

Article

Data fusion and visualization of a multi-sensor personal exposure campaign

Rok Novak ^{1,2,*}, Ioannis Petridis ³, David Kocman ¹, Johanna Amalia Robinson ^{1,2}, Tjaša Kanduč ¹, Dimitris Chapizanis ³, Spyros Karakitsios ^{3,4}, Benjamin Flueckiger ^{6,7}, Danielle Vienneau ^{6,7}, Ondřej Mikeš ⁸, Céline Degrendele ^{8,9}, Ondřej Sánka ⁸, Saul García Dos Santos-Alves ¹⁰, Thomas Maggos ¹¹, Demetra Pardali ¹¹, Asimina Stamatelopoulou ¹¹, Dikaia Saraga ¹¹, Marco G. Persico ^{5,12}, Jaideep Visave ^{5,12}, Alberto Gotti ¹², Dimosthenis Sarigiannis ^{3,4,5}

¹ Jožef Stefan Institute, Department of Environmental Sciences, Ljubljana, Slovenia

² Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

³ Environmental Engineering Laboratory, Department of Chemical Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁴ HERACLES Research Centre on the Exposome and Health, Center for Interdisciplinary Research and Innovation, Thessaloniki, Greece

⁵ University School of Advanced Study IUSS, Pavia, Italy

⁶ Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel, Switzerland

⁷ University of Basel, Basel, Switzerland

⁸ Masaryk University, Faculty of Science, RECETOX, Brno, Czech Republic

⁹ Now at Aix-Marseille University, CNRS, LCE, 13003 Marseille, France

¹⁰ Institute of Health Carlos III, Instituto de Salud Carlos III (ISCIII), Department of Atmospheric Pollution, Madrid, Spain.

¹¹ Atmospheric Chemistry and Innovative Technologies Laboratory, INRASTES, NCSR "Demokritos", Aghia Paraskevi, 15310 Athens, Greece.

¹² Eucentre Foundation, Via A. Ferrata, 1, 27100, Pavia, Italy

* Correspondence: rok.novak@ijs.si

Abstract: Use of a multi-sensor approach can provide citizens a holistic insight in the air quality in their immediate surroundings and assessment of personal exposure to urban stressors. Our work, as part of the ICARUS H2020 project, which included over 600 participants from 7 European cities, discusses data fusion and harmonization on a diverse set of multi-sensor data streams to provide a comprehensive and understandable report for participants, and offers possible solutions and improvements. Harmonizing the data streams identified issues with the used devices and protocols, such as non-uniform timestamps, data gaps, difficult data retrieval from commercial devices, and coarse activity data logging. Our process of data fusion and harmonization allowed us to automate the process of generating visualizations and reports and consequently provide each participant with a detailed individualized report. Results showed that a key solution was to streamline the code and speed up the process, which necessitated certain compromises in visualizing the data. A thought-out process of data fusion and harmonization on a diverse set of multi-sensor data streams considerably improved the quality and quantity of data that a research participant receives. Though automatization accelerated the production of the reports considerably, manual structured double checks are strongly recommended.

Keywords: data fusion, multi-sensor, data visualization, data treatment, participant reports, air quality, exposure assessment

1. Introduction

Health impacts of poor air quality have become a central point of discussion in policy development and in personal exposure studies [1]–[3]. A widening selection of low-cost sensors (LCS) that measure environmental conditions allows individuals to collect data about their own living environment and estimate their exposure to different stressors [4]–[6]. Several issues remain regarding data loss, bulkiness, design, power consumption, [7], unreliable and (unintentionally) misleading data, lack of quality control, validation and calibration [8] and user experience [9]. Providing meaningful information to individuals about their environment and stressors present in their lives is in line with the United Nations Sustainable Development Goals (SDGs), which call for participatory, integrated and sustainable human settlement planning (Target 11.3 [10]), which can only be achieved if the public is well informed. Several goals and targets in the SDGs are assessed based on the “Mean urban air pollution of particulate matter (PM) of different sizes” indicator [11]. Considering the usually low spatial resolution of PM measurements at a city level and that they only sample outdoor air pollution, the use of individual low-cost PM sensors could be useful in estimating human exposure to PM.

Airborne particulate matter concentration is only one facet of air quality, and when assessing the impact of air quality on human health, pollutants such as nitrogen dioxide (NO₂) [12], ozone (O₃) [13] and volatile organic compounds (VOC) [14], [15] should be considered. Carbon dioxide (CO₂) is not considered a pollutant, though elevated concentrations indoors can pose health risks [16]. Fusing data from different low-cost sensors has been employed in several cases, when supplementing existing data sets from environmental monitoring networks with high-resolution spatiotemporal measurements from LCS [17], [18], by using mobile LCS for air quality mapping in combination with dispersion model calculations [19] or by using stationary data with transport model results [20].

On the other hand, an increase of devices with very diverse input parameters and data collection protocols poses some unique data fusion and visualization challenges, including non-standard timestamps, data gaps, different classifications, a multitude of data logging processes, etc. While LCS generally provide a larger quantity of data, there is a lack of data on comparability from one device to another. Steps can be taken to better comprehend this prior to using the sensor, such as providing more metadata and insight into how data is recorded and presented [21]. The reliability and accuracy of LCS can come into question which necessitates a validation/calibration. This process is not standardized and can vary from device to device. The results are usually presented using the correlation coefficient, root mean square error and mean absolute error, which are useful, but have to be accompanied with information regarding the conditions under which the validation/calibration was performed [22], which can in turn make the process of data fusion and visualization more straightforward.

To facilitate data fusion and visualization, where one of the goals is to provide meaningful information to participants, there should be greater focus on assessing the characteristics of the sensor itself, providing more context and associated uncertainties (where available) [21]. A benefit of participatory approaches, where citizens use LCS is the ability to gain additional (qualitative) information from the user through interviews or surveys that could be implemented in their smartphones [23] about specific environmental conditions or how the sensors function and provide the data.

Visualizing data for lay end-users requires a balance between providing the largest amount of data in an understandable way to individuals who are not accustomed to using plots and figures in their daily lives. Selecting the proper type of visualization can have a meaningful impact on the perception of the end-user and the information that they are able to extract [24], [25], and promote better risk assessment and reduction of exposure by personal decision-making [26]. An improvement, which is already being employed in some visualization efforts, is to allow users to interact with the final data set and make their own adjustments [27].

This work is a part of the Integrated Climate forcing and Air Pollution Reduction in Urban Systems (ICARUS) H2020 project, which applied integrated tools and strategies for urban impact assessment in support of air quality and climate change governance [28], [29]. One of the central parts of the project was to develop a methodology for estimating personal exposure using air quality sensors, personal activity and GPS tracking. For this purpose, about 100 participants were recruited in seven European cities – Ljubljana, Athens, Thessaloniki, Brno, Milano, Madrid and Basel each, and providing them with all the tools to collect the necessary data [30], [31]. For this, an algorithm was developed that would clean, fuse and visualize the collected data and present it to the participants within a straightforward and understandable report. This “final report” for the participants was automatically generated in the respective local language, and included as much data as possible without making the report too long and complicated. The report aimed to give the participant enough details for them to discern relevant information related to air quality, about their living environment, and their behaviour to eventually promote more environmental conscious lifestyles.

Addressing the above-mentioned challenges, the objectives of this study are to provide information on:

- collecting multi-sensor and multi-parameter data flows,
- aggregation and visualization of data,
- compilation and communication of final report for the participants,
- lessons learnt from this data fusion exercise and recommendations for future studies.

2. Materials and Methods

Input data for data fusion and visualization was obtained from three sensor devices and other data points collected through questionnaires for households and individuals, and Time Activity Diaries (TADs) used within the ICARUS campaign. Two of the sensor devices were commercial: a Smart Activity Tracker (SAT) and an Indoor Air Quality (IAQ) sensing station. The third one was specifically constructed for the purposes of the research project, using the Arduino platform. A schematic representation of the devices and protocols used is shown in Figure 1. A detailed description of the campaign and its goals can be found in Robinson et al. [32]. All data cleaning, harmonization, fusion, visualization, and report compilation and output were done in R [33] with support from different R packages, e.g., ggplot2 [34], dplyr [35], knitr [36], rmarkdown [37] and others.

