Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Raw data collected from NO$_2$, O$_3$ and NO air pollution electrochemical low-cost sensors

Pau Ferrer-Cid*, Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal

Universitat Politecnica de Catalunya (UPC), C/Jordi Girona 31, Barcelona 08034, Spain

A B S T R A C T

Recently, the monitoring of air pollution by means of low-cost sensors has become a growing research field due to the study of techniques based on machine learning to improve the sensors' data quality. For this purpose, sensors undergo a calibration process, where these are placed in-situ nearby a regulatory reference station. The data set explained in this paper contains data from two self-built low-cost air pollution nodes deployed for four months, from January 16, 2021 to May 15, 2021, at an official air quality reference station in Barcelona, Spain. The goal of the deployment was to have five electrochemical sensors at a high sampling rate of 0.5 Hz; two NO$_2$ sensors, two O$_3$ sensors, and one NO sensor. It should be noted that the reference stations publish air pollution data every hour, thus at a rate of 2.7×10^{-4} Hz. In addition, the nodes have also captured temperature and relative humidity data, which are typically used as correctors in the calibration of low-cost sensors. The availability of the sensors' time series at this high resolution is important in order to be able to carry out analysis from the signal processing perspective, allowing the study of sensor sampling strategies, sensor signal filtering, and the calibration of low-cost sensors among others.

* Corresponding author.
E-mail addresses: pau.ferrer.cid@upc.edu (P. Ferrer-Cid), jose.maria.barcelo@upc.edu (J.M. Barcelo-Ordinas), jorge.garcia@upc.edu (J. Garcia-Vidal).

Specifications Table

Subject	Environmental science (air pollution)
Specific subject area	High-frequency measurements of nitrogen dioxide (NO_2), tropospheric ozone (O_3) and nitrogen monoxide (NO) using electrochemical low-cost sensors.
Type of data	Table data. More precisely, data is arranged in CSV files.
How data were acquired	The Captor nodes were designed and built by UPC (Universitat Politècnica de Catalunya, Spain), during the national project "IoT monitoring of air quality (IMAQ)". The nodes include the following Alphasense electrochemical sensors: Captor node labeled as 20001 includes 1x NO2-B43F (NO_2 sensor), 1x OX-B431 (O_3 sensor) and 1x NO-B4 (NO sensor) with their respective electronic boards (called ISB, individual support boards) provided by Alphasense, and Captor node labeled as 20002 includes 1x NO2-B43F (NO_2 sensor) and 1x OX-B431 (O_3 sensor) with their respective ISB electronic boards provided by Alphasense. In addition, both nodes include a DHT-22 temperature and relative humidity sensor inside the node box, which means that the temperature and relative humidity are not environmental measurements, but that these sensors include the heating of the electronics inside the measurement box.
Data format	Raw: NO_2, O_3 and NO in quantum steps of an 10-bit A/D converter (values between 0 and 1024, with 0 equivalent to 0 Volts and 1024 equivalent to 1.1 Volts), temperature in °C, relative humidity in %.
Parameters for data collection	The Alphasense electrochemical sensors have a lifetime of approximately two years [1–3], according to the manufacturer. The sensors were installed in the nodes in early October 2020, but the measurements that are made public are from January 16, 2021 to May 15, 2021. The data from October to December are not published as they correspond to laboratory tests (without pollution) and first connection tests of the node at the reference station to verify that the node (electronics, communications, software and sensors) worked properly.
Description of data collection	The nodes with the sensors installed underwent a set of tests at the UPC laboratories and at the reference station for two and a half months to validate the electronics, the software, the communication with the UPC database repositories, the sensor sampling parameters, etc. After this testing period, the nodes were located at the Palau Reial reference station for four months. The Palau Reial reference station is a governmental reference station managed by the IDAEA research group of the CSIC (Spanish National Research Council). Data from the low-cost sensors can be compared with the Palau Reial reference station, which provides accurate air pollution data.
	The electronics of the Captor node request data from the sensors (NO_2, O_3 and NO, temperature and relative humidity) every two seconds. Measurements from the NO_2, O_3 and NO electrochemical sensors are passed through a 10-bit A/D converter. Thus, these readings are values between 0-1024. The temperature and relative humidity sensors give readings in °C and % respectively.
Data source location	Institution: Universitat Politécnica de Catalunya (UPC)
	City/Town/Region: Barcelona
	Country: Spain
	GPS coordinates for collected samples/data: Palau Reial Reference Station, Latitude 41.3872889, Longitude 2.115666667
Data accessibility	Repository name: Zenodo Files names: 20001_O3_NO2_raw.csv, 20002_O3_NO2_raw.csv, 20001_NO_raw.csv, 20001_T_RH_raw.csv, 20002_T_RH_raw.csv URL: https://doi.org/10.5281/zenodo.5770589 [4]
Related research article	P. Ferrer-Cid, J. Garcia-Calvete, A. Main-Nadal, Z. Ye, J. M. Barcelo-Ordinas, J. Garcia-Vidal, Sampling trade-offs in duty-cycled systems for air quality low-cost sensors, Sensors 22 (2022) https://doi.org/10.3390/s22103964 [5].

