Data in Brief 54 (2024) 110411

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Particulate matter 1µm (PM_1) dataset collected by low-cost sensors in residential and industrial areas at the neighborhood level

Luis A. Garcia-Garza[a], Edgar Tello-Leal[b,*], Bárbara A. Macías-Hernández[b], Gerardo Romero[a], Jaciel David Hernandez-Resendiz[a]

[a] *Multidisciplinary Academic Unit Reynosa-Rodhe, Autonomous University of Tamaulipas, Reynosa 88779, Mexico*
[b] *Faculty of Engineering and Science, Autonomous University of Tamaulipas, Victoria 87000, Mexico*

ARTICLE INFO

Article history:
Received 13 March 2024
Revised 4 April 2024
Accepted 8 April 2024
Available online 15 April 2024

Dataset link: Particulate matter with aerodynamic diameters ≤1µm (PM1) dataset (Original data)

Keywords:
Air pollution
Air quality
LCS
IoT
Data science

ABSTRACT

The incursion of low-cost sensors (LCS) for monitoring particulate matter in different fractions of particles (PM_{10}, $PM_{2.5}$, and PM_1) allows the characterization of the concentration levels of specific sources or events, including the analysis of ultrafine fractions (PM_1). Several studies have documented adverse effects on human health due to exposure to PM_1, such as morbidity and mortality from respiratory, cardiovascular, and, in some cases, carcinogenic diseases. Hence, studying the concentration levels and the sources that cause PM_1 is imperative. LCS is an alternative to understanding contaminant concentration levels by considering spatial and temporal community dynamics by monitoring critical zones. Furthermore, collecting and managing large amounts of data through automatic processing and analysis generates information to support decision-making to reduce exposure and risks to people's health. The dataset presents the concentration level of PM_1 (µg/m^3) calculated from the particles of size 0.03 µm, 0.05 µm, and 1.0 µm recorded and counted by the sensor in a sample per minute for 24 h for seven continuous days. The values of the meteorological factors of relative humidity, temperature, and heat index complement these attributes.

* Corresponding author.
E-mail address: etello@docentes.uat.edu.mx (E. Tello-Leal).

https://doi.org/10.1016/j.dib.2024.110411

The dataset comprises records collected (in the same period) at four particulate matter monitoring stations, which compose an LCS network supported by Internet of Things (IoT) technologies. The data collection points were located in different areas of Reynosa, Mexico, considering strategic places for monitoring environmental pollution, such as industrial parks, residential areas, avenues with high vehicular traffic and transportation of heavy cargo, and an airport.

Specifications Table

Subject	Environmental science-pollution
Specific subject area	Air pollution, low-cost sensor, air quality monitoring network, data analytics
Type of data	Table, Figure Raw, Analysed
Data collection	Collecting data on particulate matter and meteorological factors was carried out using low-cost monitoring equipment installed in each air quality monitoring station (4) at a height of 8 m from ground level. The equipment comprises a calibrated PMS7003 low-cost sensor (LCS) for particulate matter. This sensor allows for measuring the mass concentration of particulate matter of the PM_1, $PM_{2.5}$, and PM_{10} fractions. Relative humidity and air temperature were measured using a high-precision, low-power-consumption sensor. The values were collected every minute and transmitted automatically and in real-time to a database hosted in a repository available in private cloud storage over the Internet.
Data source location	• Institution: Autonomus University of Tamaulipas • City/Town/Region: Reynosa, Tamaulipas • Country: México • Latitude and longitude: 26.02937321240764, −98.27578996495555
Data accessibility	Repository name: Particulate matter with aerodynamic diameters ≤1 µm (PM_1) dataset DOI: 10.17632/djnkmx3tzn.2 Direct URL to data: https://data.mendeley.com/datasets/djnkmx3tzn/2

1. Value of the Data

- The air pollutant PM_1 dataset can be considered the first available online from the northeastern area of Mexico.
- The data is cataloged by the areas of Reynosa, Mexico, characterized by high vehicular traffic, heavy transportation traffic, industry, and residential. This information is valuable for understanding and evaluating local air quality.
- The value data were collected and prepared at different time granularities (minute/hour), which makes it possible to carry out different analyses of air pollution through time series models and predictive models.
- This dataset can be a starting point for studies to determine how the air pollutant PM_1 affects human health.
- The dataset contains attributes on the size of particles considered within the PM_1 fraction, which can be linked to the emission sources that generate the air pollutant PM_1.

