Particulate Matter Matters—The Data Science Challenge @ BTW 2019

Holger J. Meyer[1] · Hannes Grunert[1] · Tim Waizenegger[2] · Lucas Woltmann[3] · Claudio Hartmann[3] ·
Wolfgang Lehner[3] · Mahdi Esmailoghli[4] · Sergey Redyuk[4] · Ricardo Martinez[5] · Ziawasch Abedjan[4,5] ·
Ariane Ziehn[5] · Tilmann Rabl[6] · Volker Markl[4,5] · Christian Schmitz[7] · Dhiren Devinder Serai[7] ·
Tatiane Escobar Gava[7]

Abstract
For the second time, the Data Science Challenge took place as part of the 18th symposium "Database Systems for Business, Technology and Web" (BTW) of the Gesellschaft für Informatik (GI). The Challenge was organized by the University of Rostock and sponsored by IBM and SAP. This year, the integration, analysis and visualization around the topic of particulate matter pollution was the focus of the challenge. After a preselection round, the accepted participants had one month to adapt their developed approach to a substantiated problem, the real challenge. The final presentation took place at BTW 2019 in front of the prize jury and the attending audience. In this article, we give a brief overview of the schedule and the organization of the Data Science Challenge. In addition, the problem to be solved and its solution will be presented by the participants.

1 Introduction

For the second time—after BTW 2017 in Stuttgart [29]—the Data Science Challenge took place at the BTW conference series. The participants of the Challenge had the opportunity to develop their own approach to cloud-based data analysis and to compete directly against other participants.

✉ Holger J. Meyer
hm@IEEE.org, hme@informatik.uni-rostock.de

Hannes Grunert
hg@informatik.uni-rostock.de

1 Institut für Informatik, Universität Rostock, 18051 Rostock, Germany

2 IBM R & D GmbH, Böblingen, Germany

3 Database Systems Group, Technische Universität Dresden, 01062 Dresden, Germany

4 Technische Universität Berlin, Berlin, Germany

5 Deutsches Forschungszentrum für Künstliche Intelligenz, Berlin, Germany

6 Hasso Plattner Institute, Potsdam, Germany

7 IPVS, Universität Stuttgart,
Universitätsstr. 38, 70569 Stuttgart,
Germany

In this section, we would like to briefly describe the Data Science Challenge schedule and conditions before the participants present their developed approaches from the final round in detail in the following sections.

1.1 Schedule and Organization

The starting signal for the Data Science Challenge was given on October 26th, 2018 and the challenge itself lasted until the final presentation at the BTW in Rostock on March 5th, 2019. As an orientation, we referred to the first Data Science Challenge at BTW 2017 in Stuttgart. The participants had to pass through the following four stages [9]:

Stage 1: Application Task In order to apply for the Data Science Challenge, the participants first had to solve an application task. The focus of the task was on the use of the given data set with the measured particulate matter values. To ensure that the results are as diverse as possible, participants should limit themselves to data from their University city or the surrounding area. In addition, the participants were free to choose the analysis tools and algorithms. It encompasses the choice of *what* to discover from the data in terms of interesting facts and patterns, actually. The participants then submitted a short contribution on the methods and results of their analyses. The application task was not

only used to select the participants, but also to prepare the teams for the final task.

Stage 2: Review of Contributions After the application deadline was postponed from the end of November to the pre-Christmas period, we received a total of five contributions from various research groups in Germany and additionally an informal inquiry by a senior researchers team. As the contributions were of very high quality and, above all, produced different analyses and results, all five submissions were accepted by the committee chair.

Stage 3: Final Task One month before the BTW conference, the announcement and provision of the data sources and the final task for the Data Science Challenge had been carried out. Until the live presentation, the participants had time to adapt their approach to the new situation.

Stage 4: Live Presentation Finally, the results were presented at the BTW 2019 in Rostock. The participating groups each had 30 minutes to present their results to the audience and a jury of experts consisting of representatives from research and industry. The jury: Holger Meyer (University of Rostock), Stefan Goers (TÜV Nord, Environmental Services), Daniela Nicklas (University of Bamberg), Kai-Uwe Sattler (TU Ilmenau), Holger Schwarz (University of Stuttgart), Tim Waizenegger (IBM R & D GmbH Böblingen) and Rajko Zschiegner (OKLab Stuttgart).

1.2 The Task

The topic and task for this year was particulate matter pollution in major German cities and the associated driving bans and air pollution. Since there was a lively discussion about the pros and cons of driving bans during the preparation phase and the application period, the focus of the given example tasks and the contributions submitted was close to these discussion topics. Thus, the contributions tried to predict how the fine dust pollution could look like in the near future and which "No go" areas would result from the fine dust pollution.

At the beginning of 2019, the political and social climate changed: The measurement methods of the state sampling stations were questioned. It is (still) criticized that the stations were placed incorrectly and that the consequences for health due to particulate pollution remain unclear. At this point, we as computer scientists, especially as database researchers, are confronted with one problem that is only too well known: the lack of a data basis, since the network of official measuring stations in Germany is relatively thinly staffed.

The Citizen Science project luftdaten.info is dedicated to the large-scale measurement of particulate matter in ad-

dition to the official measuring stations. So-called sensor godfathers install self-built measuring instruments worldwide. From the collected and submitted data, a constantly updated particulate matter map[1] is generated. In addition, since October 2015 the project has archived the measured data[2]. As in the application task, this archive was used as the primary data source.

For the final data source, the authors had to select a metropolitan area or a city for their analysis that is affected by driving bans and has a sufficient density of sensors (such as Stuttgart, Hamburg or the Ruhr area). First, they had to make sure they were using only trusted sensors that didn't provide false readings and had few flickers in their records. The data from luftdaten.info should then be blended with other data from the German Weather Service (DWD), open traffic datasets or the German Aerospace Center (DLR).

At the beginning of 2019, the report "Das Diesel-Desaster"[3] was published in the media. In this documentary, different questions are raised:

- How has the traffic volume in cities with driving bans changed?
- Is there a correlation between life expectancy and particulate pollution? Can the previous studies by the Helmholtz Institute [5] be confirmed?
- In addition to traffic, what other factors play a role in particulate pollution? Identify public events that lead to a short-term increase in particulate pollution.
- How can a more uniform picture of particulate pollution be created at national level by integrating several measurement data?

Based on these questions, the participants had to create their own analysis, which would answer a question of social relevance.

1.3 List of Participants

The following groups participated in this year's Data Science Challenge:

Team **TU Dresden**: Lucas Woltmann, Claudio Hartmann, Wolfgang Lehner [31]
Team **ScaDS Leipzig**: Georges Alkhouri, Moritz Wilke [1]
Team **TU Berlin & DFKI & HPI**: Mahdi Esmailoghli, Sergey Redyuk, Ricardo Martinez, Ziawasch Abedjan, Ariane Ziehn, Tilmann Rabl, Volker Markl [7]
Team **Uni Ilmenau**: Stefan Hagedorn, Kai-Uwe Sattler [10]

Team **Uni Stuttgart**: Christian Schmitz, Dhiren Devinder Serai, Tatiane Escobar Gava [23]

1.4 Findings and Summary

The evaluation and selection of the winners by the jury included the following criteria: (i) Novelty and feasibility of the results; what is the potential for avoiding particulate matter?, (ii) Completeness/ Magnitude of the results, (iii) Social relevance, (iv) Data visualization, and (v) Live presentation on March 5th, 2019 at the BTW in Rostock.

During the meeting of the jury after the live presentation, two aspects were identified: First, it seems unfair to compare student results with the advanced work of doctoral (PhD) students. Therefore, it was decided to award the prize money in two categories, *students* and *PhD students*. Secondly, very different analyses and problem solutions were submitted due to the broad scope of the task. It was not easy for the jury to classify them in a direct comparison, which led to the decision to award two first prizes in the *PhD students* category. Therefore, three equal first prizes were awarded, endowed with €500 (students) and €250 (PhD students).

In the category *students*, the team from Stuttgart won the first place. By cleverly integrating further data sources, it was possible to establish interesting relationships between particulate pollution and the games held by VfB Stuttgart. In addition, the team provided the first approach to avoiding particulate matter in residents' dwellings.

In the *PhD students* category, the groups from Dresden and Berlin took the first place. The Dresden group focused on the four aspects of data preparation, visualization, prediction and the determination of "No-go" areas around Dresden and Stuttgart. By consulting an expert from the Fraunhofer Institute for Verkehrs- und Infrastruktursysteme (Transport and Infrastructure Systems), it was possible to uncover additional false correlations and prevent wrong conclusions. The team from Berlin also dealt intensively with data preparation, but focused more on the search for explanations for particulate pollution based on air traffic, events, weather and road conditions. Furthermore, the system tested in Berlin can be applied to other cities in order to identify the causes of particulate matter.

In the following three sections, the award-winning groups describe their solution concepts and the technologies and algorithms used. We very much hope that the format of the Data Science Challenge will continue to take place in the coming years within the scope of the BTW and will be met with a positive response.

2 Prediction of Air Pollution with Machine Learning *(winner students category)*

Christian Schmitz, Dhiren Devinder Serai, Tatiane Escobar Gava

2.1 Introduction

According to the World Health Organization (WHO) [30], urban air pollution increased by more than 8% between 2008 and 2013, despite all efforts on improving air quality in many countries around the globe. Urban air pollution may lead to a number of diseases, including reduced lung function, respiratory infections, and aggravated asthma. The OK Lab [24] provides data for Particulate Matter (PM)—inhalable particles with diameters of 10 micrometers and smaller [28]—that can be used to find patterns and extract information of its distribution over time and space across Germany.