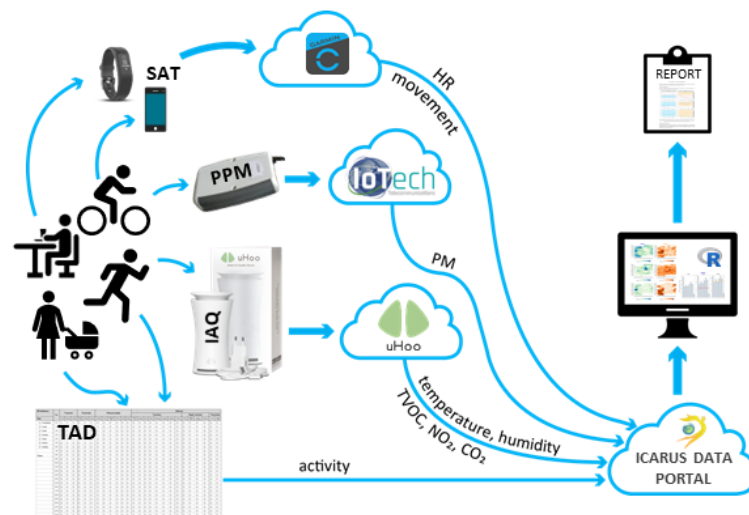


Figure 1: Schematic representation of data collection devices and protocols, transfer paths, aggregation, visualization and delivery protocols

2.1. PPM data

The Personal Particulate Matter (PPM) sensing device was designed and compiled for the purposes of the sampling campaign of the ICARUS project [38]. Apart from providing PM concentration data in three class sizes ($<1\ \mu\text{m}$ (PM_{10}), $<2.5\ \mu\text{m}$ ($\text{PM}_{2.5}$), and $<10\ \mu\text{m}$ (PM_{10})), it also provided ambient temperature and relative humidity data, and GPS/location coordinates (including speed and altitude). As the device did not have a Real-time clock (RTC) module (e.g. [39]), the timestamp was obtained by connecting it to an online server via a SIM card. Without this connection, the device did not provide data with accurate timestamps, which in turn produced several data gaps. Timestamp logging was irregular and inconsistent as evident in an example of the data set in section A of Supplementary Data (SD-A).

2.2. SAT data

A commercial SAT was used (Vivosmart 3, manufactured by Garmin [40]) to collect heart rate and movement data. The device provided several data points about physical activity and movement with minute resolution, e.g., average heart rate, stress level, sleep status, calories burned, etc. As the export of data is not freely available through the Garmin interface, an additional connection between a dedicated ICARUS data portal and the Garmin Connect portal was established to transfer the data. The SAT data had very few gaps (excluding the time while the device was charging). Some issues were present when the user did not fasten the belt tight enough, which in turn meant that the device could not record the heart rate data.

These data were not separately visualized as the focus of this research was to provide meaningful insight into the relationship between air quality and heart rate/activity. To give a brief overview of the data a summary in the form of a table was included in the final report.

2.3. IAQ data

A “uHoo Smart Indoor Air Quality sensor” (IAQ) [41], a stationary device with multiple sensors, was used in every household. At every full minute, the IAQ provided nine data points: temperature, relative humidity, CO_2 , total VOCs (TVOC), $\text{PM}_{2.5}$, NO_2 , carbon monoxide (CO), ozone (O_3) and air pressure.

Visualization of AQ parameters measured by the IAQ was limited to three parameters (CO₂, NO₂ and TVOC) that showed the best performance during the collocation experiments with validated devices, and other tests. As offsets were observed for some sensors during these experiments, visualizing this specific dataset was made using a heatmap, focusing on relative changes of each variable over time. A heatmap, in this case, consists of tiles which are colored relatively to all other tiles (lower values are lighter, higher values are darker), as implemented in Mahajan, et al. [42]. Using minute values would create a heat-map with small tiles, which would obscure the relative differences within a day. To counteract this, hourly values were calculated and used in the heat-map, reducing the number of tiles from ~10.000 to ~170.

2.4. ICARUS data portal

A dedicated data portal was constructed for the purposes of the ICARUS2020 project, and a Decision Support System (DSS) with it, which collected, compiled and stored the data. The DSS additionally had a presentation tier with a user interface and a logic tier that stored the computational models and handled their execution [43]. In this study, the data portal was mainly utilized to store and obtain the PPM and SAT data in a uniform format, which allowed further manipulation and fusion of data.

2.5. TAD data

A key data input were the TADs, which allowed the participant to input their activity, location, means of transport and other variables, for each hour of the day. This data was collected from each participant, for seven days in two seasons, for all cities, accumulating up to ~10.000 TADs.

There were two approaches to filling in the TADs, one was to select only one option for each hour, the other was to allow participants to select multiple options. The latter would seem as a more appropriate approach from a participant standpoint, as individuals mostly perform several activities within an hour. This posed a unique challenge in selecting which data point to use, which activity was more relevant or more characteristic for each hour.

Some manual corrections of the data were necessary in the final stages when it was observed that the participants filled in the wrong activity. As this information was not double-checked with the participants, the corrected activities were only the most obvious ones, e.g., checking if a non-smoking person truly smoked in just one instance the entire period.

Because the data for activities was for hourly values and the sensor data had minute resolution, the former was repeated 60 times for every hour, which proved to be a major issue when calculating averages and trying to discern if there were meaningful differences between activities [44]. However, the goal was to include as much data as possible in the final dataset.

The TAD dataset was used in three visualizations, in combination with PPM and SAT data:

- a) A scatter plot was made for every PM size class and heart rate, for both seasons. Additionally, the points were colored based on the activity at that minute, which allowed the reader to observe what activities took place at, for example, elevated levels of PM or elevated heart rate. Only the activities which the participant filled in were shown in the legend.
- b) A similar scatter plot as in a) was constructed, with an additional layer which showed vertical bands or ribbons of different colors corresponding with the participant's location and transports. As this added another layer of

complexity to the visualization, the decision was made to provide these plots only to specific individuals who expressed interest. Though activity information was missing in several TADs, the location and transport data was logged for almost the entire period of observation (for most participants). Consequently, participants could associate specific means of transport with elevated levels of PM, and corresponding activities with higher heart rate.

- c) The third plot showed the average weekly PM values for each activity. Six plots were constructed, three per season, one for each PM size class.

TAD data was not used in combination with IAQ data due to higher uncertainty associated with absolute values of CO₂, NO₂ and TVOC.

2.6. Final report compilation and production

The generation of final reports for participants was performed in three phases:

- a) Generation of plots as described in points 2.1. – 2.4., which was iterated over all the participants. These plots were saved locally in a jpeg format and labelled according to each participant ID.
- b) They were integrated in a markdown script which was constructed with several parts that were specific for each participant. The customization of each report was designated in an Excel file. Each participant had a custom greeting with their name and gender appropriate pronoun. All the plots and other graphics were inserted using the `include_graphics` function in the knitr package.
- c) Finally, the script was iterated over all participants in a separate script to allow some further customizations. Some participants had additional visualizations, while others had some omitted due to missing data. After all the reports were generated in different languages, they were manually checked for errors by local organizers in each participating city and distributed to all the participants.

In addition to the technical construction and production of the final report, the participants feedback and wishes for visualization were considered where appropriate [45].

2.7. Temporal resolution and data treatment

Minute resolution of data was deemed as sufficient to provide enough detail of PM concentrations and exposure. The SAT and IAQ also logged data with minute resolution, though these logs were at every full minute, while the PPM logged the measurements at different fractions of the minute. These were later rounded to the nearest minute.

To compare the PM data with WHO guideline limits the minute resolution data were aggregated to daily means. More uncertainty was associated with PM daily means from the first and last day the participant was involved in the campaign, as the participants did not collect data for the entire 24-hour period.

An outlier correction was made for the PM data, where all values above 180 µg/m³ were converted to 180 µg/m³, based on the maximum values provided by “Air quality in Europe” as part of CITEAIR and CITEAIR II projects [46]. This approach was used only for visualizing the data and providing the final reports to participants for clearer data representation. Some datapoints showed uncharacteristically high concentrations of PM (>3000 µg/m³), which was deemed as sensor error and subsequently corrected with the aforementioned procedure.

The PPM showed good agreement of absolute values with a reference research-grade device, a GRIMM Model 11-A, increasingly so with larger time averaging intervals [38].

3. Results and discussion

3.1. A merged dataset

The final merged dataset had 93 columns. Due to sensor failures, data gaps, incorrect TAD filling, etc., there were several instances of empty columns or in some cases completely empty datasets. This was appropriately labelled in the final reports.

SD-B presents an example of the final data set, with all the data harmonized to 1-minute resolution. Each data set includes

- specific characteristics for each participant (age and gender),
- PPM data (PM values, temperature, humidity, battery charge level, location coordinates, speed and altitude),
- SAT data (where several columns proved to be somewhat redundant and were therefore removed),
- IAQ data (which proved to be the most user-friendly as it had a correct timestamp for each recorded value, almost no missing values and a simple interface to download the data), and
- TAD data, presented the same way as they were recorded on the physical paper sheets: location of the participant (home, office, indoor, outdoor), transport data (bus, car, foot, etc.), indoor and outdoor activities (cooking, smoking, sports, etc.) and some specific conditions for the indoor space the participant was in (burning candle or fireplace, open windows and/or AC turned on).

