-55.0, 47.3]	12.6 [-55.0, 44.6]
Humidity (%)	Mean (SD)	48.73 (21.60)	60.76 (22.77)
	Median [Min, Max]	47.0 [-25, 100]	62.0 [-25, 100]
Pressure (mb)	Mean (SD)	943.3 (34.91)	940.62 (63.28)
	Median [Min, Max]	945.0 [0.0, 962.0]	945.0 [0.0, 1073.0]
Solar Irradiance (W/m^2)	Mean (SD)	220.73 (301.06)	191.95 (279.83)
	Median [Min, Max]	11.0 [0.0, 1103.0]	9.0 [0.0, 1113.0]
Precipitation (l/m^2)	Mean (SD)	0.03 (0.41)	0.03 (0.27)
	Median [Min, Max]	0.0 [0.0, 30.4]	0.0 [0.0, 13.5]
Intensity (vehicles/ h)	Mean (SD)	885863 (59.98%)	892197 (60.09%)
	Median [Min, Max]	245.69 (402.73)	161.45 (313.33)
		63.0 [0.0, 6348.0]	34.19 [0.0, 6588.0]
Occupancy time (%)	Mean (SD)	845031 (57.21%)	822652 (55.41%)
	Median [Min, Max]	3.96 (6.36)	2.57 (4.9)
		0.95 [0.0, 100.0]	0.42 [0.0, 99.0]
Load	Mean (SD)	881500 (59.68%)	884950 (59.60%)
	Median [Min, Max]	11.65 (14.91)	7.85 (11.75)
		4.0 [0.0, 100.0]	2.2 [0.0, 100.0]
Average traffic speed (km/h)	Mean (SD)	233415 (15.8%)	223052 (15.0%)
	Median [Min, Max]	4.39 (13.28)	4.04 (12.96)
		0.0 [0.0, 96.5]	0.0 [-127.0, 127.0]

stations, meteorological stations and traffic measurement points are different, our approach is to combine them in a temporal grid format, which allows capturing spatiotemporal interconnections. The initial step was to create a grid in a specified area (Fig. 1), which was defined as a section of the city of Madrid with a width and height of 1,000 metres within the following boundaries: Top – 4,486,449.725263 metres; Bottom – 4,466,449.725263 metres; Left – 434,215.234430 metres; Right –451,215.234430 metres. Regarding the projected coordinate system, EPSG: 25830, ETRS89/UTM zone 30N was used (EPSG: European Petroleum Survey Group, ETRS89: European Terrestrial Reference System 1989, UTM: Universal Transverse Mercator)[2].

Summary statistics for each type of data for the periods studied for the defined area are displayed in Table 1.

The reconstructed dataset was already used in several publications [2,5,6,7]. This work [2] used the reconstructed data to predict NO_2 in the next 6 h using previous 6 h data by implementing Bidirectional Convolutional Long Short-Term Memory. The following work [5] used the part of reconstructed dataset, including only air quality and meteorological data, to predict NO_2 in next 1, 12, 24 and 48 h based on the previous 24 h data by implementing Convolutional Long Short-Term Memory. An additional objective was to compare the performance of predictive analysis between the pandemic (January-June 2020) and non-pandemic (January-June 2019) periods to find out if restrictions applied to curb the progress of COVID-19 impacted predicted output. Another work [6] was devoted to an exploratory analysis of the reconstructed dataset and the implementation of feature selection methods in terms of NO_2 prediction. Our next work [7] used the data as an input to the Attention Temporal Graph Convolutional Network to predict

[2] Projected coordinate system: https://epsg.io/25830. Accessed February 15, 2023.