Despite data availability, the automated analysis of air pollution data is difficult, since it requires in-depth domain knowledge. Therefore, we have developed an application for visual analysis of particulate matter concentration. This application helps us and domain experts to understand and analyze the Particulate Matter data. We employ this app to derive four insights. Based on these insights, we suggest smart windows that automatically open when Particulate Matter concentration is low outdoors.

In this article, we first present the steps of data cleaning and exploration used in the project for BTW Data Science Challenge 2019. Then we describe the insights we obtained from our visual analysis application. Finally, we suggest smart windows for minimizing Particulate Matter concentration indoors.

2.2 Data Sources

We use the OK Lab data [25] as the main source for air pollution information. We collect all measurements from DHT22 sensors, which provide temperature and humidity; and SDS011 sensors' measurements for Particulate Matter, provided as PM10 (particles that generally are 10 micrometers or smaller) and PM2.5 (particles that generally are 2.5 micrometers or smaller). Both sensors collect data every 3-5 minutes.

In order to add more attributes, we use [6] data regarding wind direction, precipitation and air pressure. These values are collected on an hourly basis. Further note that the sensors measuring these values are distributed in lower quantities than the sensors from the OK Lab dataset [25].

We further use [19] to add current weather information to our application. This source is a widely used and reliable source for current and forecast weather information.

	Unnamed: 0	sensor_id	sensor_type	location	lat	lon	timestamp	P1	P2	date
0	370821	164.0	SDS011	72.0	48.773	9.174	2016-09-17 09:05:26.362	5.80	2.90	2016-09-17 09:05:26.362000+00:00
1	370822	164.0	SDS011	72.0	48.773	9.174	2016-09-17 09:02:13.125	5.30	2.63	2016-09-17 09:02:13.125000+00:00
2	370823	164.0	SDS011	72.0	48.773	9.174	2016-09-17 09:07:17.682	3.58	2.62	2016-09-17 09:07:17.682000+00:00
3	370824	164.0	SDS011	72.0	48.773	9.174	2016-09-17 09:09:57.200	4.32	2.55	2016-09-17 09:09:57.200000+00:00

		sensor_id	location	timestamp	P1	P2	
	1	164.0	72	2016-09-17 09:00:00	6.500000	2.853611	

Fig. 1 Data Standardization The colors indicate columns that were kept in the intermediate dataset. This is further joined with data from [6] to compose the final dataset

2.3 Data Preparation

For data cleaning, we first filter the data in order to narrow the results to Stuttgart area. This allows us to remove large amounts of data (approximately 75% of the original data size) that is not needed for our study case and enables us to find the results faster.

After filtering, we identify and remove incorrect data by applying the following two approaches: (i) finding measurements which were taken under certain conditions which invalidated the measurement itself; and (ii) identifying outliers in measurements that are allegedly correct, e.g., measurements of malfunctioning sensors.

Finding measurements which were taken under unreliable conditions is defined by looking into the sensors' datasheets and finding their ideal working environments. We obtain all variables to identify unreliable conditions in the collected datasets, so removing the invalid measurement values is straightforward.

To find outliers within the remaining measurements, we employ the InterQuartile-Range (IQR) [27] to identify values that obviously do not fall in a plausible range. We apply the IQR approach for measurements on a daily basis so that different environmental characteristics do not significantly influence the detection of outliers. This approach allows us to use data from malfunctioning sensors until the day before they start measuring incorrect values.

2.4 Data Fusion

In the Data Fusion step, we reduce the SDS011 and DHT22 data tables to hourly measurements, averaging the attributes such as temperature, humidity and particulate matter (PM10 and PM2.5). Afterwards, we join both tables and the additional information of air pressure, precipitation and wind direction on timestamp and location. Finally, we get a materialized and integrated dataset, which we make use of for our visual analytics application.

2.5 Visual Analytics Application

We develop a visual analytics application to help us and domain experts to understand the integrated data and obtain insights from that data.

The application consists of a cloud-based webserver developed in Python using Dash [21], a Visual Analysis Framework that is used for data communication, plot generation and user interaction. Dash supplies default plot generation embedded on itself, but we decided to integrate Folium [8], the best-fit library that supports robust heatmaps visualization integrated with Google Maps.

When using the VA application, we can pick dates and regions and generate heatmaps of Particulate Matter 10 concentration in that zone for both locations and dates. This enables us to identify the influence of events such as traffic, festivals, and sport matches in the increase of air pollution in specific regions of a city.

2.6 Findings and Insights

By using the application's tools, we were able to discover insights and suggest helpful recommendations to reduce accidents in the future. In the following parts, we introduce four of those findings that cover the period 2015 to February 2019.

Traffic. In Fig. 2, the considerably constant Particulate Matter values on Sunday and the peak values on Friday suggest that traffic, particularly during rush hours on weekdays, is one of the key factors for air quality degradation. We analyze data from further days and noticed that Air Pollution strongly correlates to traffic peak hours, meaning that at rush periods air quality can decrease to approximately half (in terms of PM10 values, as can be seen in Fig. 2 if compared to non-rush periods.

Season. Season also plays an important role in air quality. The average of PM10 values are four times higher in winter when compared to summer, and also average values of PM2.5 concentration particles is 2 times higher in winter than summer. One factor that explains the higher val-

Fig. 2 P1 index vs Hour. The plot shows two subsequent days and their distributions of Particulate Matter 10 over time. Sunday is represented by the pointed line and Monday, by the straight line
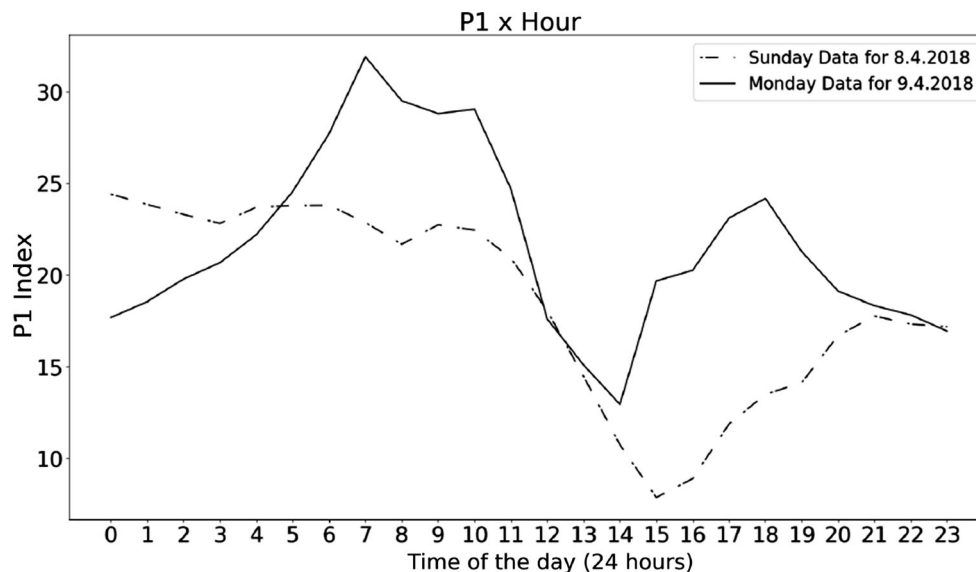
ues during the winter are the lower average temperatures. Another factor may be the heating mechanisms used inside houses, e.g., fireplaces, which liberate gases that again help decreasing air quality specially in highly populated zones (e.g., City centers).

Wind Direction. We have analyzed the impact of the wind direction on the air pollution in the Stuttgart area. Since Stuttgart is located in a valley, it is hard for air to flow through the valley when it comes from certain directions. This can be seen in Fig. 3. The first image, from [26], shows Stuttgart's topology. The second shows a compass rose indicating Particulate Matter 2.5 and 10 indices when air is flowing in the respective direction. The lower (dark pink) part of the topology represents mountains that block air to flow when going South, which entails in higher Particulate Matter stagnated in Stuttgart's city center. On the one hand, it is visible that there is a significantly lighter presence of polluted particles, when the air flows to North (light pink/yellow).

Events. Regarding events, our analysis indicates that soccer matches are also partially responsible for poor air quality. As displayed in Fig. 4, on a game day (01/09/2018) the air quality in Stuttgart is not as good as on the non-game day (25/08/2018). This holds not only for the near vicinity of the stadium itself, but also for the area roughly 8km away from the stadium.

2.7 Smart Windows

Based on the insights we collected, we propose a Smart Window approach for reducing the amount of polluted air breathed by people. The windows will open, close and calculate the aperture dynamically, based on current and past weather and Particle Matter data. This will enhance people's

life quality by decreasing pulmonary diseases and other effects caused by polluted air.

2.8 Future Work

We plan to extend our work on metropolitan areas all across Germany, where many sensors are available, enabling better understanding of air pollution distribution over bigger areas as well as finding specific insights for more cities. Also, we want to develop a prototype of the Smart Window, based on the concepts described in this article and using Internet of Things facilitators such as a Raspberry Pi. This can be the beginning of an important step towards inserting smart devices into people's homes aiming at a future with better life quality with lower efforts.

3 Assessing the Impact of Driving Bans with Data Analysis *(winner PhD students category)*

Lucas Woltmann, Claudio Hartmann, Wolfgang Lehner

Assessing the impact of suspended particular matter in city areas has become a topic of discussions over the last years. With the emerging field of data analysis, we have the opportunity to combine the two areas to generate insights from environmental data. With the help of an domain expert, we introduce an approach which generates visualizations and predictive models to give insights into the different influences on particle concentrations. Our work aims to provide a tool for transferring knowledge across domains. This information transfer process is completely data-driven.