3.2. Visualizing the data

All the visualizations are presented and described here as they were shown in the final report for participants and are collected from different reports.

Figure 2 shows temperature, relative humidity and air pressure during both seasons (IAQ data). Meteorological data showed the highest accuracy when compared to reference instruments and was in turn presented with absolute values. Although the ribbons show “optimal conditions” as per the General health and comfort guidelines (modified for the appropriate climate) [36], this information is somewhat subjective and can differ from person to person. As shown in the example in **Figure 2**, this person had very similar indoor temperatures in both seasons, and even though the summer values are mostly outside the “optimal zone”, one could argue that a constant temperature throughout the year provides more comfort to certain individuals.

Arranging the individual plots in columns according to their season makes comparisons between the seasons easier.

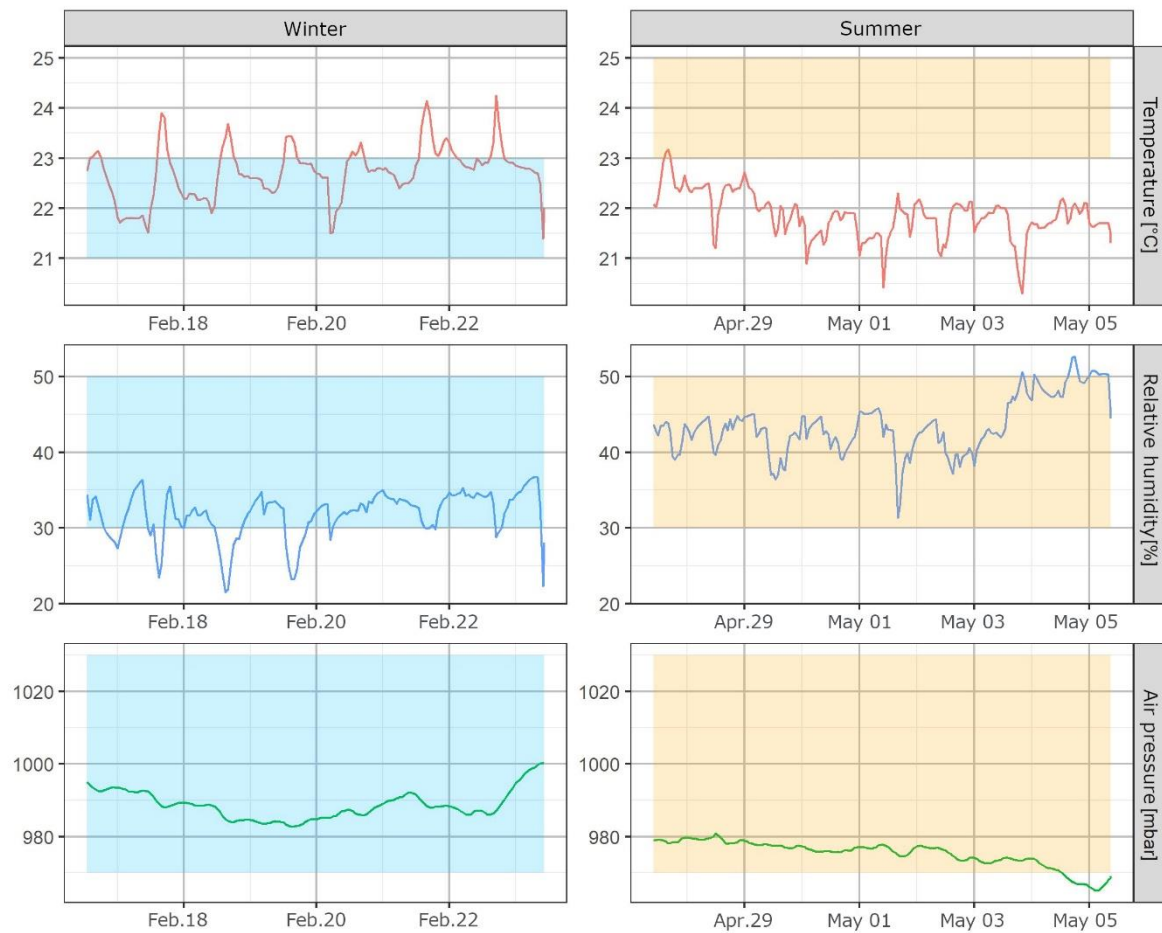


Figure 2: Faceted plots with meteorological variables - temperature, relative humidity and air pressure; data from IAQ

Figure 3 shows an example of the compiled visualizations of CO₂, NO₂ and TVOC for this particular household. In our analysis these parameters have shown to follow expected trends, e.g., decreased values of CO₂ when opening a window and in turn increasing the NO₂ values if it was in a high traffic area [47], as seen in Figure 3 on Tuesday, 19th of February 2019 at around 13:00 when CO₂ concentrations quickly fell and NO₂ increased rapidly at the same time. The plots allow an intuitive way of observing changes in these parameters, which can be more relevant to each specific household. Collocation with a reference device has shown that the absolute values of these parameters were not accurate enough to present to participants at that time [48], though newer research shows moderate to high correlation with reference instruments in laboratory conditions [49]. These relative values still give participants an insight into their indoor air quality and possible correlations with external factors such as traffic.

The setting of the visualization allows the reader to compare trends between seasons and between pollutants. Mostly, the main conclusions can only be made by the participant themselves, as only they know all the detailed activities and conditions in their household. For example, higher TVOC values during the evening and night could indicate poor ventilation in combination with a specific activity that raises the concentrations, such as cooking or smoking [50]. By putting these plots in the same figure, they can immediately observe the trends in the other two parameters and come to some conclusions.

Each date is also labelled with a language specific day of the week to facilitate better observation of specific trends.

