AI4EU

AI4IoT - Whitepaper

Sensing and Decision Making about Air Quality Data

Authors
Sigmund Akselsen, Telenor Research
Kerstin Bach, NTNU
Tiago Veiga, NTNU
Ilias Kalamaras, ITI

March 5, 2021

Comments: Draft

https://www.ai4eu.eu/

Version	Date	Description	Name
0.1	14-Apr-2020	initial version – collecting pieces	KB, TV

Contents

List of Figures

1 Introduction

The necessity of healthy air has always been of great importance. As air is vital for all living beings on earth, it is our responsibility to keep the air clean. The rapid urbanization and industrialization have led the world into a new era of air pollution and is seen as a modern-day curse. Air pollution refers to the contamination of the air by excessive quantities of harmful substances. Most air pollution occurs from energy use and production, where emissions from traffic and industry are major contributors. Air pollution is a widespread problem due to its impact on both humans and the environment. Urban cities usually have the worst air pollution due to human activities (Kampa & Castanas 2008). Clear links between pollution and health effects have been revealed, which includes both short- and long-term consequences (Brauer et al. 2011). Associations with reduced lung function and increase in heart attack (Pope III et al. 2002), direct impact on people with asthma and other types of pneumonia (Guarnieri & Balmes 2014) and once inhaled, a fine particular matter may hardly be self-purified by the immune system (Becker et al. 2002).

Ambient (outdoor) air pollution poses a major threat to both health and climate, with a steadily increasing 4.2 million[1] premature deaths per year worldwide due to stroke, heart disease, lung cancer, and chronic cardiovascular and respiratory diseases as a result of high pollution exposure. The economic impact of these health risks in the 15 countries responsible for the most pollution is estimated to be more than 4% of their GDP. This is in an essential problem and we believe that the combination of applying AI methods on a network of air quality sensors will improve the level of detail and AQ forecast.

1.1 Background: Measuring Air Quality

Providing decision makers with AI-based solutions requires to monitor the ambient AQ accurately, as AI models highly depend on the underlying data used to justify the predictions. Unfortunately, the hyper-locality of AQ, varying from street to street, makes this difficult to monitor using high-end sensors, as the cost of the amount of sensors needed for such local measurements is too high.

1.1.1 Air Quality

Air quality can be formalized as the concentration of dust particles and other pollutants in the air. The most important pollutants are particulate matter (PM), ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). Of these, PM affects by far the most amount of people, and are thus used as a proxy indicator for AQ.

Particulate matter is a complex mixture of different particles of varying sizes, most notably sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, and water. Because of this complex structure it is measured by size, under $2.5\mu m$ in diameter indicated as $PM_{2.5}$, and under $10\mu m$ in diameter indicated as PM_{10}. While PM_{10} can physically damage the lungs while breathing, $PM_{2.5}$ can bypass the lung barrier and enter the blood stream, leading to higher risk of cardiovascular and respiratory diseases, and lung cancer, as such being considerably more dangerous.

The concentrations found in highly developing countries are not uncommon to reach $300\mu g/m^3$, while some European countries might have an annual average of $8\mu g/m^3$. This means that people in some developing countries are highly likely to experience severe consequences, while people in western countries might not experience such consequences at all. WHO still estimates that the life expectancy of European countries is lowered by 8.6 months because of air pollution.

[1] https://www.who.int/health-topics/air-pollution

Because of all the things affecting AQ, it varies a lot between even close areas[2]. The gases mentioned in addition to PM tend to cover larger areas more evenly, but PM is hyper-local. PM_{10} is considerably heavy, and thus struggle to move large distances. This leaves PM_{10} to be found in high concentrations where these particles are created. $PM_{2.5}$ are lighter, and thus moving more freely, leading to concentrations moving together with e.g. wind. It's also important to note that there is considerable cross-sensitivity between the pollutants, as some specific pollutants directly contribute to $PM_{2.5}$.

1.1.2 Wireless Sensor Networks

A Wireless Sensor Networks (WSN) is a collection of mobile or static sensors that monitors the same measurand in different places. A basic architecture for illustration can be found in figure 1. The connectivity of such a network is vital, as each deployed sensor must be able to communicate and send its measurements to at least one sink node. The placement is often modeled as k-coverage, where k sensors are measuring data from any given point of interest. Because of the locality of AQ, this coverage solution is not ideal for AQ monitoring systems as it leads to deploying perhaps too many nodes.

