Explanation of Air Pollution Using External Data Sources

Mahdi Esmailoghli[1] Sergey Redyuk[2] Ricardo Martinez[3,4]
Ziawasch Abedjan[1,3] Tilmann Rabl[2,3] Volker Markl[2,3]

High concentrations of fine-grained particles in the air can adversely affect human health[5]. To control it, the European Union has undertaken several strategies, such as the introduction of certain particle concentration thresholds allowed in populated areas, or limitations for vehicle access [Ra15]. However, many cities in Germany are unable to follow this legislation[6] and control the particle emission because it is hard to attribute the pollution to a clear source. Therefore, it is important to understand the dynamic process of the fine-grained particles distribution and the reasons the emission occurs. In this project, we aim at designing a system that provides the human analyst with descriptions about polluted areas within a city, and potential causes.

During the phase of exploratory data analysis on the sensor dataset, which is provided by BTW specifically for the data science challenge[7], we selected all sensors located within a 10-kilometer radius from Berlin city center (255 sensors with 3GB of data), and provided a common data schema that fitted all the sensor types. We found that a particular subset of sensors (lat. 52.556) shows consistently higher degree of pollution. Since the original data was not enough to explain potential causes of this anomaly, we integrated the dataset with external air traffic data. Then, we established that the location of Tegel (TXL) airport airways correlated with the sensors that recorded higher pollution. We also observed seasonal fluctuations in pollution, and considered inversion during the winter as a potential cause. However, the seasonal trend near TXL turned out to be different. Air pollution is increased drastically during the summer, more likely, due to the higher number of flights to or from the airport. During the exploration phase, we discovered that air pollution might be caused by numerous local events organized in the city. For instance, we observed an instant increase in pollution ratio near the Berlin TV Tower during the New Year's Eve. Checking external web sources revealed the news feed about the New Year celebration fireworks.

[1] Technische Universität Berlin, lastname@tu-berlin.de
[2] Technische Universität Berlin, firstname.lastname@tu-berlin.de
[3] Deutsches Forschungszentrum für Künstliche Intelligenz, Berlin,
[4] ricardo_ernesto.martinez_ramirez@dfki.de
[5] https://www.pca.state.mn.us/air/fine-particles-and-human-health
[6] http://ec.europa.eu/environment/air/quality/standards.htm
[7] https://archive.luftdaten.info/

doi:10.18420/btw2019-ws-32

As demonstrated, external data can provide human analyst with comprehensible understanding of causes for the observed pollution levels. Therefore, we formulate the goal for the project as *finding explanations for air pollution through integration of external data sources, and building a new tool that provides human analysts with the explanation of potential sources of pollution*. In the project, we face several challenges: (i) streaming scenario and fast-changing data that current outlier explanation tools do not handle well [ZDM17] (e.g., MAD algorithm for univariate outlier detection is not applicable for streaming scenarios; solutions that adapt MAD for data streams, process batches instead of true streaming); (ii) heterogeneity of sensory data that leads to multiple schemata and makes data integration harder; and (iii) malfunctioning sensors that create erroneous and incomplete data [Ab16]. We address the aforementioned challenges in our project.

Progress Report and Outlook
We choose Berlin as the target region, and take two additional data sources for data integration - weather and air traffic data (airports TXL and SXF). As the data contains temporal information, we propose an event-based simulation model for our prototype that "replays" historical information as if the events are happening now, thus supporting stream processing to fit the fast-changing real-world scenario [Gr18] (Challenge (i)). In order to accommodate different schemata for external data sources, we provide a common schema that fits all external sources, and use data integration techniques [DHI12] for merging (Challenge (ii)). We utilize hexagon binning [Le11] and clustering methods to group the data spatially and integrate the readings from different neighboring sensors. Assuming that the sensors close to one another record similar data, we can fuse these data points into a single record, improving the data quality. This approach can also be used for cross-validation, in order to handle anomalies that are generated by malfunctioning sensors (Challenge (iii)). For interactive data analysis, we propose to use visualization tools, such as Thingsboard[8], and Plotly Dash[9]. To find the reason of pollution observed by aforementioned sensors, we use state-of-the-art outlier explanation systems such as Macrobase [Ba17], and integrate the correlated features with external sources, to provide reasonable interpretation of feature-wise causal relationships for interesting points [Mi13].

