Data Article

H2020 project CAPTOR dataset: Raw data collected by low-cost MOX ozone sensors in a real air pollution monitoring network

Jose M. Barcelo-Ordinas\textsuperscript{a,*}, Pau Ferrer-Cid\textsuperscript{a}, Jorge Garcia-Vidal\textsuperscript{a}, Mar Viana\textsuperscript{b}, Ana Ripoll\textsuperscript{b,c}

\textsuperscript{a} Universitat Politècnica de Catalunya, Barcelona, Spain
\textsuperscript{b} Institute of Environmental Assessment and Water Research, Spanish National Research Council (IDAEA-CSIC), Barcelona, Spain
\textsuperscript{c} 4sfera Innova, Girona, Spain

\textbf{A B S T R A C T}

The H2020 CAPTOR project deployed three testbeds in Spain, Italy and Austria with low-cost sensors for the measurement of tropospheric ozone (O\textsubscript{3}). The aim of the H2020 CAPTOR project was to raise public awareness in a project focused on citizen science. Each testbed was supported by an NGO in charge of deciding how to raise citizen awareness according to the needs of each country. The data presented in this document correspond to the raw data captured by the sensor nodes in the Spanish testbed using SGX Sensortech MICS 2614 metal-oxide sensors. The Spanish testbed consisted of the deployment of twenty-five nodes. Each sensor node included four SGX Sensortech MICS 2614 ozone sensors, one temperature sensor and one relative humidity sensor. Each node underwent a calibration process by co-locating the node at an EU reference air quality monitoring station, followed by a deployment in a sub-urban or rural area in Catalonia, Spain. All nodes spent two to three weeks co-located at a reference station in Barcelona, Spain (urban area), followed by two to three weeks co-located at three sub-urban reference stations near the final deployment.

\textsuperscript{*} Corresponding author.
E-mail addresses: joseb@ac.upc.edu (J.M. Barcelo-Ordinas), pferrer@ac.upc.edu (P. Ferrer-Cid), jorge@ac.upc.edu (J. Garcia-Vidal), mar.viana@idaea.csic.es (M. Viana), anna.ripoll@4sfera.com (A. Ripoll).

https://doi.org/10.1016/j.dib.2021.107127

\textcopyright{} 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
The nodes were then deployed in volunteers’ homes for about two months and, finally, the nodes were co-located again at the sub-urban reference stations for two weeks for final calibration and assessment of potential drifts. All data presented in this paper are raw data taken by the sensors that can be used for scientific purposes such as calibration studies using machine learning algorithms, or once the concentration values of the nodes are obtained, they can be used to create tropospheric ozone pollution maps with heterogeneous data sources (reference stations and low-cost sensors).