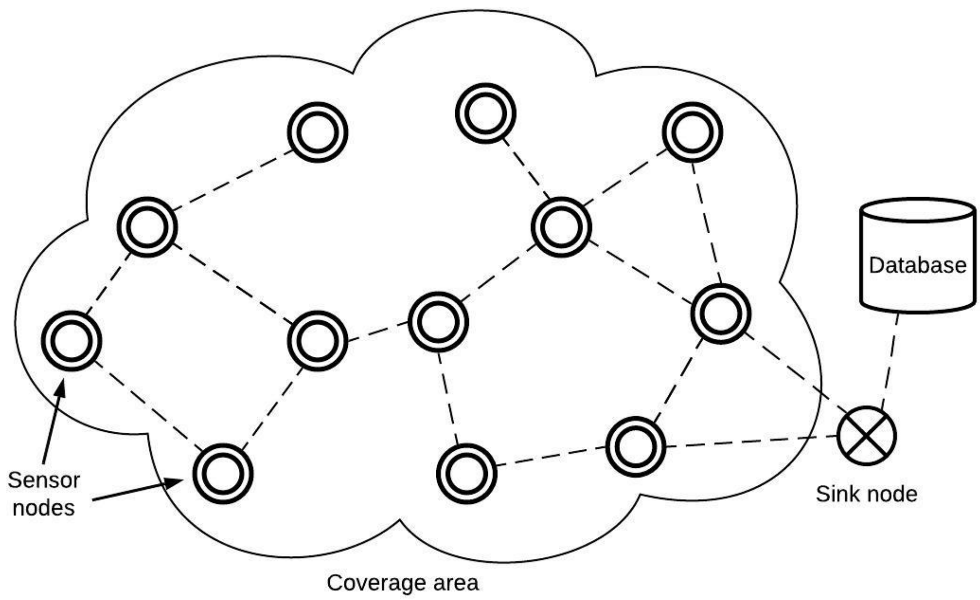


Figure 1: A basic WSN. Sensor nodes measure their measurand and forward their data to the sink node. The sink node then populates a database for data analysis.

In order to deploy a large net of sensors, cheaper or mobile sensors can be used. This can lead to lower data quality, but enables measuring AQ at a local level. We assume that such a locality leads to more informed models for AQ within a city, leading to better understanding of key factors, e.g. sources of pollution.

The sensors used in these networks vary based on the measurand. Low-cost PM sensor are almost exclusively measuring particle concentration optically. That is, they pump air into a

[2] https://www.edf.org/airqualitymaps

small chamber and use a LED, or a low-powered laser, together with a photo-diode to measure the concentration of the particles, as different concentrations scatter the light differently. The problem with these sensors are still very similar, and can be grouped into internal or external reasons [Maag et al. (2018)]. Internal reasons are errors as a result of the sensors architecture and principles while external sensors are error sources coming from the environment, and how the sensors react to this. Due to these factors the sensors will need to be calibrated and the timing can be another challenge addressed by machine learning methods.

2 Decision Making

Decision making in general involves collecting the available data for a situation, analyzing them and taking a decision that will advance the situation towards a desired goal. In the context of air pollution, the situation of interest is air pollution with its geographical distribution and evolution in time. The desired goal is the reduction of pollution and the actions towards this goal may include traffic route modifications, municipal restrictions in the use of fuels, street cleaning actions, etc. In classical, non-automatic, decision making, the decision about which actions to take is taken by the decision maker, an expert who assesses the situation and decides on the most appropriate action. In automatic decision making, the decision can be taken by the system itself, after estimating the potential impact of alternative scenarios.

2.1 Facilitated Decision Making

ICT solutions have traditionally assisted the decision making process in the first parts: data collection and analysis. A *Decision Support System (DSS)* is a system that facilitates a decision maker in taking a decision, by presenting them with the collected data, detected trends, predictions for future events, potential risks, etc., and suggesting possible actions that can have a significant impact. In the context of air pollution DSSs usually present the decision maker with several views of the data (interactive maps, tables, timelines, etc.), presenting the raw data along with the results of data analysis, e.g. meteorological model predictions (Lim et al. 2005, Elbir et al. 2010). In our study, our DSS will be based on the collection of large volumes of data using the microsensor network presented above. The fact that the employed sensors are low-cost and can thus be scattered easily in several places within a city will result in a large amount of data regarding pollutant concentrations and environmental conditions, that can fuel a variety of decision making tools. In order to make sense in this large volume of information, such decision making tools necessarily involve data analysis and visualization methods.

Data analysis The collected data are in the form of time-series, so time-series models will be employed in order to extract useful information. Prediction models, based on traditional machine learning techniques as well as deep learning techniques and trained neural networks, are being developed with the goal to provide important information to the decision maker. Predicting future states of air pollution or other related quantities (e.g. traffic) allows decision makers to better plan their actions and take preventive measures. Part of the data analysis methods we use employs computation of correlations among the collected variables (e.g. between weather conditions and pollutant concentrations, or between measurements from different geographical areas), which is also important and may provide hints to the decision maker about the causes of specific pollution patterns. We make use of matrix factorization techniques, in order to discover significant events that can explain much of the pollution behaviour within a time window. We are also examining the use of probabilistic models of pollution and traffic, which will allow the quantification of the uncertainty in our predictions and consequently the risk of certain decisions. Probabilistic modeling will also allow the simulation of alternative scenarios and the measurement of their potential outcomes, in order for decision makers to examine the impact and risks of potential decisions. This toolset of data analysis and decision support methods aims to provide high-level information to the decision maker, allowing them to see the patterns behind the noise, which is crucial for taking better decisions.