Fig. 3 Stuttgart wind direction
insights. The upper image is
a topology map of Stuttgart. On
the bottom, there are radial plots
of average PM2.5 and PM10
distributions over wind direction
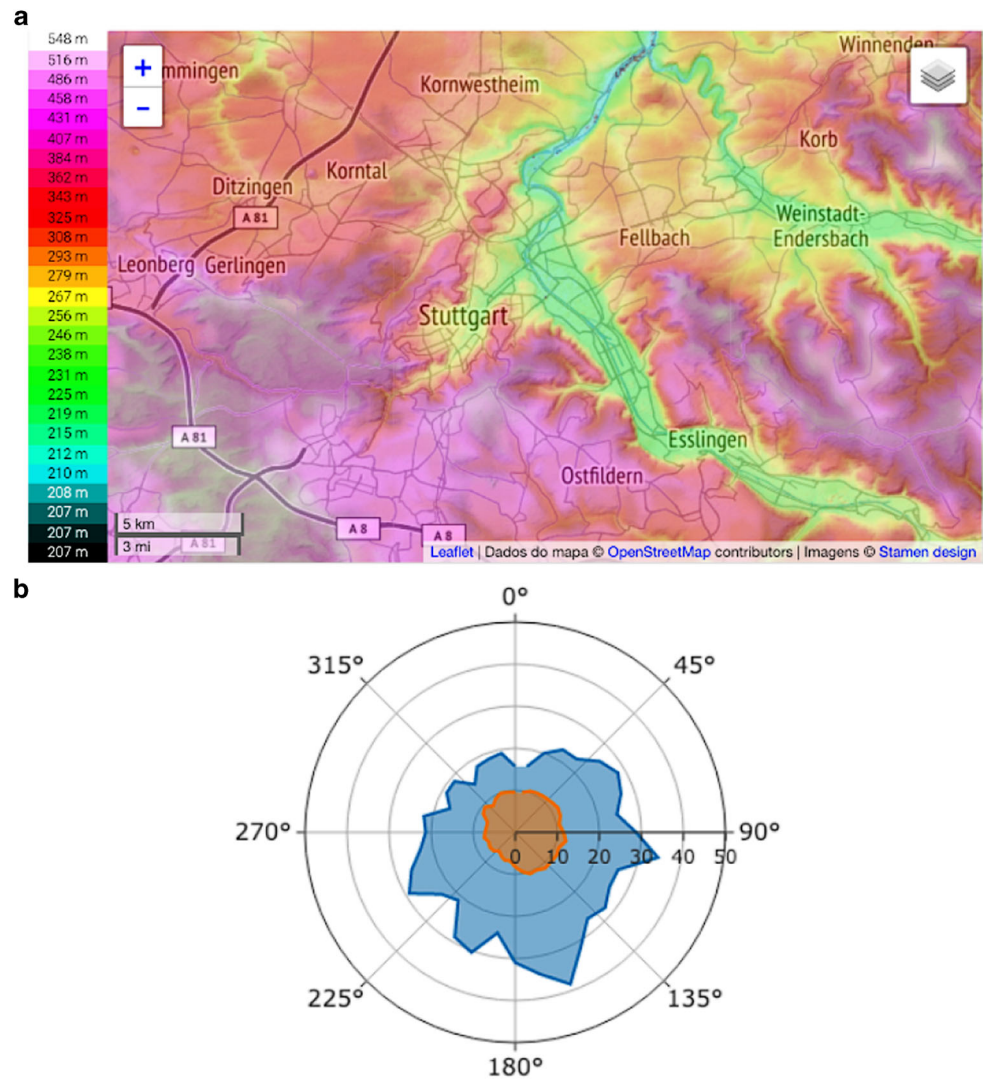
3.1 Introduction

As defined by Drew Conway[4], *data science* consists of three
parts: *hacking skills*, *math and statistics*, and *substantive
expertise*. We are confident that we can cover the first two
parts but when it comes to domain knowledge, we cannot
offer any insights about suspended particular matter and its
properties. This brings us to the most important aspect of
our approach: the use of external expert knowledge to ver-
ify our analysis results. For this, we ask a domain expert
to review our analysis results. This is why large parts of
our analysis are based on visualization. We believe, com-
munication of information via visualization is a key feature
in data science. Images and plots make it easier to transfer
information across domains. Due to this fact, our domain

expert was even able to correct mistakes made by us during
a preliminary analysis. Whenever our expert has a remark or
we correct a wrong interpretation, we will clearly state this
in the following text. We focused on the cities of Dresden
and Stuttgart because of their similar valley topology and
the introduction of a driving ban in Stuttgart in January
2019. This gives us the opportunity to assess the impact
of such a driving ban with the support of our analysis and
expert knowledge.

Our work is split into three parts. First, we show how
we integrated the different data sources in Sect. 3.2. In
Sect. 3.3, we will present two visualization techniques.
These are *time series plots* for finding recurrent patterns
and *distribution maps* for finding irregularities in the data.
Last, we give a short introduction to our SPM modeling
with forecasting in Sect. 3.4.

---

4 http://drewconway.com/zia/2013/3/26/the-data-science-venn-
diagram.

Fig. 4 Heatmap displaying Particulate Matter 10 values for two different timestamps in Stuttgart city center. The figures illustrate the influence of a VfB Stuttgart home game in the city's on the air quality. The upper map shows PM10 values for a VfB Stuttgart home game which took part on Saturday, 1st of September of 2018. The lower map shows values from exactly one week before, when there was no VfB Stuttgart home game taking place
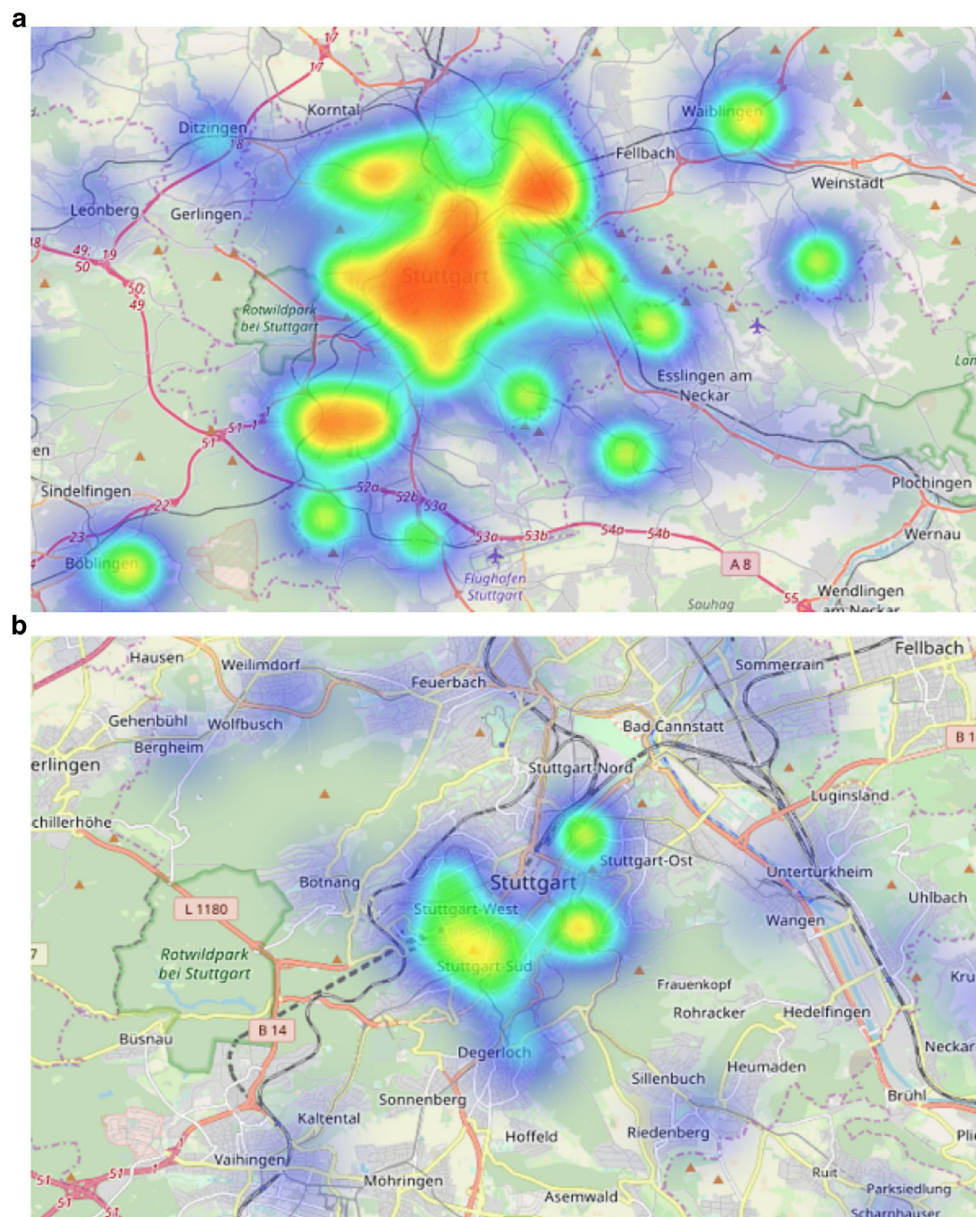

3.2 Data Cleaning and Preparation

First, the data of all sensors in Dresden and Stuttgart is collected. We use a nearest neighbor search around the center of both cities with a radius of 10km to retrieve all sensors in both city areas.

A general problem is the division of sensor types into SPM sensors and temperature/humidity sensors. To construct a common base for the analysis, we need to integrate the SPM sensors and the humidity/temperature sensors into one data set. Therefore, we aggregate the data for each sensor type to a minute-wise time scale by averaging all measurements within every minute for each sensor. This provides a common minute granularity for all sensors. The next step is to merge the temperature and humidity data

with the particle concentration data by sensor location and time. Note that this is only possible in a standardized minute granularity in both sensor type data sets.