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

### Specifications Table

| Subject | Environmental science (air pollution). |
|---------|-----------------------------------------|
| Specific subject area | Measurements of tropospheric ozone (O₃) using metal-oxide low-cost sensors. |
| Type of data | CSV files |
| How data were acquired | Captor nodes designed and built by UPC (Universitat Politècnica de Catalunya, Spain), during the H2020 CAPTOR project, using SGX Sensortech MICS 2614 ozone (O₃) metal-oxide (MOX) sensors. |
| Data format | Raw: ozone in KΩ, temperature in °C, relative humidity in %. |
| Parameters for data collection | Maximum ozone concentrations occur in summer with high temperatures and high solar radiation. Thus, the data are sensitive to environmental conditions (temperature and relative humidity). Each node first passed through a reference station in April 2017 (spring) in an urban area of Barcelona to test the nodes. Then, the nodes were moved to the deployment area with different environmental conditions during the period from May to September (late spring and all summer), 2017. Each ozone sample presented in the files is the result of an average of 100 samples taken every 5 s, to avoid outliers and spikes in the data. |
| Description of data collection | The nodes passed through four locations: reference station in Barcelona, reference station near the deployment site, deployment site and reference station near the deployment site. Each node sent every half hour the value measured by each sensor to a cloud repository. The measured values of the metal-oxide sensors, after passing through the analogue-to-digital converter, are in Ohms. The values of the temperature and relative humidity sensors are in °C and %. The values presented in this document are the raw values sent by the nodes to the repository. |
| Data source location | Catalonia, Spain. |
| Data accessibility | Repository name: Zenodo |
| Data identification number: twenty-five files called captor170xx_raw.csv with xx=01, 02, ..., 27, where two numbers (08 and 24) are left over because nodes 17008 and 17024 were not deployed on the testbed in Spain. A zip file called captor_2017_raw_files.zip is also provided which includes all CSV files. Direct URL to data: [http://doi.org/10.5281/zenodo.4570449](http://doi.org/10.5281/zenodo.4570449) |
| Related research article | A. Ripoll, M. Viana, M. Padroa, X. Querol, A.Minutolo, K-M. Hou, J. M. Barcelo-Ordinas and J. Garcia-Vidal, (2019). Testing the performance of sensors for ozone pollution monitoring in a citizen science approach. Science of the total environment (Elsevier), 651, 1166–1179, [https://doi.org/10.1016/j.scitotenv.2018.09.257](https://doi.org/10.1016/j.scitotenv.2018.09.257) |

### Value of the Data

- The sensor data complement the data provided by the AQMN (Air Quality Monitoring Network) reference stations deployed by governmental organisations. The spatial density of AQM reference stations is typically low, therefore data from nodes with low-cost sensors complements the AQM reference network. The data are raw, meaning that the data are as measured by a real sensor in the field, without being calibrated.
• As the nodes are deployed over a wide area of about 30Kmx25Km, many different studies can be carried out in various research fields: calibration, data fusion, sensor networks, interpolation of calibrated ozone concentrations, among others. Therefore, the major beneficiaries of these data are researchers in air pollution monitoring networks using low-cost sensors.
• The data can be used by researchers in several fields: i) researchers interested in building maps with ozone concentrations in areas where AQM reference stations are not deployed, ii) researchers interested in testing machine learning algorithms for low-cost sensor calibration, iii) researchers interested in studying network of heterogeneous data sources including reference stations and low-cost sensor nodes.
• Each node has up to four O₃ metal-oxide sensors mounted on it. As the deployment has twenty-five nodes, it adds up to one hundred low-cost sensors. In addition, some nodes were left in place at reference stations throughout the summer to have a large part of the dataset in the same place as the reference data, while other nodes were moved to volunteers’ homes. This allows calibration-only studies with a large number of sensors, or research of sensors forming a network.