Data visualization Visualizing the available data in a comprehensive manner allows the decision maker to have a better insight in the situation, in order to take better decisions. In the context of the air quality pilot, we have been developing the VisualBox (Aurdal 2019) and the VFlow (Kalamaras et al. 2019) systems for interactive dashboard design. VisualBox facilitates the connection to data sources through the concept of *integrations*, and allows the user to create customized dashboards, by dropping ready-to-use visualization components (maps, timelines, bar charts, etc.) on a canvas and connecting them to data sources. VFlow is a web library for dashboard design in which different types of visualization and data analysis are considered as components supporting specific functionalities, and the connections between components determine the dashboard logic and the user interaction. Both systems are currently being enhanced with advanced functionality and there is a goal to merge them to provide interactive dashboards using real-time data sources. Visualizations offer multiple views of the same data and interacting with them allows a decision maker to have a better understanding. An important challenge regarding visualizations, which is one of our major goals, is how to use visualizations not only for displaying raw data, but also as a means to better explain the results of AI models trained on the data. Explainability of the AI models is important so that the decision makers know why e.g. a specific prediction was computed, which will let them be better informed about the decisions they take. Our effort regarding the above developed systems is to also include means of visually explaining the results of our AI models.

2.2 Automated Decision Making

In a world increasingly connected and robotized, automated decision making may provide a tool to, simultaneously, improve the readings and/or calibration of mobile sensors and actuate in the environment to prevent high pollution peaks. For instance, mobile sensors may be mounted on-board public buses which provide a good coverage of an urban scenario but run on fixed routes. On the other hand, other types of vehicles might have a more flexible route scheduling, such as street cleaning vehicles and others. Moreover, automated agents might provide solution to prevention of pollution peaks. For instance, taking into account the traffic and weather patterns it might be beneficial to reroute traffic to prevent pollution accumulating at a particular region. While modelling and testing models for these scenarios can be particularly complex, simulations let us safely training and developing decision agents as to showcase the effects of those on air quality. As such, based on a traffic simulator, we may develop a world simulator that models the main sources of pollution. More detail on it is given on Section 5.2.1. In the following we will detail how to tackle automotated decision making in the pilot scenario.

Active Perception Controllable mobile sensors are, typically, noisier than static sensors. Also, the influence of other features (such as weather, traffic patterns) might be uncertain and/or change over time. Taking this into account, we plan to use decision-theoretic methods to improve pollution measuring, in particular partially observable Markov decision processes (POMDPs) (Kaelbling et al. 1998). POMDPs provide a principled framework for modelling the uncertainties associated with acting and sensing and computing optimal policies for an automated agent. The problem of controlling mobile sensors is framed under the active perception concept and has

Figure 2: AQ Sensors deployed in Trondheim.

been studied in robotics contexts. For our scenario, state of the art algorithms will have to be used, including online POMDPs, which are more flexible and scalable, at the cost of global optimality. Nevertheless, they have proven to be effective in practice.

Traffic rerouting Decision-theoretic methods also offer a solution to learn policies that map the state of a dynamic and uncertain environment to actions that influence this environment. In this scenario, we plan to test the efficiency to mitigate high pollution levels with automated decision making. In particular, reinforcement learning (RL) (Sutton & Barto 1998) techniques are able to learn policies directly from data and we have access to traffic, pollution and weather data, among possible other datasets. While it is very difficult to learn explicit models for the dynamics of the world, we plan to have an RL agent to learn preventive actions to prevent high levels of pollution. For instance, whether to close some roads and redirect traffic to prevent a pollution peak at a particular area in the city.

3 Pilot Description

3.1 Develop and test low cost sensors

Exploratory Engineering/Telenor and Trondheim Municipality are testing devices and checking data quality. Two generations of sensors were developed, with the characteristics presented in Table 1. A picture of the second generation sensors is presented in Figure 2.

Sensor platform	Rev. 1	Rev. 2
GPS	Yes	Yes
Temperature	Yes	Yes
Humidity	Yes	Yes
Particulate matter	$PM_{2.5}$, PM_{10}	PM_1, $PM_{2.5}$, PM_{10}
Chemical pollutants	CO_2	CO_2, NO_2, NO, O_3

Table 1: A list of the different sensors used in revision 1 and 2

3.2 Microsensors

The second generation board was an improved version of its predecessor with ability to detect more chemical pollutants in addition to CO_2, such as NO_2, NO and O_3. A general challenge

Figure 3: Microsensor and deployment

with chemical sensors is stability, since the performance and sensitivity degrade over time, which in turn limit the operating time before re-calibration is required Bandodkar et al. (2016).

The new prototype features an upgraded particle sensor from Alphasense (OPC-N3), that is able to differentiate between PM_1, $PM_{2.5}$ and PM_{10} particle sizes. Furthermore, it is able to detect chemical pollutants, such as CO_2, NO_2, NO and O_3 A feature overview is given in Table 1. The sensor measures particle counts in 24 bins from 0.35 micrometre to 40 micrometre by illuminating one particle at the time with a laser, and measure the intensity of the light scattered. The amount of the light scattered is a function of the particle size, which is calibrated using a proprietary algorithm from Alphasense.

The sensors have been placed around in strategic places in Trondheim. Figure 4 shows the locations of both static sensors, as well as 1st and 2nd generation low-cost sensors.