2. Background

In recent years, the use of low-cost sensors to measure the concentration of air pollutants has increased. Several studies in the state of the art describe the advantages, benefits, efficiency, and challenges of using low-cost sensors [1,2]. In this way, low-cost sensors to monitor the presence of particulate matter in the environment or inside buildings, both in fractions of 1 µm (PM_1), 2.5 µm ($PM_{2.5}$) and 10 µm (PM_{10}), have increased considerably [3,4]. The United States Environmental Protection Agency (US EPA) study concluded that prolonged exposure to air particles smaller than 2.5 µm ($PM_{2.5}$) increases cardiovascular morbidity, mortality, and respiratory diseases [5]. Ultrafine particles with aerodynamic diameters of $\leq$0.1 µm (PM_1) can cross the air-blood barrier of the lungs to enter people's blood, causing respiratory and cardiovascular problems and damaging the nervous system [6,7]. In conjunction with Internet of Things (IoT) technologies, LCS makes it possible to build real-time air quality monitoring networks, expanding spatial visualization, and facilitating the dispersion of information on measurements of specific air pollutants in your neighborhood [8]. An important aspect to consider is that the LCS requires a correction method (calibration) to improve the precision of the measured data, which can be determined from the values measured by a scientific-level reference instrument [9,10], which allows for improvement in the reliability of the measurements made by the LCS, considering the operating context of the LCS, that is, the meteorological factors recorded in the area of operation.

3. Data Description

The air pollution and meteorological parameter dataset was collected from four air quality monitoring stations in Reynosa, Tamaulipas, in northeastern Mexico (see Fig. 1). Monitoring station 1 (AQ-IoT-01) is located next to an industrial park (west of the city) and near two ample avenues with high vehicular traffic and heavy cargo transportation. Monitoring Station 2 is located in a residential area (north of the town) near the Bravo (Grande) River, which marks the border with the southern United States. The third monitoring station is in the city's southern zone, near a large industrial park and international airport, surrounded by ample avenues with high vehicular traffic and heavy cargo transportation. Finally, the monitoring station 4 is located northwest of the city, in a residential area near an avenue with high vehicular traffic.

The dataset is made up of two CSV files. The raw_pm1_full.csv file contains raw data (34,475 records) collected from February 22 to 28, 2022, from the four air quality monitoring stations.

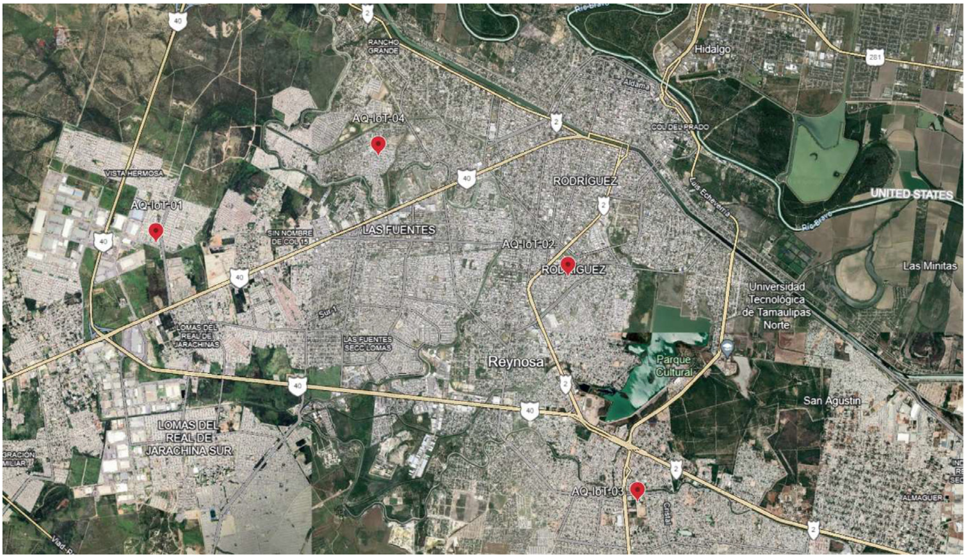

Fig. 1. Location of air quality monitoring stations (red map pointers) over a close view of Reynosa, Mexico. The main roads are shown in yellow.

