Statistical analysis of air quality data

By Kajus Merkeliūnas

Supervisors: Franck Mercier, Marie-Lise Pannier, Alain Godon

Table of Contents

Introduction

Context

As part of my studies, I was offered the opportunity to participate in a group project. The subject was chosen to be in the field of statistics. This was done, as I had not chosen anything related to statistics before in my academic studies and wanted to experiment in the field. As such, this was a good opportunity to learn more about the study of statistics.

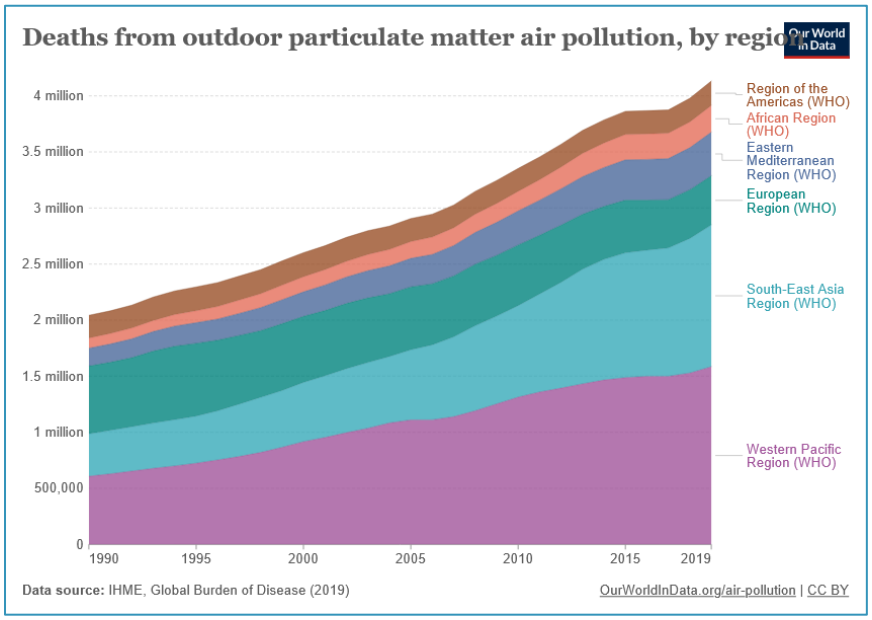


Figure 1 - Data on deaths caused by air pollution.

Furthermore, the matter of air quality is important today. This can be seen in *Figure 1*. There have been deaths rising from 1990 onwards, and my project tries to deal with the case of this rampant danger. There have been many studies on this matter, with the dangerous factors that air pollution increases:

- Respiratory and Cardiovascular Diseases: Lareg-term exposure to air pollutants, particularly PM2.5 and ozone, is associated with increased incidence of respiratory diseases like asthma and chronic obstructive pulmonary disease (COPD), as well as cardiovascular diseases (Brook et al., 2010).
- Mortality: Research indicates that air pollution significantly contributes to premature mortality. The Global Burden of Disease study estimates

millions of deaths annually are attributable to ambient air pollution (Cohen et al., 2017).

- Children's Health: Air pollution has profound effects on children's health, leading to developmental issues and exacerbation of asthma (Perera, 2017).

For this reason, low-cost sensors were employed. Advances in technology have led to the development of low-cost air quality sensors. These provide greater spatial resolution and can be deployed in dense networks, although they often require calibration against reference methods (Castell et al., 2017).

It is important to mention that coronavirus disease has greatly impacted the standards of air quality control. It is important to open windows at intervals to obtain fresh air in the case of epidemic coronavirus (Elsaid et al., 2021). Opening windows or operating air conditioning can help alleviate the problems of air pollution or diseases. As such, it is important to monitor the data in these regions: Bibliothèque Universitaire Saint-Serge, FABLAB in School Polytechnic of L'université D'Angers and my supervising professors' office.

My research will primarily consist of sensors in FABLAB, as that area is accessible to me, and it holds much equipment that releases particles. For example, 3D printers usually release particles and volatile organic compounds (VOCs) (Byrley et al., 2020). Also, FABLAB holds a total of 8 sensors, which are enough to cross-reference data to gather more accurate results. Most importantly, many students visit the laboratory to use the machines. For this reason, it is important to analyze the data and protect the students' health.

The gathered data from the sensors are in these parameters: CO2, temperature, humidity, VOC, C2H5OH, CO, NO2, NH3, sound, light, pm1.0, pm2.5, pm10. With these parameters, we can gauge many aspects. These parameters can be categorized to improve understanding of air quality. Gaseous pollutants include carbon dioxide (CO2), volatile organic compounds (VOCs) such as ethanol (C2H5OH), carbon monoxide (CO), nitrogen dioxide (NO2), and ammonia (NH3), which are all indicators of air pollution and its potential health impacts. Particulate matter is divided into PM1.0, PM2.5, and PM10, based on the size of the particles. These metrics help in assessing air quality concerning respiratory health. Environmental parameters like

temperature and humidity influence the concentration of pollutants, while sound and light levels provide additional information for environmental conditions. By organizing the data into these categories, we can conduct a more comprehensive analysis of air quality and environmental conditions in the laboratory setting.