Value of the Data

- High-frequency[1] measurements of low-cost sensors located in-situ at a reference station, deployed by authorities, are useful to carry out sensor calibration studies. In addition, having

---

[1] In this field, where reference stations provide data at a rate of 2.7×10^{-4} Hz, a sensor sampling rate of 0.5 Hz is considered high frequency.

the raw sensor data available at such a high frequency allows investigating different techniques related to sampling, filtering and further analysis of the raw sensors signals.

- The data from these electrochemical sensors comprise measurements of three different pollutants (O_3, NO_2 and NO), with the actual concentrations measured by the reference station being available. Hence, the main beneficiaries are researchers specialized in the use of low-cost sensors for the measurement of air pollution, those who focus on the calibration of these sensors, and their possible use in regulated air pollution monitoring networks.
- Most air pollution data published in repositories are post-processed data with a sampling resolution of half an hour or one hour. Using the data presented in this paper, researchers can perform studies that require a high sensor sampling frequency. These data, allow studying the impact of sampling frequency on calibration, the impact of the type of aggregation used for calibration, and even the use of signal filtering techniques to improve sensor calibration. Therefore, as they are not pre-processed or already aggregated data, they can be of great interest for future studies that require a high sampling frequency.
- Reference stations give very accurate data, but they are expensive and are used with low spatial resolution. The use of low-cost air pollution sensors, although they cannot be used to generate alarms, can be used to generate social awareness, and as an indication that there is more pollution than desired in certain locations. The presented data set will help researchers and engineers, who want to deploy their low-cost sensor nodes, to test various calibration techniques, with varying sampling frequencies, thus allowing them to better learn how to adjust the parameters of the nodes they deploy.

1. Data Description

All raw data described in this section can be downloaded at Zenodo [4]. The data set consists of five CSV (comma separated values) files; three CSV files for Captor node 20001 and two CSV files for Captor node 20002. The format of the files with the data collected by the sensors is described below.

The format of a *"20001_O3_NO2_raw.csv"* and *"20002_O3_NO2_raw.csv"* files is:

date;WE_O3;AE_O3;WE_NO2;AE_NO2
2021-01-16 00:00:00;924;914;872;869
2021-01-16 00:00:02;923;910;876;870
2021-01-16 00:00:04;927;911;871;867

The format of a *"20001_NO_raw.csv"* file is:

date;WE_NO;AE_NO
2021-01-16 00:00:00;560;580
2021-01-16 00:00:02;560;582
2021-01-16 00:00:04;563;584

The format of a *"20001_T_RH_raw.csv"* and *"20002_T_RH_raw.csv"* files is:

date;T;RH
2021-01-16 00:00:00;16.4;49.2
2021-01-16 00:00:02;17.2;48.9
2021-01-16 00:00:04:04;16.9;48.8

Every CSV file row corresponds to sensor measures taken at the timestamp indicated. However, the first row indicates the columns' names. The description of each one of the variables is the following:

- **date:** timestamp of the taken measurement with format "YYYY-mm-dd HH:MM:SS", where YYYY = year, mm = month, dd = day, HH = hour, MM = minute, and SS = second. Times-

Table 1

Files description.