1. Data Description

The data can be accessed in Zenodo (http://doi.org/10.5281/zenodo.4570449). There are twenty-five CSV files called “captor170xx_raw.csv” with xx=01, 02, ..., 27, where two numbers (08 and 24) are left over because nodes 17008 and 17024 were not deployed on the testbed in Spain. A zip file called captor_2017_raw_files.zip is also provided which includes all CSV files for easy downloading of all files. Each CSV file corresponds to a Captor node. The format of each CSV file is as follows:

```
date;s1_o3;s2_o3;s3_o3;s4_o3;s_temp;s_rh;location
07/04/2017 01:30;637.43;337.4°;314.367;128.5283;9.0;60.8;PR
07/04/2017 02:00;651.7123;354.1657;314.7233;135.8793;9.0;60.8;PR
07/04/2017 02:30;621.0247;377.149;314.367;128.5283;9.0;58.9;PR
...```

Each row is a measure, except for the first row of the file which is a header indicating the content of each column. The meaning of the columns is as follows:

• **date**: timestamp with the date of the sample taken in Universal Coordinated Time (UTC) format. The format given in the files is dd/mm/yyyy H:M, where dd = day, mm = month, yyyy = year, H = hour, and M = minute, and samples are taken every half an hour.
• **s1_o3**: value in KΩ of the first SGX Sensortech MICS 2614 metal-oxide (O₃) sensor,
• **s2_o3**: value in KΩ of the second SGX Sensortech MICS 2614 metal-oxide (O₃) sensor,
• **s3_o3**: value in KΩ of the third SGX Sensortech MICS 2614 metal-oxide (O₃) sensor,
• **s4_o3**: value in KΩ of the fourth SGX Sensortech MICS 2614 metal-oxide (O₃) sensor,
• **s_temp**: value in °C of the temperature sensor,
• **s1_rh**: value in% of the relative humidity sensor,
• **location**: label indicating the location of the sensor. The label can be of two types: a reference station label (PR, MANLLEU, TONA, VIC, MONTSENY), or a volunteer’s house (VOL).

Table 1 shows the GPS coordinates of each reference station or the name of the villages where the nodes were deployed. The GPS coordinates of the volunteers’ houses where the nodes were deployed are not included for privacy reasons. The column “Reference stations used to calibrate the sensors” shows the reference station used to calibrate the sensors using a multiple linear regression algorithm before they were deployed at the volunteers’ houses. We do not provide the calibrated data as this data can be obtained from the raw data provided here using any type of machine learning algorithm, not necessarily multiple linear regression, as discussed in the next subsection. So, although we provide the location where the sensors were calibrated, all data pertaining to the CSV files, including those labelled “VOL”, are raw data.
Table 1: Location names, location labels and GPS coordinates where the nodes were deployed.\(^1\)

| Location (Ref. Station name or the village name) where the nodes were deployed | "Location" label in the data file | Reference station used to calibrate the sensors | GPS coordinates |
|---|---|---|---|
| Ref. Station of Barcelona (Palau Reial) | PR | PR | 41.3872889, 2.11566667 |
| Ref. Station of Manlleu | MANLLEU | MANLLEU | 42.0033667, 2.287091667 |
| Ref. Station of Vic | VIC | VIC | 41.9349167, 2.239825 |
| Ref. Station of Tona (Zona Esportiva) | TONA | TONA | 41.846738, 2.217445 |
| Ref. Station of Montseny (La Castanya) | MONTSENY | MONTSENY | 41.7793889, 2.357997222 |
| Cantonigros | VOL | MANLLEU | – – |
| St. Vicenç de Torello | VOL | MANLLEU | – – |
| St. Pere de Torello | VOL | MANLLEU | – – |
| Centelles | VOL | TONA | – – |
| Gurb | VOL | MANLLEU | – – |
| Els Hostalets de Balenya | VOL | TONA | – – |
| Sta. Eulalia de Riuprimer | VOL | TONA | – – |
| Sta. Cecilia de Voltrega | VOL | MANLLEU | – – |
| St. Martí Sescorts | VOL | MANLLEU | – – |
| Taradell | VOL | TONA | – – |
| Calldetenes | VOL | TONA | – – |
| Sentmenat | VOL | PR | – – |
| Llinars del Valles | VOL | TONA | – – |
| Tona-ciutat | VOL | TONA | – – |
| Canoves | VOL | TONA | – – |
| Sta. Eugenia de Berga | VOL | TONA | – – |