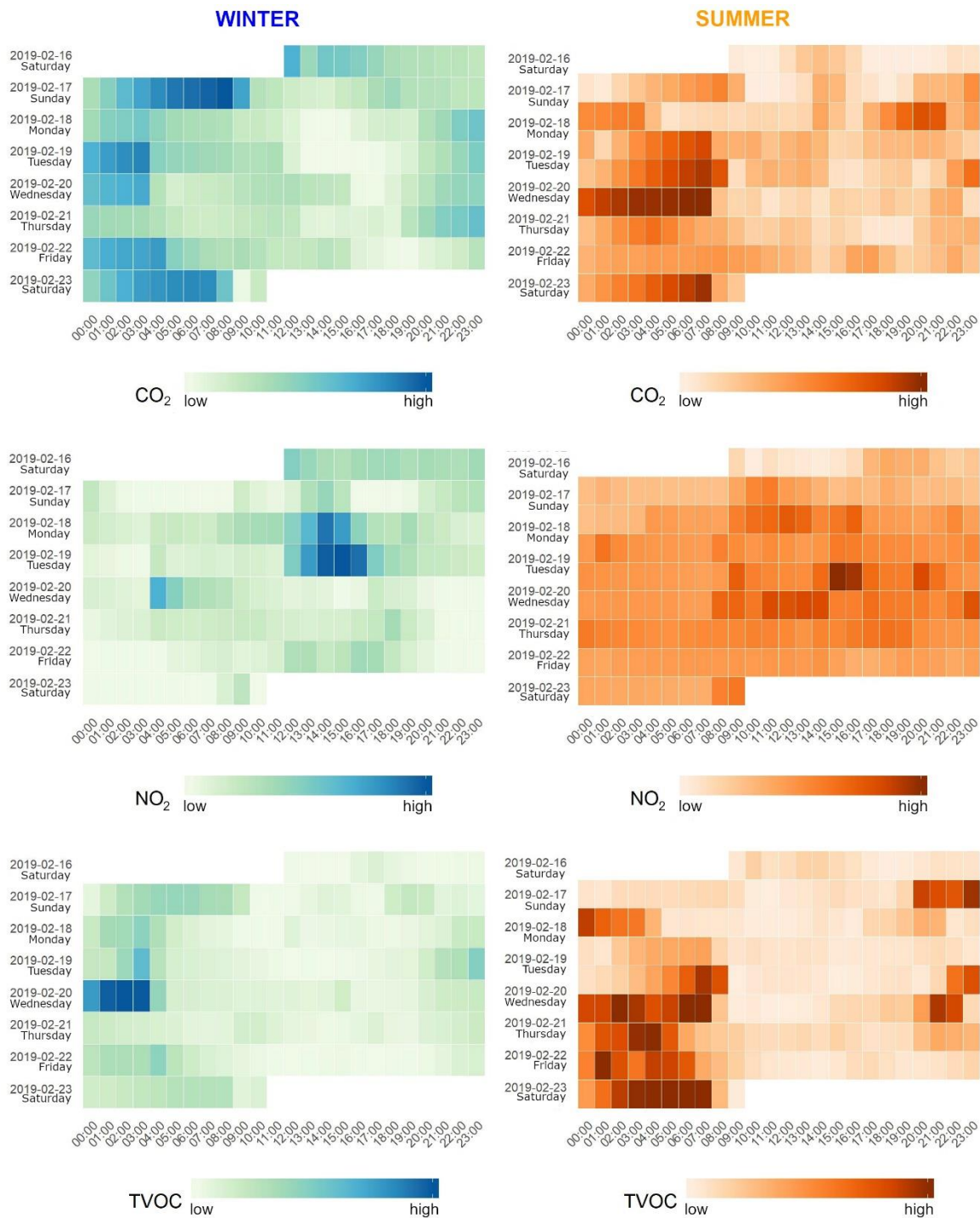


Figure 3: Faceted heatmaps of three pollutants (CO₂, NO₂, TVOC); data from IAQ

Figure 4 presents the PM concentrations, heart rate and designated activities for each minute in the time-span the participants was involved in the data collection. Shown here are only the specific activities, there is no additional information about the location of the participant, their transport or specific conditions in the household. Not including this information makes the visualizations less crowded and more likely easily readable and understandable, which was subsequently further explored in structured focus groups [45]. All the

values are also plotted with exact concentrations, due to the fact that the PPM device showed fairly accurate results compared to reference devices.

The participants could deduce by themselves some correlations and extra information from the plots, e.g., higher heart rate when running, dips during the night, a specific time of day when the PM concentrations are elevated and if they are correlated with a specific activity like smoking or cooking, etc. Again, the interpretation of the plots is mostly on the participants themselves, because they have a more complete overview of their surrounding and activities at that exact moment.

No particular difficulties were associated with constructing this visualization, with the possible exception of some alterations to the color scale and legend to also include the activities that the participant did not perform.

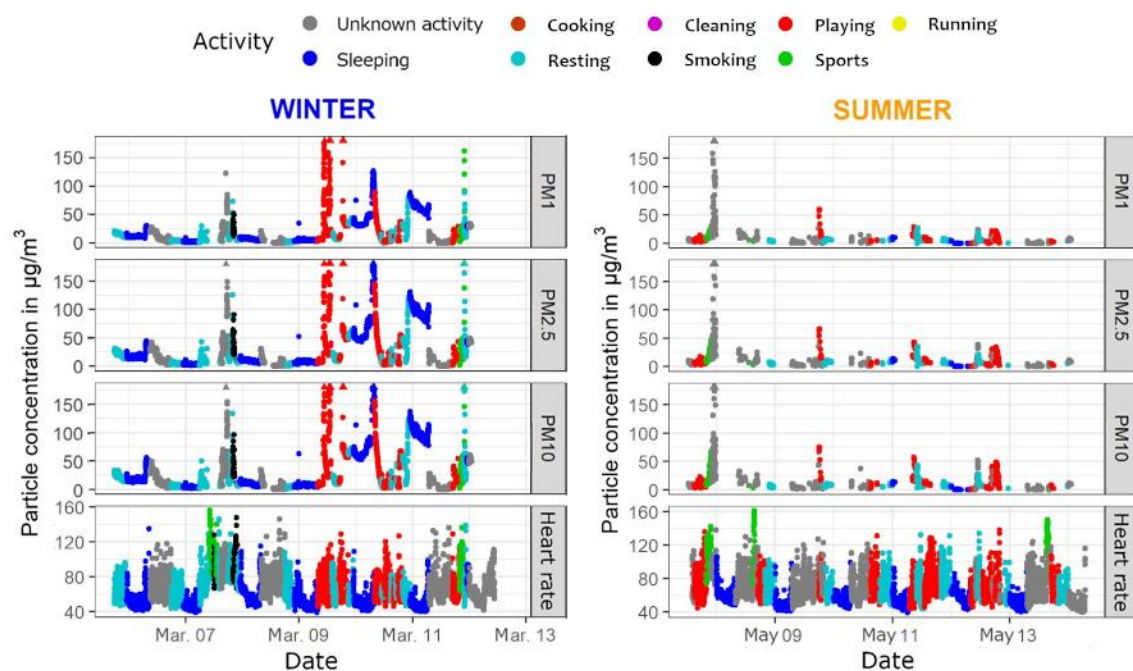


Figure 4: Three size classes of PM (PM₁, PM_{2.5}, PM₁₀) and heart rate values for both seasons with each point colored according to the associated activity; data from PPM, SAT and TAD

An additional figure was created to include location and transport of the participants in addition to PM and activity values. As Figure 5 has more information and would otherwise be illegible the plots were enlarged to be the size of a whole A4 page.

Several difficulties were encountered while constructing these plots, as the ribbons that show each activity had to have a start and end time to each location/transport as every interval. Another approach was considered to only include a vertical line at each minute in the color of the location/transport but it considerably increased the time to produce each plot. As there were hundreds of plots to construct, it was necessary to find a more time efficient process. An additional section of code was implemented to construct a separate data frame which had a start and end time with a label for each location/transport. This was used in the ggplot2 geom_rect function while compiling the plot and noticeably reduced the time it took to compile each plot.

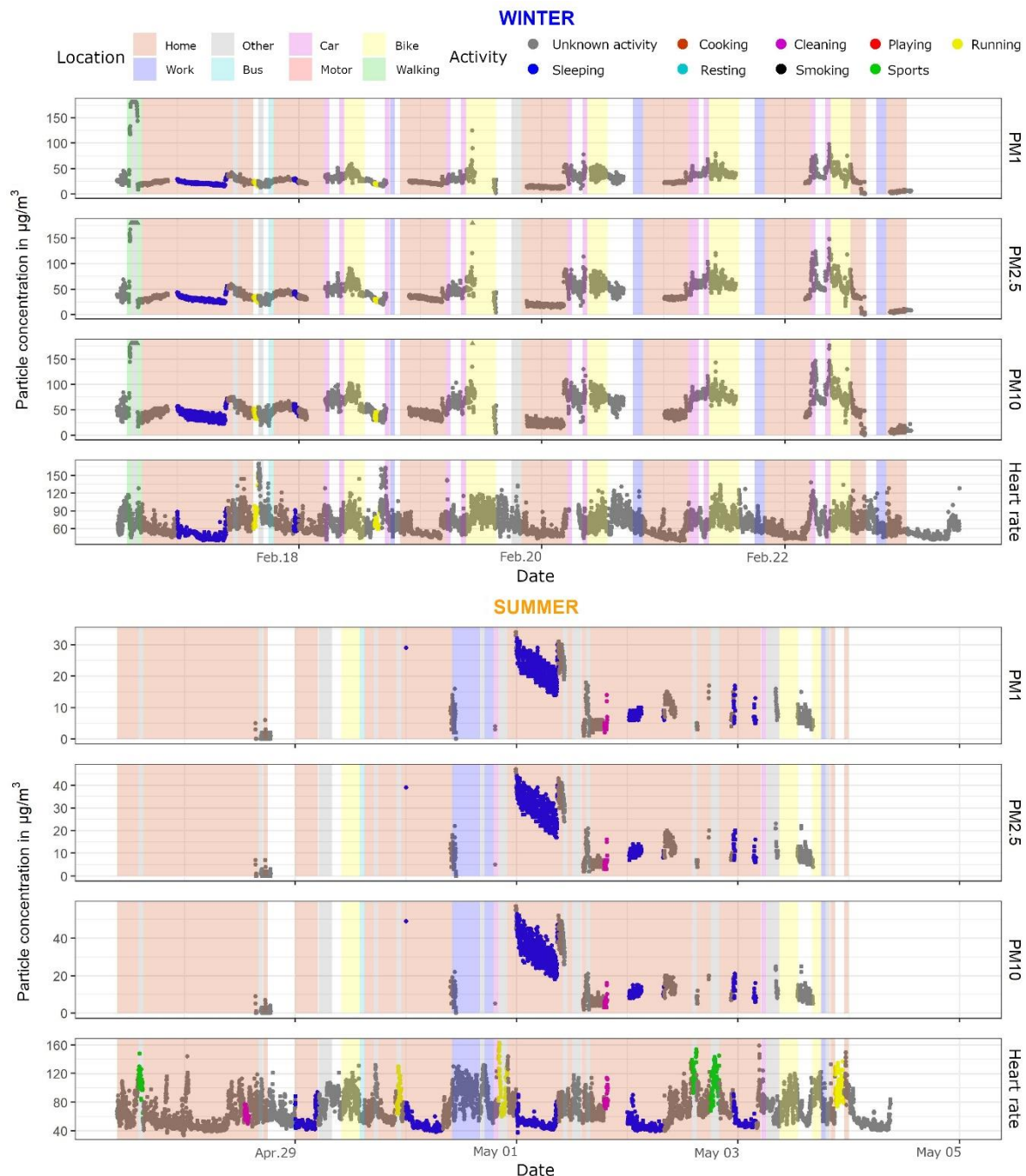


Figure 5: Three size classes of PM and heart rate values for both seasons with each point colored according to the associated activity and each ribbon representing a location or means of transport for that tie period; data from PPM, SAT and TAD

