Le pourquoi et le comment du prétraitement des données de collecte participative par micro-capteurs environnementaux

Illustration avec POLLUSCOPE

Karine Zeitouni (DAVID/UVSQ)*, Philippe Aegerter (VIMA/UVSQ)

Séminaire ACE-ICSEN le 04/12/2020

(*) en collaboration avec les partenaires du projet ANR Polluscope

Limites de l'observation de la qualité de l'air vis-à-vis de l'exposition

➢ Fournies par les AASQA

– Basées sur un réseau de stations de mesures fixes + la modélisation

Capteurs de dioxyde d'azote (NO2)

Crédit Airparif

➢ Mais on manque de données sur :

1. La quantification de **l'exposition individuelle réelle**
2. L'analyse **d'impacts** de la pollution sur la **santé individuelle**
3. La compréhension des **disparités de risque sanitaire** observées entre des groupes de population

Comment mesurer et analyser l'exposition individuelle ?

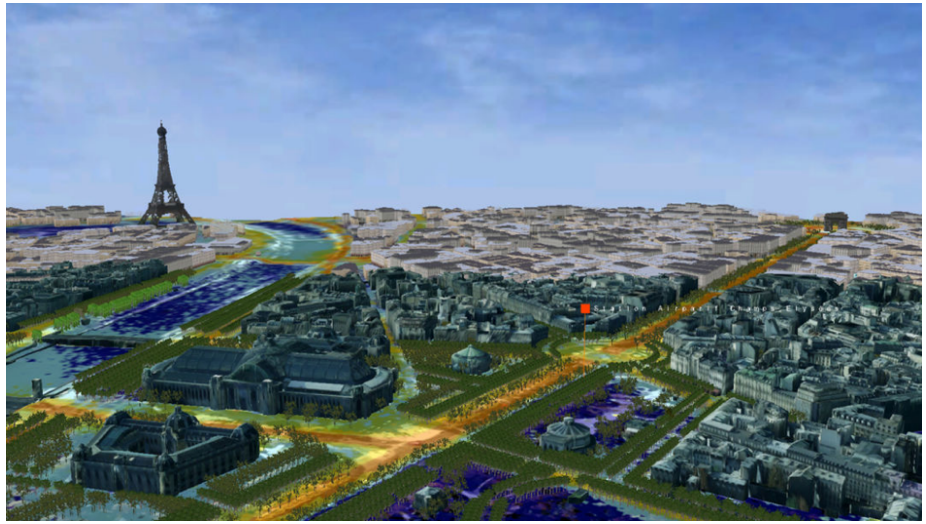
- Emergence de mini-capteurs à bas coût :
  - Une palette de **capteurs portables & connectés**
  - Couplés avec la **géolocalisation** par GPS

- Technologie prometteuse pour **mesurer en continu et partout** l'exposition individuelle et révéler les changements rapides et les pics d'exposition
  - Le problème reste la **fiabilité** de la mesure

Plan de la présentation

- ❑ Contexte
- ☑ **Présentation du projet POLLUSCOPE**
- ❑ Pré-traitement des données & Impacts sur l'analyse

Le projet POLLUSCOPE – Volet ANR



http://polluscope.uvsq.fr

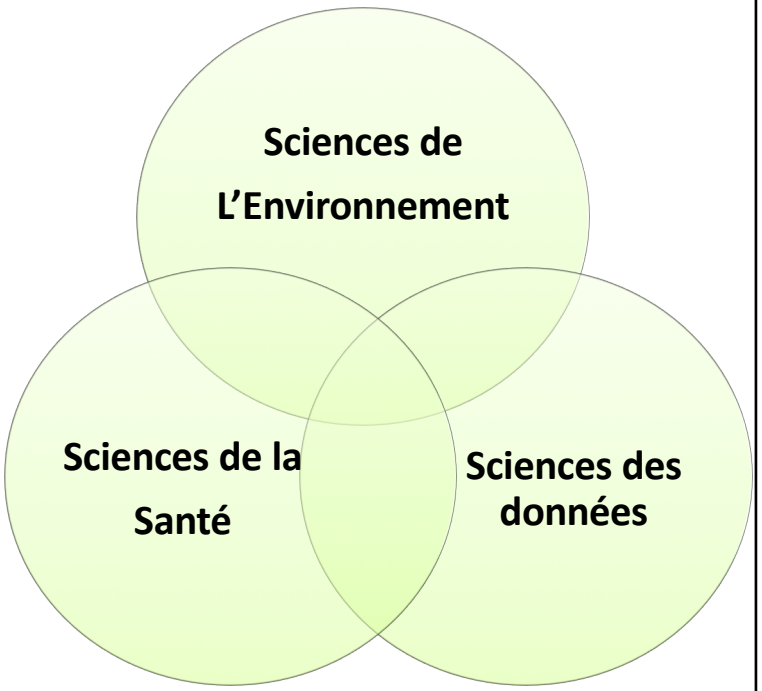
Observatoire participatif pour la surveillance de l'exposition individuelle à la pollution de l'air en lien avec la santé

Les micro-capteurs connectés émergents offrent l'opportunité de mesurer l'exposition réelle des individus à la pollution atmosphérique en tout lieu tout au long de leurs activités journalières. Le but du projet ANR Polluscope est d'évaluer sur le terrain les capacités et les limitations de ces nouveaux capteurs dans la compréhension fine de l'exposition individuelle à la pollution de l'air et de ses effets sur la santé, notamment chez des sujets asthmatiques ou BPCO. Pour y parvenir, des verrous devront être levés en termes de métrologie, de protocole de collecte, d'intégration aux données de mobilité, de traitement et d'analyse de données imparfaites, de confidentialité des données personnelles, etc. Polluscope réunit des spécialistes en sciences environnementales, en santé, en géosciences et en informatique. Il sera déployé et testé sur le terrain et ciblera plusieurs types de population, à la fois pour une étude épidémiologique et pour valider une collecte participative.