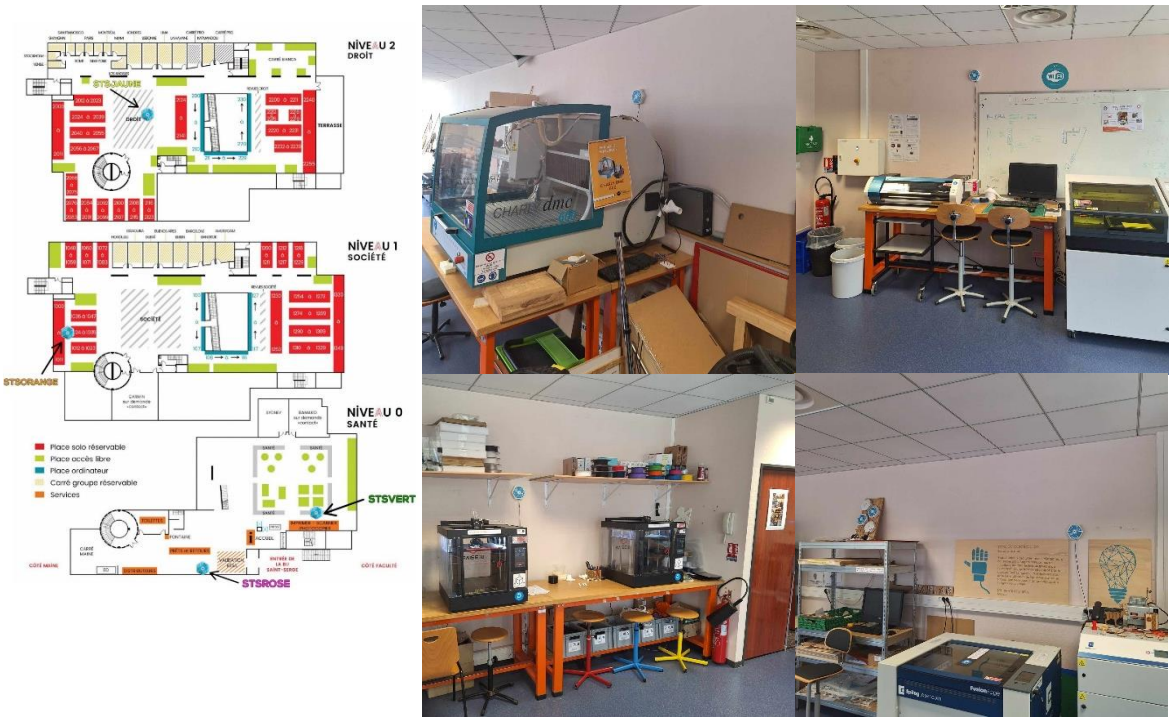


Figure 2 - Showcase of sensors. Left – map of sensors in library. Right – sensors in FABLAB, colored blue.

LARIS has developed sensors of different types, linked to air quality. The sensors have been gathering data since the end of June 2022. The earliest sensors were employed in the library, of which there were 4. Starting from April 2023 FABLAB got its own sensors sensor. There are 4 sensors around the room and 4 sensors next to each other. The 4 sensors around the room measure different equipment, as one is placed right next to a 3D printer, while another is near a laser cutter. These different areas can show what particles each machine releases and pinpoint accurately if it is working, and if the particles the machine produces have diffused over the room. The 4 sensors next to each other should be gathering the same data.

Project presentation

My project will deal with analyzing the gathered data over 2 years. I have been given much control over what methods to use to get various results. The main problem is that sensors do not provide calibrated data, which complicates the extraction of quantitative information. My main subject proposed is to create a program to analyze this data and extract the relevant information. Formatting according to several reading levels is expected.

Difficulties arise, as the data isn't calibrated, comes at non-exact intervals, sometimes is missing and has unwanted noise. With these problems in mind, it is important to calibrate data. For this derivative analysis was used as the main analysis method.

With the idea of derivative analysis, we do not need to calibrate the values the sensors are displaying, but only their derivatives. For example, if both sensors are rising rapidly, even though the values of them are different, we can assume that both sensors are detecting the same particle in the air.

By analyzing previously gathered data, it is possible to get what should be the derivative throughout the day. With this table, we can align the current data and determine if it is abnormal. As such, getting the curve of an average day is important. To get it as precise as possible statistical methods will be used, such as:

- getting rid of duplicate entries,
- getting rid of extreme noise entries,
- smoothing the data to get rid of the noise,
- calculating the mean of an interval,
- and calculating the average itself.

For this data analysis, it is required to acknowledge the tools. In Python, powerful libraries like Pandas and NumPy offer comprehensive tools for data analysis. Matplotlib lets us display the data and libraries like SciPy and DateTime allow us to simplify the code.

Literature review

This work was built upon base Python code [Annex 1], and information from previous student Ayoub HANNAD in his *Projet de Fin d'Étude "Étude de la qualité de l'air intérieur dans la bibliothèque Saint Serge: Analyse des données et détection d'anomalies."*

Previous and current code: https://github.com/Sas3y/SAAQD

Studies researched:

Brook, R. D., Rajagopalan, S., Pope, C. A., et al. (2010). Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation*, 121(21), 2331-2378.

Cohen, A. J., Brauer, M., Burnett, R., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907-1918.

Perera, F. (2017). Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: Solutions exist. *International Journal of Environmental Research and Public Health*, 15(1), 16.

Castell, N., Dauge, F. R., Schneider, P., et al. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99, 293-302.

Elsaid, A. M. M., & Ahmed, M. S. (2021). Indoor air quality strategies for air-conditioning and ventilation systems with the spread of the global coronavirus (COVID-19) epidemic: Improvements and recommendations. *Environmental Research*, 199, 111314.

Byrley, P., Wallace, M. A. G., Boyes, W. K., & Rogers, K. (2020). Particle and volatile organic compound emissions from a 3D printer filament extruder. *Science of The Total Environment*, 736, 139604.

Methods and results

Research stage

The research stage took around 2 weeks. During it, I familiarized myself with the previously written code provided by professors, got the concepts of Python and learned about statistical methods. I experimented with various concepts, like Principal Component Analysis (PCA) and wrote code to establish my learnings, but I never incorporated it into the main project.

We set out our goals, but my experience with the chosen coding language Python was subpar and we needed to change them. In this stage we decided to focus on exploring, rather than achieving results with various methods, but mainly derivative analysis.

First, it is important to address that there are problems with gathering the data. Some sensors would not be working, some were plugged out of power, and some months would not be recorded at all because of server reasons. Some to highlight:

- 'FABTEST1' has not been working since 2023-10-26 until recently.
- 'FABTEST2' has not been working since 2023-06-29 until recently.
- 'STSROSE' has not been working since 2024-04-05 until recently.
- All not working from 2024-02-27 until 2024-04-01.
- Sensors do not work during the night.

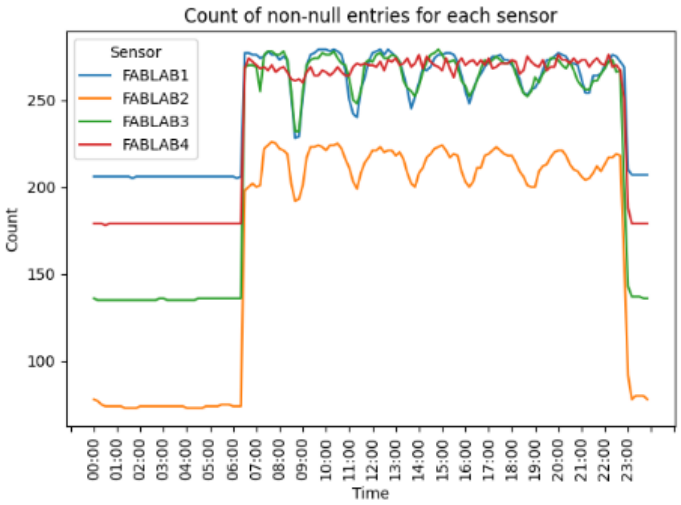


Figure 3 - Count of recorded data of FABLAB sensors during the span of the day during the year 2023.