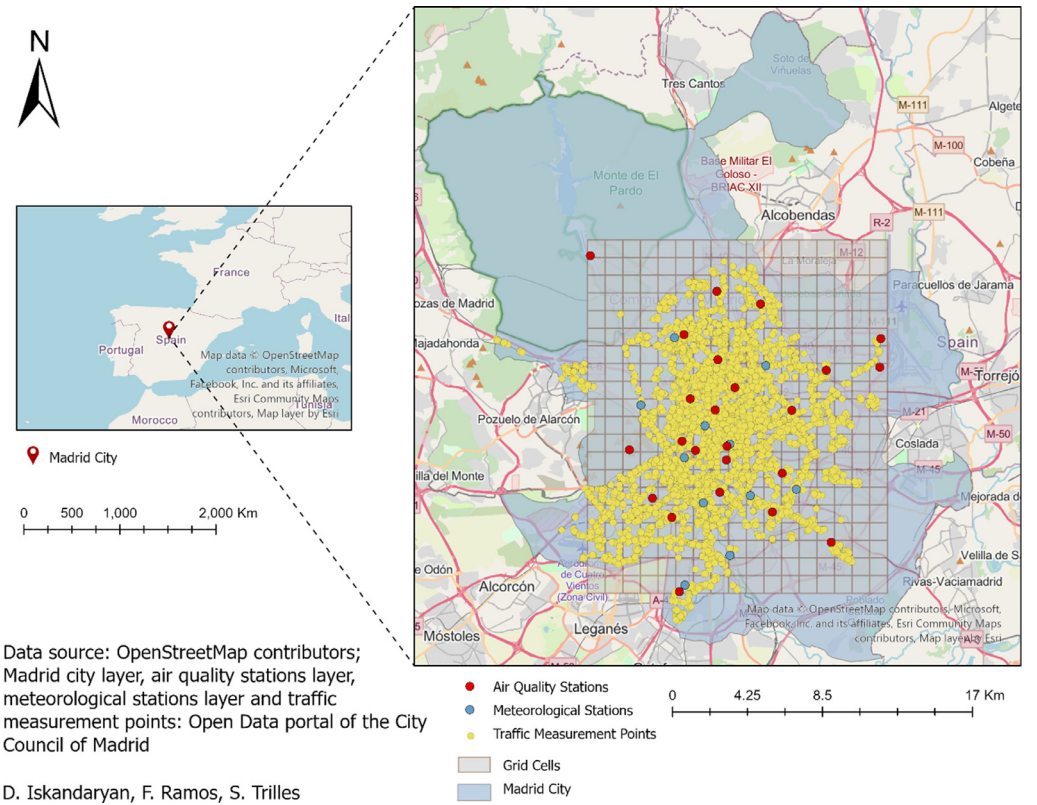

Fig. 1. Air quality stations, meteorological stations, traffic measurement points (January 2019) and grid cells segments on the defined area of the city of Madrid.

the concentration of NO_2 in the next 1–12 h, 12–24 h, 24–36 h, and 36–48 h. As it can be seen, the secondary data, due to its structure, becomes flexible and multifunctional in various machine learning and deep learning methods and allows them to be used in models with different architectures and compare the performance of the implemented models without any significant efforts.

The uniqueness of this reconstructed secondary dataset can be highlighted by the following components, including (1) the purpose of the dataset: to predict NO_2; (2) the features (NO_2, meteorological and traffic data): relevant to NO_2 with the purpose to perform predictive analysis of NO_2 with higher accuracy; (3) spatial and temporal extension: defined extent in the city of Madrid using data from the period January-June 2019 and January-June 2020.

The above studies highlight the importance of the reconstructed dataset and the value that they may have to the research community. Overall, the main contributions of reconstructed dataset are the ability to save expenses, resources, efforts and time to repeat the same procedure for different studies and analyses, it serves as a tool to disseminate primary data, and it allows to generate new insights from primary data.

3. Experimental Design, Materials and Methods

3.1. Data Creation

This section provides a detailed procedure of the reconstruction stage from raw data, i.e. the protocol which will make the work more transparent and replicable.

As already mentioned the raw data was obtained from the Open Data portal of the Madrid City Council [3]. After defining the relevant and important datasets to perform air quality prediction (air quality, meteorological and traffic data), we used the Application Programming Interface (API) in order to extract the data. Below are the APIs for each dataset and the description of the following procedures:

Air quality data: https://bit.ly/3ZLEdzm. The API enables us to access the data for each year. After extracting the .zip file of the period of interest (2019 and 2020), we combined the contents of all .csv files into the pandas dataframe. The dataframe looks like Table 2 which will be detailed later. The next step was to filter and extract information only on contamination of NO_2 for the first six months of the defined years (January-June 2019 and January-June 2020) per air quality monitoring station.

Meteorological data: https://bit.ly/3krJWKa. By the API we access the data for each year. After extracting the .zip file of the period of interest (2019 and 2020), we combined the contents of all .csv files into the pandas dataframe. The next step was to extract the data for the period January-June 2019 and January-June 2020 per meteorological monitoring station and per magnitude.

Traffic data: https://bit.ly/3waELAW. There is a .zip file for each month. Considering the fact that the location of traffic measurement points was changing every month, and also it is captured every 15 min, we created .csv file for each hour during the defined period (January-June 2019 and January-June 2020).

After acquiring data from the Open Data portal using the API for which we used Google Colab with Python, the following step was the concatenation of the datasets in spatiotemporal dimensions using ArcGIS Pro with ArcPy package.

