

PROPOSITION DE STAGE DE M2

Caractérisation des micro-environnements et de l'exposition à la pollution atmosphérique dans un contexte de données participatives – Apprentissage automatique multi-vue

Mots clés : *Science de données, Apprentissage automatique multi-vue, Base de données spatiotemporelles, Analyse de données de capteurs, Collecte participative, Santé environnementale*

Contexte et contenu scientifique :

Que ce soit en milieu intérieur ou dans un environnement extérieur, la qualité de l'air que nous respirons est l'une des principales préoccupations de la population, notamment en zones urbaines denses en raison de l'importance des risques sanitaires de la pollution atmosphérique. Cependant, l'exposition individuelle réelle à cette pollution est encore mal connue. L'émergence de micro-capteurs environnementaux portables et connectés rend envisageable la mesure de l'exposition individuelle, d'une manière continue, en tout lieu et à tout moment. C'est dans cette optique que s'inscrit le projet ANR Polluscope. Le projet se base sur des campagnes avec des volontaires équipés de capteurs personnels afin d'observer leur exposition individuelle et de collecter par la même occasion des données participatives sur les lieux qu'ils visitent au grès de leurs déplacements. Polluscope propose une plateforme informatique de gestion et d'analyse de ces données. Cette plateforme déjà bien avancée devra s'enrichir de nouvelles fonctionnalités de traitement et d'analyse avancées des données. Le stagiaire intégrera l'équipe du projet dans **l'objectif de développer une méthode de caractérisation de l'exposition à la pollution et des micro-environnements** (intérieur, extérieur, dans les gares ou les transports).

Méthodologie et/ou outils utilisés :

Ces données ont plusieurs facettes : séries temporelles, des trajectoires, des séquences annotées, voire des données contextuelles externes, essentiellement géo-spatiales. Aujourd'hui, plusieurs méthodes ont été explorées au sein du projet exploitant une de ces facettes individuellement. Par exemple, la segmentation en stops et moves s'est basée sur le *clustering* des données de localisation ; ou encore sur la détection de changement dans les mesures de capteurs. L'objectif du stage est de combiner ces facettes pour détecter automatiquement et caractériser les micro-environnements.

La méthode de reconnaissance de micro-environnement préconisée est d'appliquer ou adapter l'apprentissage multi-vue (*multi-view learning*) dont la bibliographie fournie, afin de combiner plusieurs classifieurs. La première partie du stage sera consacrée à l'étude et la prise en main de ces méthodes, puis leur application dans le contexte réel du projet. Le stagiaire sera intégré à l'équipe de recherche.

Dans un deuxième temps, le stagiaire contribuera à l'implémentation d'un indicateur de l'exposition individuelle au risque de pollution, une fois qu'il aura été défini par les partenaires du projet.

La plate-forme Polluscope en cours de développement par l'équipe ADAM, se base sur une architecture flexible à base de services, dont un gestionnaire de données avec Postgres/PostGIS, des services d'extraction, transformation et chargement de données de sources diverses, des services de prétraitement, des analyses basiques et une interface de visualisation interactive utilisant Grafana. L'implémentation intégrera des fonctions de prétraitement de données comme le débruitage ou l'imputation des valeurs manquantes.

Bibliographie et webographie :

Projet Polluscope : <http://polluscope.uvsq.fr>

Brahem M. *et al.*, Vers un observatoire participatif de l'exposition individuelle à la pollution de l'air et de ses effets sanitaires, Conférence internationale de Géomatique et d'Analyse Spatiale (SAGEO 2019), 275-280.

Muslea, I., Minton, S., & Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27, 203-233.

Xu, C., Tao, D., & Xu, C. (2013). A survey on multi-view learning. arXiv preprint arXiv:1304.5634. <https://arxiv.org/pdf/1304.5634.pdf>

Li, Sheng, Yaliang Li, and Yun Fu. "Multi-view time series classification: A discriminative bilinear projection approach." *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM 2016*. <https://static.aminer.org/pdf/fa/cikm2016/lfp0510-liA.pdf>

Profil du candidat :

Le stage est destiné aux étudiants de Master 2 ou de dernière année de cycle Ingénieur en Informatique.

Le candidat doit avoir des compétences solides en ingénierie informatique, en bases de données et en apprentissage statistiques (*machine learning*). La connaissance des SIG est un plus. Le travail en équipe et la communication avec des non informaticiens sont également nécessaires pour la réussite de ce stage.

Equipe de recherche et laboratoire d'accueil : Équipe ADAM, Laboratoire Données et Algorithmes pour une Ville Intelligente et Durable (DAVID) – UVSQ, Université Paris-Saclay, Versailles.

<https://www.david.uvsq.fr>

Responsable(s) du stage : Karine Zeitouni, Professeure, Yehia Taher, Maître de Conférences

E-mail (Contacts) : karine.zeitouni@uvsq.fr, yehia.taher@uvsq.fr

Collaborations :

Ce stage est en complémentarité avec un autre stage en sciences de l'environnement réalisé au Laboratoire LSCE (partenaire du projet Polluscope). Les stagiaires seront amenés à travailler de pair et avec leurs encadrants respectifs pour spécifier et implémenter les indicateurs et analyses pertinents qui seront appliqués sur des données réelles collectées.

Durée : 6 mois à partir de mars 2020

Financement : DIM Qi2, Réseau de recherche sur la qualité de l'air en Ile-de-France, <https://www.dim-qi2.fr>

Condition de rémunération : 600,60 € (pour 7 heures/jour x 22 jours/mois et au prorata selon la présence), auquel s'ajoute le remboursement de la moitié du prix d'abonnement aux transports, soit 37,60 €/mois