From the resulting table, we remove all particle concentrations where the humidity is larger than 70%, because, above this value, the particle sensors do not provide reliable readings [18]. This leads to sparse data for Dresden and Stuttgart because both cities have a rather humid river climate as our expert notes. Additionally, both cities are located in a valley. Therefore, the humidity stays at the bottom of the valley whenever the layers of air are incapable of mixing. This effect is known as *thermal inversion*. This means that a denser layer of air cannot exchange particles with the upper layers. Our expert says that not only the humidity is higher at ground level, but also the parti-

Fig. 5 Particle concentration PM10 compared to humidity. **a** Measurements Dresden for one day, **b** measurements Stuttgart for one day
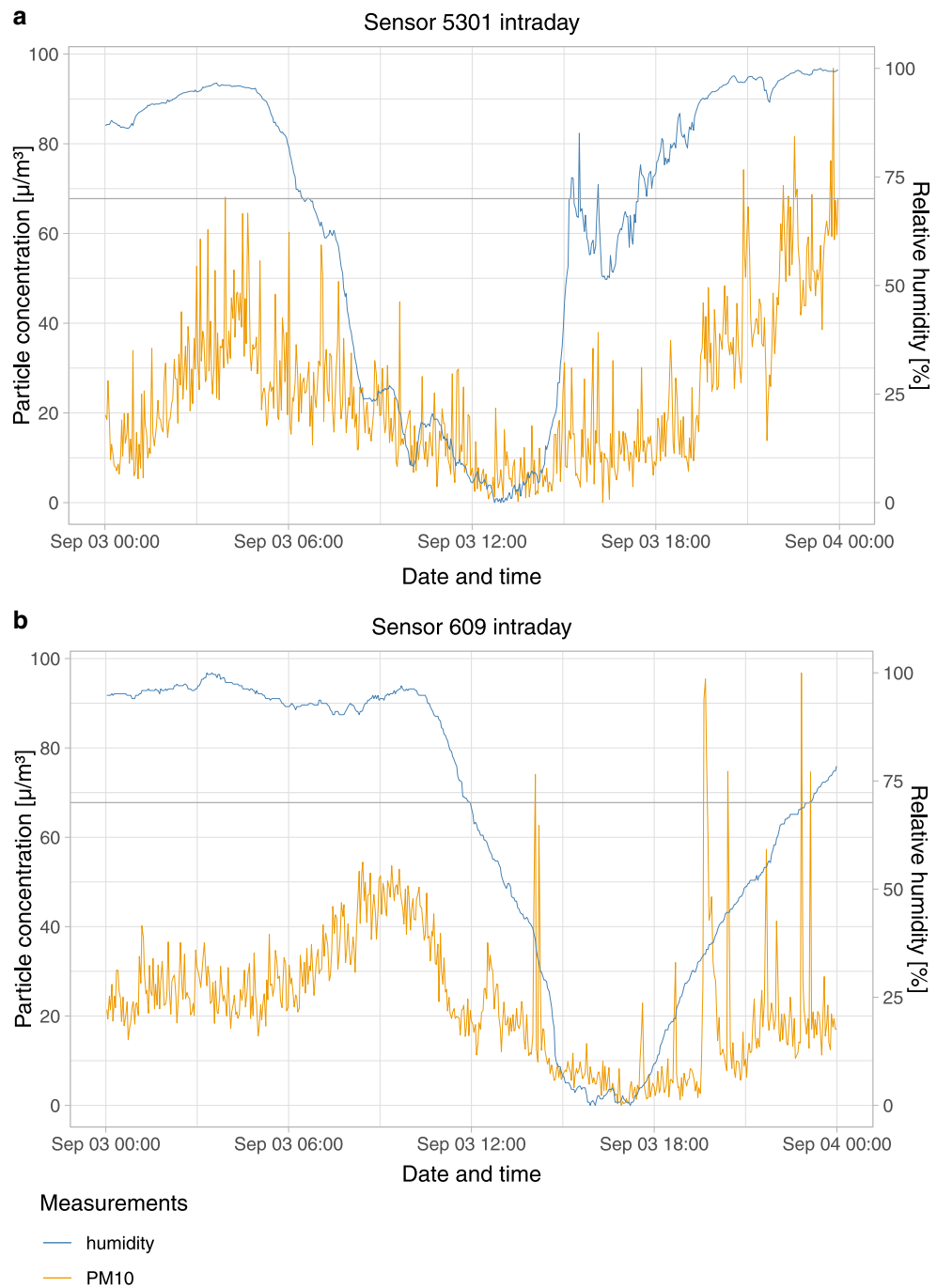
cle concentration. This problem is shown in Fig. 5 which compares the particle concentration PM10 to the humidity at one sensor in each city close to the river. The 70% limit for the humidity is marked with a black line. It is visible that approx. 50% of the data is lost.

Finally, we impute all missing values with linear interpolation if the gap between two timestamps is not too wide. Otherwise, we drop the values from the data set. This has proven to be a very robust approach in the field of time series analysis [16].

3.3 Visualization

Visualization is a common approach to find interesting patterns. We present two different kinds of visualizations helping with the identification of structure in the data. Time series plots show the value of a time-dependent variable over a given time span. This can be useful for finding patterns in data sets like seasonal spikes in shopping behavior or temperature measurements. The other form of visualization are distribution maps. They offer spatial information about
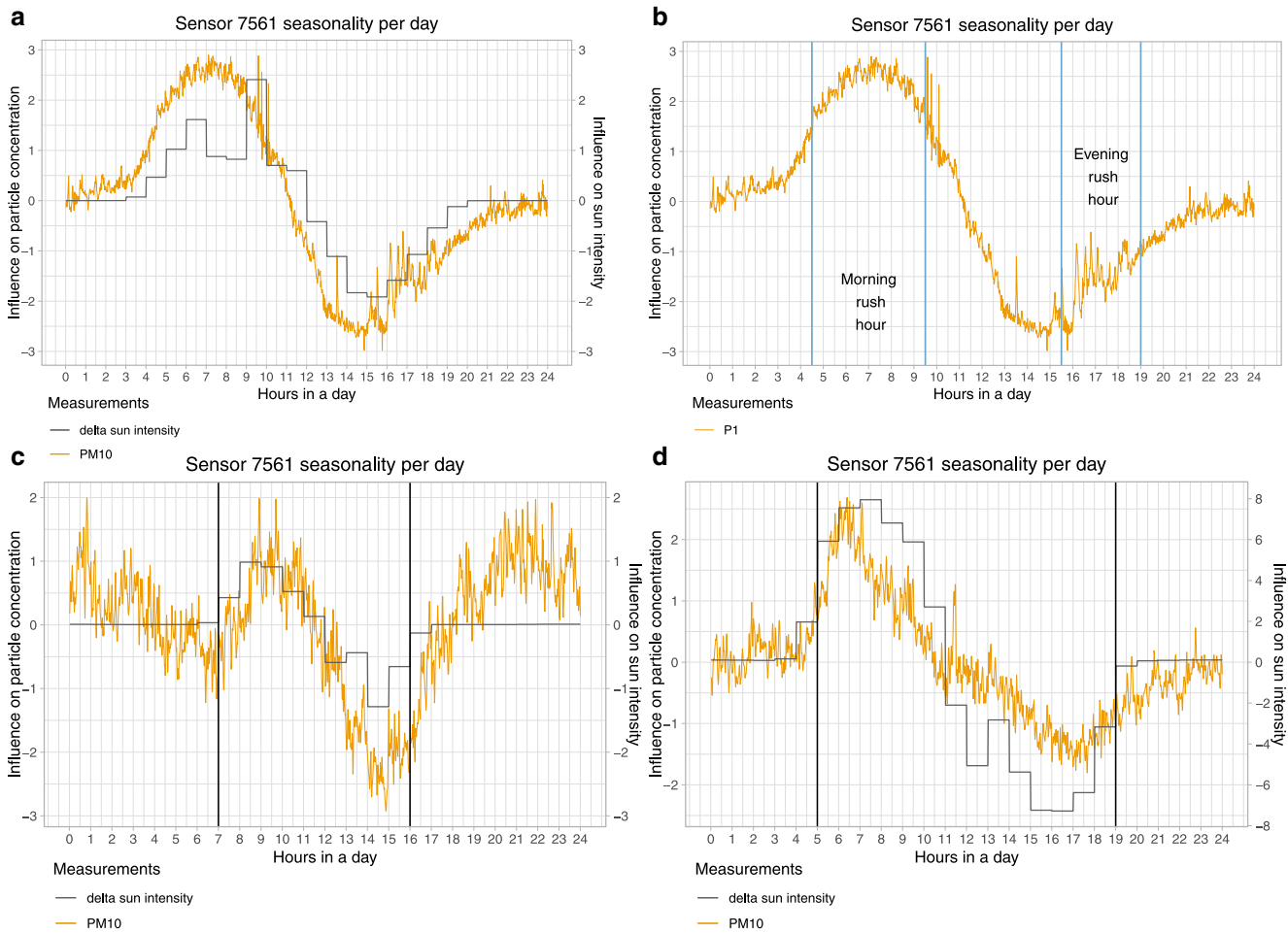
Fig. 6 Particle concentration PM10 for sensor 7561 (Stuttgart). **a** Influence of delta sun intensity, **b** wrong interpretation of peaks in PM10, **c** influence of delta sun intensity in January, **d** influence of delta sun intensity in August

a variable. Distribution maps have colored areas according to the intensity of the visualized variable. This gives the possibility to analyze the distribution of a measurement, like particle concentration, in a certain area.