Figure 6 shows the daily average PM concentrations for both seasons and is the only set of plots where guidelines or recommended values could be inserted. The WHO and the EU do not have minute or hourly guidelines for concentrations of PM, though studies show that short term exposure to elevated levels can have adverse effects on health [51], [52]. The WHO does provide daily guidelines for PM_{2.5} and PM₁₀, which are 25 $\mu\text{g}/\text{m}^3$ and 50 $\mu\text{g}/\text{m}^3$, respectively [53], revised in 2021 to 15 $\mu\text{g}/\text{m}^3$ and 45 $\mu\text{g}/\text{m}^3$, respectively [54].

Mainly there are two important pieces of information in these plots, they allow the participant to observe (1) the inter-seasonal differences and (2) the day-to-day differences, while also having the information about a specific size class of

PM. This mainly shows that the concentrations are generally higher in winter time (more indoor activities, weather patterns that trap pollution in low-lying areas, combustion of solid fuels, more use of car/buses in contrast with cycling/walking, etc.), and even when there are elevated levels of PM during the summer, they are still much lower than in winter time. The participant can also observe that some particular days have elevated levels of PM which could be associated with some specific activities performed that day (or weather patterns).

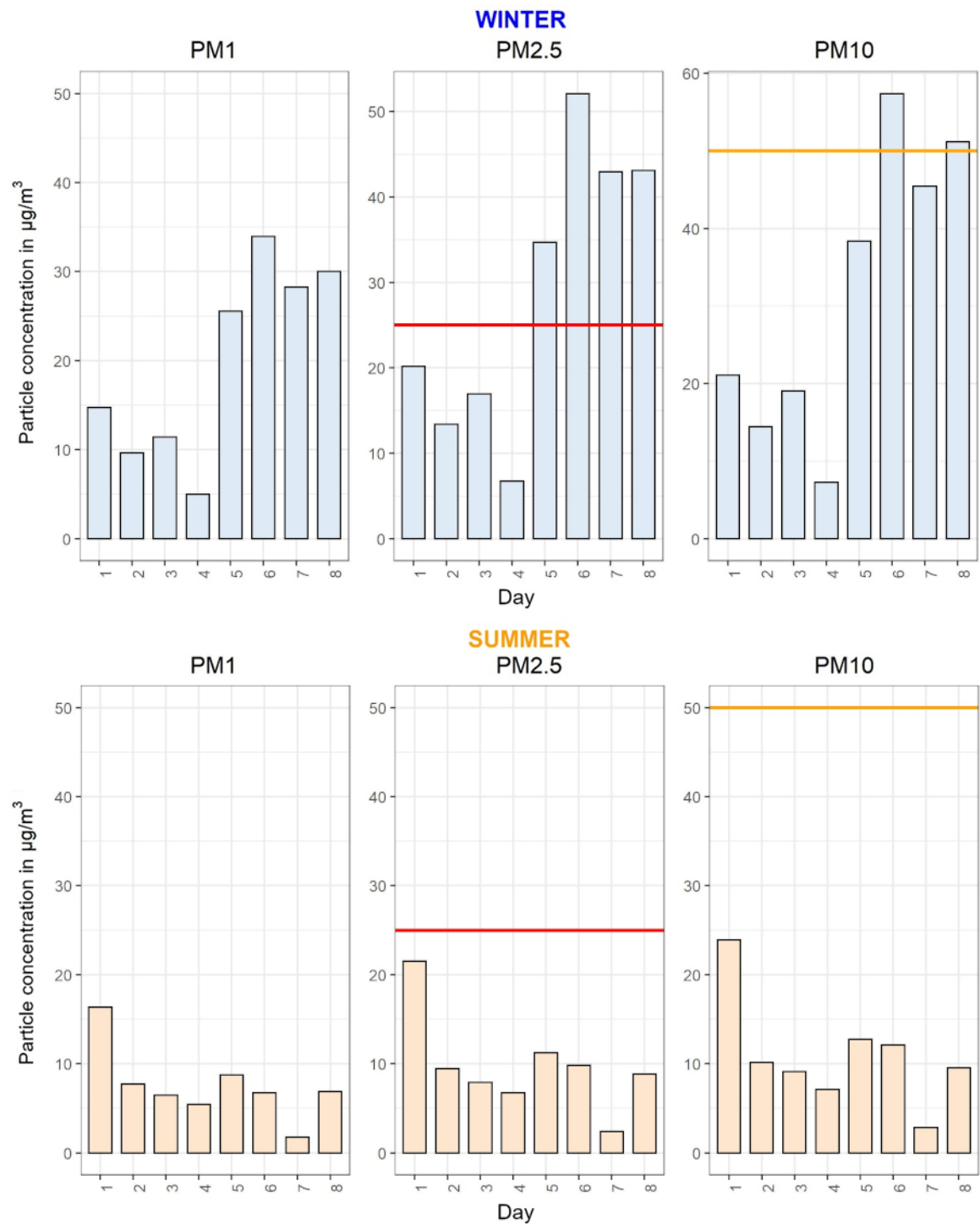


Figure 6: Faceted plots of average daily concentrations of three size classes of PM for each season, with WHO guidelines; data from PPM.

Figure 7 presents two tables which show the average values for each SAT variable for each day. No additional visualizations were made for the SAT data (apart from the heart rate plots in Figure 4 and Figure 5). There were several visualizations already available on the Garmin connect portal for each variable.

Day	Steps	Stress level	Average heart rate	Max heart rate	Sleep[h]	Kcal
1	4922	NaN	65	128	0	120
2	11702	25	66	170	9	389
3	17683	43	69	163	5	346
4	10504	26	69	142	7	194
5	8481	22	67	131	7	158
6	10871	17	67	134	6	189
7	10589	22	67	138	5	185
8	1244	8	52	128	8	7

Day	Steps	Stress level	Average heart rate	Max heart rate	Sleep[h]	Kcal
1	6240	7	63	148	0	186
2	8000	20	60	144	8	172
3	19377	16	70	132	4	312
4	16041	28	73	163	7	488
5	17717	14	63	118	8	407
6	19561	25	69	154	7	592
7	30675	26	74	159	4	443
8	520	12	52	102	7	9

Figure 7: Aggregated data from SAT

Figure 8 shows the average PM values for each activity that was indicated by the participant in the TADs. There are certain shortcomings to this visualization as it does not provide any data about the number of instances for each activity, e.g., there can only be one hour of smoking in the entire week, but 50 hours of sleeping. Although the caption under the plots clearly states that the empty columns mean that there were no recorded instances of that specific activity, there can still be some confusion where the reader could assume that the average concentration is 0 $\mu\text{g}/\text{m}^3$.

Primarily this plot should communicate differences between the activities in each respective reason. In the example provided in **Figure 8**, the PM values for smoking are higher than all other activities during the summer season, but not that different from all other activities during winter. A possible explanation would be that there is less natural air circulation during the winter (opening windows or doors), though there could be other explanations. This is another prime example where the detailed information about their surroundings would give the individual the most accurate assessment of what the source of the elevated concentrations of PM could be.

