Noname manuscript No.
(will be inserted by the editor)

Explanation of Air Pollution Using External Data Sources

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

the date of receipt and acceptance should be inserted later

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 neighbourhoods [7]. Serious measures have been taken in order to sustain 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 [8].

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 [2]. 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 [3].

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 (Section 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 (Section 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 (Section 4).

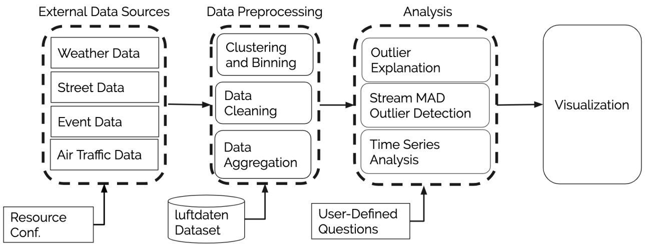


Fig. 1 Overall view of the architecture of our system

2 Data

The main source of the data for this project is provided by the Luftdaten service[1]. 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. 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 [5]. 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

Mahdi Esmailoghli
Technische Universität Berlin, E-mail: esmailoghli@tu-berlin.de

[1] http://luftdaten.info

the project, we incorporate the main dataset with additional descriptive features.

2.1 External Data Sources

As specified in Fig. 1, we use four external data sources: weather[2], geo data[3], air traffic[4], 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 neighbourhood of each sensor help to determine what factors might explain the source of pollution [4]. We obtain the data from the Open Street Map platform and the GeoFabrik service[5] that regularly publishes up-to-date data on geometries and other features[6] that describe 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 such as New Year's Eve and Easter. Besides, non-holiday events are selectively picked from online services for instance VisitBerlin[7]. The non-holiday events are the events or activities that attract crowds in a particular neighbourhood, 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.

3 Architecture

Fig. 1 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.

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 neighbourhoods, 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.

3.2 Data preprocessing

The data preprocessing component consists of the following operations: clustering, binning, cleaning, and aggregation (Fig. 2). 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 neighbourhood 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) [6] 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

[2] ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate

[3] https://www.openstreetmap.org

[4] https://www.flightradar24.com

[5] http://download.geofabrik.de/

[6] https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[7] www.visitberlin.de

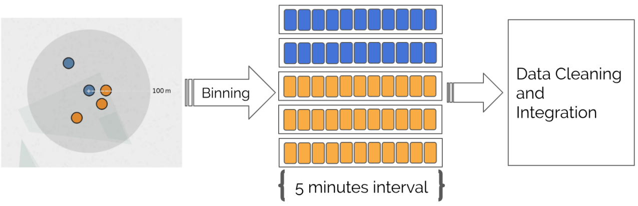


Fig. 2 Clustering and binning techniques

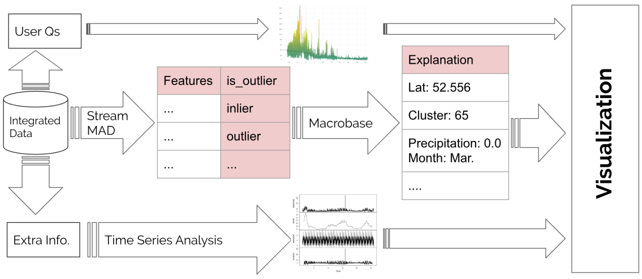


Fig. 3 Analysis phase

conditions, such as high humidity. That is done due to recommendations from the sensor specification[8].

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.

3.3 Analysis

Fig. 3 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 (Median Absolute Deviation) 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, a small amount of air pollution that is caused by the movie festival event. Stream MAD algorithm uses Min-Max heap to update the median value by checking every new value in the dataset. After applying Stream MAD algorithm on our dataset, we use the Macrobase [1] outlier explanation tool 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 oncentration.

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.

[8] `https://www.watterott.com/media/files_public/reiknvyoc/SDS011.pdf`

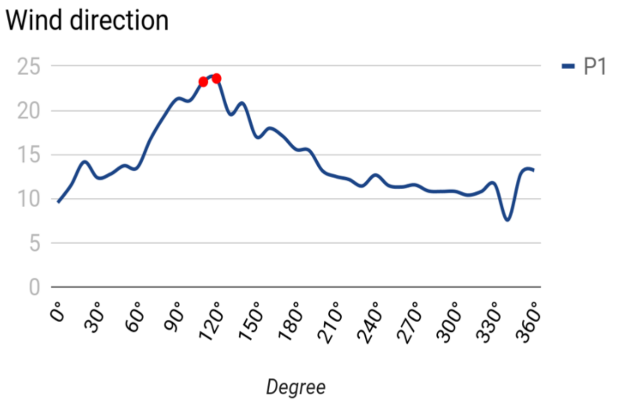


Fig. 4 Correlation between the direction of wind and pollution levels for the example cluster

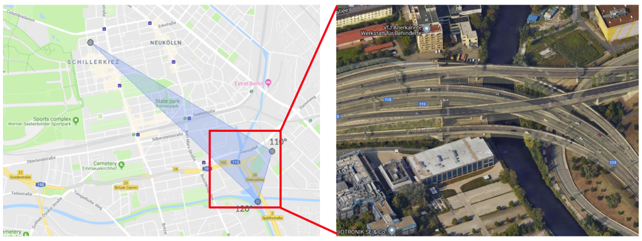


Fig. 5 Location of the example cluster and the highway located between directions 110 and 120 degrees