As mentioned, the raw data went through a reconstructed grid-based transformation process, allowing it to be used in predictive geospatial analysis. The grid was created with the help of *ArcPy* package [8], specifically with the *CreateFishnet* function [9]. Within the required extent, the output generated a grid with 340 cells (20 by 17) covering 340 km^2 or 56.27% of the total area of the city of Madrid. The rationale for selecting this area was to have a minimum extent to encompass all air quality control stations. The value of each cell consists of the values of NO_2, meteorological and traffic attributes obtained from assigned stations covered by that cell at a certain time. The value of the cell that does not contain any station was set to zero and in the case of

Table 2
Air quality data

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	H02	V02
(PROVINCE)	(MUNICIPALITY)	(STATION)	(MAGNITUDE)	(POINT_SAMPLE)	(YEAR)	(MONTH)	(DAY)	(H01)	(V01)	(H02)	(V02)
28	79	4	1	28079004 1 38	2019	1	1	23	V	17	V

Algorithm 1
Data preparation

Input: Data - [Hourly NO_2, Meteorological and Traffic data]; Period - [01.01.2019-30.06.2019; 01.01.2020-30.06.2020]

1: **for** each hour ∈ Period **do**
2: Create grid with Fishnet tool (ArcPy library)
3: Add field to the Fishnet
4: **for** each item $i \in$ Data **do**
5: i spatial join with grid: arcpy.management.AddField, arcpy.analysis.SpatialJoin, arcpy.da.SearchCursor, arcpy.da.UpdateCursor
6: input the mean of the values of each corresponding cell to the field
7: end for
8: end for
Output: *.csv* files for each hour including NO_2, Meteorological and Traffic data

several stations, an average value was calculated and assigned to the cell. The above procedure was repeated for each hour of the selected period. The following functions were used to execute the aforementioned process, including *arcpy.management.AddField* [10], *arcpy.analysis.SpatialJoin* [11], *arcpy.da.SearchCursor* [12], arcpy.da.UpdateCursor [13]. The output was exported as *.csv* files. Overall, 4,344 and 4,368 *.csv* files were generated corresponding to every hour during January-June 2019 and January-June 2020, respectively (Fig. 2). The input data X can be defined as follows:

$$X = X_{no2} + X_{uv} + X_{ws} + X_{wd} + X_{temp} + X_{hum} + X_{press} + X_{sr} + X_{prec} + X_{intens} + X_{ocup} + X_{load} + X_{ats}$$

where "+" is a vector concatenation operator, $X_{no2} \in R_{no2}^{sxmxn}$ is the NO_2 input data, R_{no2} is the NO_2 domain; $X_{uv} \in R_{uv}^{sxmxn}$ is the UV input data, R_{uv} is the UV domain; $X_{ws} \in R_{ws}^{sxmxn}$ is the wind speed input data, R_{ws} is the wind speed domain; $X_{wd} \in R_{wd}^{sxmxn}$ is the wind direction input data, R_{wd} is the wind direction domain; $X_{temp} \in R_{temp}^{sxmxn}$ is the temperature input data, R_{temp} is the temperature domain; $X_{hum} \in R_{hum}^{sxmxn}$ is the relative humidity input data, R_{hum} is the relative humidity domain; $X_{press} \in R_{press}^{sxmxn}$ is the barometric pressure input data, R_{press} is the barometric pressure domain; $X_{sr} \in R_{sr}^{sxmxn}$ is the solar irradiance input data, R_{sr} is the solar irradiance domain; $X_{prec} \in R_{prec}^{sxmxn}$ is the precipitation input data, R_{prec} is the precipitation domain; $X_{intens} \in R_{intens}^{sxmxn}$ is the intensity input data, R_{intens} is the intensity domain; $X_{ocup} \in R_{ocup}^{sxmxn}$ is the occupancy time input data, R_{ocup} is the occupancy time domain; $X_{load} \in R_{load}^{sxmxn}$ is the load input data, R_{load} is the load domain; $X_{ats} \in R_{ats}^{sxmxn}$ is the average traffic speed input data, R_{ats} is the average traffic speed domain, s is the number of samples: 4,344 and 4,368 for January-June 2019 and January-June 2020, respectively, m is equal 20, and n is equal 17. The final input $X \in R^{sx340xf}$, where s is the number of samples: 4,344 and 4,368 for January-June 2019 and January-June 2020, respectively, 340 is the multiplication of m and n (20*17), and f is the number of features equal to 13 ($X \in R^{4344 \times 340 \times 13}$ for January-June 2019 and $X \in R^{4368 \times 340 \times 13}$ for January-June 2020).

Overall, the data extraction and preparation workflow can be defined as follows:

(1) Retrieve requirements datasets from the Open Data portal of the Madrid City Council using their API.
(2) Filter and extract only the data that will be needed for prediction of NO_2 over a specific period (January-June 2019 and January-June 2020).
(3) Generate a grid within the required extent over the study area.
(4) Combine the datasets along with the generated grid by merging the datasets in spatiotemporal dimensions.

A formal description of the data preparation process is given by Algorithm 1.