File	Period	Location	# samples
20001_O3_NO2_raw.csv	2021-01-16 00:00:00 to 2021-05-15 00:00:00	Palau Reial	5,786,415
20001_NO_raw.csv	2021-01-16 00:00:00 to 2021-05-15 00:00:00	Palau Reial	5,810,598
20001_T_RH_raw.csv	2021-01-16 00:00:00 to 2021-05-15 00:00:00	Palau Reial	5,396,195
20002_O3_NO2_raw.csv	2021-01-16 00:00:00 to 2021-05-15 00:00:00	Palau Reial	5,627,601
20002_T_RH_raw.csv	2021-01-16 00:00:00 to 2021-05-15 00:00:00	Palau Reial	5,336,654

tamps are Barcelona (Spain) local time (UTC+1 until 2021-03-28 02:00:00, UTC+2 from 2021-03-28 02:00:00 to 2021-10-31 03:00:00).

- **WE_O3:** value in quantum steps of a 10-bit A/D converter for the working electrode (WE) of the OX-B431 O_3 sensor.
- **AE_O3:** value in quantum steps of a 10-bit A/D converter for the auxiliary electrode (AE) of the OX-B431 O_3 sensor.
- **WE_NO2:** value in quantum steps of a 10-bit A/D converter for the working electrode (WE) of the NO2-B43F NO_2 sensor.
- **AE_NO2:** value in quantum steps of a 10-bit A/D converter for the auxiliary electrode (AE) of the NO2-B43F NO_2 sensor.
- **WE_NO:** value in quantum steps of a 10-bit A/D converter for the working electrode (WE) of the NO-B4 NO sensor.
- **AE_NO:** value in quantum steps of a 10-bit A/D converter for the auxiliary electrode (AE) of the NO-B4 NO sensor.
- **T**: value in Celsius degrees ($°C$) of the temperature sensor.
- **RH**: value in percentage (%) of the relative humidity sensor.

Both nodes (Captor 20001 and Captor 20002) have been deployed at the Palau Reial reference station, in the city of Barcelona, Spain, with GPS coordinates; latitude 41°23′14.2″N and longitude 2°06′56.4″E. The reference air pollution concentrations collected by this reference station are available at Catalonia Open Data web page[2]. The data set corresponds to samples from January 16 at 00:00 to May 15 at 00:00. Table 1 shows a brief description of the different files that make up the data set, with the initial and final date of the measurements, and the total number of samples per file. From the deployment period shown in Table 1 there is a data gap of 3 days, from 2021-03-13 00:00:00 to 2021-03-15 23:59:59.

2. Experimental Design, Materials and Methods

Captor nodes have been developed at the Polytechnic University of Catalonia (UPC), Barcelona, Spain, under the national project "IoT monitoring of air quality (IMAQ)". The objective of this project is a deployment of air pollution low-cost sensor nodes in the metropolitan area of Barcelona, Spain. During the development of the node, called Captor, and as a preliminary study before the deployment of the final nodes, we wondered what were the optimal sampling frequencies of the low-cost sensors, so that we could at the same time perform an optimal calibration of the sensors, and be able to develop a node that consumes little energy. This is very important since the node needs to run on batteries because in the locations where we want to deploy the nodes we do not have a point of electricity to plug the node.

An early version of the Captor nodes deployed ozone sensors with MOX technology, Barcelo-Ordinas et al. [6], Ferrer-Cid et al. [7]. These nodes were deployed in testbeds in Spain and Italy, and allowed the development of sensor calibration studies, in addition to publishing the data openly for other researchers to perform their studies [8]. As a continuation of the nodes that

---

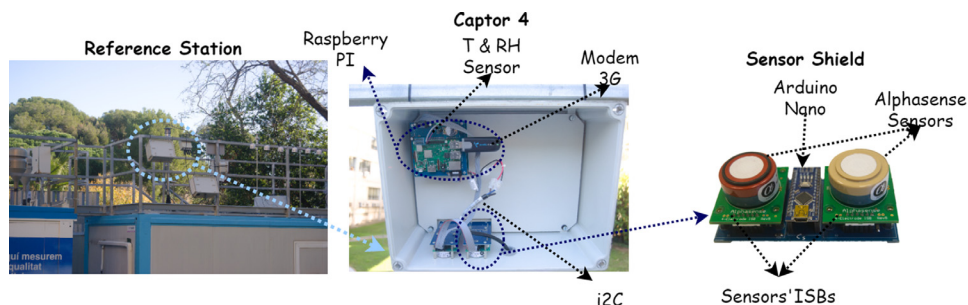[2] Catalonia air pollution open data web page