To measure how much data is missing I designed *Figure 3*. It showcases how many non-null entries are for each 10-minute interval. As most calculations deal with this interval and the average designed interval between the sending of data is 5 minutes, each 10-minute interval should have 2 entries. The given graph does not count each entry, but whether there was a data entry in the 10-minute interval. The tests were done with 'FABLAB' sensors, which are in the FABLAB of Angers Polytech faculty and are placed around the laboratory. Interestingly though, most of them dip under specific time frames. *Figure 4* shows a sample of a 6.5-hour timeframe with missing results highlighted in yellow.

It is important to mention that this is only a temporary problem, that does not appear during 2024 as displayed in *Figure 5*, where during the first half of 2024, there have not been similar dips in the gathered data.

date	FABLAB1	FABLAB2	FABLAB3	FABLAB4
2023-04-20 08:00:00	450.5	402	476.5	471.5
2023-04-20 08:10:00	450.5	413	478	466.5
2023-04-20 08:20:00	451	416	477	473
2023-04-20 08:30:00		420.5		476
2023-04-20 08:40:00				471.5
2023-04-20 08:50:00	452		473	472
2023-04-20 09:00:00	452.5		470	472
2023-04-20 09:10:00	449	407	467	467.5
2023-04-20 09:20:00	450.5	406	464.5	472
2023-04-20 09:30:00	458	405.5	458	468
2023-04-20 09:40:00	458	406	468	464
2023-04-20 09:50:00	457.5	411.5	469.5	462.5
2023-04-20 10:00:00	448	409	463.5	465
2023-04-20 10:10:00	450	414	464	460.5
2023-04-20 10:20:00	451.5	411.5	460.5	459
2023-04-20 10:30:00	447	418	468	
2023-04-20 10:40:00	453	405	464	466
2023-04-20 10:50:00	453.5	405.5		459
2023-04-20 11:00:00		405		467.5
2023-04-20 11:10:00		405		460
2023-04-20 11:20:00		402	458	457.5
2023-04-20 11:30:00	444		461	459
2023-04-20 11:40:00	445		458.5	460
2023-04-20 11:50:00	440.5	399	457.5	
2023-04-20 12:00:00	442	400	456	466.5
2023-04-20 12:10:00	442.5	395	458.5	464.5
2023-04-20 12:20:00	442.5	401	456	461.5
2023-04-20 12:30:00	440	402	453.5	457
2023-04-20 12:40:00	439	413.5	456.5	458
2023-04-20 12:50:00	443.5	399	454	451
2023-04-20 13:00:00	442.5	400.5	458.5	458.5
2023-04-20 13:10:00	444.5	396.5		461
2023-04-20 13:20:00		397	453	
2023-04-20 13:30:00		398		462
2023-04-20 13:40:00		398		457
2023-04-20 13:50:00	443.5			450
2023-04-20 14:00:00	440		449	448.5
2023-04-20 14:10:00	435.5		447.5	450.5
2023-04-20 14:20:00	433	389	438.5	455.5
2023-04-20 14:30:00	436	389	438.5	454

Figure 4 - Random time frame of a day (2023-04-20) displaying the missing data.

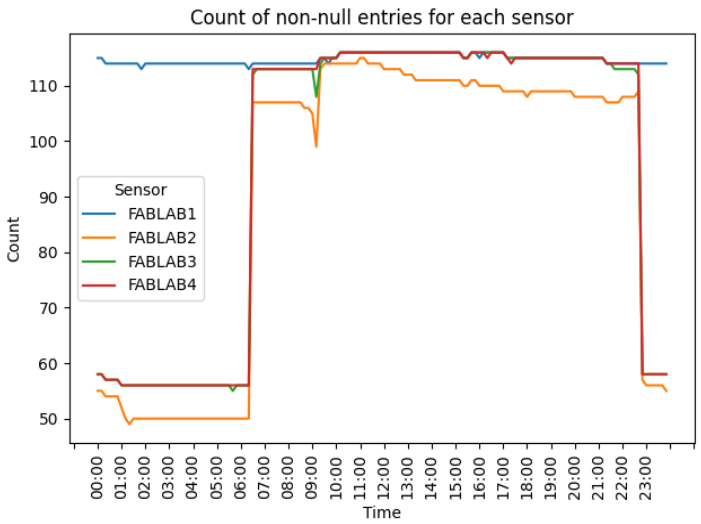


Figure 5 - Count of recorded data during the span of the day during the first half of 2024.

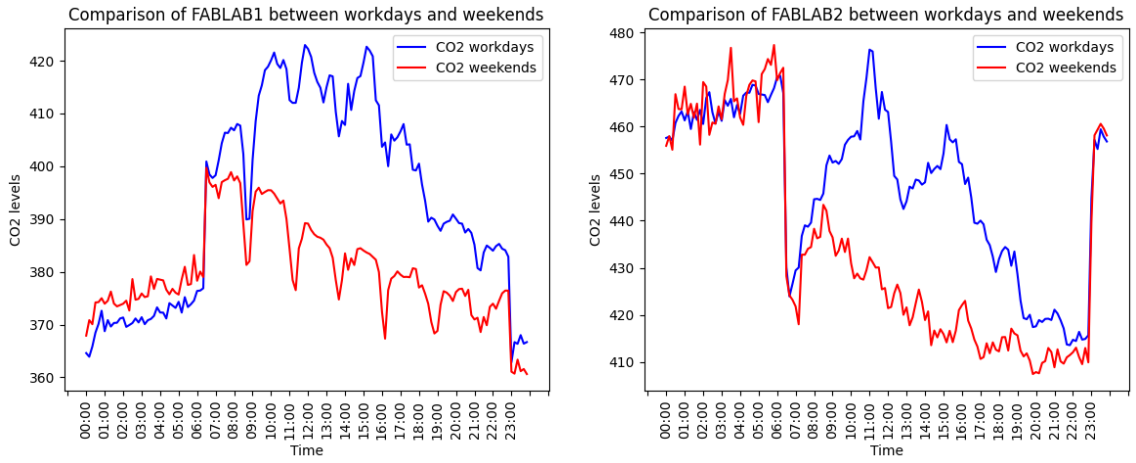


Figure 6 - Average curve of the day throughout 2023-04 to 2023-10 in FABLAB1 and FABLAB 2 sensors.

These discussed problems have greatly impacted the gathered data. The dips in *Figure 6* are noticeable and the lack of data during the night provides rare results which impact the average day curve to be non-specific and as such, not usable for comparisons to analyze future data. However, it does get better with time, as by 2024 the sensors are working more stable now.

With this, my course of action will be transforming this data to be more accurately applicable to future sensor data, by gathering its derivative. As the data should not be cross-referenced by value, as the sensors are uncalibrated, it will be necessary to obtain a valuable derivative.

