PhD Proposal

# *Big Data Series Analytics in the Context of Environmental Crowd Sensing*

DAVID Lab – University of Versailles Saint-Quentin – Paris-Saclay University

## 1. Context and motivation

Upon the recent development of advanced computing and communication technologies, the world is witnessing the rise of the so-called Internet of Things (IoT). IoT envisions a world where everything is connected - from humans and computing devices to animals, vehicles, and even the smallest appliances. Sensors and actuators are fetched on things enabling them to sense, generate data, communicate, act, and share information. This is leading to the generation of massive amount of data, now regarded as Big Data or Big Sensing Data in the IoT context. With great embedded potential in this data, both industry and academia are rushing to develop methods and technologies that not only can handle this large amount of data but can also exploit them in order to mine new knowledge and insights.

One application of IoT is monitoring of air pollution. Several research initiatives have used fixed air pollution sensors to monitor air quality [1]. However fixed sensors have been facing shortcomings in modeling air quality because of the high spatiotemporal variability nature of air pollutants. That is why the community is shifting toward a new monitoring paradigm, namely *mobile crowd sensing*, that empowers volunteers to contribute data acquired by their personal sensor-enhanced mobile devices [2]. This is enabled by the use of emerging low-cost and lightweight air pollution sensor boxes, which can be fetched on pedestrians, cyclists, or on vehicles. Opportunistic air quality monitoring takes advantage of existing mobile infrastructure or people common daily routines to perform monitoring [3].

This paradigm has several advantages compared to conventional monitoring techniques. First, it promotes personalization where each individual will be able to gain insights on his/her exposure. Second, it measures indoor and outdoor environments (Home, Work, Transportation, Streets, Parks, etc.) and expands the spatial coverage, depending on the participants whereabouts. Finally, it enables insights at a higher resolution along the participants trajectories, thereby allowing to capture local variability and peaks of pollution. Nevertheless, the main limitation of opportunistic sensing arises from its uncontrolled sampling nature, leading to highly uneven data density across regions and times of the day. Mining such inhomogeneous samples inherently raises unique challenges that we intend to tackle in this thesis. From the perspective of the study of daily exposure, typical exposure profiles could be mined from the longitudinal data set. However, there is a gap to fill between the raw sensor data series and high-level profiles.

## 2. Objectives & Challenges

While Mobile Crowd Sensing paradigm has opened the door for new possibilities, it has also generated some challenges [2]. Indeed, the nomadic nature of sensors, and their combination (air pollution is often monitored using multi-sensor devices) lead to revisit the traditional methods of data mining and knowledge extraction. These sensors typically produce multivariate time series where one variable is the geographical position of the device (we call it *complex data series*). Nevertheless, exploiting such complex data series for analytical purpose, such as exploratory analysis using data mining techniques, is far from straightforward. Since raw sensor data are mostly noisy and acquired at irregular (and asynchronous) frequencies, direct use of the state-of-the-art methods, such as time series analysis and mining, is insufficient. Besides, to take full advantage of these data, it should not be only analyzed in isolation, but rather by matching them with the context, and analyzing them under multiple dimensionality and scale (e.g., spatial, user, micro-environment, time dimensions). Here comes one of the challenges on how to transit from raw and heterogeneous complex data series into such a type of high-level models.

Moreover, going further in exploiting the personalization aspect of opportunistic mobile sensing enables individuals to relate air pollution to themselves [4], and to act upon gained insights. For example, an individual may change his/her daily routes, transportation means, even his/her activities in sake of lesser exposure and lesser health effects. Nonetheless, this requires building individual profiles, and correlating them with personal health data and activities. This correlation opens the way for highlighting potential relations of causality, or inferring the exposure based on an activity profile or a planned route.

## 3. Methodology

In this thesis, we aim at developing data mining methods adapted to opportunistic samples of geodated series along with associated contextual data on the one hand, and studying multi-dimensional exploratory analysis and aggregation of such data on the other hand. To achieve the aforementioned objectives, and effectively make sense of the collected data series, the Ph.D. will consider several challenges in data processing and mining such as:

> ***Data Acquisition:*** This includes the calibration phase of sensors, the adjustment of asynchronous data, as well as spatial data fusion, while taking into consideration the data quality aspects. Functional Data Analysis (FDA) is a possible candidate approach [5].
>
> ***Data Enrichment:*** To cope with multidimensional analysis requirement, methods to transit from raw data series into more semantic views should be searched and developed. An example of semantic view is time annotated sequence of traversed zones. Data enrichment can make use Intelligent Data Analysis (IDA) approaches, which fill the gap between data generation and data comprehension [6].

***Data Analysis:*** The enriched data opens the way to learn meaningful patterns out of it. This includes exposure estimate for uncovered place and time, by correlating pollution level with other observable phenomena (traffic, weather, events), or categorization of different indoor and outdoor micro-environments, mining exposure profiles by segmenting longitudinal data, etc. This may call for multivariate time series data mining in the context of opportunistic sensing where the data are geodated and sparse [7].

The proposed solutions will be implemented and evaluated in a perspective of large-scale collection of complex data series, and applied to a real scenario of opportunistic air quality monitoring in the framework of the ANR project Polluscope [8].

## References


[1] Y. Zheng, L. Furui, and H. Hsun-Ping. "U-air: When urban air quality inference meets big data." *19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013.

[2] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R Huang, X. Zhou, (2015). Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM Computing Surveys (CSUR), 48(1), 7.

[3] OpenSense project: http://opensense.epfl.ch

[4] A. Sîrbu *et al.*, "Participatory Patterns in an International Air Quality Monitoring Initiative," *PloS one*, vol. 10, no 8, p. e0136763, 2015.

[5] A. Mustapha, K. Zeitouni, Y. Taher, (2018) Towards Rich Sensor Data Representation - Functional Data Analysis Framework for Opportunistic Mobile Monitoring. GISTAM 2018: 290-295.

[6] F. Roda and E. Musulin, "An ontology-based framework to support intelligent data analysis of sensor measurements," Expert Syst. Appl., vol. 41, no. 17, pp. 7914–7926, 2014.

[7] G. K. Kang,,J. Z. Gao, S. Chiao, S. Lu, G. Xie (2018). Air quality prediction: Big data and machine learning approaches. International Journal of Environmental Science and Development, 9(1), 8-16.

[8] POLLUSCOPE project (grant ANR-15-CE22-0018): http://polluscope.uvsq.fr


## Prerequisite skills:

The applicant should hold a Master diploma in Computer science, or equivalent:
- Good background in data mining and machine learning
- Strong programming, system, and database skills
- Good oral communication and technical reading and writing skills in English
- Proficiency in French is desirable.

## Application & Contacts:

Candidates should submit a full CV, an interest letter, the copy of their diploma and the transcripts of the last two years of their studies, and contact information for two references. The email should be addressed to all the contacts below, with "PhD Application" in the subject.


DAVID Lab: Prof. Karine Zeitouni: Karine.Zeitouni@uvsq.fr and Dr. Yehia Taher Yehia.Taher@uvsq.fr

In collaboration with IRENav: Dr. Cyril Ray: cyril.ray@ecole-navale.fr


## Hosting laboratory:

DAVID Lab/ADAM Team, University of Versailles St-Quentin / Paris-Saclay University: www.david.uvsq.fr

Doctoral school: https://www.universite-paris-saclay.fr/en/doctorate