Fig. 1. The Captor node location on the left, the location of the components on the center, and the sensing shield on the right. Modified from Ferrer-Cid et al. [5].

measured with MOX sensor technology, we have developed two prototype nodes in which we could connect the electrochemical (EQ) sensors in various conditions that would allow us to decide the working parameters of a node in the final deployment. In these prototypes there are two parameters that we find interesting to investigate and that are the focus of this data set that we present: i) the sampling frequency at which the sensors take air pollution samples, which has a very high impact on the duty cycle of the node (and therefore on the energy consumption of the node) and on the calibration quality of the sensors, and ii) the amplification that we had to perform on the data taken by the electrochemical sensors, which are very sensitive to interference, and therefore have an impact on the quality of the data taken by the electrochemical sensors.

The new Captor node, Fig. 1, with O_3, NO_2 and NO EQ sensors has four main blocks; the processing unit, the 3G modem, the power supply and the sensing shields. It has a Raspberry PI 3 B+ as processing unit, which is responsible for collecting the samples from the sensing shields, and sending these samples to a centralized database via the 3G modem. The master node (Raspberry) is attached to an I2C bus where all the sensing shields are attached together, so that all measurements can be requested an collected using this bus. In addition, the node has two Arduino Nanos, one per sensing shield, that are responsible for collecting samples directly from the sensors and sending them to the Raspberry Pi via the I2C bus. In fact, there is one Arduino Nano per sensing shield, so one Arduino every two sensors. Finally, the nodes have a NO2-B43F sensor (NO_2 sensor [2]), an OX-B431 sensor (O_3 sensor [3]), a NO-B4 (NO sensor [1]), with their respective ISBs [9] electronic boards provided by Alphasense, and a DHT-22 temperature and relative humidity sensor directly attached to the Raspberry Pi.

The ISBs' output values range from 200 to 250 mV (few ppb to 200 ppb, part per billion) for NO_2 [2], from 200 to 280 mV (few ppb to 200 ppb) for O_3 [3], and from 400 to 480 mV (few ppb to 200 ppb) for NO [1]. To minimize the A/D conversion error, we amplify the signal before it is converted by the A/D converter. We tested two amplifications at the nodes, one amplification by 2 and another by 4. Table 2 shows the amplification of each of the sensors that make up the data set. For instance, the O_3 and NO_2 sensor data captured by the Captor 20001 node were amplified by a factor $\times 4$, while the O_3 and NO_2 sensors in Captor 20002 were amplified by a factor $\times 2$.

Table 2
Sensor amplifications.

$\times 2$	$\times 4$
20002_O3	20001_O3
20002_NO2	20001_NO2
20001_NO	

The Alphasense B4 sensor family has four electrodes; the working electrode (WE), the reference electrode, the counter electrode, and the auxiliary electrode (AE). When the gas specie enters the sensor membrane an oxidation or reduction reaction occurs in the WE. The electric current generated is proportional to the gas concentration [10]. The value of the WE and AE electrodes must be subtracted since the AE is invariant to the conditions to which the WE is subjected, therefore it serves as a corrector for the WE. More precisely, the raw sensor measures s_x can be computed as [11]:

$$s_{NO} = s_{NO_{WE}} - s_{NO_{AE}}$$
$$s_{NO2} = s_{NO2_{WE}} - s_{NO2_{AE}}$$
$$s_{O3} = (s_{O3_{WE}} - s_{O3_{AE}}) - (s_{NO2_{WE}} - s_{NO2_{AE}})$$

Where the NO_2 measures need to be subtracted from the O_3 sensor measures according to the Alphasense manufacturer, as the latter also measures the NO_2 concentrations. In any case, we provide the raw data taken by the WE and by the AE separately for each sensor. In this way, users can choose to use the raw WE and AE data separately, or to use the difference of these values in their investigations.