| Barcelona | VOL | PR | – – |
| Badalona | VOL | PR | – – |
| Vilanova del Valles | VOL | PR | – – |
| St. Cebria de Vallalta | VOL | PR | – – |

Table 2 shows the different locations where each node was deployed during the 2017 campaign. In the first column, "Captor Label" shows the captor name that appears in the CSV file. The second to fifth columns, show the “location” of each node followed by the “starting date”, “end date” and “number of samples”. The label “location” in this table corresponds to the same label in the CSV file. When the label is (PR, MANLLEU, VIC, TONA or MONTSENY) means that the node was placed during the indicated period of time at a reference station. When the label is VOL (village name) means that the node was placed in a volunteer’s house located in the “village name”. The village name does not appear in the CSV data file. Therefore, when the label VOL appears in the CSV file, the village name has to be obtained by consulting Table 2.

2. Experimental Design, Materials, and Methods

During the H2020 CAPTOR project (2016–2018), tropospheric ozone sensor nodes were deployed in three countries: Spain, Italy and Austria [1]. The objective of the project was to raise citizens’ awareness of the pollution they suffer from and to involve them, through a citizen science process, in solving or relating to the problem of this pollutant. During the project, two types of nodes were developed using the Do-It-Yourself (DiY) philosophy: thirty-five Captor nodes and twenty-five Raptor nodes. Captor nodes were used in the Spanish and Italian testbeds, while Raptor nodes were used in the Austrian and Italian testbeds. The nodes in Austria and Italy were spread across the respective countries, while the nodes in Spain were concentrated in a small area of about 25Kmx30Km. This paper focuses on the twenty-five Captor nodes deployed in the Spanish testbed that formed a network together with the reference stations present in the area.
Table 2
Locations and dates in which each Captor node was deployed.

| CaptorLabel | Location            | Starting date | End date    | Number of samples |
|-------------|---------------------|---------------|-------------|-------------------|
| Captor17001 | PR                  | 06/04/2017 12:00 | 08/05/2017 06:00 | 1276              |
|             | MANLEU              | 08/05/2017 12:00 | 26/05/2017 07:00 | 851               |
|             | VOL (Cantanigros)   | 31/05/2017 12:00 | 18/09/2017 10:30 | 5215              |
|             | MANLEU              | 18/09/2017 13:00 | 24/09/2017 06:30 | 275               |
| Captor17002 | PR                  | 06/04/2017 12:00 | 08/05/2017 06:00 | 1292              |
|             | MANLEU              | 08/05/2017 12:00 | 26/05/2017 07:00 | 853               |
|             | VOL (St. Vicenc de | 31/05/2017 12:00 | 18/09/2017 09:30 | 5248              |
|             | Torello)            | 18/09/2017 12:30 | 05/10/2017 09:00 | 809               |
| Captor17003 | PR                  | 06/04/2017 12:00 | 08/05/2017 06:00 | 1291              |
|             | MANLEU              | 08/05/2017 09:00 | 26/05/2017 07:00 | 858               |
|             | VOL (St. Pere de    | 30/05/2017 13:00 | 18/09/2017 09:00 | 5194              |
|             | Torello)            | 18/09/2017 12:30 | 05/10/2017 09:00 | 807               |
| Captor17004 | PR                  | 06/04/2017 12:00 | 08/05/2017 06:00 | 1430              |
|             | TONA                | 26/05/2017 12:00 | 14/06/2017 07:00 | 899               |
|             | VOL (Centelles)     | 14/06/2017 12:00 | 22/09/2017 09:30 | 4798              |
|             | TONA                | 22/09/2017 11:00 | 05/10/2017 10:00 | 622               |
| Captor17005 | PR                  | 06/04/2017 12:00 | 08/05/2017 09:00 | 1343              |
|             | MANLEU              | 08/05/2017 14:00 | 26/05/2017 07:00 | 850               |
|             | VOL (Gurb)          | 31/05/2017 12:00 | 18/09/2017 09:00 | 5233              |