Table 1
Extract from the dataset collected on February 22, 2022, at the four air quality monitoring stations.

idsample	Place	Date	Hour	PM_1 (μg/m^3)	Particle-03um	Particle-05um	Particle-10um	RH (%)	T (°C)	Heat_index (°C)
751,188	AQ-IoT-01	22/02/2022	00:02:34	24	4680	1332	220	86.9	21.6	22.08
751,189	AQ-IoT-02	22/02/2022	00:02:35	23	5046	1387	254	88.3	21.1	21.57
751,190	AQ-IoT-04	22/02/2022	00:02:39	15	2793	788	157	86.6	21.8	22.3
751,191	AQ-IoT-03	22/02/2022	00:02:55	22	4482	1228	258	99.9	21.6	22.42
751,196	AQ-IoT-01	22/02/2022	00:03:31	23	4719	1334	264	86.7	21.6	22.08
751,197	AQ-IoT-02	22/02/2022	00:03:37	24	5181	1442	242	88.3	21.2	21.68
751,198	AQ-IoT-04	22/02/2022	00:03:44	12	2439	693	125	86.8	21.7	22.19
751,199	AQ-IoT-03	22/02/2022	00:03:58	22	4662	1281	262	99.9	21.6	22.42

This file contains values in the following attributes: *idsample* (unique record identifier), *Place* (name of the monitoring station), *date, time, PM1* (concentration of the air pollutant, measured in μg/m^3), *particle03um, particle05um, particle10um* show the quantities of particles recorded of sizes 0.3, 0.5, and 1.0 μm (from these values the concentration of the PM_1 pollutant is calculated). The *RH* and *T* values contain the recorded values of the percentage of relative humidity and air temperature (°C) and the *Heat_index* variable, which is calculated from the *T* and *RH* values. The PM1_dataset_Average_Hour.csv file contains the hourly mean of each of the above variables in the same period and for the four monitoring stations. This data range was selected because it presents meteorological conditions with high variability, which can be observed in the characteristic weather of each of the four seasons in the study area. These conditions occurred within the ninth week of the winter season. For example, on the first day, a relative humidity between 35 % and 90 % and a temperature between 20 °C and 40 °C were recorded, typical summer weather conditions. On the second day, there was an ambient temperature between 6 °C and 30 °C with a relative humidity between 60 % and 99 %. On the other hand, the fourth and fifth days are characteristic of the winter period, with temperatures between 5 °C and 10 °C with a range in relative humidity values between 75 % and 99 %. Finally, the seventh day presents a behaviour typically observed in the first weeks of the spring season with cold dawn (around 6 °C) and relative humidity with a value of 95 %, and in the early afternoon with a maximum temperature of 28 °C and relative humidity around 20 %. Continuing with a sunset with an increase in relative humidity and a decrease in the temperature value. These drastic changes in relative humidity are due to constant high wind speeds during the day, which causes a rapid decrease in the percentage of ambient relative humidity due to less water evaporation.

An example of the records contained in the raw_pm1_full.csv file is shown in Table 1. These data are displayed as collected by low-cost sensors installed at air quality monitoring stations. The records are sorted by the time and date they were stored in the repository by the software system that automatically manages the operation of the monitoring station.