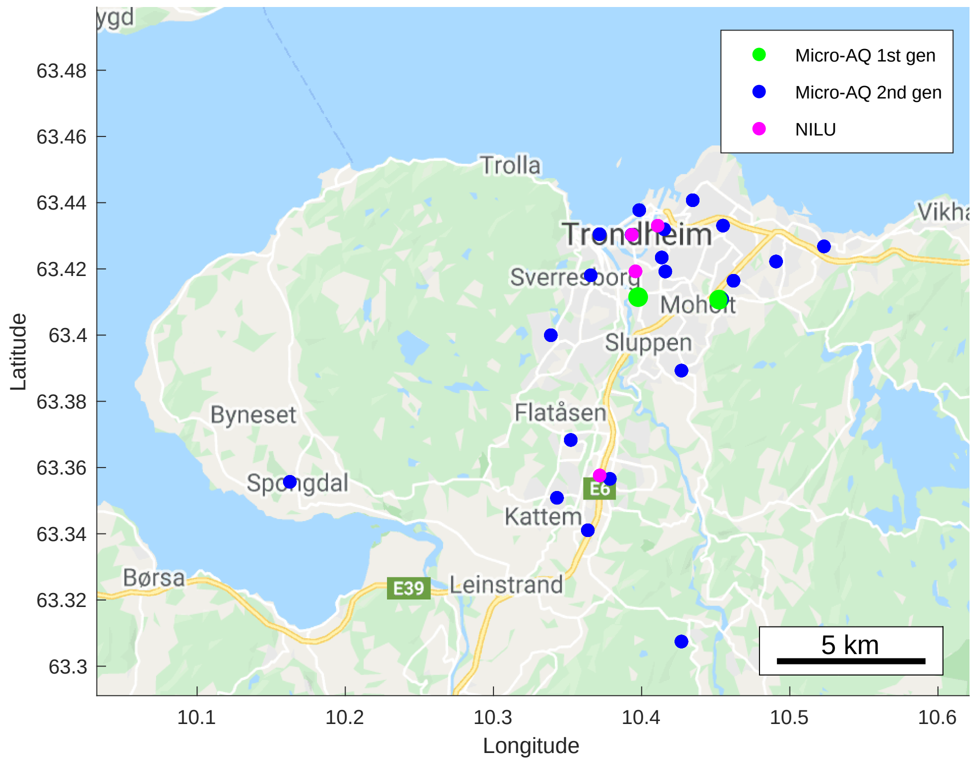


Figure 4: Placement of the micro-sensors 1st and 2nd generation. NILU reference stations are marked in purple.

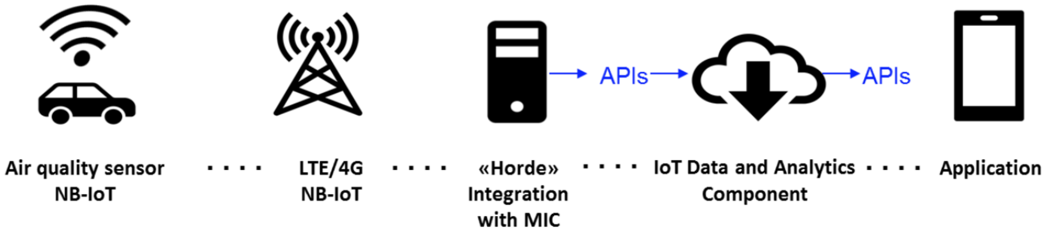


Figure 5: Overview data pipeline

3.3 User Stories

3.3.1 Citizen

3.3.2 Decision Maker

4 Architecture

4.1 Pipeline

4.2 AI4EU Platform Services

5 Data & Services

5.1 Air Quality Sensors

5.1.1 Simulated Data

5.1.2 Real-World Data

5.1.3 Analysis

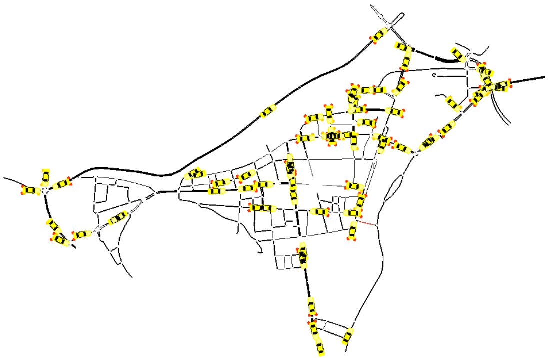
5.2 Car traffic

5.2.1 Simulated Data

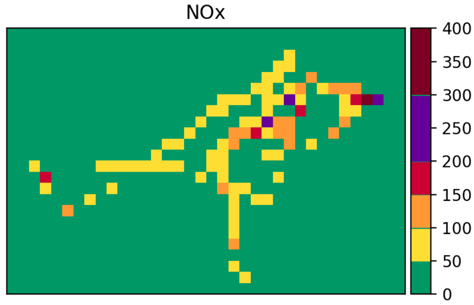
The advantages of simulated scenarios are two fold: first, as previously mentioned, they are useful to model and test automated decision making without the risks of performing it in the real world, as they allow to analyze the effects of actions on the future dynamics of the environment. On the other hand, and for this last reason, they might also provide an useful tool in Decision Support Systems as they allow a human decision maker to change some features of the world and analyze its effect on the air quality as well as in the citizen's life. In the following, we will detail how we implement a traffic simulation scenario based on the real traffic data and obtain pollutant emission data from there.