Test stage – primary focus on CO2 data

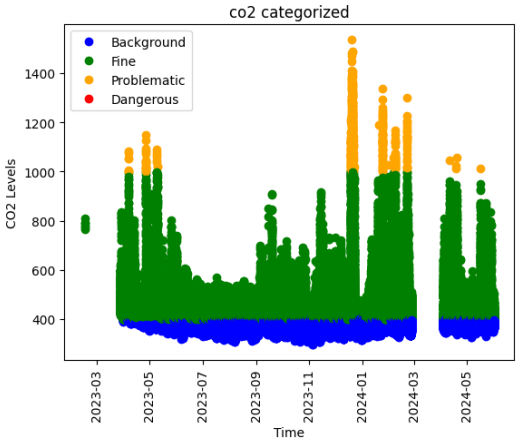
This phase was primarily focused on data from CO2 (carbon dioxide) with FABLAB sensors. As many living organisms use it to create energy molecules, it is released as a byproduct of breathing. Many companies have found ways to create effective and precise sensors for it. In turn, this sensor was deemed to be the most precise, and working with it will give the best results for comparisons.

Most of the information portrayed and done in the report is on FABLAB main sensors. While it is possible to change the sensors easily in the code, I believe they give the best results for observing trends and analyzing the information. Their data is preserved quite well with them being online in tandem for most of the observed period.

CO2 data is a great gauge of how many people occupy a space. In *Figure 6* it is noticeable that there is a high difference between days when people occupy the room (workdays in blue) and when there aren't any people (weekends in red). As such, all my analysis started with this sensor's data.

Coronavirus is also an important factor in this. As the pandemic struck, it was important to review precautions to stop disease spreading. If it is possible to distinguish the number of people in the room it is possible to send an alarm to open windows – a safe way to reduce the chance of disease spreading (Elsaid et al., 2021). This is one of the main reasons for installing the sensors, so we should investigate the possibility of determining the number.

After gathering the results which are akin to *Figure 6*, we understood it would be too imprecise. The idea for determining the number from CO2 sensors would need a reference point, i.e. how many people on average are present in the room by hour. While we could guess, for example, 3, the effect of the results would largely change the data we did get. While we could examine that, if it would be double the value of the workday sample minus the weekend sample, which would correspond to double the amount of CO2 where there are people present, we could assume that there are 6 people, but this value is once again estimated on the initial assumption of 3 people. As such we had to abandon this idea.



250-400ppm	Normal background concentration in outdoor ambient air
400-1,000ppm	Concentrations typical of occupied indoor spaces with good air exchange
1,000-2,000ppm	Complaints of drowsiness and poor air.
2,000-5,000 ppm	Headaches, sleepiness and stagnant, stale, stuffy air. Poor concentration, loss of attention, increased heart rate and slight nausea may also be present.
5,000	Workplace exposure limit (as 8-hour TWA) in most jurisdictions.
>40,000 ppm	Exposure may lead to serious oxygen deprivation resulting in permanent brain damage, coma, even death.

Figure 7 - CO2 categorized into levels of dangerous concentrations.

Another idea was determining the healthy amount of CO2 ppm in the room. This was possible under the assumption that CO2 sensor values are precise enough to estimate its zone in *Figure 7*. While this graph portrays too much

information for the size of the given graph, we can still observe some tendencies. I believe the lack of problematic zones during summer was because the summer holidays. The same could be said right before and after New Year's. As I have not got a full grasp of what kind of projects are done there and when are their deadlines, I can not conclude on that, but I believe investigation under this assumption could provide structure mandatory brakes and supervised air conditioning during the season in which many of the deadlines are being met.

The work with this data also provided me with an insight into the noise. It was important to distinguish it early, as abnormal values caused unexplainable spikes in the graphs. As such, it was easy, yet important, to filter the data by dropping entries with values, i.e. ±2147484 or 0. These values appeared rarely, so not much data was lost and future purposes of smoothing the data and calculating averages will be more precise. Different sensors also had different ranges of data presented, and for some, like temperature, the range was specific and the filtering of rare cases of 0 and 100 degrees were discarded, as they could not naturally occur. For other parameter data, it needed to be observed to measure the range the results should be recorded, and then set accordingly.

The tools of Python's exhaustive library were also explored. The aforementioned libraries were used to process and display data easily. With documentation provided by Python users and forums, it was easier than I thought to create the code I wanted.

There was also exploration into moving workdays which aligned with holidays to weekends, as the reason for the lack of people would be even more distinct, but after doing it for the time frame of 1 year the results were negligible and the process of writing future-proof holidays complicated the process.

Derivative analysis stage

This step was the heart of my project - derivative analysis. This is used, as even uncalibrated data can be compared.

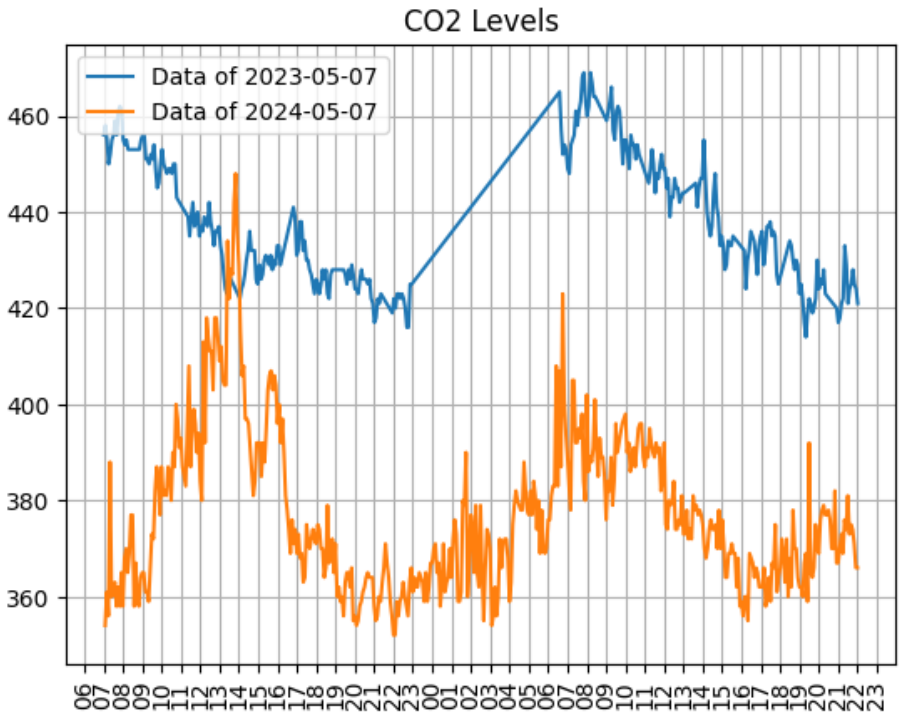


Figure 8 - Raw data of CO2 of a timeframe of 2 days, a year apart.
Note: 2023 data wasn't recorded periodically during the night.