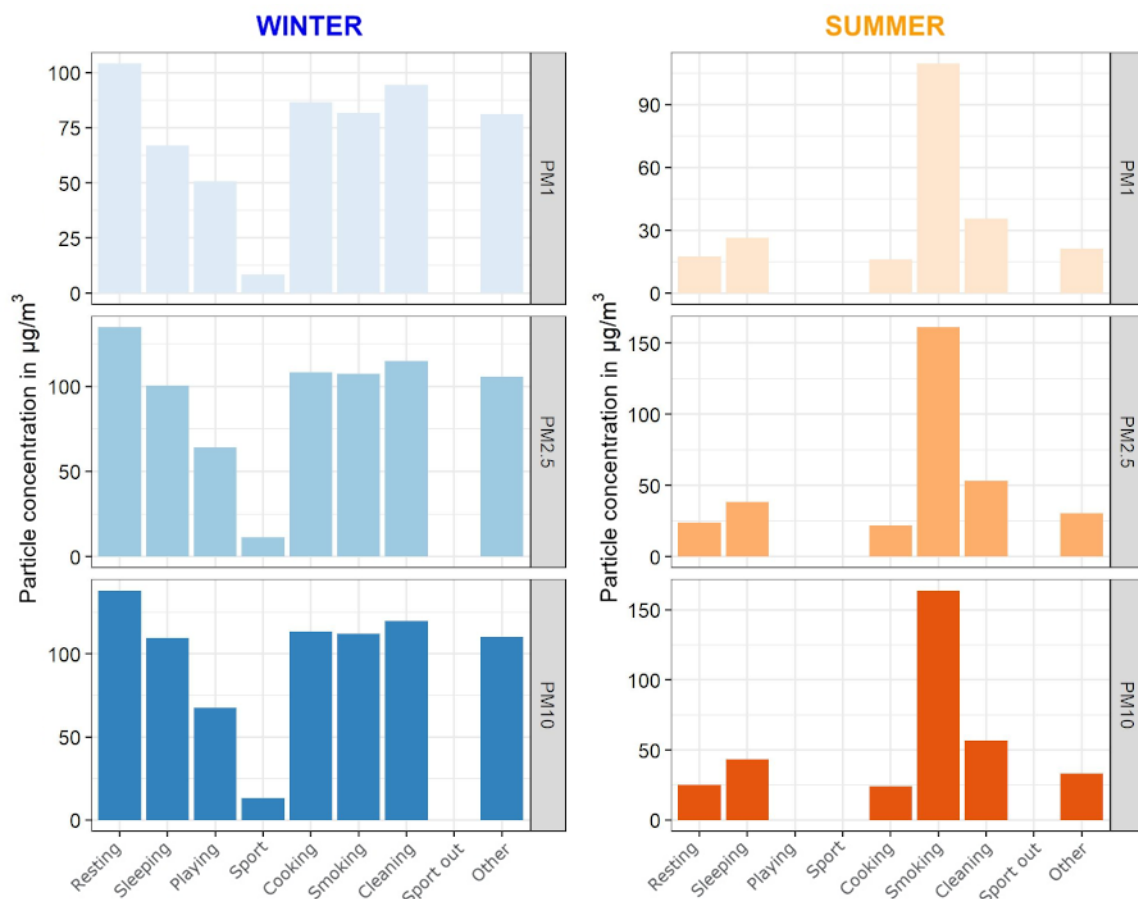


Figure 8: Faceted plots of average values of three size classes of PM (PM₁, PM_{2.5}, PM₁₀) for each specific activity and each season; data from PPM and TAD

3.3. The final report

An example of the final report that was provided to all participants is shown in section SD-C of the supplementary data. The report began with a personalized greeting and a general description of the project and what were the contents of the report. There were also disclaimers about the nature of the low-cost sensors and the uncertainty associated with them. The next page was the “Section A” of the report and had a more detailed description of the study, the devices and approaches that were used and what specifically should the reader of the report focus on. Following this was the “Section B” which described the household conditions, focusing mainly on the data from the IAQ with Figure 2 and Figure 3, accompanied with appropriate captions.

“Section C” contained the plots concerning personal exposure to air pollutants, beginning with PM data, shown in Figure 4, Figure 5 and Figure 6. Figure 5 was provided only to a handful of participants who had more recorded data and requested a more thorough overview for the entire duration of their involvement. The physical activity information, shown in Figure 7, was presented next and accompanied with a more detailed description of each variable, including some measures for low, average and elevated heart rate, to aid the reader in their interpretation of the data. Figure 8, showing the average PM values for each activity, was the last plot included, with a specific disclaimer that the scales on the y-axis are free.

Some general recommendations on “How to improve indoor air quality” were provided at the end of each report together with two tables extracted from the uHoo sensor device recommendations and descriptions [41].

3.4. *Issues faced and recommendations for future studies*

Several issues were encountered while compiling, cleaning and visualizing the data collected from LCS. While the PPM proved to be the most accurate when compared with a reference instrument, it also had the most issues regarding data gaps and inconsistent time stamps. Two relatively small improvements to the device would have made this issue much easier to deal with, due to the device being independent of the GPRS signal: (a) installing a Real-time clock (RTC) module which would provide consistent time stamps, and (b) larger internal storage and buffer that would allow the device to record PM values without a connection to the server. Several optimizations to reduce energy usage would be possible, e.g., less frequent GPS recordings while stationary, option to only upload the data when the device is charging, etc.

On the other hand, the IAQ had very consistent data streams, accurate time stamps, and a very intuitive interface. Two improvements would make the device function more independently: (a) a small internal storage for times when there was no wi-fi signal, which would allow the device to store the data in an internal buffer and upload it when the connection was re-established, and (b) possibly a small battery to allow the device to function during power outages.

The SAT was very reliable, had an internal storage capacity for 14 days of data and had a battery that lasted between 5 and 7 days. An improvement would be to provide a way to observe if the data is being logged correctly. At certain times, the device was not placed properly on the wrist or had some other errors with data logging, which could be observed only at the end of the sampling campaign. Though the SAT did provide a uniform data-set it had to be extracted by a separate process in collaboration with the company that produced the device. Accessing data from commercial devices proved to be somewhat complicated and preconditioned on setting up exclusive deals with companies. Even when the deal is set, the entire data retrieval process still relies on the company to cooperate. A preferred way to collect data would be if the raw data streams were open and access to them not preconditioned.

A key improvement for the TADs would be to allow more granular activity logging during the day, e.g. every 15 minutes. TADs could also be somewhat customized to different participants or days of the week, e.g., participants who perform only one activity, such as work, during morning hours could have a different TAD during workdays than during weekends. Hourly resolution of TADs caused some issues when presenting and visualizing data for the participants as the average values were skewed, due to the fact that most activities do not have a duration of exactly one hour and mostly don't start and end at full hours. This meant that, for example, someone went for a run for 40 minutes and after they got home, they lit a cigarette and smoked for a few minutes, but only recorded “running” for that hour. Even though the person could have checked both activities there would still be no information about in which part of that full hour “smoking” and “running” occurred. Recording activities minute by minute would be a heavy burden for participants, so future research should focus on other solutions, such as complex activity recognition using machine learning, smartphones or other tools [44], [55].

Visualizing the data proved challenging at times and required unique solutions. The main challenge in producing the plots in **Figure 2** proved to be the ribbon with “optimal values” which had to be referenced in a way that would allow this value to be presented for each individual hour, while also enabling faceting of the plots. Additional variables with minimum and maximum data for each season were introduced, which shortened the script for the final construction on the plots.

Several difficulties were encountered while constructing the plots in **Figure 5**, as the ribbons that show each activity had to have a start and end time to each location/transport as every interval. Another approach was considered to only include a vertical line at each minute in the color of the location/transport, but it considerably increased the time to produce each plot. As there were hundreds of plots to construct, it was necessary to find a more time efficient process. An additional section of code was implemented to construct a separate data frame which had a start and end time with a label for each location/transport. This was used in the ggplot2 `geom_rect` function while compiling the plot and noticeably reduced the time it took to compile each plot.

A rather easy, though important improvement for the plots in **Figure 6** would be to show specific dates and days of the week instead of the number of days since the participant joined the sampling campaign. Participants do not always remember which day they have started the campaign and would have to go back to the IAQ figure to find out. As a significant amount of time can elapse between the campaign and the distribution of reports to participants, it could be good for future studies to always indicate in the figures the date and day of the week.

Figure 8 could be improved by indicating the number of instances for each activity by coloring the bars according to a color scale reflecting the frequency of activities or by changing the width of each bar accordingly. The activities without data should be clearly marked with a symbol or a text. A requirement for a minimal amount of data should be considered to remove activities with only a few instances. The color schemes should also be intuitive, such as coloring winter blue (cold color) and summer orange (warm color) or smoking black, which instinctively guides the reader

Manually collected data from TADs was double checked by the researchers as there were some non-obvious errors, e.g., smoking selected for a person who designated that they do not smoke, which, in the end, sometimes showed that it was a user error and other times that the person does indeed smoke but very infrequently. As with any data set, these inconsistencies and all permutations can be very time consuming to implement into the report generating code. A large number of reports (and the associated data) also necessitates that there is a careful process when deciding what functions to use, and how much time and processing power will they use.

5. Conclusions

Data fusion and visualization was made on data obtained in personal exposure campaigns performed in 7 European cities within the ICARUS project. By using a diverse set of devices (wearable and static, commercially available and custom-made) with different temporal and spatial resolutions, a significant amount of data was obtained for each participant. Data fusion was performed by using complex algorithms in order to provide a report to the >600 participants.