|             | MANLEU              | 18/09/2017 09:30 | 05/10/2017 09:00 | 804               |
| Captor17006 | PR                  | 06/04/2017 14:00 | 08/05/2017 06:00 | 381               |
|             | TONA                | 08/05/2017 15:00 | 26/05/2017 09:00 | 850               |
|             | VOL (Sta. Eulalia de| 02/06/2017 17:00 | 19/09/2017 08:00 | 4580              |
|             | Riuprimer)          | 19/09/2017 12:00 | 05/10/2017 10:00 | 762               |
|             | TONA                |                |              |                   |
| Captor17007 | PR                  | 30/04/2017 01:30 | 08/05/2017 06:00 | 373               |
|             | TONA                | 08/05/2017 15:00 | 26/05/2017 09:00 | 741               |
|             | VOL (El Hostalets  | 02/06/2017 12:00 | 19/09/2017 07:00 | 5095              |
|             | de Balanya)         | 19/09/2017 12:00 | 05/10/2017 10:00 | 764               |
|             | TONA                |                |              |                   |
| Captor17009 | PR                  | 06/04/2017 17:00 | 11/10/2017 08:00 | 7599              |
| Captor17010 | PR                  | 06/04/2017 14:00 | 08/05/2017 06:00 | 1259              |
|             | MANLEU              | 08/05/2017 11:00 | 26/05/2017 07:00 | 838               |
|             | VOL (Sta. Cecilia de| 30/05/2017 10:00 | 18/09/2017 08:00 | 4960              |
|             | Volterra)           | 18/09/2017 12:30 | 05/10/2017 08:30 | 787               |
|             | MANLEU              |                |              |                   |
| Captor17011 | PR                  | 07/04/2017 01:30 | 08/05/2017 06:00 | 1228              |
|             | MANLEU              | 08/05/2017 11:00 | 26/05/2017 07:00 | 836               |
|             | VOL (St. Marti      | 30/05/2017 11:00 | 18/09/2017 11:00 | 4621              |
|             | Sescorts)           | 19/09/2017 02:30 | 05/10/2017 08:30 | 374               |
|             | MANLEU              |                |              |                   |
| Captor17012 | PR                  | 06/04/2017 15:00 | 08/05/2017 06:00 | 1100              |
|             | TONA                | 08/05/2017 14:00 | 26/05/2017 09:00 | 830               |
|             | VOL (Taradell)      | 02/06/2017 09:00 | 17/09/2017 00:30 | 4982              |
|             | TONA                | 19/09/2017 12:30 | 05/10/2017 10:00 | 741               |
| Captor17013 | PR                  | 06/04/2017 14:00 | 08/05/2017 06:00 | 1000              |
|             | TONA                | 08/05/2017 13:00 | 26/05/2017 03:30 | 837               |
|             | VOL (Calldetenes)   | 03/06/2017 12:00 | 19/09/2017 08:30 | 4337              |
|             | TONA                | 19/09/2017 12:30 | 05/10/2017 10:00 | 717               |
| Captor17015 | PR                  | 28/04/2017 14:00 | 08/05/2017 06:00 | 1146              |
|             | VOL (Sentmenat)     | 09/06/2017 11:00 | 29/09/2017 10:30 | 5256              |
|             | PR                  | 29/09/2017 15:30 | 11/10/2017 08:00 | 547               |
| Captor17016 | PR                  | 28/04/2017 14:00 | 08/05/2017 06:00 | 497               |
|             | MONTSENY            | 09/05/2017 10:00 | 22/05/2017 10:00 | 604               |
|             | VIC                 | 26/05/2017 10:00 | 05/10/2017 09:00 | 6035              |

(continued on next page)
Table 2 (continued)

| CaptorLabel | Location         | Starting date | End data  | Number of samples |
|-------------|------------------|---------------|-----------|-------------------|
| Captor17017 | PR               | 06/04/2017 14:00 | 08/05/2017 06:00 | 1311            |
|             | TONA             | 08/05/2017 13:00 | 05/10/2017 10:00 | 6966            |
| Captor17018 | PR               | 28/04/2017 14:30 | 30/06/2017 07:30 | 1134            |