Objectif général du projet POLLUSCOPE

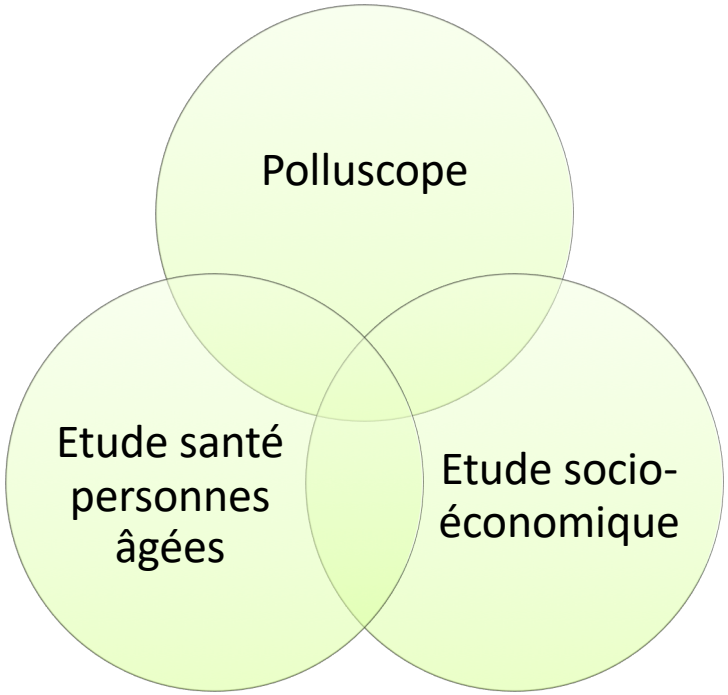
- Le projet Polluscope étudie les **opportunités** et les **limites** des micro-capteurs dans l'analyse de **l'exposition individuelle** à la pollution de l'air et de **ses effets sanitaires**.

- Propose une **plateforme** de **collecte**, de **gestion** et d'**analyse de données** issues de **capteurs environnementaux**, **d'activité** et **de santé**.