As can be seen in *Figure 8*, even though the data during the first day at around 13:00 to 14:00 matches, the levels of actual CO2 might be very different. This is because 2024 data (orange) has increased to a high level in a short time. This contrasts with the other day, where while the values are different between the specified days, the direction of their increase is around the same rate. This means that both days have experienced around the same rate of CO2 emissions. It is also important to highlight that 2023-05-07 was Saturday, which likely had no people in the room, whereas other shown days were workdays.

As such, this data could be used to:

- Calculate the derivative of the previous data and save it in a file.
- Compare the current data's derivative to the previously calculated data.
- Recognize if it is a dangerous rise to the respective parameters.

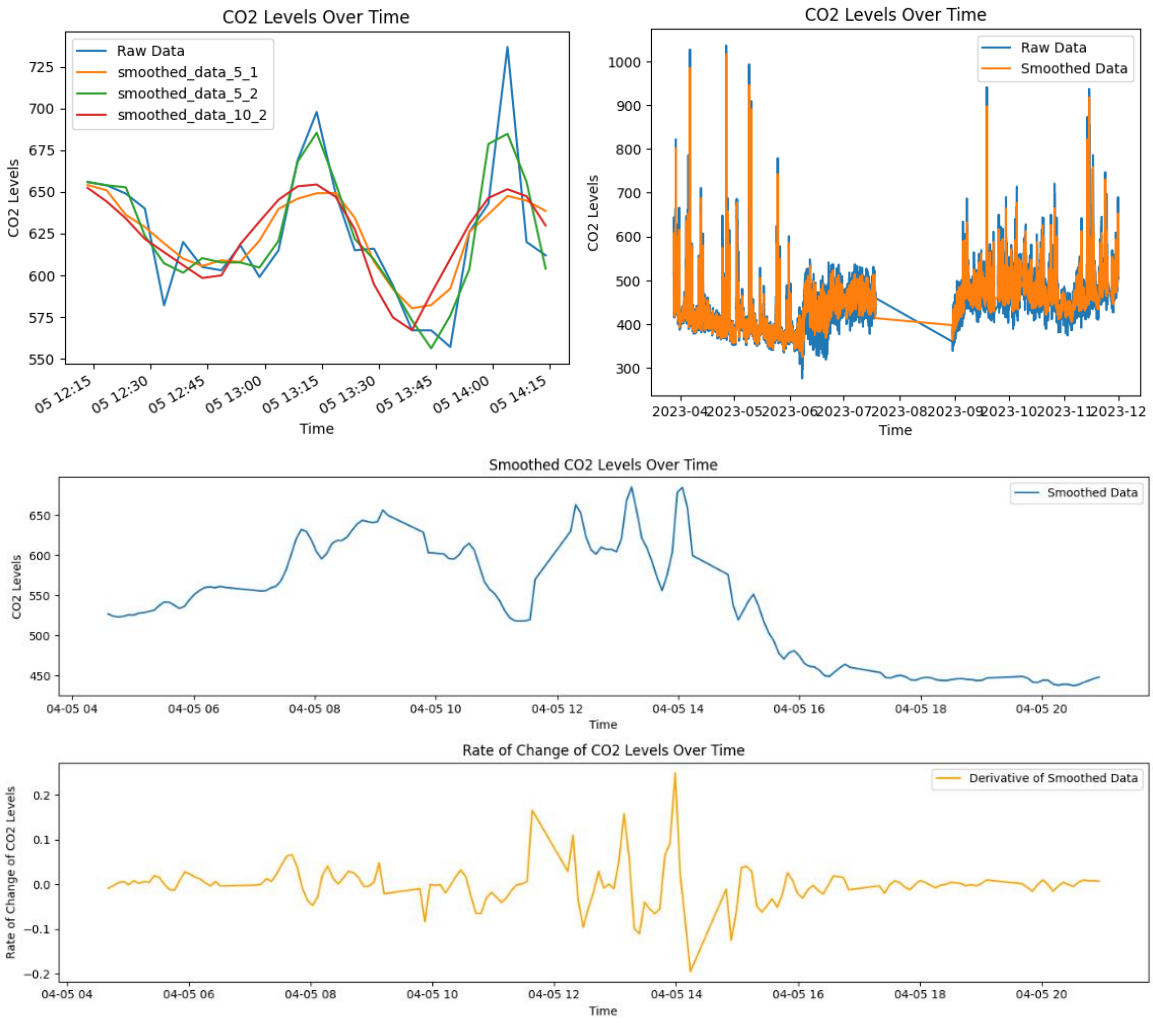


Figure 9 - Graphs showcasing smoothing.

It is also important to smooth the curve. As seen in the previous figures, the data is not smooth enough and can calculate too noisy derivatives. In *Figure 9* three graphs are displayed. First showcases various degrees of smoothing. I chose the Savitzky–Golay filter, as it was fast and appropriate for my work, had a library in Python, and for all future smoothing decided to go with parameters: window length of 5 and polynomial order of 1 (color yellow in the first graph). This decision was made as the data was appropriately smoothed unlike 5-1 (green) and did not lag behind the graph unlike 10-2 (red) from the raw data (blue). I did some tests of data, and the second graph of *Figure 9*, looked appropriate for all the future graphs I used. Lastly, the third graph showcases the derivative of a randomly chosen day, and that it is much less noisy and applicable for future analysis.

The data is also inconsistent with its timing. Calculating the derivative requires a difference in value and time of the last recorded entry. Then, the calculation of the mean will require to arrange the data in segments. The 10-minute interval is appropriate, as usually the data arrives every 5 minutes and will create a mean of 2 values, or a different amount if the sensors send a different frequency of entries. With the 10-minute intervals placed evenly, it is possible to calculate its average over the period much easier.

With this, I wrote the code that:

- gathered raw data throughout the whole period the sensors were active,
- dropped duplicate entries,
- filtered for extreme noise,
- smoothed the data with Savitzky–Golay filter,
- calculated the derivative for each entry,
- calculated the mean of the derivative for a time frame of 10 minutes,
- separated the data into workdays and weekends,
- calculated the *stabilized* curve,
- displayed the data.

With this, I obtained a graph for each sensor with 3 lines, representing the usual curve of the derivative throughout the day. *Figure 10* shows one such example. We can see the separation between workdays (in blue) and weekends (in orange) during work hours. For example, after 8:00 the curve jumps high, indicating an increase of CO2 in the room that the sensor is observing.