Now, since the two nodes were placed at a reference station for four months, different in-situ sensor calibration techniques can be studied. Hence, it can be checked how good the sensors are for the estimation of air pollutants. In particular, once the data from the sensors have been synchronized with those of the reference station, we obtain a set of tuples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i \in \mathbb{R}^P$ are the values collected by the sensors, P is the size of the sensor array, and $y_i \in \mathbb{R}$ are the reference values obtained by the reference station. Then, the in-situ calibration consists of inferring the function $f : \mathbb{R}^P \rightarrow \mathbb{R}$ using machine learning techniques:

$$y_i = f(\mathbf{x}_i) + \epsilon_i \qquad ; i = 1, \ldots, N \qquad (1)$$

Where the error term ϵ_i is independent and identically distributed. In this way, the function $f(\cdot)$ can be estimated using different machine learning algorithms. Usually, governmental reference instruments directly provide the concentrations of a certain gas pollutant. In this way, the raw values measured by the sensors can be compared with the reference values and both can be used to train a calibration model to translate from low-cost sensor measurements to concentrations for a specific gas specie and to correct the sensor measurements to improve their accuracy. Linear models such as the multiple linear regression (MLR) have been widely used in the literature [11,12]. Lately, nonlinear models have also been applied to the low-cost sensor calibration problem [11,12]. In this way, air pollution measurements can be obtained from the raw values of the sensors.

This data set is also interesting because of its high sampling rate, 0.5 Hz, that not only allows to investigate the final sensor calibration, but also to study all the pre-processing steps to obtain representative measurements synchronized with the reference measurements. Reference stations deployed by regulatory agencies typically provide hourly samples, which are the result of the aggregation of continuous samples taken during every hour. Thus, the data periodicity of low-cost sensors is usually smaller than the reference station periodicity. In this line, and as an example of use of this data set, Ferrer-Cid et al. [5] have studied the trade-offs present between the sensor sampling frequency, the quality of the resulting sensor calibration, and the energy consumption of a sensing subsystem. Basically, it is shown how with a larger sampling period a good data quality can be obtained in exchange for significant energy savings. This is of special importance, since in most of the literature the efforts have been focused on obtaining the highest data quality (calibration quality) without taking into account possible restrictions that an internet of things (IoT) node may have, specially in the case of being powered by batteries. Ultimately, this is useful for researchers who intend to deploy battery-powered nodes with maximum energy savings while obtaining good sensor data quality.

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 paper.

CRediT Author Statement

Pau Ferrer-Cid: Conceptualization, Data curation, Writing – original draft, Writing – review & editing; **Jose M. Barcelo-Ordinas:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing; **Jorge Garcia-Vidal:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Acknowledgments

References

[1] Alphasense ox-b431 sensor datasheet (2019a). [online] https://www.alphasense.com/products/nitric-oxide-safety/. Accessed November 20, 2021.
[2] Alphasense no2-b43f sensor datasheet (2019b). [Online] https://www.alphasense.com/products/nitrogen-dioxide/. Accessed November 20, 2021.
[3] Alphasense ox-b431 sensor datasheet (2019c). [Online] https://www.alphasense.com/products/ozone/. Accessed November 20, 2021.
[4] Raw captor data zenodo website (2021). [Online] doi:10.5281/zenodo.5770589.
[5] P. Ferrer-Cid, J. Garcia-Calvete, A. Main-Nadal, Z. Ye, J.M. Barcelo-Ordinas, J. Garcia-Vidal, Sampling trade-offs in duty-cycled systems for air quality low-cost sensors, Sensors 22 (2022) 3964.
[6] 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 (2019).
[7] P. Ferrer-Cid, J.M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, M. Viana, Multi-sensor data fusion calibration in IoT air pollution platforms, IEEE Internet Things J. 7 (2020) 3124–3132.
[8] J.M. Barcelo-Ordinas, P. Ferrer-Cid, J. Garcia-Vidal, M. Viana, A. Ripoll, H2020 project captor dataset: raw data collected by low-cost MOX ozone sensors in a real air pollution monitoring network, Data Brief 36 (2021) 107127.
[9] Support circuits (ppb): Isb individual sensor board datasheet (2019). [Online] https://www.alphasense.com/products/support-circuits-air/. Accessed November 20, 2021.
[10] Zauli-Sajani, S., Marchesi, S., Pironi, C., Barbieri, C., Poluzzi, V., & Colacci, A. (2021). Assessment of air quality sensor system performance after relocation. Atmos. Pollut. Res. 12(2), 282–291.
[11] 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 (2019) 9563–9571.
[12] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part b: NO, CO and CO_2, Sens. Actuators B 238 (2017) 706–715.