Table 2 shows the statistical analysis of the data of each variable considered in the PM1_dataset_Average_Hour.csv dataset of the four-monitoring station. This analysis allows us to know the data's distribution and central tendency and the values' dispersion through the interquartile range or the standard deviation according to the presence or non-presence of atypical data values. These atypical data are also called outliers and are characterized by values that are noticeably different from the rest of the data contained in the variable or attribute of the dataset. The descriptive analysis is presented split by an air quality monitoring station (see station column in the Table 2), making it possible to observe the concentration of the pollutant by area of the city, as well as the number of particles classified by their size and that were collected by the sensor in the same period.

In the following figures (Figs. 2–5) shows the relationship between particulate matter PM_1 (μg/m^3) and meteorological variables. This matrix of graphs (see Fig. 2) allows you to visualize the data dispersion of the relationship between the air pollutant PM_1 and each variable independently, considering the four filtering conditions for the data.

Table 2
Descriptive statistics of meteorological factor data and particulate matter concentration.

Station	Variable	Mean	SD	Median	IQR	Min	25 %	75 %	Max
AQ-IoT-01	PM_1	18.36	12.96	17	9.25	3	11	20.25	91
	Particle03um	3878	3399.69	3139	1857	726	2144	4000	22415
	Particle05um	1098.2	985.44	885.5	525.5	200	594.8	1120.2	6544
	Particle10um	189.05	218.66	140.5	104.25	16	82.75	187	1468
	RH	80.5	17.8	87	13	21	79	92	96
	T	13.14	8.84	9	14	5	7	21	40
	Heat_index	12.84	9.97	8	15	4	6	21	45
AQ-IoT-02	PM_1	10.08	7.11	9	11	0	4	15	26
	Particle03um	2050.3	1415.25	1737.5	2110.5	199	882.5	2993	5855
	Particle05um	559.1	391.25	467.5	563	53	235.2	798.2	1615
	Particle10um	86.33	77.17	60	100.25	3	24.75	125	319
	RH	79.21	17.32	86	13.25	20	76.75	90	92
	T	13.28	8.22	10	13	5	7	20	37
	Heat_index	12.86	9.15	9	15	4	6	21	41
AQ-IoT-03	PM_1	16.42	11.6	15	11	0	9	20	73
	Particle03um	3499	2862.93	2999	2295	236	1852	4147	18993
	Particle05um	959.2	816.81	811.5	619	60	504.5	1123.5	5744
	Particle10um	172.84	208.06	133.5	113.75	4	71.75	185.5	1852
	RH	91.34	20.18	100	0	24	100	100	100
	T	13.75	8.1	10	13	6	8	21	37
	Heat_index	13.79	9.19	10	15	5	7	22	42
AQ-IoT-04	PM_1	6	4.7	5	7	0	2	9	26
	Particle03um	1272.5	892.96	1088.5	1151	90	606.8	1757.8	5567
	Particle05um	351.4	249.39	300.5	323.75	23	166.2	490	1550
	Particle10um	51.78	45.80	38	59.25	2	17	76.25	266
	RH	78.05	17.77	85	13.25	18	75.75	89	92
	T	13.69	8.38	10	14	5	7	21	37
	Heat_index	13.25	9.24	9	16	4	6	22	41