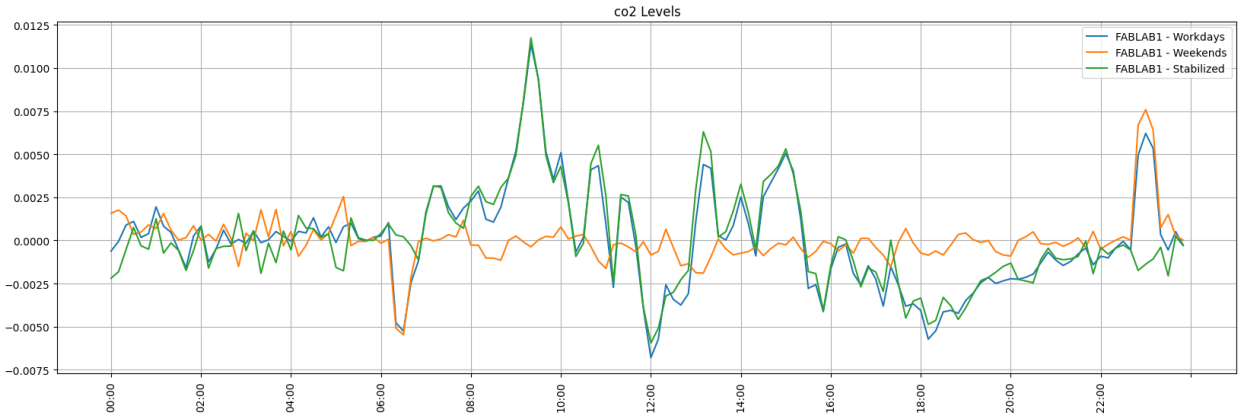


Figure 10 - Average derivative throughout the day of FABLAB1 CO2 sensor. Data sampled from 2023-04 to 2024-06.

The green curve in the graph indicates a *stabilized* value. It is obtained by subtracting the weekend derivative value from the workday derivative. As such, we can analyze the data more reliably by seeing the graph with less noise. For example, the dip around 6:30 is observed in both sensors, yet the stabilized curve shows that it is minimal. The same can be observed with the spike around 23:00.

Such dips and spikes are possible because of the missing values which were discussed in *Figure 3* and *Figure 4*, the missing data values create the illusion of CO2 levels rapidly dropping. As such the sensor data could be curated to not include those days, or, as we are interested in work hours, could only look at the appropriate period.

Also, the stabilized value should represent the actual human emissions derivative. This is because, if we assume that natural CO2 emissions are coming from outside, the green curve showcases the observed humans and the operating machines in FABLAB activity. However, it is important to mention that the curve should only be used for analysis and not data processing. As sensors detect the background emissions from outside, they are most likely to repeat the same dips during their recording of the data.

Lastly, I believe that the gathered data showcased in *Figure 10* could be used in the real world, as opposed to some other gathered graphs. While it is possible to curate days where the missing values are not apparent, e.g. *Figure 11*, the data as it stands showcases an average cycle of the day. With this data, we can measure higher than wanted jumps of CO2 and alert the administration to open the windows or increase air conditioning and reduce the effects of air pollution.

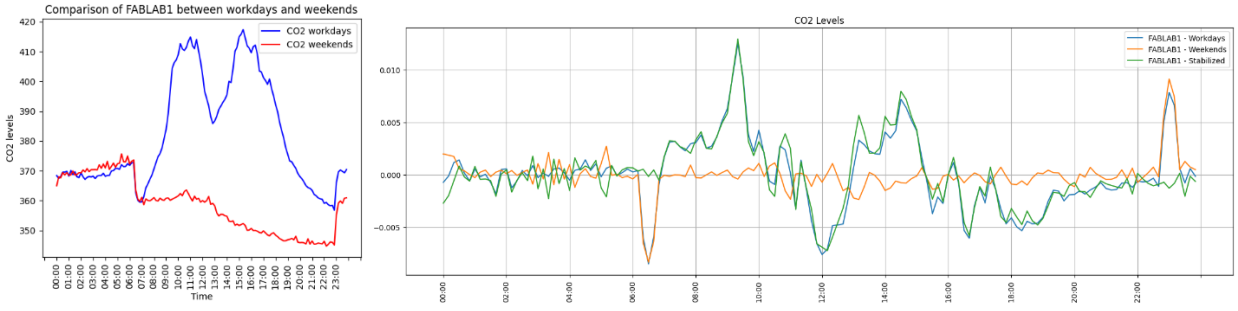


Figure 11 - More curated look at FABLAB1 CO2 data. Data sampled from 2023-06-10 to 2024-03-01.