3.3.1 Time Series Plots

First, all time series get split into trend, season, and residual components with time series decomposition [4]. Here, one time series represents one PM10 sensor in either Dresden or Stuttgart. The seasonal component details peaks and valleys in the time series which occur in a regular pattern. This component shows the recurrent influence of each point in time on the time series, as shown on the y-axis. We have chosen 24 hours as the seasonality of every time series because we are interested in patterns occurring every day. This leaves us with a seasonal influence for the particle concentration for each hour of a day.

Figure 6a shows the seasonal component of the particle concentration at one sensor in Stuttgart for different hours in a day. We thought, we can clearly show the morning and

evening rush hours as two peaks in the seasonal component of the PM10 sensor data. The sensor is located in the center of Stuttgart, so this conclusion seems obvious [31]. However, our domain expert says that this is a spurious correlation and a common misinterpretation. Because the average seasonal peaks of the particle concentration occur at the same time as the rush hours, one could assume that traffic is the major cause for the high particle concentration. Extensive research [12] has shown that the major influence is in fact the change in sun intensity. Through a large change, also referred to as delta, in sun intensity, particles are set free from a solid object like the ground. The highest deltas are achieved during sunrise and sunset every day. So, the seasonal component of the particle concentration should only follow these changes. This can be shown by comparing a bright and a dark month. We choose January 2018 as a dark month and August 2018 as a bright month. Figure 6c and d show the shift of the seasonal component of the particle distribution if it is calculated only from the data in this month. The same applies to the delta sun intensity. With the average time for sunrise and sunset per month marked with
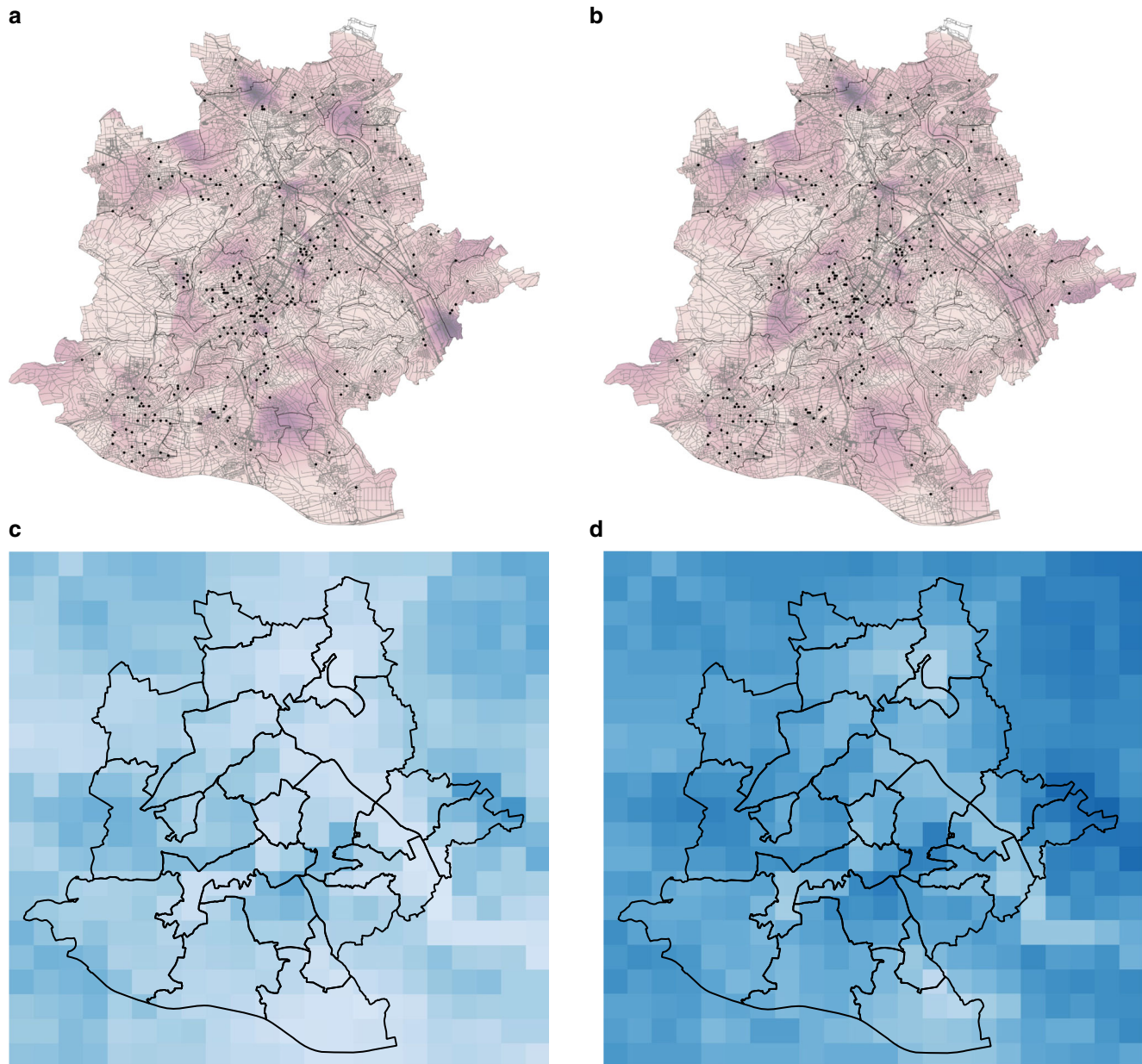
a

b

c

d

Fig. 7 Particle concentration PM10 and precipitation over Stuttgart in different months. **a** Particle concentration PM10 in January, **b** particle concentration PM10 in August, **c** precipitation in December, **d** precipitation in January

vertical black lines, the figure details the direct impact of different times of large deltas in sun intensities on particle concentration. The particle concentration directly follows the monthly pattern of brightness and therefore models the times for sunrise and sunset. The particle concentration only follows the rush hour patterns if one looks at the averaged seasonal component over a whole year. Note that the particle concentration still follows seasonally the average delta sun intensity in the annual data (see Fig. 6b). Even the outliers in both lines correlate. Therefore, we can confirm that not traffic, as assumed before, but the changes in sun intensity have a strong influence on particle concentration.

3.3.2 Distribution Maps

As a second visualization, we will map the distribution of particles in city areas. The darker the shade of an area, the higher the particle concentration. For visualization, the concentration is smoothed using cubic splines [20]. The black points on the map are the locations of the SPM sensors. We will use these maps for assessing the impact of the driving ban in Stuttgart by comparing the distribution before and after its introduction. Furthermore, we will use this plot as one of the tools to identify interesting areas of high particle concentration in Dresden.

Fig. 8 LSTM forecast for aggregated PM10 in Stuttgart
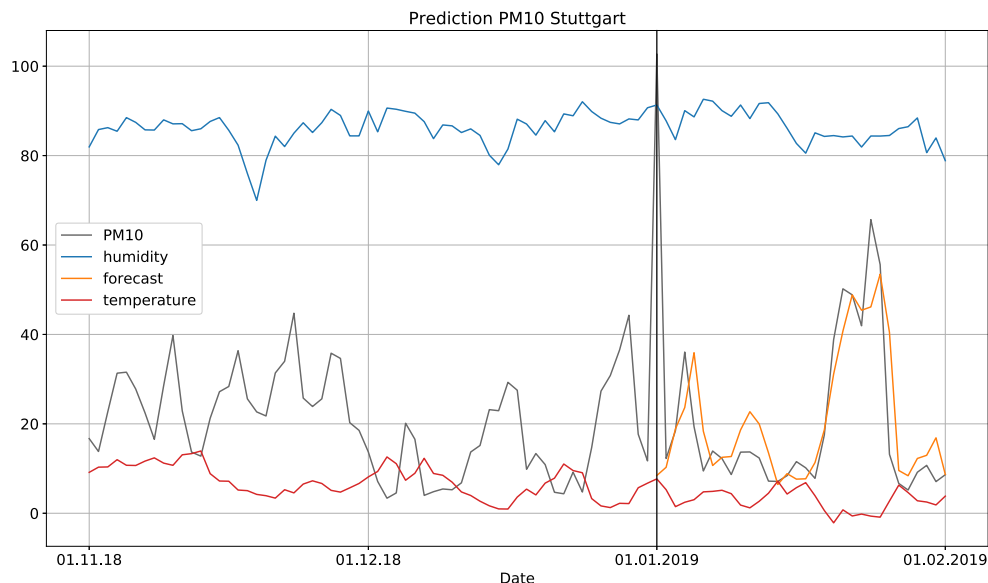
For now, we can compare the distribution of particles in Stuttgart before and after the ban. The plot in Fig. 7 highlights particle hotspots in Stuttgart for December 2018 and January 2019. Figure 7b compared to Fig. 7a might lead to the same assumption as our initial mislead interpretation. We assumed, the general pollution decreased with the ban [31], because Fig. 7b displays a lighter hue for the complete city area and smaller areas with dark hue. Again, this interpretation was wrong and our expert says that this improvement is not due to less traffic in the city, but due to the influence of precipitation. Precipitation has the property to bind particles from the air and bring them back down to the ground. This results in much cleaner air after either a strong rain shower or a long period of drizzle. So, the most important influence is the amount of rain per month and the number of rainy days per month. To get a combined measure we multiply the amount of rain on a square kilometer in a particular month in Stuttgart with the number of rainy days in Stuttgart in the same month. Figure 7c and d show this factor for the whole city of Stuttgart for December 2018 and January 2019. These figures are based on the data from the Deutscher Wetterdienst (DWD)[5]. Here, the change in hue is inverse to the one for the particle concentration. There is more precipitation in January 2019 than in December 2018. The decrease in hue for the particle concentration indirectly correlates with the increase in hue for the precipitation factor. This shows that more precipitation can bind more particles and therefore reduces particle concentration. Given all the other meteorological influences, it shows how little the impact of a driving ban is regarding this aspect.