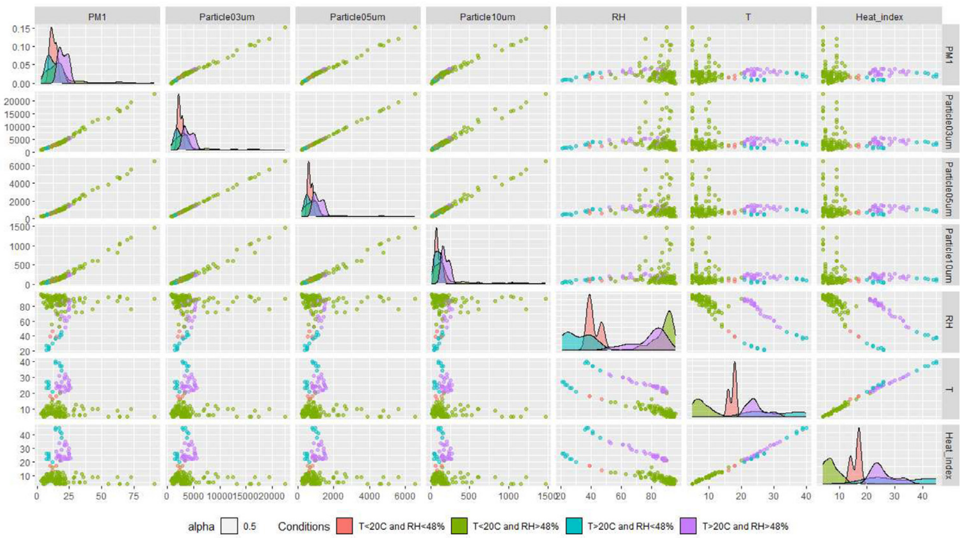

Fig. 2. Scatter and density plots were generated with the values recorded at the AQ-IoT-01 monitoring station and represent the association between the concentration of PM_1 (μg/m^3), particle sizes, and the meteorological variables measured by the PMS7003 low-cost sensor located in Reynosa, Mexico. Source: Author. Dataset: PM1_dataset_Average_Hour.csv.

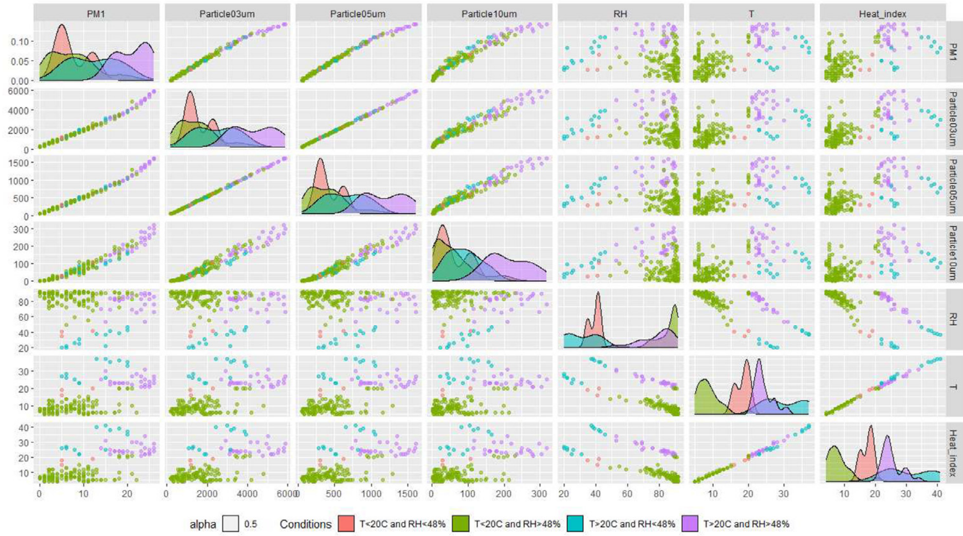

Fig. 3. Scatter and density plots were generated with the values recorded at the AQ-IoT-02 monitoring station and represent the association between the concentration of PM_1(µg/m^3), particle sizes, and the meteorological variables measured by the PMS7003 low-cost sensor located in Reynosa, Mexico. Source: Author. Dataset: PM1_dataset_Average_Hour.csv.