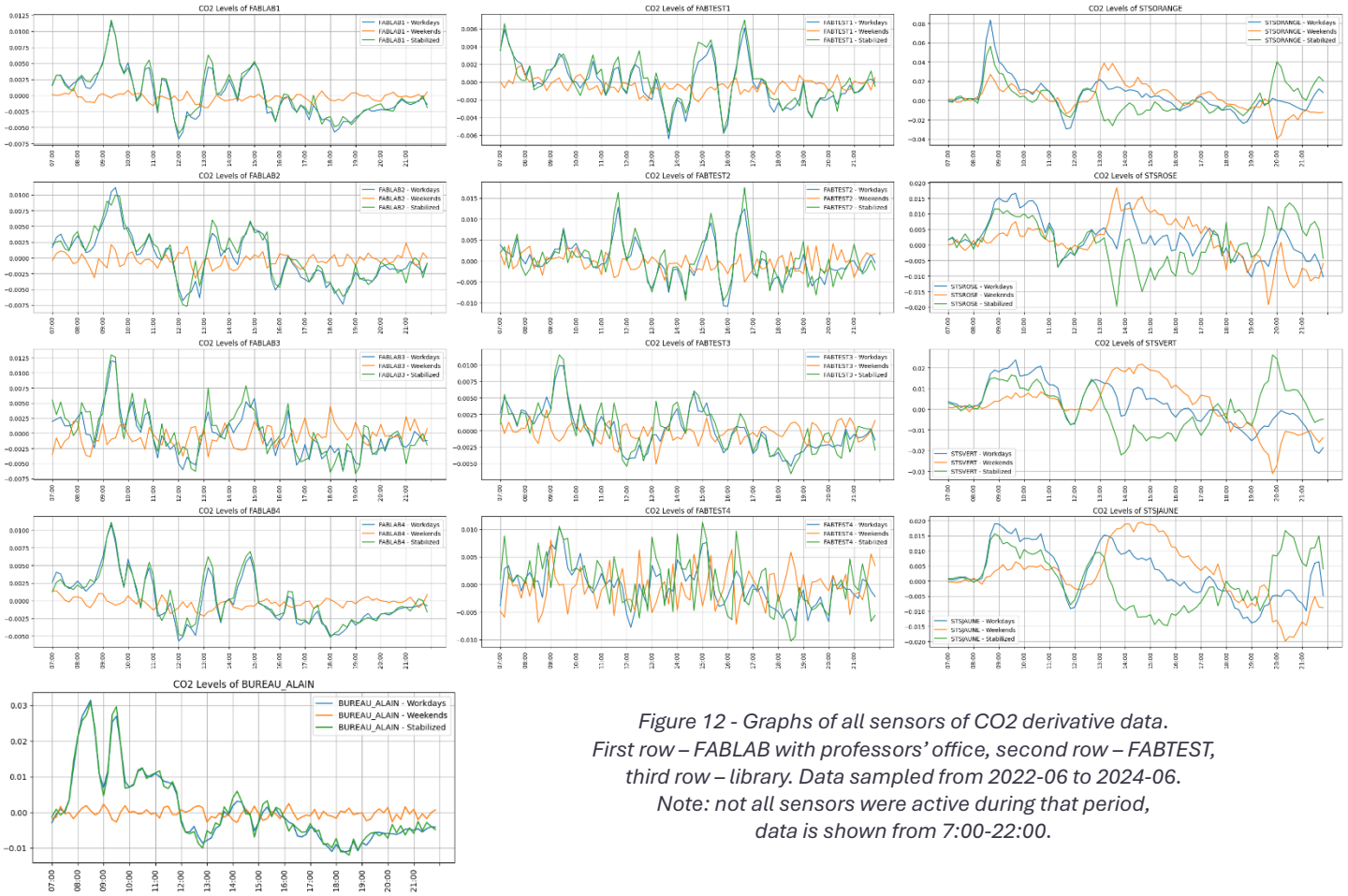


Figure 12 - Graphs of all sensors of CO2 derivative data. First row – FABLAB with professors' office, second row – FABTEST, third row – library. Data sampled from 2022-06 to 2024-06. Note: not all sensors were active during that period, data is shown from 7:00-22:00.

Figure 12 showcases all the gathered data on CO2 and shows their derivative throughout the day. Some appear well and usable, like FABLAB and especially professors' office sensors. They correlate with the average flow of people. It is important to highlight that professors' office sensor background or weekend curve is very close to being close to close to 0. This indicates that the sensor is not picking anything, besides the higher levels of CO2, which in this case corresponds to professors being in their office.

Library sensors are also recorded quite well, though it is important to emphasize that they work every day of the week (8:30 to 22:30, with Sundays being 13:00 to 20:00). As such, the weekend curve (orange) follows the same tendencies as workday curve (blue), and the stabilized curve (green) does not provide the information as it did before, as we do not have a point of reference for no people in the building. Now it shows the difference we can expect from

people entering on weekends, which is still important to differentiate and analyze.

Lastly, the FABTEST graphs are very hectic. As such I believe drawing results from them is close to useless. It is strange, as the sensors are very close to each other, but the fact that some of them were not working for a long time, especially not in tandem, produced such graphs. Maybe curating data manually could create somewhat usable results, but this will be left for the future.

Though with all these graphs it is possible to conclude, that all of them spike at 8:30 or 9:00, decreasing rapidly at 12:00 and then again increasing around 12:30 and lastly decreasing slowly around 17:00. This aligns with the typical workday start of 8:30 and lunchbreak at around 12:00, lastly ending the workday around 17:00. The drop around 17:00 is not that significant, as many people do not necessarily all go out of the lab instantaneously, but when they finish, whenever early that is for their worked, they organized. This is the case for the more reliable sensors, yet it is not the case for the professors' office. I believe it is because of the various other work they must do, like going to meetings, teaching students, and performing other work-related chores.