In the first phase of this project, we apply MAD on pollution data. We choose MAD as outlier detection algorithm because (i) pollution ratios are correlated and outlier in P1 means an outlier in P2, and vice versa; so we can use MAD which is a univariate outlier detection technique, and (ii) MAD is used in many state-of-the-art systems such as Macrobase. We introduce an online version of MAD that can treat the data as stream. Then, we acquire and prepare both weather and flight data for further integration into the prototype (fusing by the compound timestamp-location key). After data integration, we apply ranking metrics to select external data features that "explain" potential causes of anomalous pollution levels.

As fine-grained particles have many potential sources (factories, transport, cultural events, power stations, agriculture, plants' pollen, forest fires etc.), in the future we aim to generalize

[8] https://thingsboard.io/docs/user-guide/rule-engine-2-0/tutorials/aggregate-latest-data
[9] https://plot.ly/products/dash/

our solution and add more external sources. We also aim to provide a solution that selects external information automatically, by integrating web tables and web forms with the detected features [Ab15].

Relevant Experience. Coming from the DIMA and BIGDAMA research groups at TU Berlin, we cover the necessary expertise in data management, distributed computing [Al14], data integration [De17], and machine learning [Mo17]. Our previous applied projects included analysis of sensory data for the metal industry (production line optimization, hot rolling mills [St18]), urban development (traffic analysis), graph-based fraud detection in healthcare, and outlier explanation.

Acknowledgements. We thank Felix Neutatz, Batuhan Tüter, Felipe Gutierrez and Dimitrios Giouroukis for their constructive comments and help.

References

[Ab15] Abedjan, Z.; Morcos, J.; Gubanov, M. N.; Ilyas, I. F.; Stonebraker, M.; Papotti, P.; Ouzzani, M.: Dataxformer: Leveraging the Web for Semantic Transformations. In: CIDR. 2015.

[Ab16] Abedjan, Z.; Chu, X.; Deng, D.; Fernandez, R. C.; Ilyas, I. F.; Ouzzani, M.; Papotti, P.; Stonebraker, M.; Tang, N.: Detecting data errors: Where are we and what needs to be done? VLDB 9/12, pp. 993–1004, 2016.

[Al14] Alexandrov, A.; Bergmann, R.; Ewen, S.; Freytag, J.-C.; Hueske, F.; Heise, A.; Kao, O.; Leich, M.; Leser, U.; Markl, V., et al.: The stratosphere platform for big data analytics. VLDB 23/6, pp. 939–964, 2014.

[Ba17] Bailis, P.; Gan, E.; Madden, S.; Narayanan, D.; Rong, K.; Suri, S.: MacroBase: Prioritizing Attention in Fast Data. In: SIGMOD. Pp. 541–556, 2017.

[De17] Deng, D.; Fernandez, R. C.; Abedjan, Z.; Wang, S.; Stonebraker, M.; Elmagarmid, A. K.; Ilyas, I. F.; Ouzzani, S. M. M.; Tang, N.: The Data Civilizer System. In: CIDR. 2017.

[DHI12] Doan, A.; Halevy, A. Y.; Ives, Z. G.: Principles of Data Integration. Morgan Kaufmann, 2012, ISBN: 978-0-12-416044-6.

[Gr18] Grulich, P. M.; Saitenmacher, R.; Traub, J.; Breß, S.; Rabl, T.; Markl, V.: Scalable Detection of Concept Drifts on Data Streams with Parallel Adaptive Windowing. In: EDBT. Pp. 477–480, 2018.

[Le11] Lewin-Koh, N.: Hexagon binning. Online: `http://cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf/`, 2011.

[Mi13] Micenková, B.; Ng, R. T.; Dang, X.-H.; Assent, I.: Explaining outliers by subspace separability. In: ICDM. Pp. 518–527, 2013.

[Mo17] Monte, B. D.; Karimov, J.; Mahdiraji, A. R.; Rabl, T.; Markl, V.: PROTEUS: Scalable Online Machine Learning for Predictive Analytics and Real-Time Interactive Visualization. In: EDBT/ICDT 2017 Joint Conference. 2017.

[Ra15] 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. EDP Sciences, p. 07008, 2015.

[St18] Ståhl, N.; Falkman, G.; Mathiason, G.; Karlsson, A.: A Self-Organizing Ensemble of Deep Neural Networks for the Classification of Data from Complex Processes. In: IPMU. Pp. 248–259, 2018.

[ZDM17] Zhang, H.; Diao, Y.; Meliou, A.: EXstream: Explaining Anomalies in Event Stream Monitoring. In: EDBT. 2017.