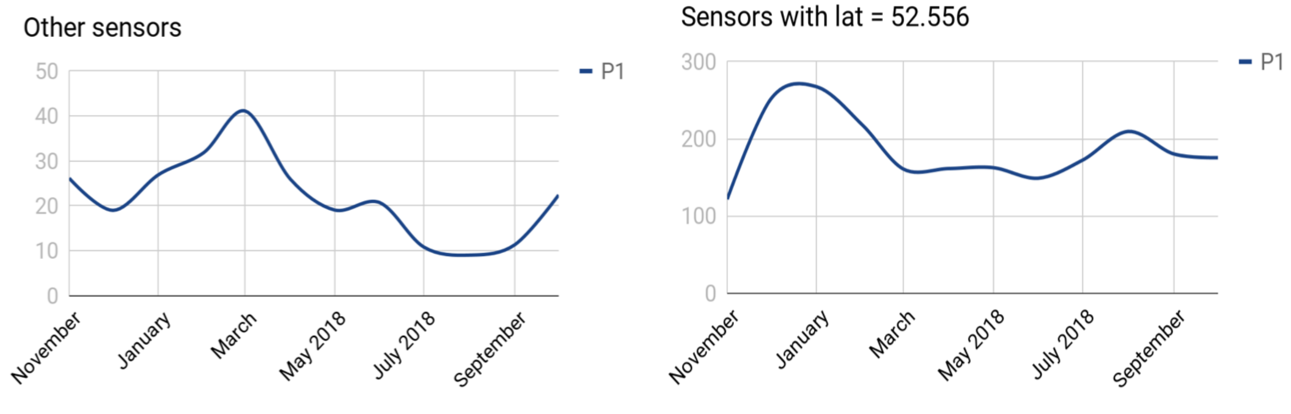
4 Findings

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. Fig. 4 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. 4, 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. 5. 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. 6(a). 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 during [9].
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 the new year because this time correlates with high tourist seasons (Fig. 6(b)).
3. *Public Events*: In general, it can be stated that events play an important role for air pollution and fine particulate matter concentration.
   Fig. 7 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. 7 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 whether the introduction of diesel bans for vehicles that use diesel engines may negatively 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 Stresemannstrae)[9]. 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).

[9] `https://www.umwelt-plakette.de/de/info-zur-deutschen-umwelt-plakette/umweltzonen-in-deutschland/deutsche-umweltzonen`



(a) Common pollution pattern

(b) Pollution pattern for sensors located in latitude 52.556

Fig. 6 Comparison of Pollution Patterns

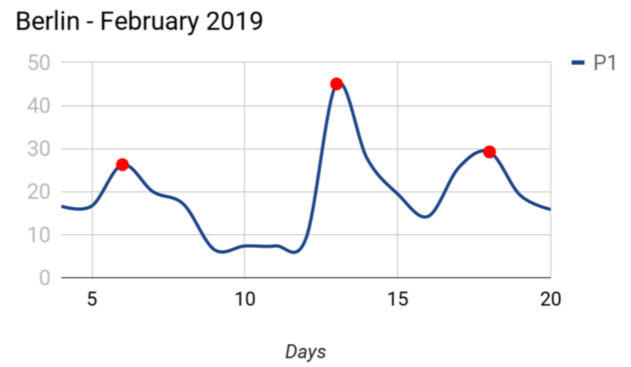


Fig. 7 Pollution trend for February 2019 sensor cluster close to Potsdamer Platz

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.

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 with any other external data sources that domain expert consider them as informative. 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 to 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.

Acknowledgements We would like to acknowledge the German Federal Ministry of Transport and Digital Infrastructure (BMVI) in the context of the research initiative mFUND, for funding the project DAYSTREAM under grant number 19F2031D, in which some of the tools and techniques used in this research are based or inspired. The work is also supported by the BZML, ECDF, and HEIBRiDS graduate school.

References

1. Bailis, P., Gan, E., Madden, S., Narayanan, D., Rong, K., Suri, S.: Macrobase: Prioritizing attention in fast data. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 541–556. ACM (2017)
2. Bougoudis, I., Demertzis, K., Iliadis, L.: Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. Integrated Computer-Aided Engineering **23**(2), 115–127 (2016)

3. Esmailoghli, M., Redyuk, S., Martinez, R., Abedjan, Z., Rabl, T., Mark, V.: Explanation of air pollution using external data sources. BTW 2019–Workshopband (2019)
4. Klingner, P.D.I.M.: Stellungnahme von Prof. Dr. Matthias Klingner zur ffentlichen Anhrung am 25. Juni 2018 (2018 (last access 2019-04-25)). https://www.bundestag.de/resource/blob/561430/42f387a20eef0041e81502cd5092b271/014_sitzung_fraunhofer-data.pdf
5. Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature **525**(7569), 367 (2015)
6. Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L.: Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology **49**(4), 764–766 (2013)
7. Mukherjee, A., Agrawal, M.: World air particulate matter: sources, distribution and health effects. Environmental Chemistry Letters **15**(2), 283–309 (2017)
8. Rausch, A., Werhahn, O., Witzel, O., Ebert, V., Vuelban, E.M., Gersl, J., Kvernmo, G., Korsman, J., Coleman, M., Gardiner, T., et al.: Metrology to underpin future regulation of industrial emissions. In: 17th International Congress of Metrology, p. 07008. EDP Sciences (2015)
9. Xiao, Q., Ma, Z., Li, S., Liu, Y.: The impact of winter heating on air pollution in china. PloS one **10**(1), e0117311 (2015)