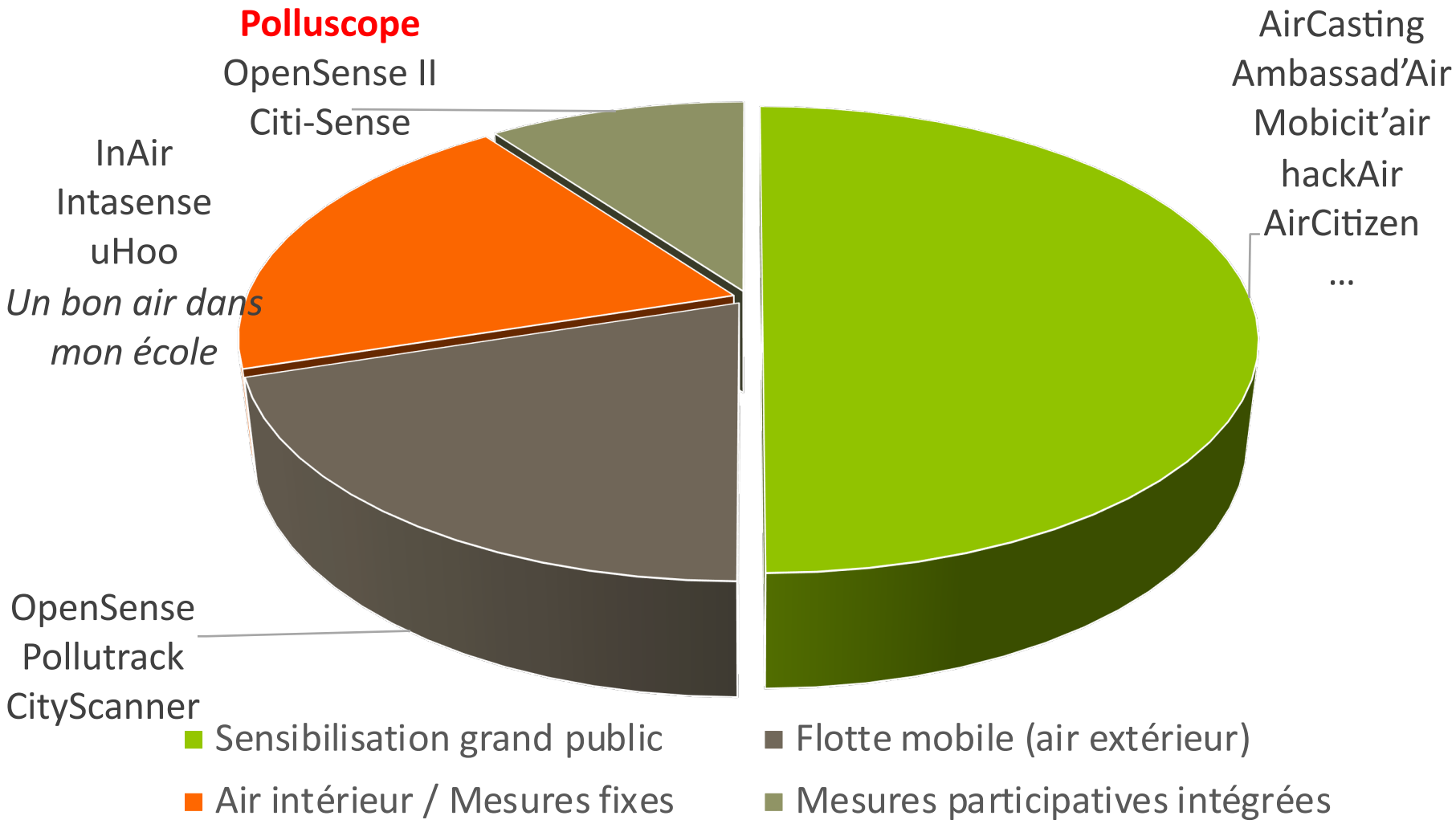


Synergie Polluscope – ACE-ICSEN

- **Renforcement de la dimension Santé et ajout d'une dimension socio-économique :**
  - **Analyse d'impacts sur la santé :**
    - **VIMA**: Focus sur l'exposition des personnes âgées
    - **INSERM-Sorbonne Université** : Exposologie & étude épidémiologique
  - **Impacts économique et social :**
    - **CEARC :** Changement de comportement, modèle économique de santé environnementale
  - **Une plateforme de données :**
    - **DAVID** : Accès aux données et aux outils Pollucope, ajout si besoin de fonctionnalités



Polluscope versus l'état de l'art



Mode opératoire

➢ Qui et combien de porteurs volontaires impliqués ?

- **Campagne** sur **2 ans**
- **60 participants vus une fois** sur le territoire de Versailles Grand Parc (**VGP**)
  –> 2ème saison reportée à Février 2021
- **20 à 25 vus deux fois** à Paris (de la cohorte **RECORD** de Basile Chaix)
- **20 patients prévus au printemps-été 2021 –** recrutement hôpitaux parisiens
- **+ 30 personnes âgées** par **VIMA** prévues dans le cadre d'**ACE-ICSEN**

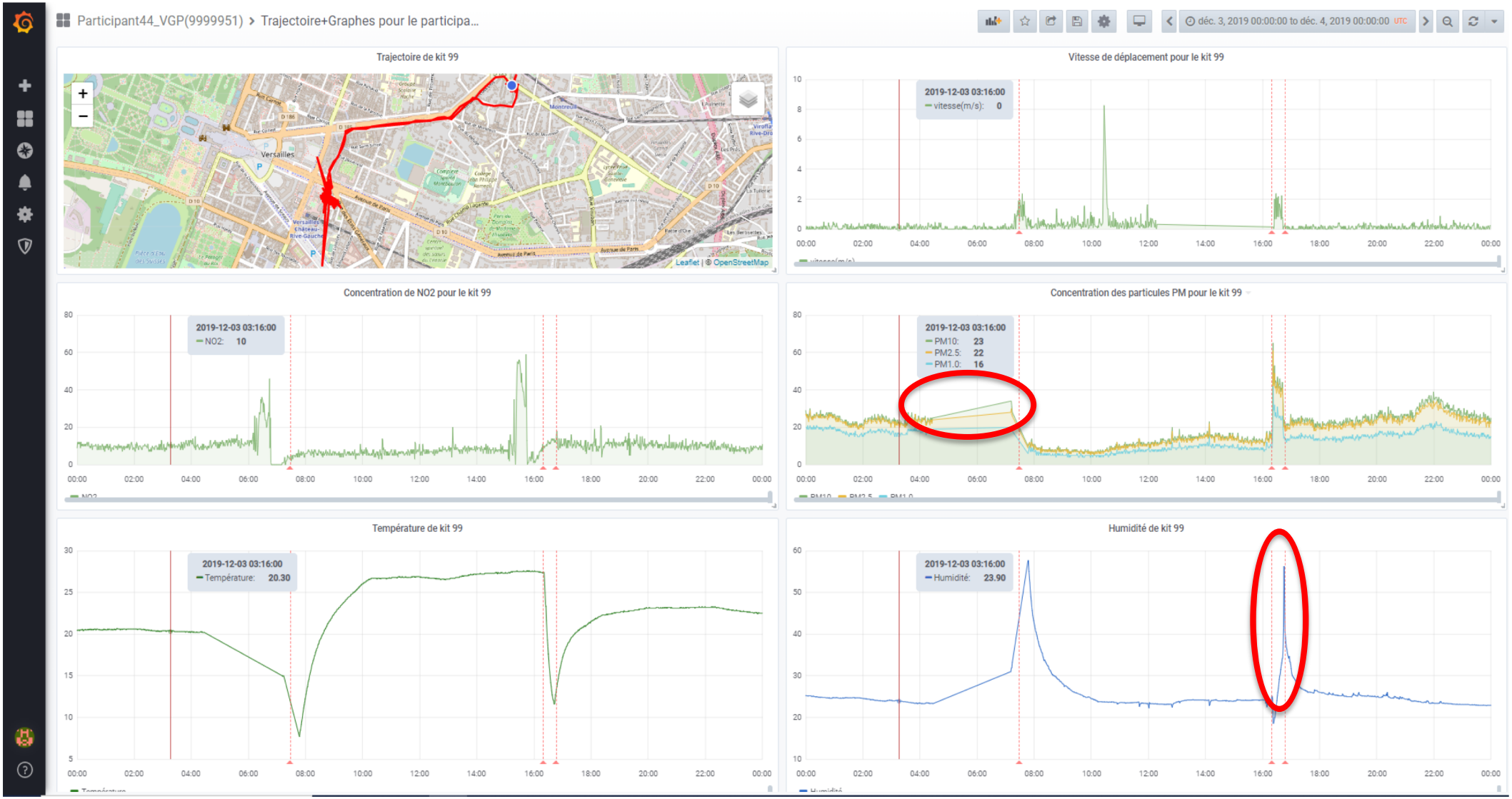
➢ Protocole :

- Portage en **continu une semaine** * 2 fois / an (hiver + été)
- Remise des **capteurs envi & santé** + Formation des participants
- **questionnaires**
- **Récupération des capteurs et entretien qualitatif**