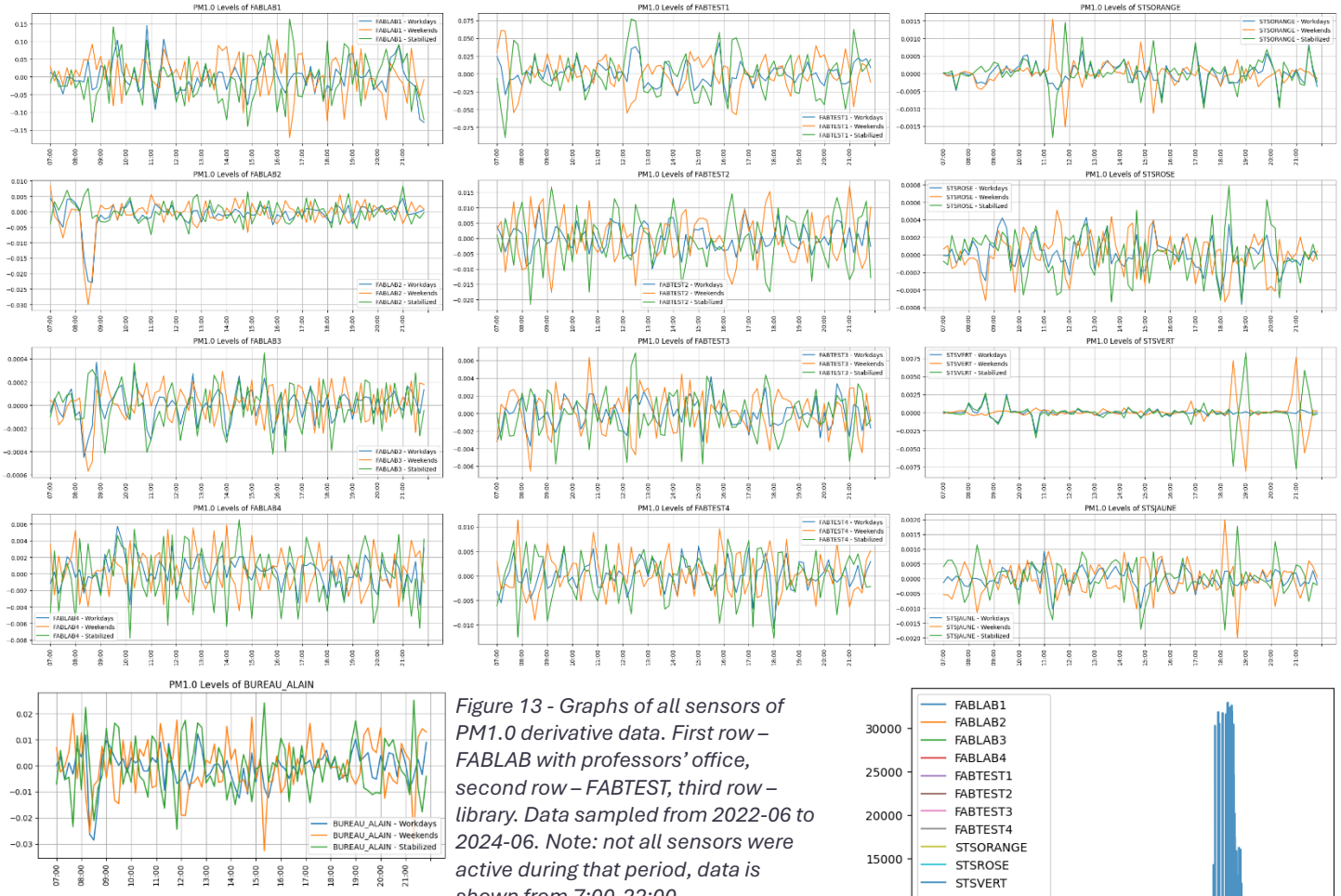


Figure 13 - Graphs of all sensors of PM1.0 derivative data. First row – FABLAB with professors' office, second row – FABTEST, third row – library. Data sampled from 2022-06 to 2024-06. Note: not all sensors were active during that period, data is shown from 7:00-22:00.

Figure 14 - Raw data of PM1.0 of all sensors. Notice the different range of values produced by each sensor.

Unfortunately, it cannot be applied to all parameters. For example, *Figure 13* shows a graph of PM1.0. The graphs of the derivatives are too noisy to conclude any reasonable guesses. This can be for many reasons, but I believe the main one is that the quality of the sensors is not the best.

It can also be the case that not a lot of machines produce noticeable data to analyze a significant change. The only possible result from *Figure 13* could be gathering the maximum and minimum of the graph, and alerting, whenever that threshold is passed. With this, a safeguard of natural noise can be obfuscated, and the real problematic rise of PM1.0 can be highlighted. As such, the value is not the only thing we can rely on. This choice is obvious after discovering in *Figure 14* that the values of sensors are wildly different from each other.

Figure 15 – FABLAB1 sensor data across all parameters.
Note: NH3 is excluded as it did not record any data besides values of 0.

Lastly, I want to examine all parameters of 1 sensor in particular, to understand recorded parameter quality. As my project focuses on FABLAB mainly, I chose its first sensor. This is because this sensor has recorded most entries. This sensor also recorded the widest range of values, as seen in *Figure 14*. As can be seen in the figure, this sensor needs calibration of values the most, which is once again confirmed in *Figure 8*, as over the years the value of the sensor has shifted, yet during a workday remains around the same shape.

Figure 15 showcases CO2, temperature, humidity, VOC, C2H5OH, CO, NO2, sound, light, pm1.0, pm2.5, pm10 in that order. NH3 results were not provided, as only a few sensors picked any kind of data, and drew a flat line on derivative graph.

From these parameters, it is important to distinguish which could be useful. The graphs for these parameters came out the best: CO2, temperature, humidity, VOC, CO, and light. These results are acceptable: C2H5OH, NO2. These are insufficient or applicable only in other ways: sound, PM1.0, PM2.5, PM10, NH3.

To analyze, it is important to distinguish that temperature, humidity, and light are environmental sensors as opposed to air quality. As such, they record data differently, and as such can belong to one group that is being well observed. Surprisingly, sound data, which is in the same category did not yield positive results, as its graph is too noisy and imprecise to draw any conclusions.

A strange observation appears in the light graph. The yellow line being data for weekends shows a neutral change during the day. This could be because the sunlight from the sun does not reach and affects the sensor reading, or the window shuts are closed during weekends. This reading can be expanded to study whether there is a strong enough correlation between light and a person being in the room. With this information data could be analyzed more meticulously.

CO2, VOC, and CO are important molecules in today's world, and as such the sensors for them developed are advanced and highly sensitive. The graph in the figure showcases that. With the gathered graphs it is very possible to develop code to track whether these particles have exceeded the established limit or

not. As mentioned before, this is one of the reasons CO2 was chosen to perform all the analysis first, as its data is of a higher rating.

The data regarding C2H5OH and NO2 presents a challenge. While the noise in the graph is minimal, the data itself appears unusual. Specifically, both parameters exhibit a rapid decrease around 7:30 and a corresponding increase around 20:00. Although I believe that a phenomenon akin to that shown in *Figure 6* may be influencing these patterns, other graphs do not follow this trend. Still, with this question answered the graph should present itself as usable for cross-referencing.

Lastly, parameters for sound and all particulate matter (PM) are too noisy to conclude. As discussed before, the only usage right now could be gathering the maximum and minimum of the graph, and alerting, whenever that threshold is passed. As for the sound particularly, it is important to distinguish that this sensor might be the most erratic one in its nature. It is easy to envision a conversation starting and ending rapidly, machines letting an alert for a minute and so on. As this is the case, I believe it would be irrational to use its derivative data as is. A safeguard minimum could be employed to filter out natural noises and examine the data for people conversing. If it is alerted, it can be an important indicator to find if there is a person or not. However, the nature of its short-lived signal can be tricky and misleading if machines are the ones producing a similar value of noise. As particulate matter was discussed, and NH3 did not capture a significant number of values, these are the only graphs I believe to be of rare usage.

Conclusion

In summary, the statistical analysis of air quality within Angers University has highlighted several aspects of data exploration using sensors. Despite the challenges with sensor calibration and data noise, the derivative analysis proved effective in extracting significant patterns from the data collected over two years.

The analysis of various parameters, notably CO2, VOC, and CO, demonstrated the capability of these sensors to provide valuable insights into air quality trends. Conversely, there were other parameters, like particulate matter, that showed significant noise and did not yield the results that were expected.

The project successfully identified the strengths and limitations of currently employed sensors, emphasizing the need for various statistical tools to use to understand the gathered data.

Ultimately, this research helps improve indoor air quality monitoring, laying the groundwork for future studies and the development of more reliable, accurate, and user-friendly air quality monitoring systems.

Personal insights and acknowledgement

To describe my feelings on the project, I believe I achieved what I set out – to explore the field of statistics through the coding language Python. There were problems, but with them being properly dealt with provided a valuable learning opportunity. While I hoped to employ other statistical methods to see various results, such as PCA, focusing on the derivative analysis was deeply satisfying and allowed for a thorough understanding of this approach.

My internship experience has been incredibly fulfilling, largely due to the support and guidance from my supervisors. They were ready to help whenever I needed it, which made a significant difference. I'm deeply grateful for the opportunity to work on this project and to contribute to the field of air quality analysis. I am very thankful for the opportunity to have had the chance to study and complete this project on statistical analysis of air quality data.