A lot of the interpretations of data from Stuttgart cannot be transferred to the data from Dresden, mainly due to the sparsity of sensors. We still let our expert label the dark spots with high particle concentration in Dresden. Figure 9 shows the labeled map. Dresden has some interesting areas which we explain in a short overview. The mountain symbols show areas with steep hillsides being the typical terrain for high particle concentration due to inversion. This effect is amplified by the industrial sites near the hillsides where pollution from the industrial chimneys is pushed towards the ground because of the inversion. Another negative effect is produced by tunnels. Their exits, located in the south west of Dresden, emit all the particles from the tunnels' interior. This leads to high measurements in these areas. An-

Fig. 9 Areas of interest in Dresden

other problem area is the river harbor. Due to unclean ship engines and industry, there are high particle concentrations around the harbor. The large dark area in the north west of Dresden is the depot for a historical steam railway having a high smoke output.

Given our analysis results, we can show that visualization is a powerful tool for assessing influences of particle concentrations. We see high potential for our visualizations to be applied to other cities and use cases. This also helps domain experts from other areas of research making decisions and drawing conclusions from data.

3.4 Forecasting with Meteorological Influences

Forecasting is an important technique to assess the impact of different influences on a target variable. This also applies to the influence of meteorological data on particle concentrations. To partially justify the results of previous forecasting models of PM10 values with neural networks [12], we build our own predictive model. Unlike the original publication, we used a Long Short Term Memory (LSTM) neural network. The network is conceived in an autoregressive way, where the inputs are the previous PM10 values, the current humidity, and the current temperature. This will show the meteorological influence of some factors on the particle concentration. Figure 8 details the data for training from December 2018 and the data for testing from January 2019 divided by the vertical black line. We used the overall or aggregated measurements on a daily basis for particle concentration PM10, humidity, and temperature for the whole city of Stuttgart. The predicted PM10 in orange follows the source PM10 in gray for January 2019. This means that a model purely relying on meteorological influences can already represent cause and effect for particle concentration. This is identical to the findings in [12].

3.5 Conclusion

We have introduced several techniques for presenting analysis results to experts and for improving one's own analysis with domain knowledge. We have shown a multi-tool workbench for assessing the influences on particle concentrations. We used visualization, like time series plots and distribution maps, and forecasting to find interesting patterns. This is done with a purely data-driven approach. This is why we think our work can be extended to different cities, but also to a different set of problems. Any data which can be represented both as a concentration distribution and a time series will be assessable with this approach. Topics closely aligned to this work would be weather and climate research, market research, and energy research. We see our approach as an important tool for knowledge transfer between persons from different areas of research. The

main goal would be to give domain experts a robust analysis platform for their decision making. Given the importance of data science and data analysis, we argue that it is fundamental to communicate results to experts and use their knowledge to enhance the analysis.

4 Explanation of Air Pollution Using External Data Sources *(winner PhD students category)*

Mahdi Esmailoghli, Sergey Redyuk, Ricardo Martinez, Ziawasch Abedjan, Ariane Ziehn, Tilmann Rabl, Volker Markl

4.1 Introduction

During the last years, high emission of fine-grained particles into the atmosphere and its negative impact on people's health and well-being has attracted the attention of researchers and governmental agencies to look for the causes of air pollution in different neighborhoods [17]. Serious measures have been taken in order to reduce the levels of air pollution, such as the introduction of fine-grained particle concentration thresholds or driving bans for vehicles that use diesel engines in several European cities [22].

When it comes to current approaches on predictive modeling in the area of air pollution, many focus on estimating the concentration of fine particulate matter in the nearest future in a particular area [3]. However, identifying the cause of high emission of fine particulate matter, as well as finding its potential sources can provide decision makers with valuable information for the design of counter measures. Detecting the sources of air pollution and treating them is a big step toward better air quality [7].

The problem we observe is that historical records from air quality sensors that are used to forecast the concentration of fine particulate matter are not sufficient for inference of factors that are likely to cause air pollution. Intuitively, we can assume that traffic, factories and production facilities, agriculture etc. might negatively affect the air quality. To test these assumptions, we need to incorporate external data sources into the main dataset of air quality sensory readings (Sect. 4.2). For this project, we aim at designing a prototype system that (a) provides a data-driven approach for detection of potential causes of air pollution, and (b) supports data integration and merging of external data sources (Sect. 4.3). The insights that we collected by using the prototype demonstrate competitive advantages of our approach (detection of potential causes of air pollution) w.r.t. existing work on prediction of future levels of air pollution (Sect. 4.4).

4.2 Data

The main source of the data for this project is provided by the Luftdaten service[6]. Important features that we use are the P1 concentration of fine particulate matter and spatio-temporal data from the sensors. However, these features are not sufficient to identify the sources of decreased air quality. A substantial body of research shows that traffic, factories, production facilities, human activity and many other factors contribute to the air quality and emission of particle dust [13]. These factors are not reflected in the Luftdaten core dataset. The lack of descriptive information for further analysis is a critical challenge. As one of the main contributions of the project, we incorporate the main dataset with additional descriptive features.

4.2.1 External Data Sources

As specified in Fig. 10, we use four external data sources: weather[7], geo data[8], air traffic[9], and public events.

Geo Data contains the information about urban infrastructure, streets, and parks that surround the air quality sensors. Descriptive features that depict the neighborhood of each sensor help to determine what factors might explain the source of pollution [11]. We obtain the data from the Open Street Map platform and the GeoFabrik service[10] that regularly publishes up-to-date data on geometries and other features[11] that describe the real-world geo-spatial objects (buildings, roads, etc.). For this project, we use the following features: the number of streets and street crossings within the distance of 100, 200, and 500 meters from the sensor location.

Weather Data contains readings from various weather sensors scattered across the cities. The main features that we use include temperature, precipitation, humidity, pressure, wind speed and wind direction (the direction that wind hits the sensor). The dataset used in this project is collected and published by the German Weather Service.

Air Traffic Data consists of the routes of airplanes flying to/from the airports. We look for any overlaps between the usual routes of airplanes and highly polluted areas in Berlin, in order to detect particular zones that are affected by the air traffic pollution.

Events Data contains the dates from various public events and holidays that occur in Germany (Google Calendar of Berlin). These events are mainly public holidays
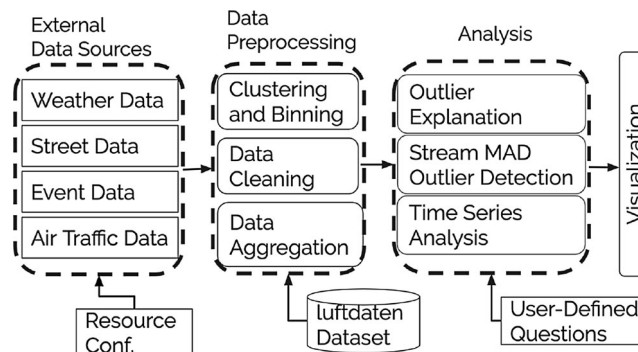

Fig. 10 Overall view of the architecture of our system

such as New Year's Eve and Easter. Besides, non-holiday events are selectively picked from online services, for instance, VisitBerlin[12]. The non-holiday events are the events or activities that attract crowds in a particular neighborhood, yet are not related to public holidays (e.g., Berlin International Film Festival). The features contain the name of the event and its starting and ending dates.

4.3 Architecture

Figure 10 depicts an overview architecture of the prototype. It consists of four main components: integration of external data sources, data preprocessing, analysis (explanation) and visualization. In this project, we focus on the first three components. As for visualization, we provide basic charts to depict discovered insights. The end-user can customize the charts to fit arbitrary use cases, which we refer to as the user-defined questions.

4.3.1 Integration of External Data Sources

The prototype can be configured to integrate various external datasets. To achieve this, the end-user specifies the foreign keys for every external data source to be joined with the core dataset For instance, in order to incorporate the temperature feature from the weather dataset to the Luftdaten core dataset, the end-user should specify the column names that contain the timestamp and geo coordinates for both datasets. The system then can use this configuration to join the datasets with the provided features as foreign keys. For the Luftdaten use case, we use spatio-temporal coordinates, aggregated to the 5-minute time intervals and 100-meter radius neighborhoods, as foreign keys for further integration. We apply aggregation techniques in order to reduce the granularity of timestamps and geographic coordinates, and achieve exact matches of these attributes as foreign keys in between data sources.