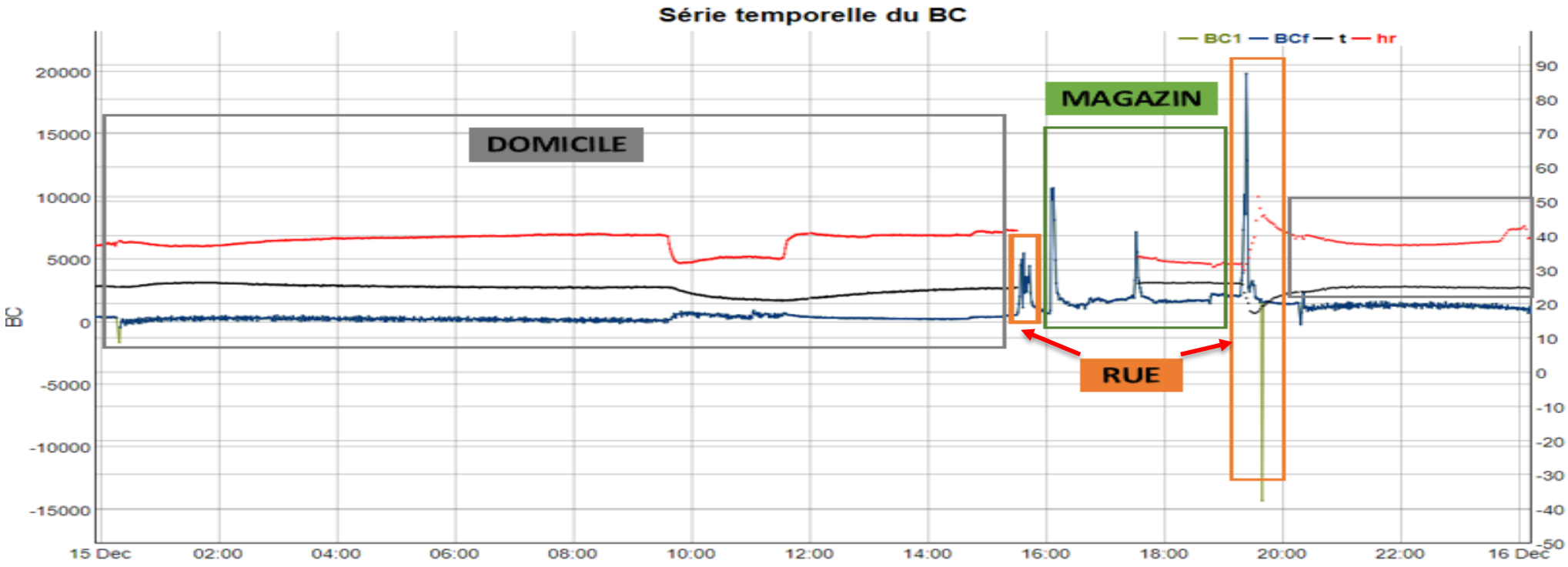
Plan de la présentation

- ❑ Contexte
- ❑ Présentation du projet POLLUSCOPE
- ☑ **Pré-traiement des données & Impacts sur l'analyse**