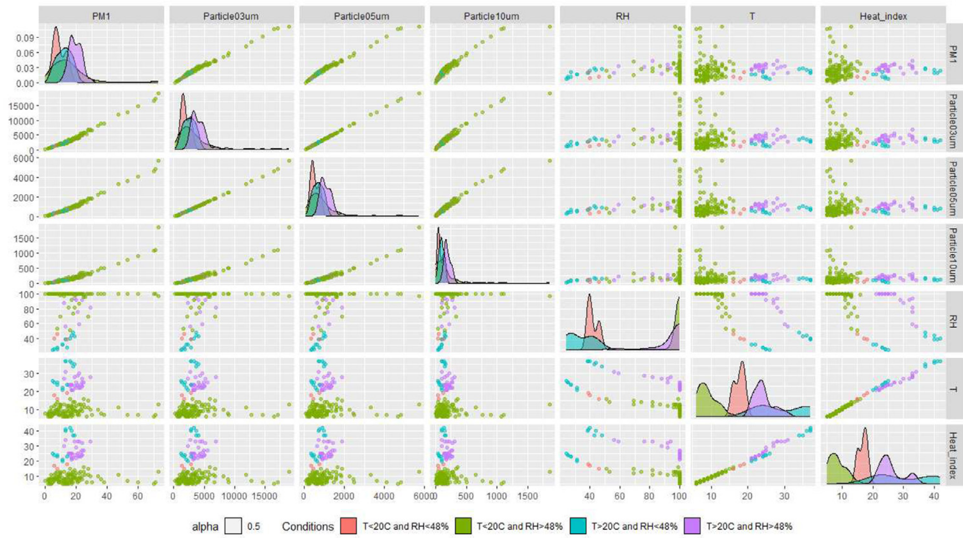

Fig. 4. Scatter and density plots were generated with the values recorded at the AQ-IoT-03 monitoring station and represent the association between the concentration of PM_1 (µg/m^3), particle sizes, and the meteorological variables measured by the PMS7003 low-cost sensor located in Reynosa, Mexico. Source: Author. Dataset: PM1_dataset_Average_Hour.csv.

The scatter diagrams are shown at the top and bottom of the Figs 2–5, which illustrate the relationship between two continuous variables through a representation of points; each point represents the intersection between the values of both continuous variables. The diagonal panel of the figure shows the density plots of the continuous variables (see Fig. 3). These diagrams show the data distribution in a continuous time interval, using the intervals on the abscissa axis and the density of the variable on the ordinate axis (see Fig. 4). Additionally, in these figures,

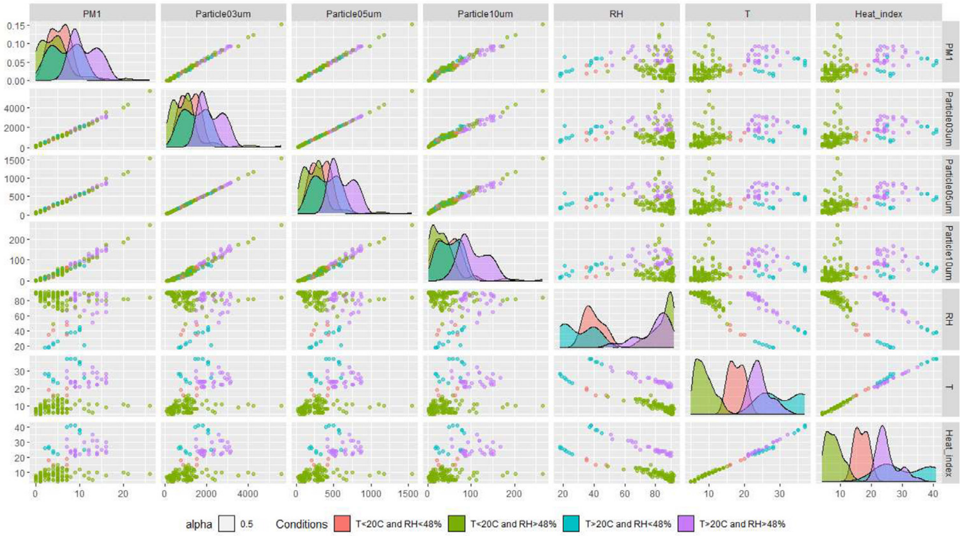

Fig. 5. Scatter and density plots were generated with the values recorded at the AQ-IoT-04 monitoring station and represent the association between the concentration of PM_1 (µg/m^3), particle sizes, and the meteorological variables measured by the PMS7003 low-cost sensor located in Reynosa, Mexico. Source: Author. Dataset: PM1_dataset_Average_Hour.csv.

the data were grouped by applying a set of conditions that allow the behaviour of the pollutant PM_1 to be identified from the combination of recorded values of RH and T. The dispersion and density diagrams are filled according to the colour of the group to which that data belongs (see Fig. 5).