|             | VOL (Badalona)   | 30/06/2017 16:00 | 25/09/2017 04:30 | 2921            |
| Captor17019 | PR               | 05/10/2017 13:00 | 09/19/2017 09:00 | 180             |
|             | VOL (Barcelona)  | 29/06/2017 14:30 | 13/07/2017 12:00 | 485             |
|             | PR               | 19/07/2017 11:00 | 27/07/2017 14:00 | 338             |
|             | VOL (Barcelona)  | 27/07/2017 16:00 | 05/10/2017 14:30 | 3116            |
| Captor17020 | PR               | 07/04/2017 13:30 | 28/06/2017 10:00 | 1995            |
|             | PR               | 29/04/2017 02:00 | 09/05/2017 06:00 | 339             |
|             | MONTSENY         | 09/05/2017 11:00 | 20/05/2017 21:00 | 538             |
|             | PR               | 23/06/2017 10:00 | 30/06/2017 07:00 | 1745            |
|             | VOL (St. Cebria de Vallalta) | 30/06/2017 12:00 | 09/10/2017 18:00 | 4748            |
| Captor17021 | PR               | 06/04/2017 14:00 | 26/05/2017 06:00 | 1829            |
|             | TONA             | 30/05/2017 08:00 | 21/06/2017 08:00 | 837             |
|             | MONTSENY         | 21/06/2017 14:00 | 06/10/2017 08:30 | 4927            |
| Captor17022 | PR               | 06/04/2017 13:30 | 08/05/2017 06:00 | 1267            |
|             | TONA             | 08/05/2017 13:00 | 30/05/2017 07:00 | 1006            |
|             | VOL (Llinars del Valles) | 09/06/2017 09:00 | 22/09/2017 06:30 | 4921            |
|             | TONA             | 22/09/2017 11:30 | 05/10/2017 10:00 | 609             |
| Captor17023 | PR               | 06/04/2017 14:00 | 26/05/2017 06:00 | 1628            |
|             | TONA             | 26/05/2017 10:00 | 14/06/2017 00:00 | 869             |
|             | VOL (Tona-ciutat) | 14/06/2017 10:00 | 20/09/2017 07:30 | 4124            |
| Captor17025 | PR               | 06/04/2017 14:00 | 26/05/2017 06:00 | 866             |
|             | TONA             | 22/09/2017 11:00 | 05/10/2017 10:00 | 623             |
|             | VOL (Canoves)    | 26/05/2017 11:00 | 14/06/2017 07:00 | 876             |
| Captor17026 | PR               | 06/04/2017 14:00 | 26/05/2017 06:00 | 4255            |
|             | MONTSENY         | 21/06/2017 13:00 | 22/09/2017 07:00 | 609             |
|             | PR               | 22/09/2017 11:00 | 05/10/2017 10:00 | 609             |
|             | MONTSENY         | 09/05/2017 11:00 | 21/05/2017 12:00 | 717             |
|             | PR               | 23/05/2017 11:00 | 04/07/2017 12:00 | 426             |
|             | VOL (Vilanova del Valles) | 04/10/2017 10:00 | 24/09/2017 07:00 | 1733            |
| Captor17027 | PR               | 07/04/2017 09:30 | 09/05/2017 06:00 | 3841            |
|             | TONA             | 09/05/2017 11:00 | 21/05/2017 12:00 | 270             |
|             | VOL (Santa Eugenia de Berga) | 02/06/2017 11:00 | 19/09/2017 09:30 | 2103            |
|             | TONA             | 19/09/2017 12:00 | 05/10/2017 10:00 | 747             |

Captor nodes have been developed at the Polytechnic University of Catalonia (UPC, Barcelona, Spain). These nodes use an Arduino microprocessor that is connected to a sensor board that can incorporate up to five metal-oxide sensors, plus a temperature and relative humidity sensor. The project decided to use four SGX Sensortech MICS 2614 metal-oxide (O₃) sensors, plus a temperature and relative humidity sensor from Grove.

The reason for using four MOX sensors was to increase the robustness of the equipment in case of sensor failure due to their low cost and the variability in the quality of their measurements. Through a calibration process, the sensor that gave the best quality was chosen as the representative of that node. The node was connected via a cellular connection to a repository where the ozone values (in μg/m³) representative of the node were stored. If a sensor failed, the second best sensor was chosen to represent the node. The raw data were also stored in the repository for research purposes.