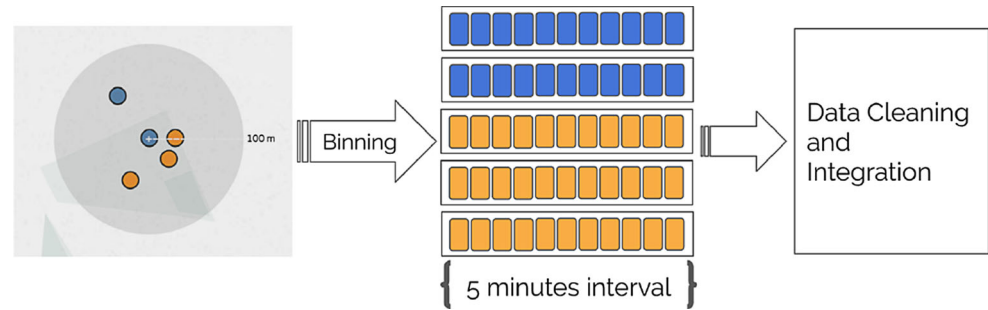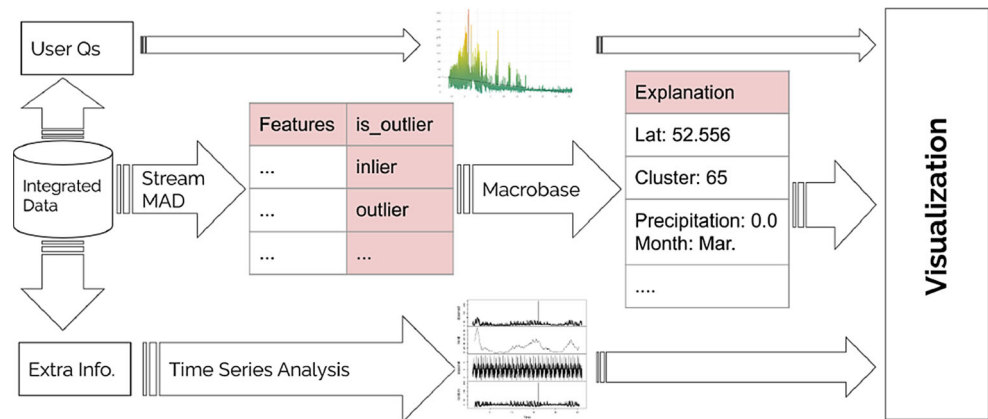**Fig. 11** Clustering and binning techniques

Fig. 12 Analysis phase

4.3.2 Data Preprocessing

The data preprocessing component consists of the following operations: clustering, binning, cleaning, and aggregation (Fig. 11). We apply spatial clustering and temporal binning in order to (1) synchronize sensory readings and (2) cross-validate them, discarding untrustworthy sensors. For instance, defining a 100-meter radius neighborhood that contains several sensors lets us compare the readings of these sensors and detect deviating patterns. We choose the radius empirically under the assumption that sensors located close to one another record similar signals.

In the data cleaning phase, we use the MAD (Median Absolute Deviation) [14] algorithm for univariate outlier detection, to remove noise from the data. Values with high variance w.r.t. other readings inside each cluster and each time interval are removed as outliers (e.g., errors or untrustworthy readings). We also remove data points that are recorded under particular weather conditions, such as high humidity. That is done due to recommendations from the sensor specification[13].

The cleaned data is then used during the data aggregation phase. In this phase, we create one data point for each cluster and each time interval. This data point contains the

min, max, average, standard deviation and median values for every attribute present in the dataset. For the Luftdaten use case, these attributes are P1 concentration of fine particulate matter, temperature, precipitation, humidity, wind speed and wind direction. External data sources that depend on neither location nor time, such as public holidays (e.g., New Year's Eve) or the number of roads, are stored in a separate data frame to reduce data redundancy.

4.3.3 Analysis

Figure 12 shows the overall architecture of the analysis component that incorporates outlier detection and explanation, time series analysis, and user-defined questions. The main purpose of this component is to automatically detect deviating patterns that differentiate data points with high concentration of fine particulate matter from normal readings, and propose an explanation based on highly correlated descriptive features.

Outlier explanation is the most important part of this component. First, the dataset is labeled (inlier/outlier) by using our implementation of the Stream MAD algorithm that enables real-time stream data processing. The current prototype processes CSV (comma separated value) files yet supports data streaming scenarios. The original, batch-mode version of MAD detects global outliers only. It is important to find local pollution peaks instead, for example,
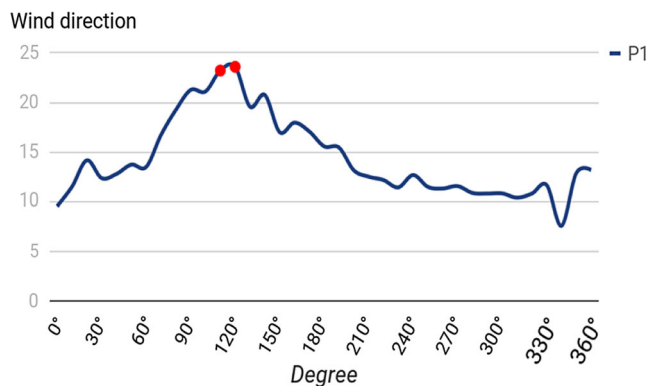
Wind direction

Fig. 13 Correlation between the direction of wind and pollution levels for the example cluster

a small amount of air pollution that is caused by the movie festival event. The Stream MAD algorithm uses a Min-Max heap to update the median value by checking every new value in the dataset. After applying the Stream MAD algorithm to our dataset, we use the Macrobase outlier explanation tool [2] to detect and "explain" outliers in the enriched feature set based on the integrated data.

Integrating the data with external data sources enables Macrobase to provide explanations for outliers with more details (i.e., additional features) available. The accuracy of Macrobase is as high as the correlation between the enriched feature set and the air pollution ratio. By adding external information and increasing the correlation, we can claim that Macrobase is able to provide explanations of higher discriminative power.

Evaluation To prove the hypothesis of increased correlation due to information gain, we train an XGBoost Regression model on the Luftdaten dataset before and after adding weather data. The experiment shows that RMSE decreased from 8.41 to 6.83 after adding weather data. These values correspond to the 18% error loss and, eventually, higher correlation between features and labels. It is worth noting that the air pollution ratio for the sensor under evaluation is in the range of 0 and 140. The average difference between the real value of air pollution and the predicted value is 6.83, which is acceptable taking into account the 0-to-140 range.

We also apply time series analysis to detect pollution patterns. The analysis considers different time spans (e.g., days, months, etc.). The detected patterns of different areas and/or time periods are compared to each other in order to facilitate the finding of explanations for lower and upper peaks of particulate matter concentration.

We built a prototype of the proposed system that allows us to add external data sources to the Luftdaten core dataset and integrate them together, to achieve more discriminative explanation of air pollution. For this use case, we focus mainly on the Berlin area, using tabular data with clearly specified foreign keys. We create a JSON configuration file that is used to store the file paths to data sources and underlying foreign keys for further joins.

4.4 Findings

Applying the prototype on the core Luftdaten dataset enriched with external data sources (Berlin area), we accumulate the following insights.

1. *Weather Impact:* In general, weather affects air quality heavily. For instance, higher wind speed leads to lower levels of air pollution as fine particular matter will be carried away. Figure 13 depicts the air pollution ratio in the example cluster and how it fluctuates w.r.t. the wind direction. As it is specified in Fig. 13, higher amount of fine particulate matter is observed when the wind direction is between 110 and 120 degrees. Further analysis of this specific example cluster is visualized in Fig. 14. We observed a highway that is located within the 110-120 segment and is likely to be the cause of air pollution.Also, temperature affects air pollution in the city heavily. To generalize the weather impact, we detected the pattern on an annual cycle and on multiple sensors, as shown in Fig. 15. This annual pattern shows that air pollution is higher during the winter and our interpretation is that this pattern is observed due to the household heating systems and inversion phenomena [32].

2. *Traffic Flows and Public Transportation:* Based on the open street map data, we figured out that most polluted areas in Berlin are around the Berlin Ring where people usually park their cars to use public transportation in order to avoid driving in the restricted environmental zone. Big train stations (cycle line) are located on the Berlin Ring area. The population density in these stations is very high. Further, the system outputs that particulate matter concentration is very high in latitude of 52.556. A specific latitude without longitude would depict a horizontal line. Interestingly, the Tegel (TXL, Berlin) airport is located on this latitude. Tracking airplanes that fly to/from the airport shows that many airplanes fly around the city before taking latitude 52.556 for landing. Thus, they cause high pollution in that area. Comparing the pollution trend between sensors located in latitude 52.556 and the other sensors is notable. Air pollution is commonly higher in winter but sensors that are affected by air traffic show higher amounts of particle concentration during the summer and around New Year, because these times correlate with high tourist seasons (Fig. 15).

Fig. 14 Location of the example cluster and the highway located between directions 110 and 120 degrees
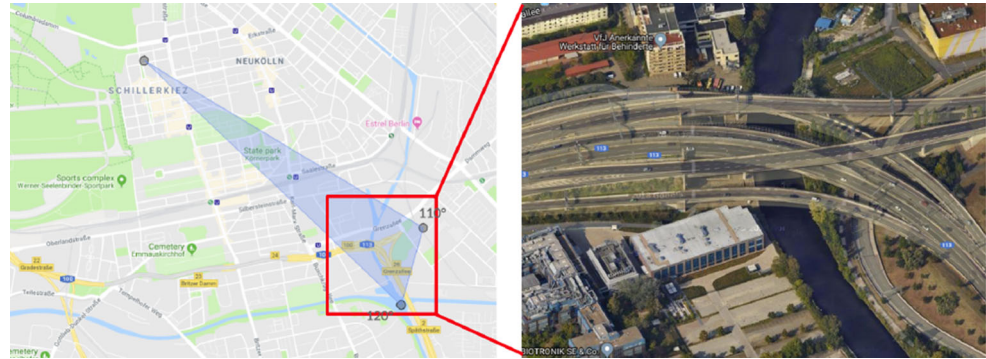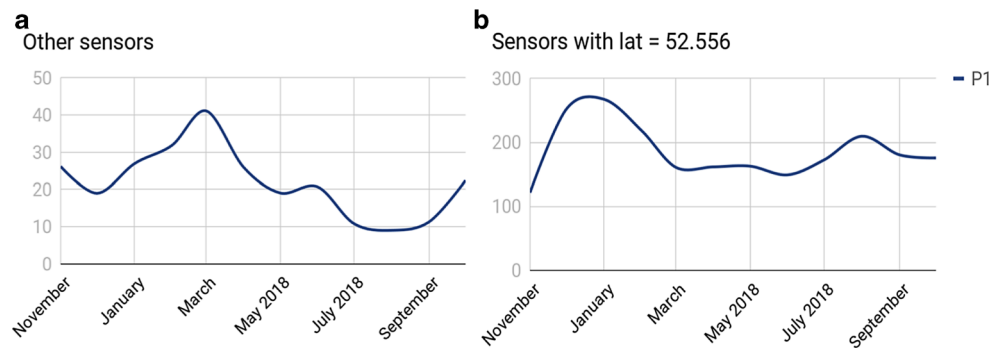

