Stage de Master1/2
IBISC-lab EA 4526

Internship subject: Multimodal learning for the calibration of low-cost air pollutant sensors.

Keywords: Deep learning, multi-modal data, sensor calibration, air pollutant, domain shift, model uncertainties, data augmentation.

Description:

Since COP21, atmospheric pollutant measurement systems have rapidly grown, which, combined with *crowdsourcing*, make it possible to represent air quality spatially (Xie, 2017). This mapping initiated by the traditional actors of surveillance, local communities, is in its infancy. Nevertheless, it raises the question of data uncertainty, their exploitation, and the possibilities new AI technologies offer, particularly by DL (Goodfellow, 2016) for (regulatory) air quality surveillance (M. K⍰unzli, 2005). It also questions the drift of sensors between laboratory calibration and their use in the _eld. Many parameters can affect this drift. The accuracy of these sensors is relative, and the measurement uncertainties are still poorly known.

Our work is based on making these estimates more accurate by overcoming simple linear regression (Tibshirani, 1990) and reconsidering critical issues such as (i) optimization of the deployment of sensors (Rivano H. Boubrima A., 2017), (ii) the management of the spatial representativeness of the measurements, (iii) the treatment of heterogeneous data in space and in time.

Many questions arise: how to make the measurements more reliable by exploiting the available data? Can we predict the measure $\hat{x}(t)$ from the history $x(t), x(t-1), x(t-2), x(t-3), \dots, x(t-T)$ or by geostatistical interpolation of other sensor measurements (Goovaerts, 1997)? how to make sure that the parameters do not diverge from the actual conditions? Should we also use the topographic data?

Our preliminary results have proven that micro-sensors suffer from (i) the dependence on meteorological variables and (ii) interferences with other pollutants. Therefore, we use multiple input-multiple outputs convolutional and recurrent (Goodfellow, 2016) models with unique time-series data processing. This internship could lead to writing a research article for a conference/journal.

Tasks:

- Master the proposed models and multimodal data.
- Run experiments on learning and evaluating air pollutant calibration models over different databases.
- Evaluate the drift of sensors and models over different periods, sensors, and databases.
- Propose a method to include the processing of missing data.

Profile and skills required:

- 2nd/3rd engineering year or M1/M2 student.
- Experience with Machine/Deep learning.
- Programming skills: Python.
- A practice of Pytorch and/or Tensorflow would be a plus.
- Autonomy, sense of relationship to interact with teams of research.
- Creativity to set up an innovative solution. Interest in the environment would be a plus.

Contract: Internship 4-6 months, asap.

Contact: Vincent Vigneron, Hichem Maaref, Aymane SOUANI

{vincent.vigneron,aymane.souani,hichem.maaref}@univ-evry.fr

Phone: +33769766386

Bibliographie

Goodfellow, I. B. (2016). *Deep Learning.* MIT Press.

Goovaerts, P. (1997). *Geostatistics for natural resource evaluation.* Technometrics.

M. K�unzli, N. J. (2005). Ambient air pollution and atherosclerosis in los angeles. *Environ Health Perspec*, 201-206.

Rivano H. Boubrima A., B. W. (2017). A new wsn deployment approach for air pollution monitoring. *The 14th IEEE Consumer Communications & Networking Conference*.

Tibshirani, T. H. (1990). *Generalized Additive Models.* Monographs on Statistics and Applied Probability. Chapman and Hall/CRC.

Xie, X. a. (2017). A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS International Journal of Geo-Information*.