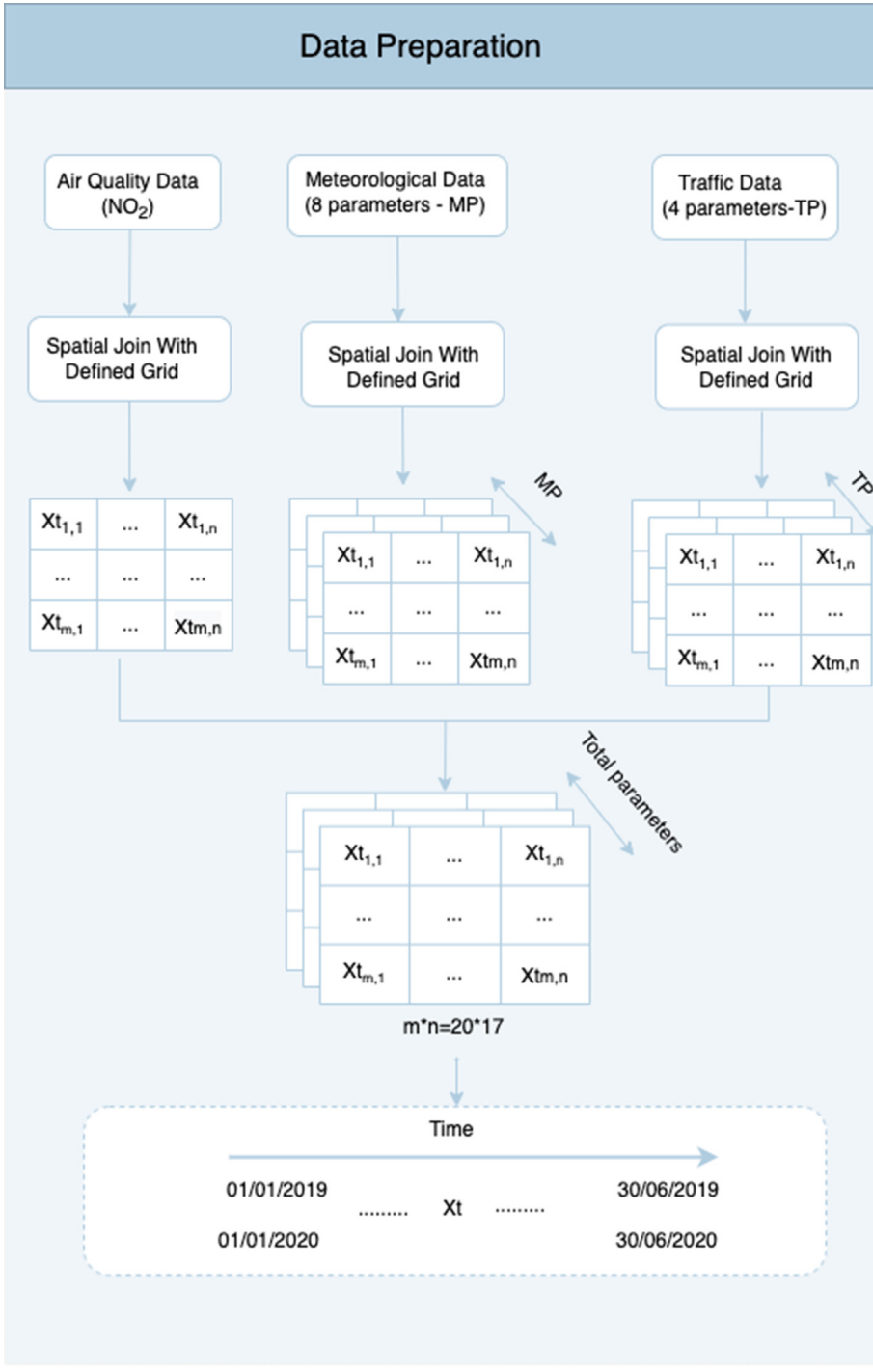

Fig. 2. The workflow of data preparation.

3.2. Data Publication

The raw data processed data and the code implemented to process raw data are displayed in Fig. 3 as a root directory named *Air_Quality_Prediction*, which is composed of two main sub-directories *Data* and *Code*.

Data sub-directory consists of *Raw-Data* and *Processed-Data*. *Raw-Data* includes *AirQuality-Data, Meteorological-Data*, and *Traffic-Data; Processed-Data* includes *AirMetTraffic_2019_2020_firstSixMonths.zip*, *Madrid_wind_yyyy.csv, Madrid_Stations_yyyy.csv, Madrid_Exploration.zip*, and *distanceNodes.txt*.