Aperçu des données pour un participant

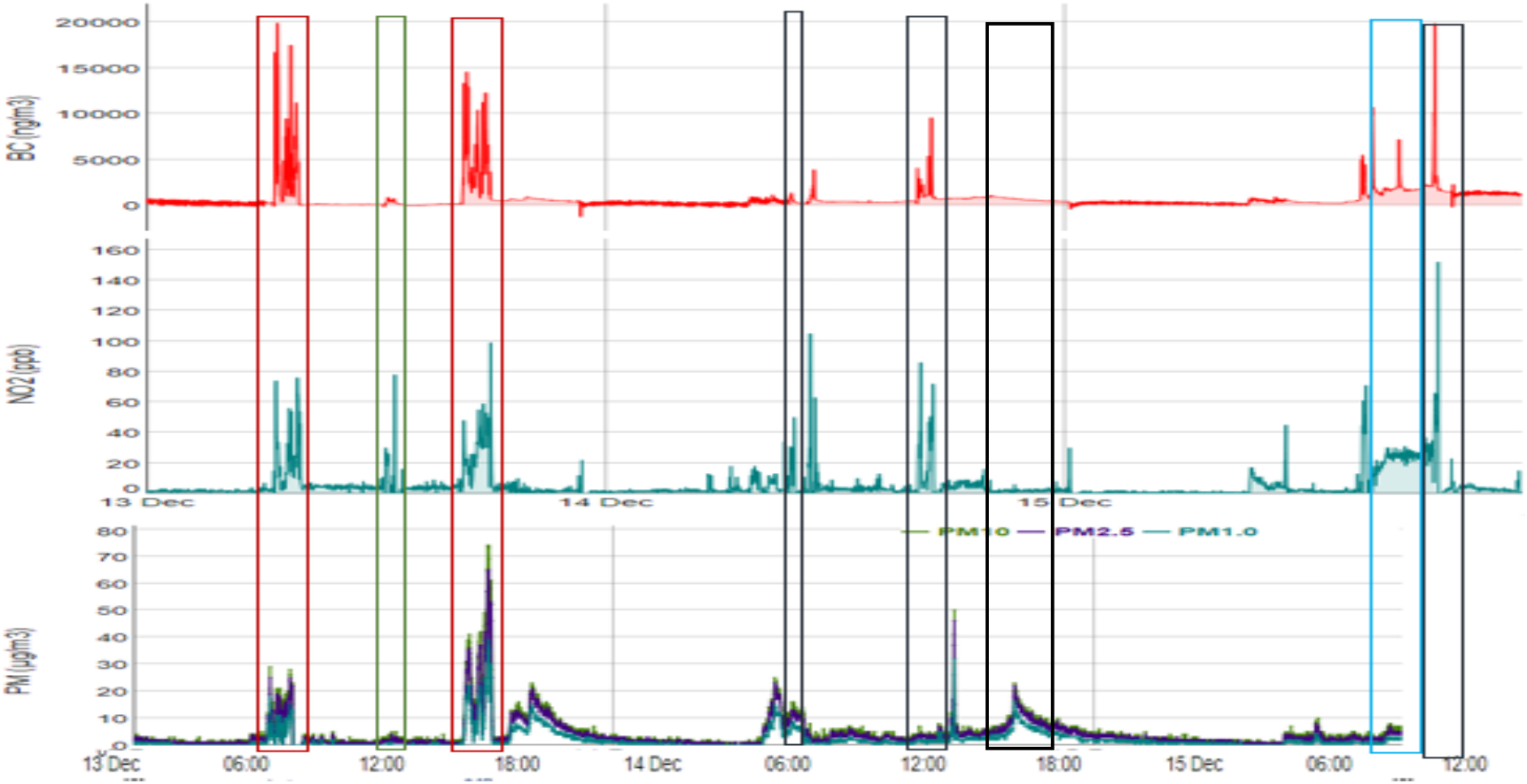


Corrélation avec le contexte



Corrélation Inter-capteurs et avec le contexte

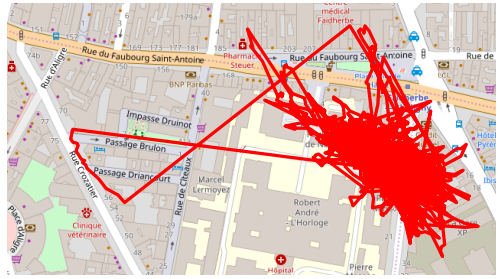
Voiture : Rouge ; Parc : Vert ; ***Rue : Noir*** *; Magasin : Bleu ;* ***Le reste : Intérieur (Domicile/Bureau).***

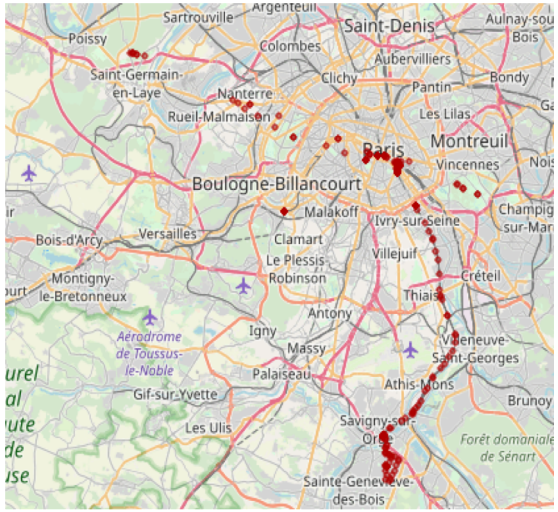


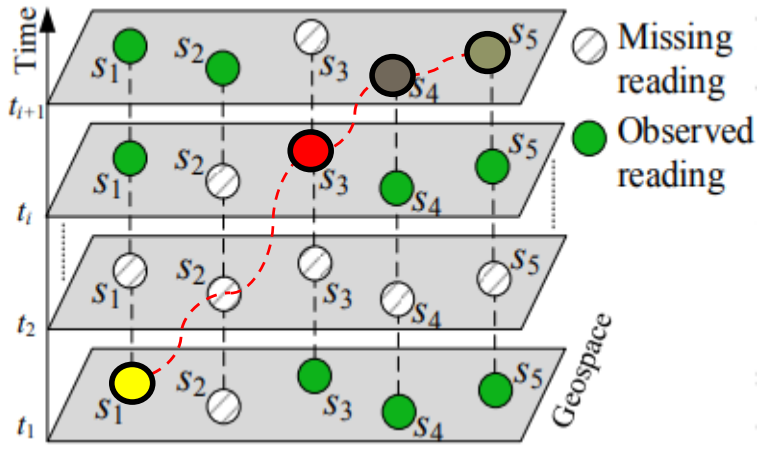
Problèmes de qualité des données ☹

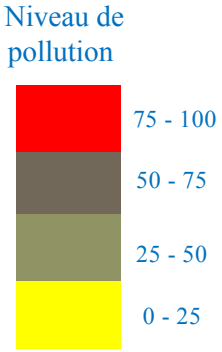
Données bruitées et données manquantes

- Perte de signal pour le GPS surtout en milieu fermé !
- Perte de données pour certains capteurs de QA
- ⇨ Nécessite **nettoyage** et imputation des données manquantes