Fig. 15 Comparison of Pollution Patterns. **a** Common pollution pattern, **b** pollution pattern for sensors located in latitude 52.556

a
Other sensors

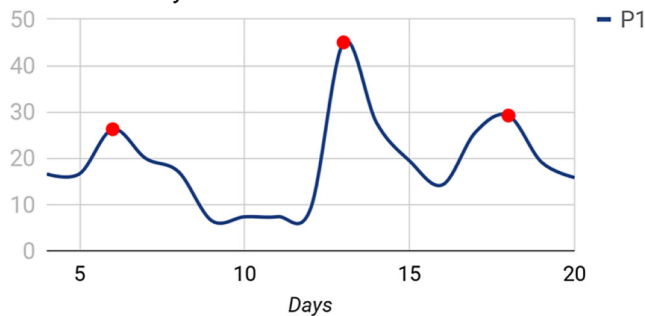**b**
Sensors with lat = 52.556

Berlin - February 2019

Fig. 16 Pollution trend for February 2019 sensor cluster close to Potsdamer Platz

3. *Public Events:* In general, it can be stated that events play an important role for air pollution and fine particulate matter concentration. Figure 16 depicts the pollution ratio of the sensor cluster close to the Potsdamer Platz (the location the Berlin International Film Festival). The first and the last red points on the Fig. 16 represent the dates of opening and closing the Berlin International Film Festival. The curve shows the increased concentration of fine particulate matter for both days. The middle red point represents the pollution on the 14th of February (Valentine's day), with multiple peaks of sudden pollution increase that are observed for many other public events, such as Easter holidays or the New Year's Eve.

Hamburg: Another question we addressed is how the introduction of driving bans for vehicles that use diesel engines affect air quality. We did this analysis for Hamburg, a city that has the *blue zone* since June 2018 for two roads (Max-Brauer-Allee and Stresemannstraße)[14]. Both roads handle the main traffic in Hamburg and are thus known for high pollution values. We grouped the sensors that are close to this area, and derived the mean values of the P1 concentration for two time periods—June 2017 to Jan 2018 (no diesel ban), and June 2018 to Jan 2019 (with diesel ban). The pollution in this area has decreased by 10% (from 19.56 to 17.40), while the overall pollution in Hamburg for same time period remained constant (20.49 without and 20.24 with diesel ban). Thus, we concluded that the introduction of diesel bans reduces the pollution locally but has no global impact on the entire city. Based on the aforementioned observations, we suggest the diesel bans to be applied in areas of higher levels of air pollution, to reduce it in the most efficient way.

4.5 Conclusion

In this project, we aimed to provide a general solution for finding explanation for air pollution, which is able to be applied to every other city. The system that we built can work

---

14 https://www.umwelt-plakette.de/de/info-zur-deutschen-umwelt-plakette/umweltzonen-in-deutschland/deutsche-umweltzonen.

with any other external data sources which are considered informative by domain experts. As we mentioned in this paper, our general solution could find interesting correlation between external sources (e.g., air traffic, weather, and public events) and particulate matter concentration as well as helped us to reach an insight regarding where to apply driving diesel bans.

We would like to mention that our system is biased towards the features we integrated. By replacing external data, we can see how the explanation changes. The point to stress is that the prototype detects high correlation between the feature set and target values (fine particulate matter concentration) yet does not prove causality.

It is also notable that some of the integrated features did not correlate with the target values (e.g., the number of cross roads), highlighting the importance of feature engineering for the provided settings.

As for future work, we look for a solution to incorporate external data sources for further enrichment without particular domain knowledge. To this end, instead of using specified external sources, we will use information on the Web, such as Web tables, to bring highly correlated features to the main dataset.

References

1. Alkhouri G, Wilke M (2019) Deep Learning zur Vorhersage von Feinstaubbelastung. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 305–308 https://doi.org/10.18420/btw2019-ws-35

2. Bailis P, Gan E, Madden S, Narayanan D, Rong K, Suri S (2017) Macrobase: prioritizing attention in fast data. ACM International Conference on Management of Data. ACM, Chicago, pp 541–556 (Proceedings)

3. Bougoudis I, Demertzis K, Iliadis L (2016) Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. Integr Comput Aided Eng 23(2):115–127

4. Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition. J Off Stat 6(1):3–73

5. Cyrys J, Eeftens M, Heinrich J, Ampe C, Armengaud A, Beelen R, Bellander T, Beregszaszi T, Birk M, Cesaroni G et al (2012) Variation of NO2 and NOx concentrations between and within 36 European study areas: results from the ESCAPE study. Atmos Environ 62:374–390. https://doi.org/10.1016/j.atmosenv.2012.07.080

6. Deutscher Wetterdienst (2019) Climate data center. ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/. Accessed 6 Feb 2019

7. Esmailoghli M, Redyuk S, Martinez R, Abedjan Z, Rabl T, Markl V (2019) Explanation of air pollution using external data sources. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 297–300 https://doi.org/10.18420/btw2019-ws-32

8. Folium (2019) Folium documentation. https://python-visualization.github.io/folium/. Accessed 6 May 2019

9. Grunert H, Meyer H (2019) Die Data Science Challenge auf der BTW 2019 in Rostock. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 281–284 https://doi.org/10.18420/btw2019-ws-30

10. Hagedorn S, Sattler K (2019) Peaks and the influence of weather, traffic, and events on particulate pollution. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 301–302 https://doi.org/10.18420/btw2019-ws-33

11. Klingner M (2018) Stellungnahme von Prof. Dr. Matthias Klingner zur öffentlichen Anhörung am 25. Juni 2018. https://www.bundestag.de/resource/blob/561430/42f387a20eef0041e81502cd5092b271/014_sitzung_fraunhofer-data.pdf. Accessed 25 Apr 2019

12. Klingner M, Sähn E (2008) Prediction of PM10 concentration on the basis of high resolution weather forecasting. Meteorol Z 17(3):263–272. https://doi.org/10.1127/0941-2948/2008/0288

13. Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A (2015) The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525(7569):367

14. Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49(4):764–766

15. Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) (2019) Datenbanksysteme für Business, Technologie und Web (BTW 2019). 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.-8. März 2019 Gesellschaft für Informatik, Bonn

16. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J (2015) Comparison of different methods for Univariate time series imputation in R. ArXiv 2015(10):arXiv:1510.03924 [stat.AP]. https://arxiv.org/abs/1510.03924

17. Mukherjee A, Agrawal M (2017) World air particulate matter: sources, distribution and health effects. Environ Chem Lett 15(2):283–309

18. Nova Fitness Co, Ltd (2015) SDS011 laser PM2.5 sensor specification. http://ecksteinimg.de/Datasheet/SDS011laserPM2.5sensorspecification-V1.3.pdf. Accessed 8 Feb 2019

19. OpenWeather (2018) Weather API – OpenWeatherMap. https://openweathermap.org/api. Accessed 28 Nov 2018

20. Alfeld P (1984) A trivariate Clough-Tocher scheme for tetrahedral data. Comput Aided Geom Des 1(2):169–181. https://doi.org/10.1016/0167-8396(84)90029-3
21. Plotly (2019) Build beautiful, web-based analytics applications with Dash. https://plot.ly/products/dash/. Accessed 20 Apr 2019
22. Rausch A, Werhahn O, Witzel O, Ebert V, Vuelban EM, Gersl J, Kvernmo G, Korsman J, Coleman M, Gardiner T et al (2015) Metrology to underpin future regulation of industrial emissions. 17th International Congress of Metrology. EDP Sciences, Paris, p 7008
23. Schmitz C, Serai DD, Gava TE (2019) Prediction of air pollution with machine learning. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 303–304 https://doi.org/10.18420/btw2019-ws-34
24. Stuttgart OL (2015) Luftdaten Info. https://luftdaten.info/. Accessed 28 Nov 2018
25. Stuttgart OL (2015) Luftdaten Info. https://archive.luftdaten.info/csv_per_month/. Accessed 28 Nov 2018
26. topographic-mapcom (2019) Topografische Karte Stuttgart. http://de-de.topographic-map.com/places/Stuttgart-8132395/. Accessed 26 Feb 2019
27. Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading
28. Environmental Protection Agency (2019) Particulate Matter (PM) basics. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM. Accessed 27 Apr 2019
29. Waizenegger T (2017) BTW 2017 data science challenge (SDSC17). In: Mitschang B, Ritter N, Schwarz H, Klettke M, Thor A, Kopp O, Wieland M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2017) 17. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Stuttgart, Germany, 6.-10. März 2017, pp 405–406
30. WHO (2016) Air pollution levels rising in many of the world's poorest cities. http://www.who.int/en/news-room/detail/12-05-2016-air-pollution-levels-rising-in-many-of-the-world-s-poorest-cities. Accessed 24 Nov 2018
31. Woltmann L, Hartmann C, Lehner W (2019) Assessing the impact of driving bans with data analysis. In: Meyer H, Ritter N, Thor A, Nicklas D, Heuer A, Klettke M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2019) 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Rostock, Germany, 4.–8. März 2019 Gesellschaft für Informatik, Bonn, pp 287–296 https://doi.org/10.18420/btw2019-ws-31
32. Xiao Q, Ma Z, Li S, Liu Y (2015) The impact of winter heating on air pollution in China. PLoS ONE 10(1):e117311