AirQuality-Data: consists of *Anio201912.zip, Anio202012.zip* and *informacion_estaciones_red_calidad_aire.geo*. The first two *.zip* files[3] contain hourly air quality data for 2019 and 2020, respectively. The data from January to June 2019 and from January to June 2020 were used in the current work. Each *.zip* file contains data in three formats: *.txt, .csv, .xml*. For our analysis, we used *.csv* files. Each record is structured as follows (Table 2; in the brackets English version of the columns):

The **POINT_SAMPLE** field includes the complete station code (province, municipality, station, magnitude, and technique of sampling); **H01** corresponds to the data of 1 a.m. of that day; **V01** is the validation code; **H02** at 2 a.m.; **V02** and so on. **Magnitude** refers to the pollutants that were recorded by the stations, of which we only focused on NO_2, which is mentioned under magnitude 8[4].

The location of the air quality monitoring stations is available in *.csv, .xlsx*, and *.geo* format[5]. This work used the *.geo* format: *informacion_estaciones_red_calidad_aire.geo*.

Meteorological-Data: consists of *mmm_meteo20.csv, mmm_meteo19.csv* and *Estaciones_control_datos_meteorologicos.geo*. The *mmm* of the names of *mmm_meteo20.csv* and *mmm_meteo19.csv* refers to the name of the corresponding month[6]. Each record of these *.csv* files is structured as follows (Table 3; in the brackets English version of the columns):

The **POINT_SAMPLE** field includes the complete station code (province, municipality, station, magnitude, and technique of sampling); **H01** corresponds to the data of 1 a.m. of that day; **V01** is the validation code; **H02** at 2 a.m.; **V02** and so on. **Magnitude** refers to the codes of the meteorological features (features with corresponding codes: UV (Mw/m^2)-80, wind speed (m/s)-81, wind direction-82, temperature (°C)-83, relative humidity (%)-86, barometric pressure (mb)-87, solar irradiance (W/m^2)-88, precipitation (l/m^2)-89)[7].

The location of the meteorological monitoring stations is available in *.csv, .xlsx,* and *.geo* format[8]. This work used the *.geo* format: *Estaciones_control_datos_meteorologicos.geo*.

Traffic-Data: consists of *mm-yyyy.zip* and *pmed_ubicacion_mm-yyyy.zip*. *mm-yyyy.zip* is available for each month, which contains *.csv* file[9]. The name of each *.csv* file contains the name of the corresponding month with the corresponding year. Each record is structured as follows (Table 4; in the brackets English version of the columns):

The SICTRAM database records and integrates all the vehicle detectors' data of the control measurement points over periods of 15 min. This current work used the following data: date (it was used to create hourly *.csv* files), intensity, occupancy time, load, and average traffic data[10].

The location of the traffic measurement points is available for every month in *.csv, .xlsx*, and *.zip* format[11]. This work used a *.zip* file: *pmed_ubicacion_mm-yyyy.zip*, each of them contains *.dbf, .prj, .shp*, and *.shx* files.

[3] Air quality. Hourly data since 2001: https://bit.ly/2IeGcrs. Accessed February 15, 2023.
[4] Interpreter of air quality data files: https://bit.ly/3Utz9g5. Accessed February 15, 2023.
[5] Air quality. Control stations: https://bit.ly/2Kp8TlV. Accessed February 15, 2023.
[6] Meteorological data. Hourly data from 2019: https://bit.ly/3DIkLLk. Accessed February 15, 2023.
[7] Interpreter of meteorological data files: https://bit.ly/3LzX8qb. Accessed February 15, 2023.
[8] Meteorological data. Control stations: https://bit.ly/3S3ZP5x. Accessed February 15, 2023.
[9] Traffic. Historical traffic data since 2013: https://bit.ly/3BBUxHs. Accessed February 15, 2023.
[10] Description of traffic dataset: https://bit.ly/3qTUJwZ. Accessed February 15, 2023.
[11] Traffic. Location of traffic measurement points: https://bit.ly/2rOkHCX. Accessed February 15, 2023.