Sensortech metal-oxide sensors measure O₃ using a voltage divider circuit that has a load resistor and a variable resistor. Each time the O₃ concentration changes, the variable resistor changes. The resistance value represented by the O₃ sample is obtained by measuring the voltage \( V_L \) across the load resistor \( R_L \) after quantifying the signal with an A/D converter and converting this voltage to a raw resistance measurement (in Ω):

\[
R_{o3} = R_L \left( 1 - \frac{V_{cc}}{V_L} \right).
\]

Where \( V_{cc} \) is the reference voltage (5 Volts). The temperature sensor gives the values in °C and the relative humidity sensor in percentage (%). The ozone sensor manufacturer does not calibrate the sensors, and the calibration has to be done in the field [2]. This means that the value obtained by the voltage divider must be converted to ozone concentrations by a calibration process, since the values obtained by the ozone sensor are in Ω and not in µg/m³. The calibration process consisted of placing the sensors on an AQM reference station for two or three weeks and performing a machine learning algorithm against the reference values. Examples of machine learning algorithms are multiple linear regression (MLR), random forest (RF) or support vector regression (SVR), among others. For example, using a multiple linear regression and introducing temperature and relative humidity as corrective features,

\[
y_{o3} = \beta_0 + \beta_1 x_{o3} + \beta_2 x_{Temp} + \beta_3 x_{RH}.
\]

Where \( y_{o3} \) is the ozone concentration in µg/m³ of the AQM reference station, \( (x_{o3}, x_{Temp}, x_{RH}) \) are the raw sensor measurements in (KΩ, °C, %), and \( \beta^T = (\beta_0, \beta_1, \beta_2, \beta_3) \) are the calibration or regression coefficients that the multiple regression model must obtain. Once the regression coefficients are obtained, the sensor node is placed elsewhere, and then new O₃ concentration values can be derived by measuring the raw data \( (x_{o3}, x_{Temp}, x_{RH}) \) using the calibration \( \beta \) coefficients. In the case of using other machine learning algorithms, the idea is the same: use the data gathered during the calibration phase to obtain the hyper-parameters of the machine learning algorithm, and then use these hyper-parameters to estimate new values when the sensors are deployed elsewhere.

As can be seen, the sensors were calibrated by placing the nodes at a reference station. Table 1 indicates which reference station was used to calibrate the sensors depending on where they were deployed. To calibrate the sensors, therefore, it is necessary to have the reference values of the reference stations during the period when the nodes were placed next to these reference stations. These values are open and can be obtained at the URL:

http://mediambient.gencat.cat/es/05_ambits_dactuacio/atmosfera/qualitat_de_laire/vols-saber-que-respires/descarrega-de-dades/index.html managed by the local government of Catalonia, Spain. This Website is in Catalan and Spanish. Therefore, we give a brief indication on how to download the reference values. There is a box indicated by the word “Filtros” (Filters) where you can indicate the reference station name, the type of pollutant (e.g. O₃) and the start and end period of the data to be downloaded. For the data shown in this document, the names of the reference stations are given in Table 1, and the time period in which the nodes were at these reference stations is given in Table 2. The reference data can then be accessed by clicking on the CSV button in the bottom right corner of the web page.