Following these large-scale campaigns, several lessons were drawn and recommendations for future studies were provided.

Using low-cost sensors to assess air quality on an individual level presented some unique challenges, e.g., fusing data by rounding, duplicating and removing certain parts of the timestamps, which allowed a uniform presentation on several plots. Mostly simple modifications were enough to provide some clarity and make data fusion more straightforward. Providing appropriate guidelines have to be considered carefully to not confuse the participant or give false impressions on otherwise non-harmful concentrations of pollutants.

Participants should receive an amount of data which does not overwhelm them, but provides enough data for them to obtain as much meaningful information as possible. While the SAT provided a large amount of data, the decision to not visualize all of it and include it in the report was essential for effective communication i.e. allowed the report to be more readable and understandable. Apart from the number of visualizations, the appropriate kind must also be carefully selected and curated. Our approach with relative values for NO₂, CO₂ and TVOC provided enough data to clearly see some trends and react to them, without providing unreliable absolute data values. On the other hand, higher reliability and accuracy of PM concentrations and meteorological parameters enabled us to provide absolute values. A clear option should be included to observe trends between days, seasons and activities. These visualizations must also reflect the results of collocations and validations made prior to deploying these devices. Citizens must be made aware of the accuracy (and shortcomings) of the device they are using and to what extent can they rely on the results. A properly structured report will guide them through the report itself and give them enough support to extract the most data they can.

A well-informed public can collaborate with and react to changes in their environment, be it by influencing policy decisions or making changes to their individual life styles. Using LCS provides a conduit for citizens to be empowered by data that they can collect, observe and interpret. We, as researchers, must provide the necessary tools and options for them and guide them through the process. Change in policy can come from the top down or from the ground up. In both cases, the citizens that are affected by these policy changes must be active participants in designing and implementing these solutions.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1 (SD-A: Example of timestamp gaps in PPM dataset, SD-B: Example of dataset used for visualizations and reports, SD-C: Example of entire report for participants)

Author Contributions: Conceptualization, R.N., D.K., B.F., D.V. and C.D.; methodology, R.N., B.F. and C.D.; software, R.N. and I.P.; validation, R.N., D.K., J.A.R. and T.K.; formal analysis, R.N.; investigation, R.N., D.K., J.A.R., T.K., D.P., A.S., D.S., M.G.P., O.M., O.S., C.D., J.V., S.G., D.C. and S.K.; resources, R.N., D.K., J.A.R., T.K., T.M., D.C. and S.K.; data curation, R.N.; writing—original draft preparation, R.N.; writing—review and editing, R.N., D.K., J.A.R., T.K., D.S., B.F., D.V., C.D. and A.G.; visualization, R.N. and J.A.R.; supervision, D.K.; project administration, D.S. and A.G.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work has received funding from the European Union's Horizon 2020 Programme for research technological development and demonstration under grant agreement No 690105 (Integrated Climate forcing and Air Pollution Reduction in Urban Systems (ICARUS)). This work reflects only the authors' views and the European Commission is not responsible for any use that may be made of the information it contains. Funding was received from the Young Researchers Program and P1-0143 program "Cycling of substances in the environment, mass balances,

modelling of environmental processes and risk assessment», both funded by the Slovenian Research Agency.

The authors thank RECETOX Research Infrastructure (No LM2018121) and ACTRIS-CZ Research Infrastructure (No LM2018122) financed by the Ministry of Education, Youth and Sports and the Operational Programme Research, Development and Innovation (CZ.02.1.01/0.0/0.0/16_013/0001315 and CZ.02.1.01/0.0/0.0/16_013/0001761), and the CETOCOEN Excellence project supported from the Teaming Action of the EU Horizon 2020 programme (857560) and the Operational Programme Research, Development and Innovation (No CZ.02.1.01/0.0/0.0/17_043/0009632)

Institutional Review Board Statement: Ethical approval for the ICARUS project in Slovenia was obtained from the National Medical Ethics Committee of the Republic of Slovenia (approval nr. 0120-388/2018/6 on 22 August, 2018), the data in this paper was selected only from participants in Slovenia.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] 68 World Health Assembly, 'Health and the environment: addressing the health impact of air pollution: draft resolution proposed by the delegations of Albania, Chile, Colombia, France, Germany, Monaco, Norway, Panama, Sweden, Switzerland, Ukraine, United States of America, Uruguay and Zambia'. World Health Organization, 2015.
- [2] D. C. Payne-Sturges *et al.*, 'Healthy Air, Healthy Brains: Advancing Air Pollution Policy to Protect Children's Health', *Am J Public Health*, vol. 109, no. 4, pp. 550–554, Feb. 2019, doi: 10.2105/AJPH.2018.304902.
- [3] P. Sicard, E. Agathokleous, A. De Marco, E. Paoletti, and V. Calatayud, 'Urban population exposure to air pollution in Europe over the last decades', *Environmental Sciences Europe*, vol. 33, no. 1, p. 28, Mar. 2021, doi: 10.1186/s12302-020-00450-2.
- [4] M. Jerrett *et al.*, 'Validating novel air pollution sensors to improve exposure estimates for epidemiological analyses and citizen science', *Environmental Research*, vol. 158, pp. 286–294, Oct. 2017, doi: 10.1016/j.envres.2017.04.023.
- [5] B. J. Hubbell *et al.*, 'Understanding Social and Behavioral Drivers and Impacts of Air Quality Sensor Use', *Sci Total Environ*, vol. 621, pp. 886–894, Apr. 2018, doi: 10.1016/j.scitotenv.2017.11.275.
- [6] G. Miskell, J. Salmond, and D. E. Williams, 'Low-cost sensors and crowd-sourced data: Observations of siting impacts on a network of air-quality instruments', *Science of The Total Environment*, vol. 575, pp. 1119–1129, Jan. 2017, doi: 10.1016/j.scitotenv.2016.09.177.
- [7] L. Morawska *et al.*, 'Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?', *Environ Int*, vol. 116, pp. 286–299, Jul. 2018, doi: 10.1016/j.envint.2018.04.018.
- [8] A. C. Rai *et al.*, 'End-user perspective of low-cost sensors for outdoor air pollution monitoring', *Science of The Total Environment*, vol. 607–608, pp. 691–705, Dec. 2017, doi: 10.1016/j.scitotenv.2017.06.266.
- [9] J. A. Robinson, D. Kocman, M. Horvat, and A. Bartonova, 'End-User Feedback on a Low-Cost Portable Air Quality Sensor System—Are We There Yet?', *Sensors*, vol. 18, no. 11, Art. no. 11, Nov. 2018, doi: 10.3390/s18113768.
- [10] 'Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable – Indicators and a Monitoring Framework'. <https://indicators.report/goals/goal-11/> (accessed Mar. 02, 2021).
- [11] '69. Mean urban air pollution of particulate matter (PM10 and PM2.5) – Indicators and a Monitoring Framework'. <https://indicators.report/indicators/i-69/> (accessed Mar. 02, 2021).
- [12] D. J. Jarvis, G. Adamkiewicz, M.-E. Heroux, R. Rapp, and F. J. Kelly, *Nitrogen dioxide*. World Health Organization, 2010. Accessed: Mar. 03, 2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK138707/>
- [13] D. Nuvolone, D. Petri, and F. Voller, 'The effects of ozone on human health', *Environ Sci Pollut Res*, vol. 25, no. 9, pp. 8074–8088, Mar. 2018, doi: 10.1007/s11356-017-9239-3.