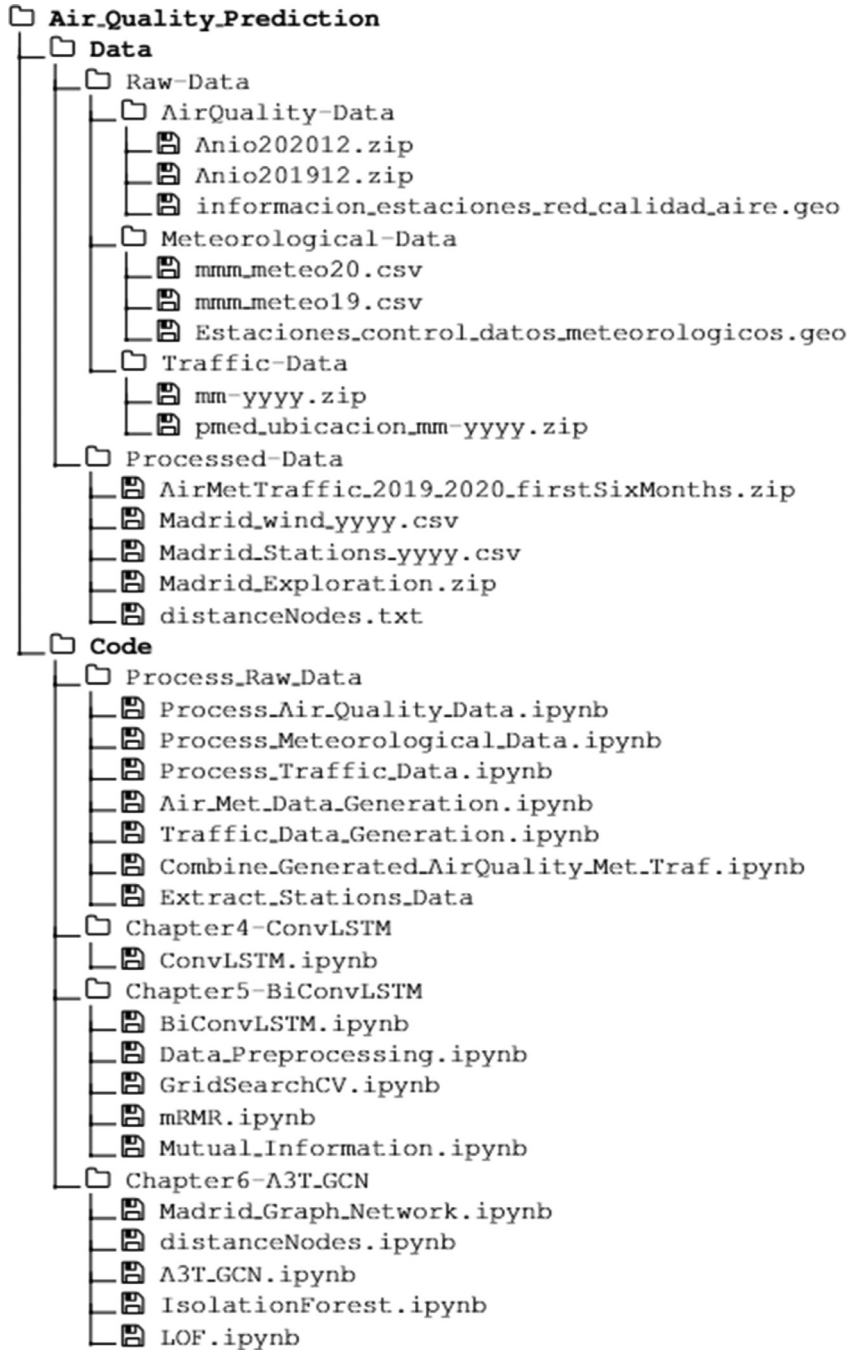

Fig. 3. Directory tree illustrating the data and implemented code.

Table 3
Meteorological data.

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	H02	V02
(PROVINCE)	(MUNICIPALITY)	(STATION)	(MAGNITUDE)	(POINT_SAMPLE)	(YEAR)	(MONTH)	(DAY)	(H01)	(V01)	(H02)	(V02)
28	79	104	82	28079004 82 98	2019	1	1	23	V	17	V

Table 4
Traffic data.

id	fecha	tipo_elem	intensidad	ocupacion	carga	vmed	error	periodo_integracion
(id)	(date)	(element_ type)	(intensity)	(occupancy_time)	(load)	(average traffic speed)	(error)	(integration_ period)
1001	01/01/2019 00:00	M302	2340	11	0	63	N	5

AirMetTraffic_2019_2020_firstSixMonths.zip: contains *.csv* files generated for each hour from January to June 2019 and from January to June 2020. Each *.csv* file name has the following structure: *fishnetAirMetyyyy_m_dd_h.csv*. There are 4344 and 4368 *.csv* files corresponding to every hour during January-June 2019 and January-June 2020, respectively. Each *.csv* file consists of 340 rows and 14 columns (*#FID, NO2, UV, windSpeed, windDir, Temp, Humidity, Pressure, SolarRad, Prec, intensidad, ocupacion, carga, vmed*).