Several research papers have made use of the dataset presented in this document. Ref. [1] describes the testbeds developed during the H2020 CAPTOR project. Ref. [2] is a survey that describes mechanisms to calibrate low-cost sensors in an uncontrolled environment. Ref. [3] compares the calibration performance of the Captor sensors when calibrated using various machine learning algorithms such as multiple linear regression, random forest and support vector regression. Ref. [4] shows the impact of environmental conditions on Captor sensors calibration. Refs. [5] and [6] study the fusion of data taken from the four Captor sensors placed at the same node as a way to reduce the estimation error with respect to the use of a single sensor. Ref. [7] propose a graph sensing framework to re-calibrate sensors, impute missing data, and reconstruction techniques in an heterogenous monitoring air pollution network with reference stations and low-cost sensor nodes. The research papers submitted used the Captor testbed
to evaluate the performance of the techniques analysed using a real IoT sensor framework. We believe that the dataset is useful for other studies that may need real sensor data to test the performance of techniques and models investigated in air pollution monitoring networks, such as the use of other machine learning algorithms, multi-hop or distributed calibration in a sensor network, graphical signal processing analysis, spatial studies using interpolation algorithms, creation of virtual sensors, etc.

CRediT Author Statement

Jose M. Barcelo-Ordinas: Conceptualization, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing; Jorge Garcia-Vidal: Conceptualization, Supervision, Project administration; Pau Ferrer-Cid: Data Curation, Software, Formal analysis; Mar Viana: Conceptualization, Supervision, Project administration; Anna Ripoll: Formal analysis, Validation, Investigation.

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.

Acknowledgments

This work was funded by H2020 project CAPTOR, and supported by National Spanish project PID2019–107910RB-I00, and regional project 2017SGR-990, and with the support of Secretaria d’Universitats i Recerca de la Generalitat de Catalunya i del Fons Social Europeu. The authors gratefully acknowledge the collaboration of the staff at the Department of the Environment of the Generalitat de Catalunya who kindly provided support for the deployment of the sensing nodes at reference stations for calibrating the sensors and provided access to the reference data.

References

[1] A. Ripoll, M. Viana, M. Padrosa, X. Querol, K-M.H A.Minutolo, J.M. Barcelo-Ordinas, J. Garcia-Vidal, Testing the performance of sensors for ozone pollution monitoring in a citizen science approach, Sci. Total Environ. 651 (2019) 1166–1179 Elsevier, doi:10.1016/j.scitotenv.2018.09.257.
[2] J.M. Barcelo-Ordinas, M. Doudou, J. Garcia-Vidal, N. Badache, Self-calibration methods for uncontrolled environments in sensor networks: a reference survey, Ad. Hoc. Netw. 88 (2019) 142–159 Elsevier, doi:10.1016/j.adhoc.2019.01.008.
[3] P. Ferrer-Cid, J.M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, M. Viana, A comparative study of calibration methods for low-cost ozone sensors in IoT platforms, IEEE Internet Things J. 6 (6) (2019) 9563–9571 NDecember 2019, doi:10.1109/JIOT.2019.2929594.
[4] J.M. Barcelo-Ordinas, P. Ferrer-Cid, J. Garcia-Vidal, A. Ripoll, M. Viana, Distributed multi-scale calibration of low-cost ozone sensors in wireless sensor networks, Sensors 19 (11) (2019) 2503, doi:10.3390/s19112503.
[5] JM. Barcelo-Ordinas, J. Garcia-Vidal, M. Doudou, S. Rodrigo-Muñoz, A. Cerezo-Llávero, Calibrating low-cost air quality sensors using multiple arrays of sensors, in: IEEE Wireless Communications and Networking Conference - (WCNC’18), Barcelona, Spain, 2018 April 15–18, 2018.
[6] P. Ferrer-Cid, J.M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, M. Viana, Multi-sensor data fusion calibration in IoT air pollution platforms (2020), IEEE Internet Things J. 6 (4) (2020) 3124–3132 IEEE IoT-JNApril, doi:10.1109/JIOT.2020.2965283.
[7] P. Ferrer-Cid, J. M. Barcelo-Ordinas, J. Garcia-Vidal, Graph learning techniques using structured data for IoT air pollution monitoring platforms, IEEE Internet Things J. (2021), doi:10.1109/JIOT.2021.3067717.