4. Experimental Design, Materials and Methods

Each air monitoring station has a Plantower PMS7003 sensor (LCS) installed for particulate matter, which is used to obtain the number of particles suspended in the air. This sensor is based on the laser dispersion of light; the laser hits the particles, and the dispersed light is collected and transmitted to the microprocessor. The microprocessor classifies the particles (reporting six size bins: $>$0.3, $>$0.5, $>$1, $>$2.5, $>$5, and $>$ 10 µm) by calculating their size and counting the number of particles by their size, using an algorithm based on the MIE theory, as described in [11].

The PM LCS was calibrated through a placement process with a regulatory grade reference instrument or equivalent monitor (FRM/FEM) under real-world conditions over a defined evaluation period by current regulations and methodologies for LCS [12,13]. Then, considering local meteorological factors before the experiment period, the PM LCS was calibrated, reaching a coefficient of determination (R^2) of 0.96 for PM_1 about the reference instrument. The calibration procedure implemented was similar to that described in our previous work [11]. The LCS monitoring equipment was installed on the buildings' roofs at a height of 8 m above ground level, with 6 feet free between the roof and the sensor and without obstacles (e.g., tree or shrub barrier) that prevent the free flow of air towards the sensor. This will make it possible to obtain more representative measurements of air quality at the neighborhood level.

Then, at each LCS air monitoring station, data on particulate matter with diameters less than 1.0 µm (PM_1), 2.5 µm ($PM_{2.5}$), and 10 µm (PM_{10}) can be sensed, as well as meteorological factors of relative humidity and temperature, from these factors the heat index is calculated internally. The data is automatically transmitted to storage in a private cloud mounted on Internet services.

Data is sensed every minute and transmitted in real-time. On the private cloud side, software components were implemented that deploy several web services (for more in-depth detail, see [14]), which receive the data transmitted by the monitoring stations, validate the consistency of the data using a set of predefined business rules, selecting the values of each attribute for insertion into the corresponding database table, automatically assigning a unique identifier and a timestamp. If the value does not meet the requirements for storage (for example, a null value), it is discarded; that is, it is not added to the database. The cloud storage was designed using a SQL-based database.

Limitations

The dataset represents the concentration of the PM_1 pollutant in a specific period in Reynosa, Mexico, which presents a geographical limitation. Furthermore, due to the city's geographic and territorial extension, it is advisable to install additional monitoring stations in other sectors at the neighborhood level. The dataset was collected in real-time with low-cost IoT devices, which may have had some Internet connectivity failure due to the service provider's intermittency or interference in WI-FI network communication. Therefore, some gaps in the pollutant concentration time series are possible.

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

Particulate matter with aerodynamic diameters ≤1µm (PM1) dataset (Original data) (Mendeley Data)

CRediT Author Statement

Luis A. Garcia-Garza: Conceptualization, Methodology, Formal analysis, Resources, Investigation, Writing – original draft; **Edgar Tello-Leal:** Conceptualization, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition; **Bárbara A. Macías-Hernández:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Visualization; **Gerardo Romero:** Writing – review & editing, Supervision, Project administration, Funding acquisition; **Jaciel David Hernandez-Resendiz:** Software, Validation, Data curation, Formal analysis.

Acknowledgements

This research was partially funded by Universidad Autónoma de Tamaulipas (México) under grant number PEI2018-UAT-2022.

Declaration of Competing Interests

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.

References

[1] I. Zeba, L. Zheng, Low cost air pollution monitoring systems: a review of protocols and enabling technologies, J. Indus. Info. Integr. 17 (2020) 100123, doi:10.1016/j.jii.2019.100123.

[2] X. Liu, R. Jayaratne, P. Thai, T. Kuhn, I. Zing, B. Christensen, R. Lamont, M. Dunbabin, S. Zhu, J. Gao, D. Wainwright, D. Neale, R. Kan, J. Kirkwood, L. Morawska, Low-cost sensors as an alternative for long-term air quality monitoring, Environ. Res. 185 (2020) 109438, doi:10.1016/j.envres.2020.109438.