Traffic simulation Our world simulation is based on the Simulation of Urban Mobility (SUMO) traffic simulator (Lopez et al. 2018). SUMO is a traffic simulator that is able to generate traffic routes from input data. Traffic data readings from public roads in Norway are publicly available by the Norwegian Public Roads Administration[3] with traffic data aggregated per time for each detector, with the minimum shortest time interval of one hour. We are now at the stage of analyzing the output of the simulator with the aggregated real data as input

[3]https://www.vegvesen.no/trafikkdata/start/om-trafikkdata



(a) Traffic screenshot



(b) Pollution screenshot

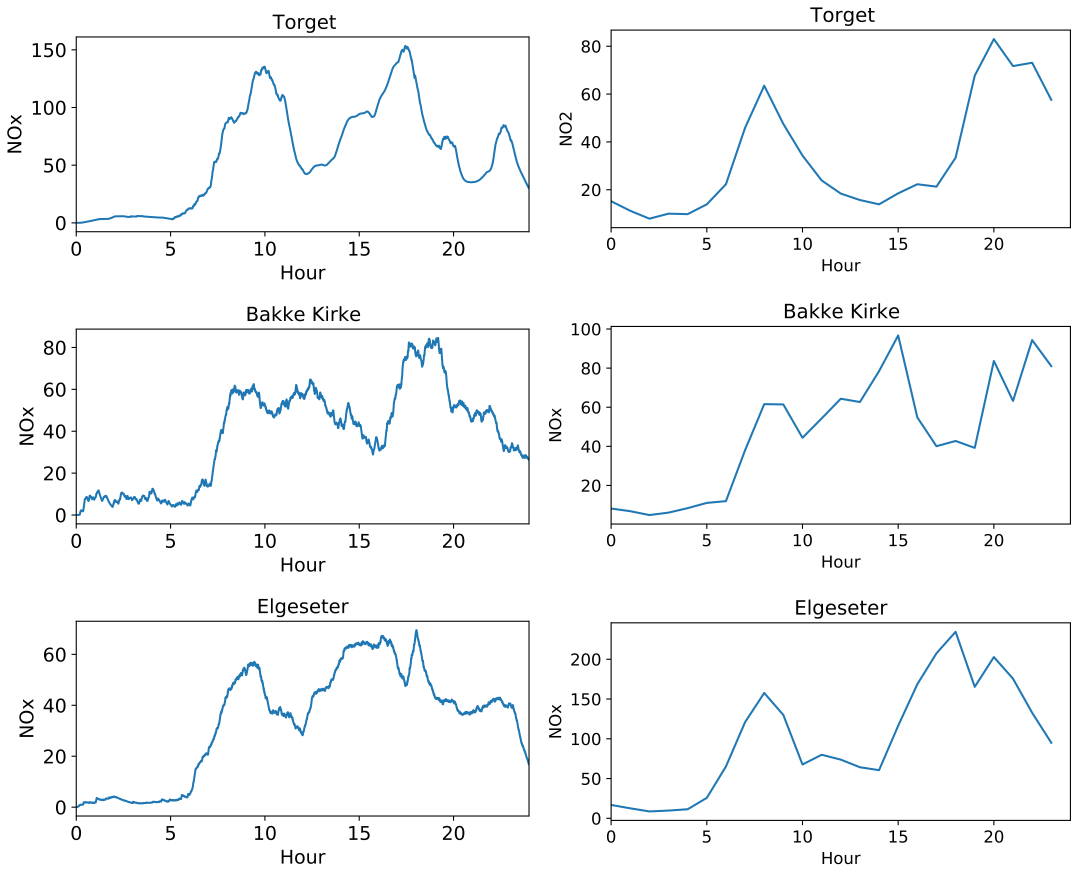
Figure 6: Screenshots of visualizations for the traffic simulator outputs: traffic and pollution levels.

and understanding whether the outputed traffic patterns are realistic or if some work on top of that is needed.

Traffic emissions SUMO also implements pollutant emission models for the passing traffic, which allows us to map pollution measures to traffic patterns. Since it performs microscopic traffic simulation it models, for each vehicle in the network and according to its speed, acceleration and slope, the amount of fuel consumed and pollutants emitted by that particular vehicle. Pollutants covered by the emission model are: CO_2, CO, HC, NO_x, PM_x.

SUMO built-in methods output emissions either by vehicle or aggregated by edges (SUMO's nomenclature for segments of road), and by timestep. We add an extra layer on top of that to obtain an aggregation of data by regions of the map. We create a grid cell, but using SUMO method to create polygons within the map any shape can be used.

Dispersion model The current SUMO implementation only outputs immediate vehicle pollutant emission, therefore we need to implement a dispersion and persistence model for the pollutants. For this, we use aggregate emitted pollutants with the grid cell previously mentioned and implemented a naive decay and dispersion model. It considers that the aggregated pollution level at a given cell is a weighted sum of the pollutants emitted during the last time period, the pollution levels in the previous period and the pollution levels in the neighbour cells. As future work, we intend to further develop this model to include some more realistic data such as, for instance, wind direction and speed as it will influence the spreading of pollutants over time.



(a) Simulation pollution emission

(b) Real measured pollution

Figure 7: Comparison between SUMO emissions and real measurements from static sensors for a full day. Traffic and pollution data from February 25, 2020.