Madrid_wind_yyyy.csv: is the modified data of the content of *AirMetTraffic_2019_2020_ firstSixMonths.zip*. The modification was applied to the wind direction. It was transformed in two ways: 1) converting wind direction into categorical data (north, east, south, west, southwest, northeast, southeast, and northwest), and passing through One Hot Encoder; 2) converting wind direction into *u* and *v* components. The *Madrid_wind_yyyy.csv* contains records for every hour during January-June 2019 and January-June 2020, and for every cell of the defined area of the city of Madrid. The columns are *NO2, windSpeed, Temp, Humidity, Pressure, SolarRad, intensidad, ocupacion, carga, vmed, v_comp, u_comp, windDir_Categ_east, windDir_Categ_north, windDir_Categ_northeast, windDir_Categ_northwest, windDir_Categ_south, windDir_Categ_southeast, windDir_Categ_southwest, windDir_Categ_west*.

Madrid_Stations_yyyy.csv:: is part of *Madrid_wind_yyyy.csv*, which includes only data from cells containing air quality monitoring stations. The data are appended in one column with the following order: NO2, intensidad, ocupacion, windSpeed, Pressure, SolarRad, Temp, Humidity, carga, vmed, v_comp, u_comp, windDir_Categ_east, windDir_Categ_north, windDir_Categ_northeast, windDir_Categ_northwest, windDir_Categ_south, windDir_Categ_southeast, windDir_Categ_southwest, windDir_Categ_west.

Madrid_Exploration.zip: contains the result of an exploratory analysis that identifies the relationship between NO_2 and additional features (meteorological and traffic data). distanceNodes.txt: includes the distance between the air quality monitoring stations placed in the city of Madrid (24 stations, 276 edges each edge is placed 2 times depending on the node order: origin, destination).

Code sub-directory consists of *Process_Raw_Data, Chapter4-ConvLSTM, Chapter5-BiConvLSTM,* and *Chapter6-A3T_GCN*.

Process_Raw_Data: is composed of *Process_Air_Quality_Data.ipynb, Process_Meteorological_ Data.ipynb, Process_Traffic_Data.ipynb, Air_Met_Data_Generation.ipynb, Traffic_Data_Generation. ipynb, Combine_Generated_AirQuality_Met_Traf.ipynb*, and *Extract_Stations_Data.ipynb.* The first three files are dedicated to processing the raw data for each dataset, respectively. *Air_Met_Data_Generation.ipynb* combines processed air quality and meteorological data in a spatiotemporal dimension. *Traffic_Data_Generation.ipynb* combines processed traffic data in a spatiotemporal dimension. *Combine_Generated_AirQuality_Met_Traf.ipynb* combines generated air quality, meteorological and traffic data for each hour in a separate *.csv* file. *Extract_Stations_Data.ipynb* contains the procedure to extract cells or rows including NO_2, Meteorological and Traffic data, where air quality monitoring stations exist.

3.3. Analyses Performed

We applied the reconstructed dataset in further analysis to predict air quality, particularly, to predict NO_2. Two main approaches were implemented: grid-based and graph-based. Grid-

based approaches are Convolutional Long Short-Term Memory and Bidirectional Convolutional Long Short-Term Memory, which use reconstructed data as an input. The input of grid-based approaches has the following shape: *samples, time_steps, channels, rows, cols*. In our analysis *samples* were assigned to 4,344 and 4,368 for January-June 2019 and January-June 2020, respectively; *time_steps* was assigned to 6 h (Bidirectional Convolutional Long Short-Term Memory) and 24 h (Convolutional Long Short-Term Memory); *channels* were assigned to 1; *row* assigned to 20; *cols* assigned to 17 (depending the number of variables/features included in the analysis the certain number of grids were concatenated along the row axis). Depending on the analysis the samples were splitted with different percentages in training, validation, and testing sets. Regarding graph-based approach, Attention Temporal Graph Convolutional Network was implemented. The input of the graph-based approach was extracted from grid-based input. Of the 340 grid cells, only 24 cells containing air quality monitoring stations were selected, which served as nodes for constructing the graph (24 nodes), and the connecting nodes with edges were assigned a weight equal to $1/d_{ij}$, where d_{ij} is the distance between i and j stations. Below are listed the subdirectories from Fig. 3 dedicated to the analyses performed:

Chapter4-ConvLSTM: includes ConvLSTM.ipynb, which develops and tests the Convolutional Long Short-Term Memory method by implementing it in two different periods: pandemic and non-pandemic.

Chapter5-BiConvLSTM: includes BiConvLSTM.ipynb, *Data_Preprocessing.ipynb, GridSearchCV.ipynb, mRMR.ipynb,* and *Mutual_Information.ipynb. BiConvLSTM.ipynb* develops and tests the Bidirectional Convolutional Long Short-Term Memory method. *Data_Preprocessing.ipynb* refers to the data pre-processing step, including implementation of neural network, outlier detection based on the statistical summary of the dataset, and the conversion of the wind direction (converting it to categorical data (north, east, south, west, southwest, northeast, southeast, northwest) and passing through One Hot Encoder). GridSearchCV.ipynb refers to parameter optimisation of the proposed model performed by applying GridSearchCV with Blocking Time Series Split. *mRMR.ipynb* and *Mutual_Information.ipynb* execute two feature selection techniques: Maximum Relevance — Minimum Redundancy and Mutual Informatio, respectively.