[3] J. Bi, A. Wildani, H.H. Chang, Y. Liu, Incorporating low-cost sensor measurements into high-resolution $PM_{2.5}$ modeling at a large spatial scale, Environ. Sci. Technol. 54 (4) (2020) 2152–2162, doi:10.1021/acs.est.9b06046.

[4] M.R. Giordano, C. Malings, S.N. Pandis, A.A. Presto, V.F. McNeill, D.M. Westervelt, M. Beekmann, R. Subramanian, From low-cost sensors to high-quality data: a summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors, J. Aerosol. Sci. 158 (2021) 105833, doi:10.1016/j.jaerosci.2021.105833.

[5] United States Environmental Protection Agency (US EPA). Particle Pollution exposure. United States Environmental Protection Agency, 2023. URL: https://www.epa.gov/pmcourse/particle-pollution-exposure. Accessed 05 February 2024.

[6] R. Xu, J. Wei, T. Liu, Y. Li, C. Yang, C. Shi, G. Chen, Y. Zhou, H. Sun, Y. Liu, Association of short-term exposure to ambient PM_1 with total and cause-specific cardiovascular disease mortality, Environ. Int. 169 (2022) 107519, doi:10.1016/j.envint.2022.107519.

[7] I.L. Schneider, E.C. Teixeira, G.L. Dotto, D. Pinto, C.X. Yang, L.F.O. Silva, Geochemical study of submicron particulate matter (PM_1) in a metropolitan area, Geosci. Front. 13 (1) (2022) 101130, doi:10.1016/j.gsf.2020.12.011.

[8] J.V. Jagatha, A. Klausnitzer, M. Chacón-Mateos, B. Laquai, E. Nieuwkoop, P. van der Mark, U. Vogt, C. Schneider, Calibration method for particulate matter low-cost sensors used in ambient air quality monitoring and research, Sensors 21 (12) (2021) 3960, doi:10.3390/s21123960.

[9] A.L. Holder, A.K. Mebust, L.A. Maghran, M.R. McGown, K.E. Stewart, D.M. Vallano, R.A. Elleman, K.R. Baker, Field evaluation of low-cost particulate matter sensors for measuring wildfire smoke, Sensors 20 (17) (2020) 4796, doi:10.3390/s20174796.

[10] Y. Hashmy, Z.U. Khan, F. Ilyas, R. Hafiz, U. Younis, T. Tauqeer, Modular air quality calibration and forecasting method for low-cost sensor nodes, IEEE Sens. J. 23 (4) (2023) 4193–4203, doi:10.1109/JSEN.2023.3233982.

[11] B.A. Macías-Hernández, E. Tello-Leal, O. Barrios S, M.A. Leiva-Guzmán, R. Toro A, Effect of environmental conditions on the performance of a low-cost atmospheric particulate matter sensor, Urban Climate 52 (2023) 101753, doi:10.1016/j.uclim.2023.101753.

[12] R. Williams, V. Kilaru, E. Snyder, A. Kaufman, T. Dye, A. Rutter, A. Russell, H. Hafner. Air Sensor Guidebook. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-14/159 (NTIS PB2015-100610), 2014. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=277996&simpleSearch=1&searchAll=air+sensor+guidebook. (accessed 05 February 2024).

[13] R. Duvall, A. Clements, G. Hagler, A. Kamal, V. Kilaru, L. Goodman, S. Frederick, K.Johnson Barkjohn, I. VonWald, D. Greene, T. Dye. Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors: use in Ambient, Outdoor, Fixed Site, Non-Regulatory Supplemental and Informational Monitoring Applications. U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-20/280, 2021. https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=CEMM&dirEntryId=350785&Lab=CEMM. (accessed 05 February 2024).

[14] J.D. Hernandez-Resendiz, E. Tello-Leal, U.M. Ramirez-Alcocer, B.A. Macías-Hernández, Semi-Automated approach for building event logs for process mining from relational database, Appl. Sci. 12 (21) (2022) 10832, doi:10.3390/app122110832.