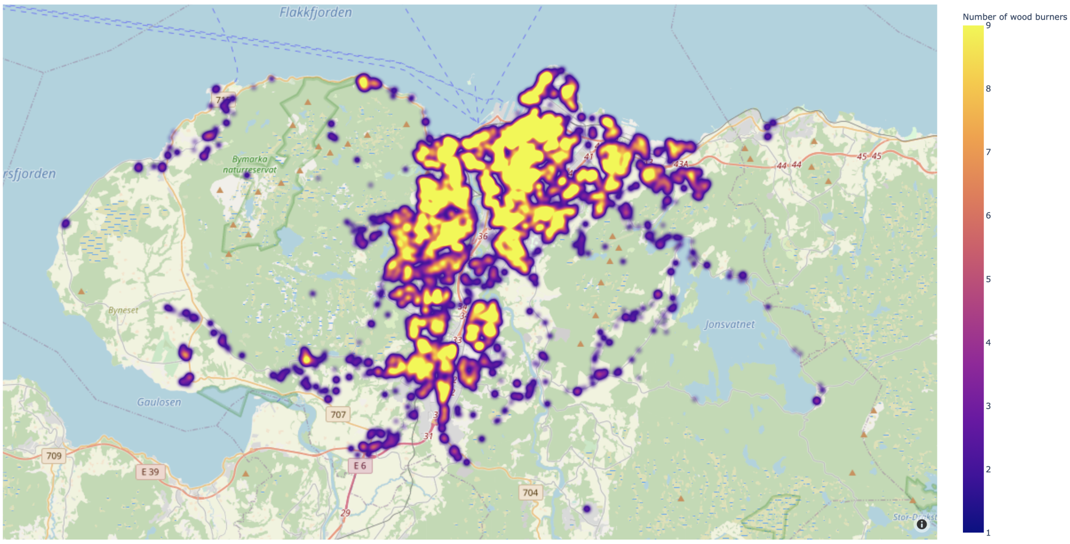


Figure 8: Fireplaces and locations in Trondheim

5.2.2 Real-World Data

5.2.3 Analysis

5.3 Mobility Data

5.3.1 Real-World Data

5.3.2 Analysis

5.4 Fireplaces

Fireplaces (more than 60 000 records, among other on type and location)

5.5 Event calendar

Municipality event calendar (including cruise ships)

5.6 Actions improving air quality

Record of cleaning actions (sweeping, sanding, . . .)

5.7 Overview of construction sites

Overview of construction sites

Drones for tracking movements during rush hours

6 Visualization

6.1 VFlow

The VFlow library (Kalamaras et al. 2019) has been developed for creating custom interactive dashboards for data exploration. VFlow is a JavaScript library available online [4] with which

[4]https://github.com/EliasKal/vflow

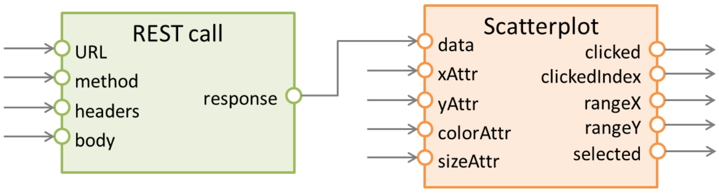


Figure 9: Component linking in the VFlow library.

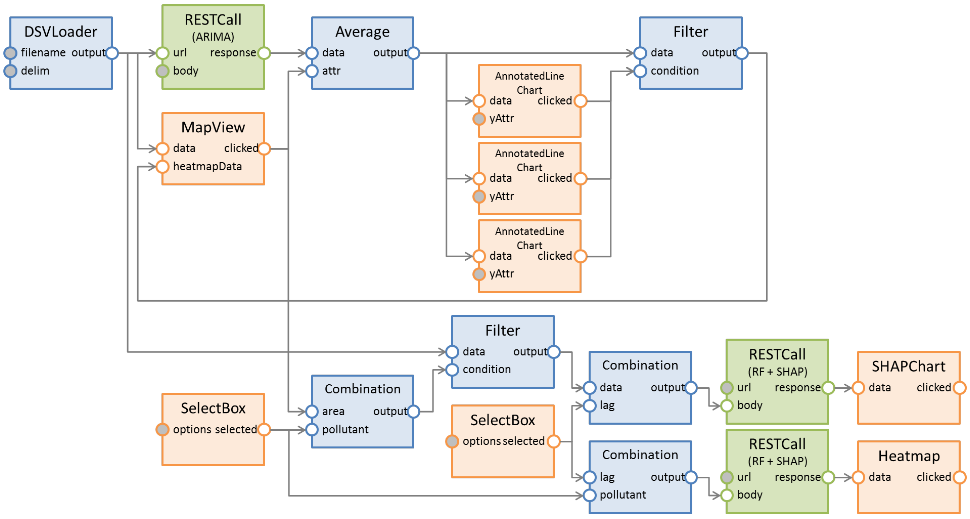


Figure 10: Example of a workflow created in VFlow. The workflow consists of visualization components (orange boxes), data manipulation components (blue boxes) and REST call components (green boxes). This workflow results in the dashboard of Figure 11.

the user can add data visualization and analysis components in a webpage. The library supports common visualization types (scatterplots, barplots, maps, etc.) as well as graph-based visualizations for visualizing similarities among entities (e.g. geographical areas, pollutants, time instances, etc.). The library includes visualizations for explaining the results of AI models, using approaches based on LIME (Ribeiro et al. 2016) and SHAP (Lundberg & Lee 2017). Data analysis beyond simple manipulations is supported by connections with CERTH's Data Analytics API or any other existing data analysis API. CERTH's data analytics API is a set of web services for data analytics (statistics, clustering, classification, prediction, anomaly detection, etc.) that is being extended within the AI4EU project.