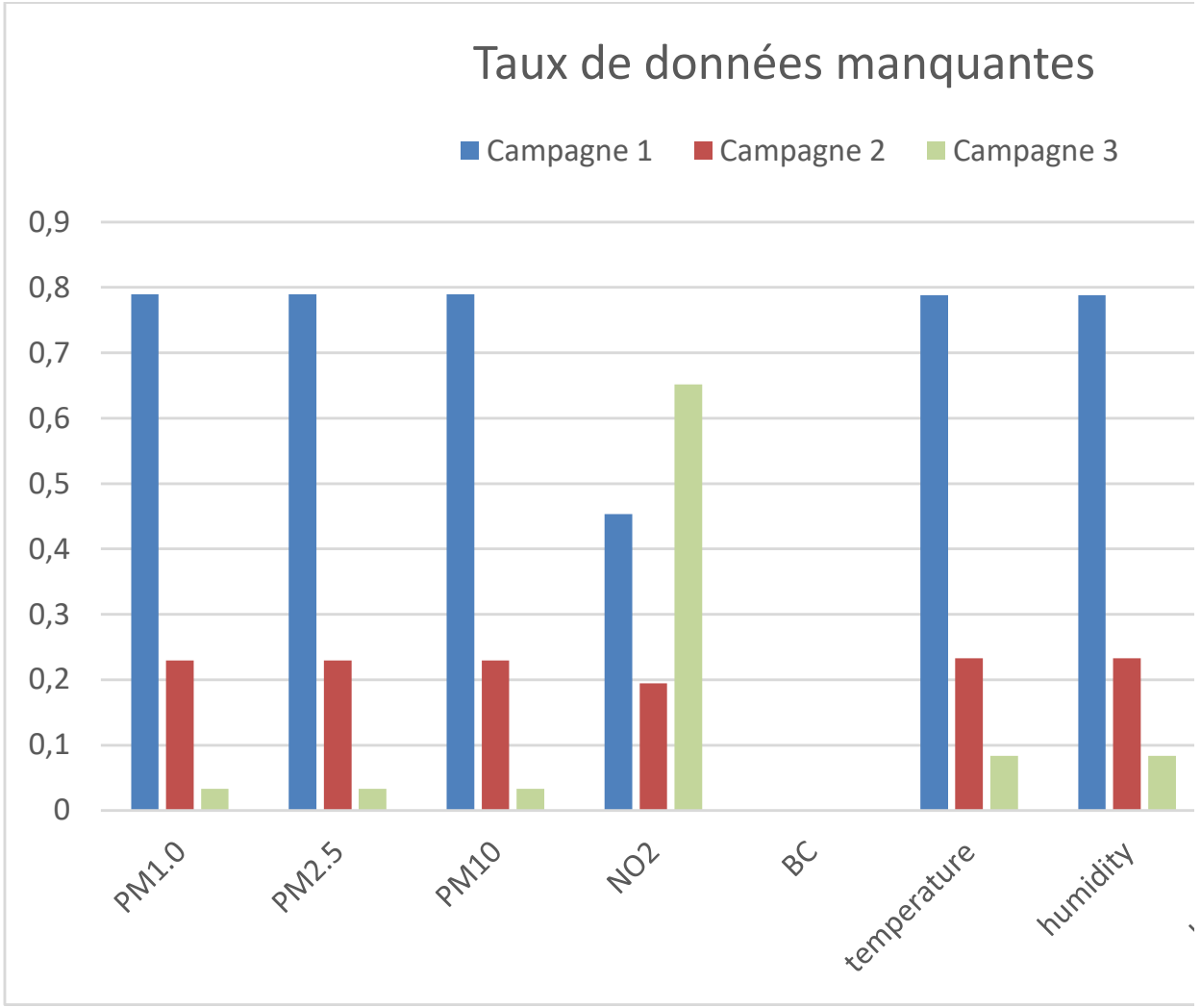






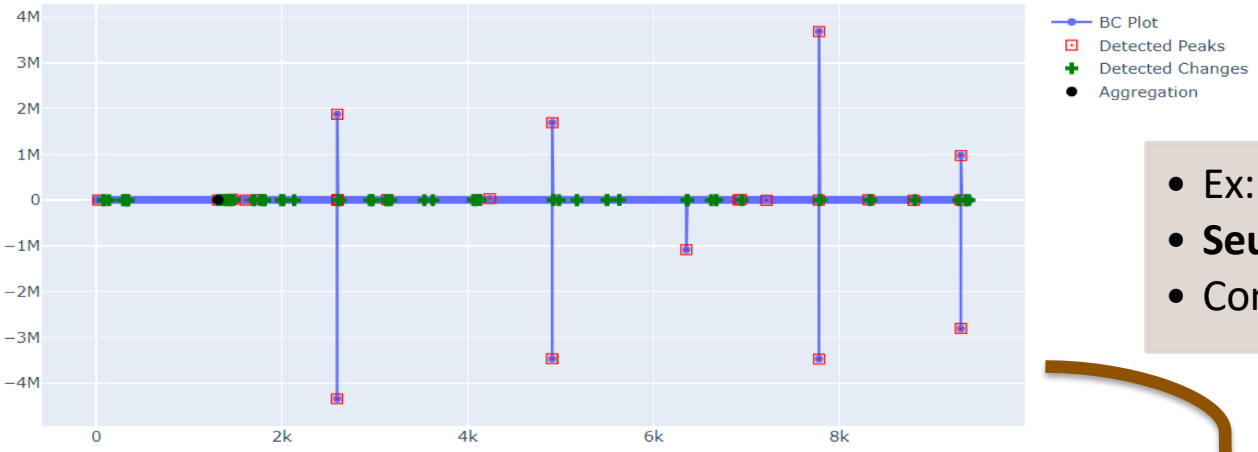


Estimation des pertes de données

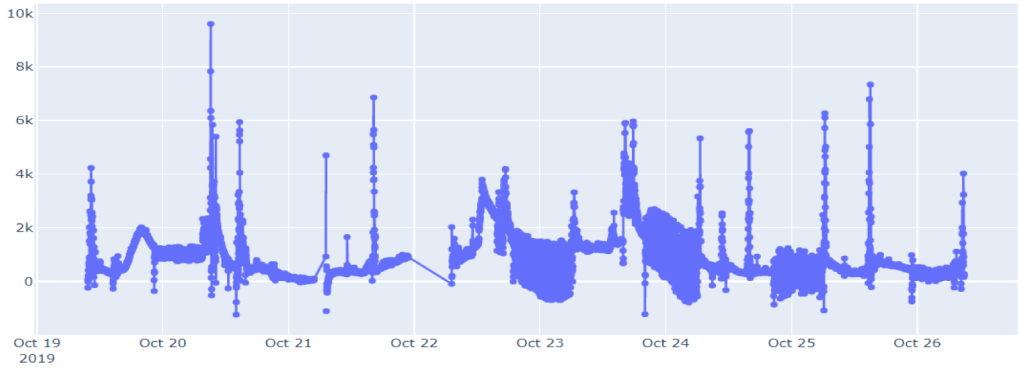


Exemple de traitement d'anomalies (1)