- [14] J. Shuai *et al.*, 'Health risk assessment of volatile organic compounds exposure near Daegu dyeing industrial complex in South Korea', *BMC Public Health*, vol. 18, no. 1, p. 528, Apr. 2018, doi: 10.1186/s12889-018-5454-1.
- [15] A. Casset and F. de Blay, 'Health effects of domestic volatile organic compounds', *Rev Mal Respir*, vol. 25, no. 4, pp. 475–485, Apr. 2008, doi: 10.1016/s0761-8425(08)71587-0.
- [16] T. A. Jacobson, J. S. Kler, M. T. Hernke, R. K. Braun, K. C. Meyer, and W. E. Funk, 'Direct human health risks of increased atmospheric carbon dioxide', *Nature Sustainability*, vol. 2, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41893-019-0323-1.
- [17] N. U. Okafor, Y. Alghorani, and D. T. Delaney, 'Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach', *ICT Express*, vol. 6, no. 3, pp. 220–228, Sep. 2020, doi: 10.1016/j.ict.2020.06.004.
- [18] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, 'Mapping urban air quality in near real-time using observations from low-cost sensors and model information', *Environment International*, vol. 106, pp. 234–247, Sep. 2017, doi: 10.1016/j.envint.2017.05.005.
- [19] A. Gressent, L. Malherbe, A. Colette, H. Rollin, and R. Scimia, 'Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value', *Environment International*, vol. 143, p. 105965, Oct. 2020, doi: 10.1016/j.envint.2020.105965.
- [20] N. Senthikumar, M. Gilfether, F. Metcalf, A. G. Russell, J. A. Mulholland, and H. H. Chang, 'Application of a Fusion Method for Gas and Particle Air Pollutants between Observational Data and Chemical Transport Model Simulations Over the Contiguous United States for 2005–2014', *International Journal of Environmental Research and Public Health*, vol. 16, no. 18, Art. no. 18, Jan. 2019, doi: 10.3390/ijerph16183314.
- [21] A. L. Clements *et al.*, 'Low-Cost Air Quality Monitoring Tools: From Research to Practice (A Workshop Summary)', *Sensors*, vol. 17, no. 11, p. 2478, Nov. 2017, doi: 10.3390/s17112478.
- [22] A. C. Lewis, E. von Schneidermesser, and R. E. Peltier, 'Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications', *World Meteorological Organization*, 2018.
- [23] J. D. Paul and W. Buytaert, 'Chapter One - Citizen Science and Low-Cost Sensors for Integrated Water Resources Management', in *Advances in Chemical Pollution, Environmental Management and Protection*, vol. 3, J. Friesen and L. Rodríguez-Sinobas, Eds. Elsevier, 2018, pp. 1–33. doi: 10.1016/bs.apmp.2018.07.001.
- [24] Y. Wang, F. Han, L. Zhu, O. Deussen, and B. Chen, 'Line Graph or Scatter Plot? Automatic Selection of Methods for Visualizing Trends in Time Series', *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 2, pp. 1141–1154, Feb. 2018, doi: 10.1109/TVCG.2017.2653106.
- [25] B. Saket, A. Endert, and Ç. Demiralp, 'Task-Based Effectiveness of Basic Visualizations', *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2505–2512, Jul. 2019, doi: 10.1109/TVCG.2018.2829750.
- [26] R. Garcia-Retamero and M. Galesic, 'Who profits from visual aids: Overcoming challenges in people's understanding of risks', *Social Science & Medicine*, vol. 70, no. 7, pp. 1019–1025, Apr. 2010, doi: 10.1016/j.socscimed.2009.11.031.
- [27] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert, 'Evaluating Interactive Graphical Encodings for Data Visualization', *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 3, pp. 1316–1330, Mar. 2018, doi: 10.1109/TVCG.2017.2680452.
- [28] 'ICARUS2020.eu', *icarus*. <https://icarus2020.eu/> (accessed Oct. 12, 2018).
- [29] D. Kocman, 'Multi-sensor data collection for personal exposure monitoring: ICARUS experience'.
- [30] D. Sarigiannis, D. Chapizanis, and A. Arvanitis, 'D4.1 Report on the methodology for estimating individual exposure', Deliverable. [Online]. Available: https://icarus2020.eu/wp-content/uploads/2018/03/ICARUS-Deliverable-D4.1_FINAL.pdf
- [31] D. Sarigiannis, S. Karakitsios, D. Chapizanis, and R. Hiscock, 'D4.2_ICARUS_Methodology for properly accounting for SES in exposure assessment.pdf'. Accessed: Nov. 30, 2019. [Online]. Available: https://icarus2020.eu/wp-content/uploads/2019/02/ICARUS_D4.2.pdf
- [32] J. A. Robinson, R. Novak, T. Kanduč, D. Sarigiannis, and D. Kocman, 'Articulating user experience of a multi-sensor personal air quality exposure study (manuscript in preparation)', 2021.
- [33] 'R: The R Project for Statistical Computing'. <https://www.r-project.org/> (accessed Dec. 05, 2019).
- [34] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: <https://ggplot2.tidyverse.org>
- [35] H. Wickham, R. François, L. Henry, and K. Müller, *dplyr: A Grammar of Data Manipulation*. 2018. [Online]. Available: <https://CRAN.R-project.org/package=dplyr>

- [36] Y. Xie, *knitr: A General-Purpose Package for Dynamic Report Generation in R*. 2021. [Online]. Available: <https://yihui.org/knitr/>
- [37] J. J. Allaire *et al.*, *rmarkdown: Dynamic Documents for R*. 2021. [Online]. Available: <https://github.com/rstudio/rmarkdown>
- [38] R. Novak, D. Kocman, J. A. Robinson, T. Kanduč, D. Sarigiannis, and M. Horvat, 'Comparing Airborne Particulate Matter Intake Dose Assessment Models Using Low-Cost Portable Sensor Data', *Sensors*, vol. 20, no. 5, Art. no. 5, Jan. 2020, doi: 10.3390/s20051406.
- [39] A. Industries, 'Adafruit PCF8523 Real Time Clock Assembled Breakout Board'. <https://www.adafruit.com/product/3295> (accessed Sep. 30, 2020).
- [40] Garmin and G. L. or its subsidiaries, 'Garmin vivosmart® 3 | Fitness Activity Tracker', *Garmin*. <https://buy.garmin.com/en-US/US/p/567813> (accessed Sep. 03, 2019).
- [41] 'uHoo | Product', *uHoo*. <https://uhooair.com/product/> (accessed Nov. 16, 2018).
- [42] S. Mahajan *et al.*, 'A citizen science approach for enhancing public understanding of air pollution', *Sustainable Cities and Society*, vol. 52, p. 101800, Jan. 2020, doi: 10.1016/j.scs.2019.101800.
- [43] T. Nikolakopoulos, A. Gotti, E. Tsiros, and E. Siora, 'D7.2: Report on the design of technical framework and system architecture of the ICARUS DSS', Deliverable. [Online]. Available: https://icarus2020.eu/wp-content/uploads/2017/08/D.7.2_ICARUS_Design_of_%20technical_framework_and_system_architecture_of_the_ICARUS_DSS_FINAL.pdf
- [44] R. Novak *et al.*, 'Low-Cost Environmental and Motion Sensor Data for Complex Activity Recognition: Proof of Concept', *Engineering Proceedings*, vol. 2, no. 1, Art. no. 1, 2020, doi: 10.3390/ecsa-7-08194.
- [45] J. A. Robinson *et al.*, 'User-Centred Design of a Final Results Report for Participants in Multi-Sensor Personal Air Pollution Exposure Monitoring Campaigns (manuscript in preparation)', 2021.
- [46] 'Air Quality Now - Indices definition'. http://airqualitynow.eu/about_indices_definition.php (accessed Jan. 20, 2021).
- [47] X. Zhang, Z. Zhao, T. Nordquist, and D. Norback, 'The prevalence and incidence of sick building syndrome in Chinese pupils in relation to the school environment: a two-year follow-up study', *Indoor Air*, vol. 21, no. 6, pp. 462–471, 2011, doi: 10.1111/j.1600-0668.2011.00726.x.
- [48] 'Field Evaluation - uHoo PM2.5, ozone, and CO sensor', AQ-SPEC. Accessed: Apr. 17, 2019. [Online]. Available: <http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/uhoo---field-evaluation.pdf?sfvrsn=12>
- [49] A. Baldelli, 'Evaluation of a low-cost multi-channel monitor for indoor air quality through a novel, low-cost, and reproducible platform', *Measurement: Sensors*, vol. 17, p. 100059, Oct. 2021, doi: 10.1016/j.measen.2021.100059.
- [50] V. V. Tran, D. Park, and Y.-C. Lee, 'Indoor Air Pollution, Related Human Diseases, and Recent Trends in the Control and Improvement of Indoor Air Quality', *Int J Environ Res Public Health*, vol. 17, no. 8, p. 2927, Apr. 2020, doi: 10.3390/ijerph17082927.
- [51] Y. Wei *et al.*, 'Short term exposure to fine particulate matter and hospital admission risks and costs in the Medicare population: time stratified, case crossover study', *BMJ*, vol. 367, p. l6258, Nov. 2019, doi: 10.1136/bmj.l6258.
- [52] P. Orellano, J. Reynoso, N. Quaranta, A. Bardach, and A. Ciapponi, 'Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis', *Environment International*, vol. 142, p. 105876, Sep. 2020, doi: 10.1016/j.envint.2020.105876.
- [53] 'WHO guidelines: Ambient (outdoor) air pollution (prior to 2021)'. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health) (accessed Mar. 03, 2021).
- [54] 'WHO guidelines: Ambient (outdoor) air pollution (revised in 2021)'. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health) (accessed Sep. 27, 2021).
- [55] J. Park and S. Kim, 'Machine Learning-Based Activity Pattern Classification Using Personal PM2.5 Exposure Information', *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, Art. no. 18, Jan. 2020, doi: 10.3390/ijerph17186573.