The visualization and analysis methods in the library are provided as reusable and reactive components, which can be linked to each other in custom workflows, as shown in Figure 9. Visualization or control components have a connected graphical user interface, which can be displayed on the website. A change in a component's output is propagated to the inputs of its connected components, facilitating real-time updates. Input to a component can be provided from other components or from interaction with the user, resulting in interactive dashboards.

An example is shown in Figures 10 and 11. Figure 10 demonstrates an example workflow, where several components are connected to each other. There are components with corresponding visual elements, such as visualizations and input controls (orange boxes), data manipulation components for data loading and transformations between components (blue boxes), and components that make REST calls to the Data Analytics API (green boxes). Processing this workflow with the VFlow library results in the dashboard of Figure 11.

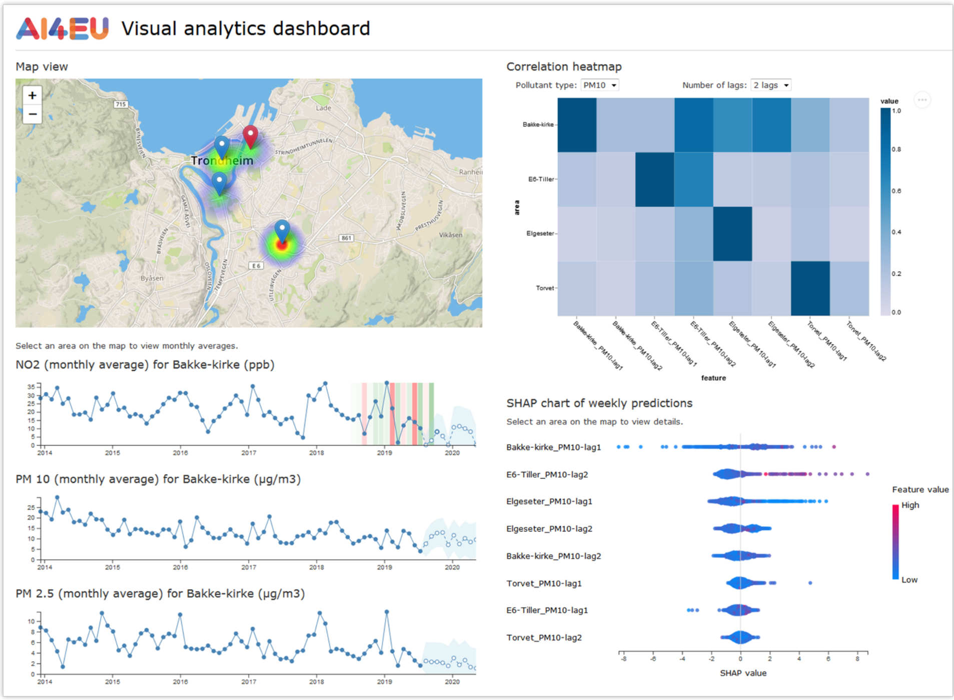


Figure 11: Example of a dashboard created in VFlow. This dashboard is generated by the workflow of Figure 10.

6.2 VisualBox

We've created a map including the location and latest measurements and predictions from official sources (https://luftkvalitet.miljostatus.no and https://api.met.no).

Link to the demo: https://s.ntnu.no/aq-trondheim

6.3 Grafana

A screenshot of the Grafana dashboard can be seen in Figure **??**, and includes a historical line chart and a corresponding map with geographical information for three sensors in the experiment. Note the time selector in the top right corner, where the user can define the start and end time for the current view. The view is dynamically updated as new readings arrived. Navigation by selecting a closer section is also possible with a click-and-drag gesture in the line graph, or with the zoom buttons in the map view. The map marker colour threshold for PM_{10} and PM_{10} values reflect the legal pollution limit levels. The exact value to a data point can be seen by clicking it. All sensor parameters were plotted in the same graph. This design selection was made to make it easy to compare e.g. CO_2 readings with particulate matter. A hover panel with raw values becomes visible when the user moves the cursor across the timeline. The labels are located under the timeline and include colour and the average value for each parameter in the open window.