Avant / Après prétraitement



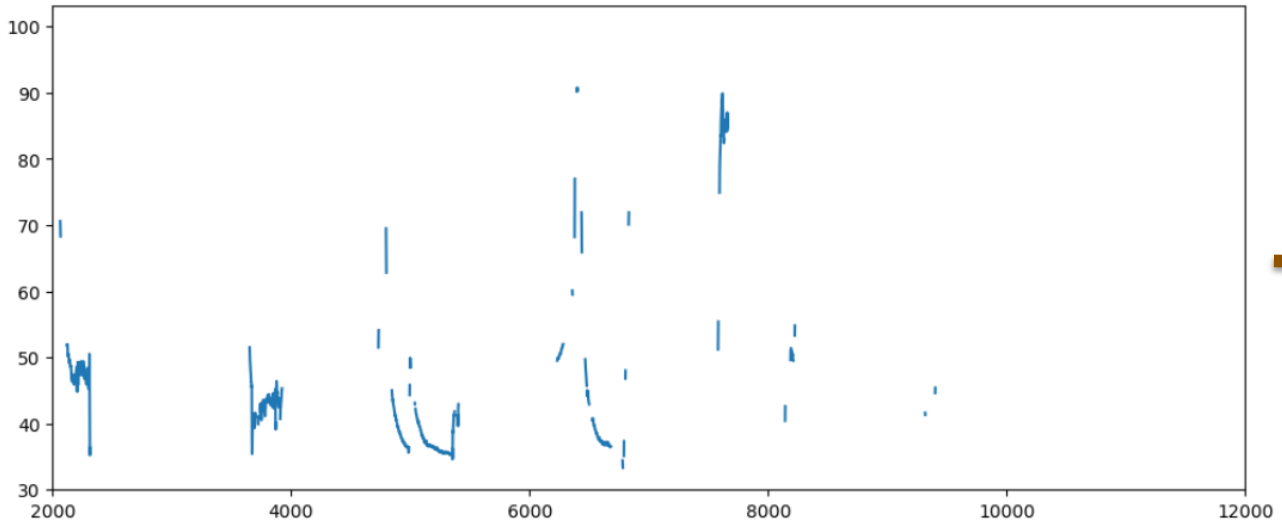
- Ex: Pics et valeurs aberrantes (Ici le Black Carbon)
- **Seuils pas évidents**
- Comment **dissocier les vrais des anomalies** ?

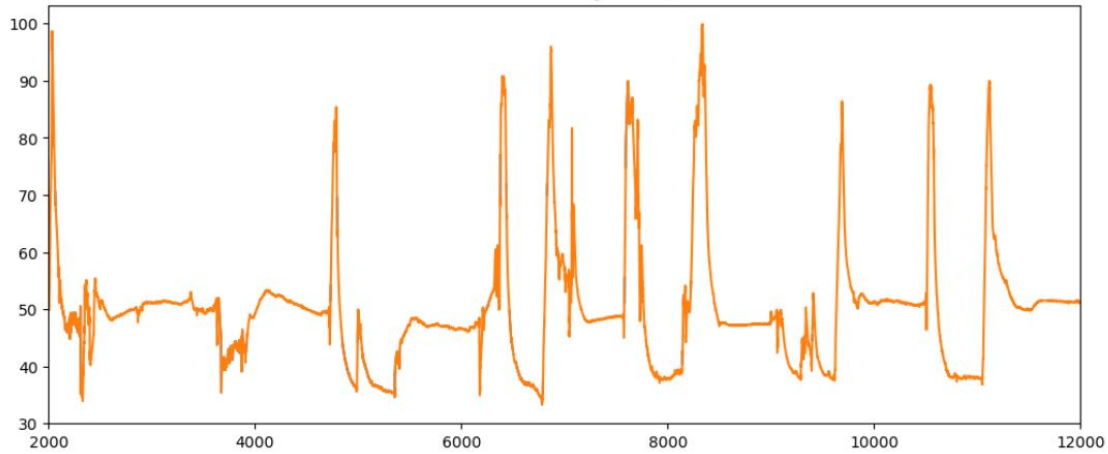


- ✓ Introduction d'un **seuil adaptatif**
- ✓ Identifier les pics causés par le contexte par **corrélation aux changements abrupts** selon d'autres mesures

Exemple de traitement d'anomalies (2)