Chapter6-A3T_GCN: includes *Madrid_Graph_Network.ipynb, distanceNodes.ipynb, A3T_GCN.ipynb, IsolationForest.ipynb*, and *LOF.ipynb. Madrid_Graph_Network.ipynb* contains the procedure for constructing a graph network of the air quality stations placed in the city of Madrid. *distanceNodes.ipynb* includes the procedure for calculating the distance between the air quality stations placed in the city of Madrid (24 stations). *A3T_GCN.ipynb* develops and tests the Attention Temporal Graph Convolutional Network method. *IsolationForest.ipynb,* and *LOF.ipynb* execute two outlier detection techniques: Isolation Forest and Local Outlier Factor, respectively.

Ethics Statement

The raw data of this study is provided by open in full compliance with ethical requirements for publication in the journal of Data in Brief. This study does not involve any modern human or animal subject.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Data Availability

Spatiotemporal Prediction of Air Quality Using Machine Learning Techniques (Reference data) (Zenodo).

CRediT Author Statement

Ditsuhi Iskandaryan: Conceptualization, Methodology, Writing – review & editing; **Francisco Ramos:** Conceptualization, Methodology, Supervision, Writing – review & editing; **Sergio Trilles:** Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition.

Acknowledgments

Grant PID2019-104065GA-I00 funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe", by the European Union. **Ditsuhi Iskandaryan** has been funded by the predoctoral programme PINV2018 - Universitat Jaume I (PREDOC/2018/61) and **Sergio Trilles** has been funded by the Juan de la Cierva - Incorporación postdoctoral programme of the Ministerio de Ciencia e Innovación - Spanish government (IJC2018-035017-I) funded by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe", by the European Union.

References

[1] Iskandaryan, Ditsuhi, Ramos, Francisco and Trilles, Sergio. (2022). Supplementary Materials for 'Spatiotemporal Prediction of Air Quality Using Machine Learning Techniques' [Data set]. Zenodo. doi:10.5281/zenodo.7351424.
[2] D. Iskandaryan, F. Ramos, S. Trilles, Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid, PloS one 17 (6) (2022) e0269295.
[3] Portal de datos abiertos del Ayuntamiento de Madrid. Catálogo de datos. Conjuntos de datos. https://bit.ly/3FFRiQM. Accessed February 15, 2023.
[4] S. Khomenko, M. Cirach, E. Pereira-Barboza, N. Mueller, J. Barrera-Gómez, D. Rojas-Rueda, K. de Hoogh, G. Hoek, M. Nieuwenhuijsen, Premature mortality due to air pollution in European cities: a health impact assessment, Lancet Planet. Health 5 (3) (2021) e121–e134, doi:10.1016/S2542-5196(20)30272-2.
[5] D. Iskandaryan, F. Ramos, S. Trilles, Comparison of nitrogen dioxide predictions during a pandemic and non-pandemic scenario in the city of madrid using a convolutional LSTM network, Int. J. Comput. Intell. Appl. 21 (02) (2022) 2250014.
[6] D. Iskandaryan, S. Di Sabatino, F. Ramos, S. Trilles, Exploratory analysis and feature selection for the prediction of nitrogen dioxide, AGILE GIScience Ser. 3 (6) (2022), doi:10.5194/agile-giss-3-6-2022.
[7] D. Iskandaryan, F. Ramos, S. Trilles, Spatiotemporal prediction of nitrogen dioxide based on graph neural networks, Advances and New Trends in Environmental Informatics, 7, Springer, Cham, 2022 ENVIROINFO 2022. Progress in IS, doi:10.1007/978-3-031-18311-9.
[8] Esri. ArcGIS Pro. Python. What is ArcPy? https://bit.ly/3UPYKjy. Accessed February 15, 2023.
[9] Esri. ArcGIS Pro. Tool Reference. Create Fishnet (Data Management). https://bit.ly/3Yv4vVx. Accessed February 15, 2023.
[10] Esri. ArcGIS Pro. Tool Reference. Add Field (Data Management). https://bit.ly/3LPo1GE. Accessed February 15, 2023.
[11] Esri. ArcGIS Pro. Tool Reference. Spatial Join (Analysis). https://bit.ly/3M6SC2J. Accessed February 15, 2023.
[12] Esri. ArcGIS Pro. Python. SearchCursor. https://bit.ly/3y3tcNz. Accessed February 15, 2023.
[13] Esri. ArcGIS Pro. Python. UpdateCursor. https://bit.ly/3y0txjU. Accessed February 15, 2023.