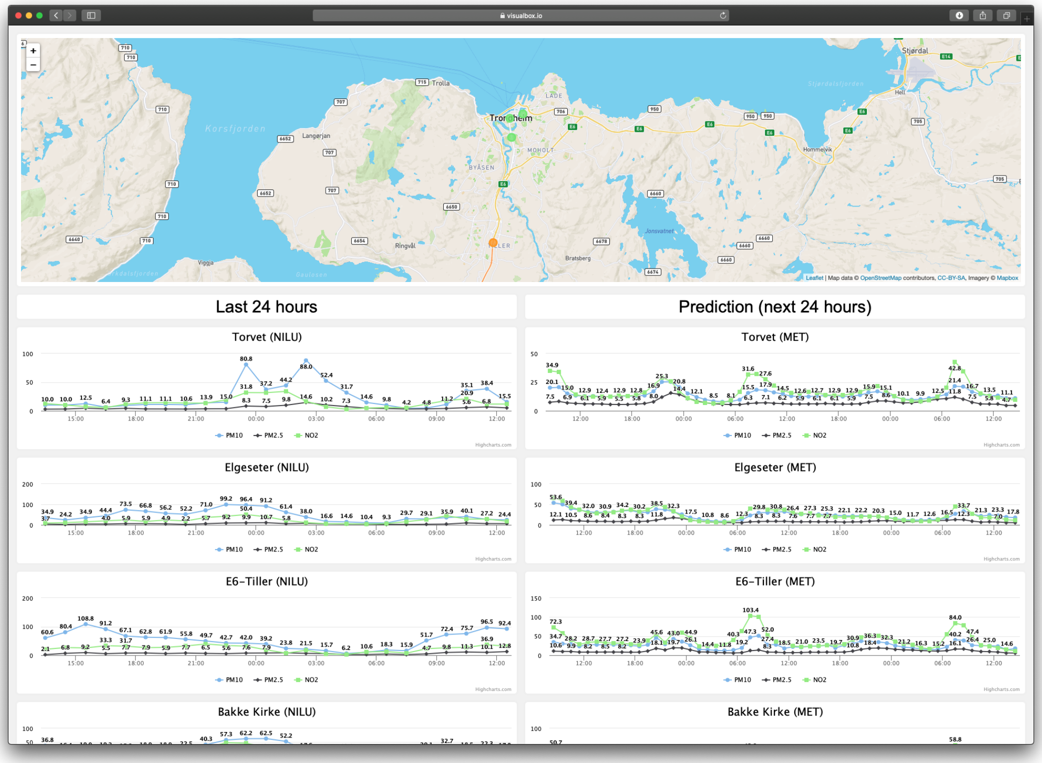


Figure 12: VisualBox

Figure 13: Grafana

7 Discussion

7.1 Impact

7.1.1 Privacy issues when collecting data

7.1.2 Inequalities tied to digital literacy and health issues

7.1.3 Legal issues tied to city planning and popularity of urban areas

8 Conclusion

Acknowledgements

This work has been done in collaboration with students and colleagues. The authors would like to thank XXX for their work.

References

Aurdal, P. E. (2019), VisualBox – A Generic Data Integration and Visualization Tool, Master's thesis, The Arctic Univeristy of Norway (UiT), Tromsø, Norway.

Bandodkar, A. J., Jeerapan, I. & Wang, J. (2016), 'Wearable chemical sensors: Present challenges and future prospects', *ACS Sensors* **1**(5), 464–482.

Becker, S., Fenton, M. & Soukup, J. (2002), 'Involvement of microbial components and toll-like receptors 2 and 4 in cytokine responses to air pollution particles', *American journal of respiratory cell and molecular biology* **27**, 611–8.

Brauer, M., Amann, M., Burnett, R. & Cohen, A. (2011), 'Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution', *Environmental science & technology* **46**, 652–60.

Elbir, T., Mangir, N., Kara, M., Simsir, S., Eren, T. & Ozdemir, S. (2010), 'Development of a gis-based decision support system for urban air quality management in the city of istanbul', *Atmospheric Environment* **44**(4), 441–454.

Guarnieri, M. & Balmes, J. R. (2014), 'Outdoor air pollution and asthma', *The Lancet* **383**(9928), 1581 – 1592.
URL: *http://www.sciencedirect.com/science/article/pii/S0140673614606176*

Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. (1998), 'Planning and acting in partially observable stochastic domains', *Artificial Intelligence* **101**(1-2), 99–134.

Kalamaras, I., Xygonakis, I., Glykos, K., Akselsen, S., Munch-Ellingsen, A., Nguyen, H. T., Lepperod, A. J., Bach, K., Votis, K. & Tzovaras, D. (2019), Visual analytics for exploring air quality data in an ai-enhanced iot environment, *in* 'Proceedings of the 11th International Conference on Management of Digital EcoSystems', pp. 103–110.

Kampa, M. & Castanas, E. (2008), 'Human health effects of air pollution', *Environmental pollution (Barking, Essex : 1987)* **151**, 362–7.

Lim, L. L., Hughes, S. J. & Hellawell, E. E. (2005), 'Integrated decision support system for urban air quality assessment', *Environmental Modelling and Software* **20**(7), 947–954.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P. & Wießner, E. (2018), Microscopic traffic simulation using sumo, *in* 'The 21st IEEE International Conference on Intelligent Transportation Systems', IEEE.
URL: *¡https://elib.dlr.de/124092/*

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in neural information processing systems', pp. 4765–4774.

Maag, B., Zhou, Z. & Thiele, L. (2018), 'A Survey on Sensor Calibration in Air Pollution Monitoring Deployments', *IEEE Internet of Things Journal* **5**(6), 4857–4870.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E. & Krew (2002), 'Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution', *JAMA* **287**(9), 1132–1141.
URL: *https://doi.org/10.1001/jama.287.9.1132*

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), " why should i trust you?" explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 1135–1144.

Sutton, R. S. & Barto, A. G. (1998), *Introduction to Reinforcement Learning*, 1st edn, MIT Press, Cambridge, MA, USA.