Avant / Après prétraitement



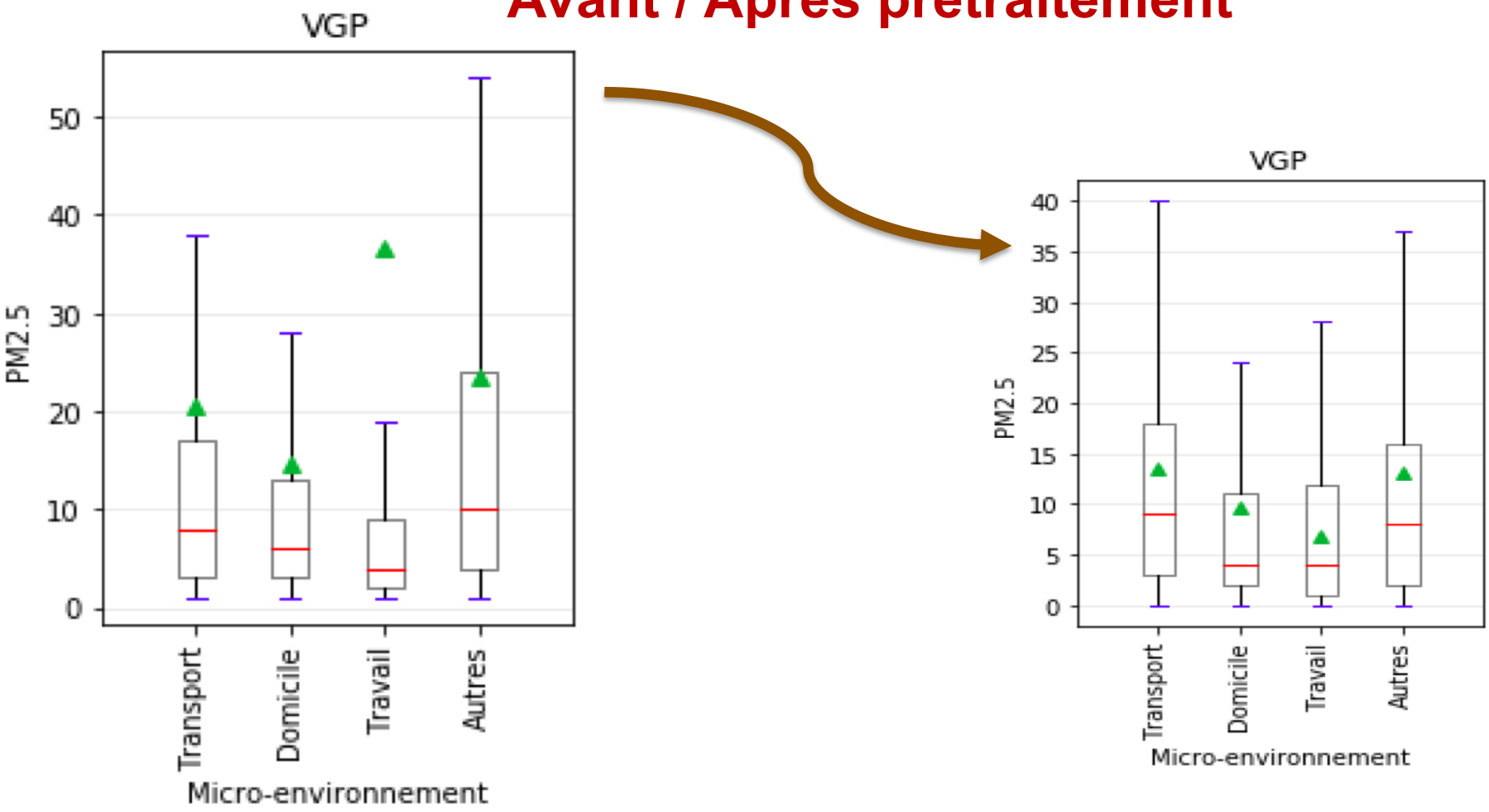


✓ Mise en oeuvre de méthodes **d'interpolation** & hybrides

Impact sur l'analyse statistique

Analyse par microenvironnement

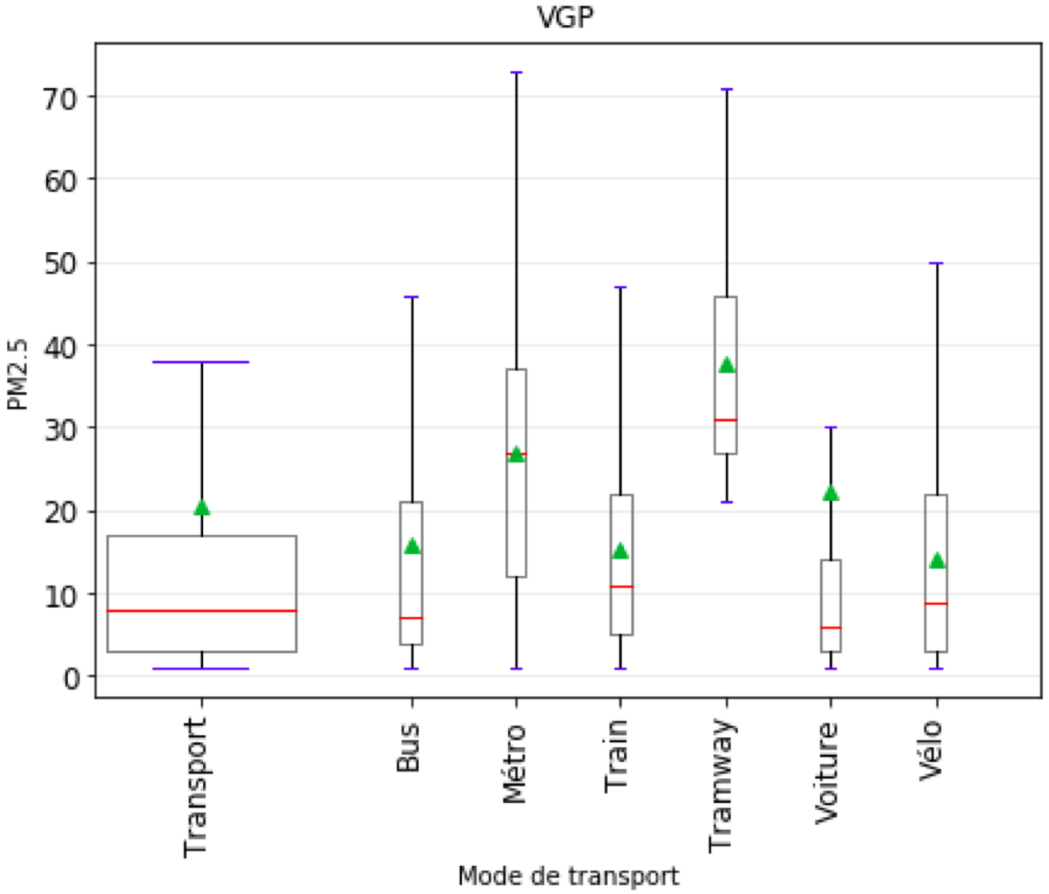
Avant / Après prétraitement

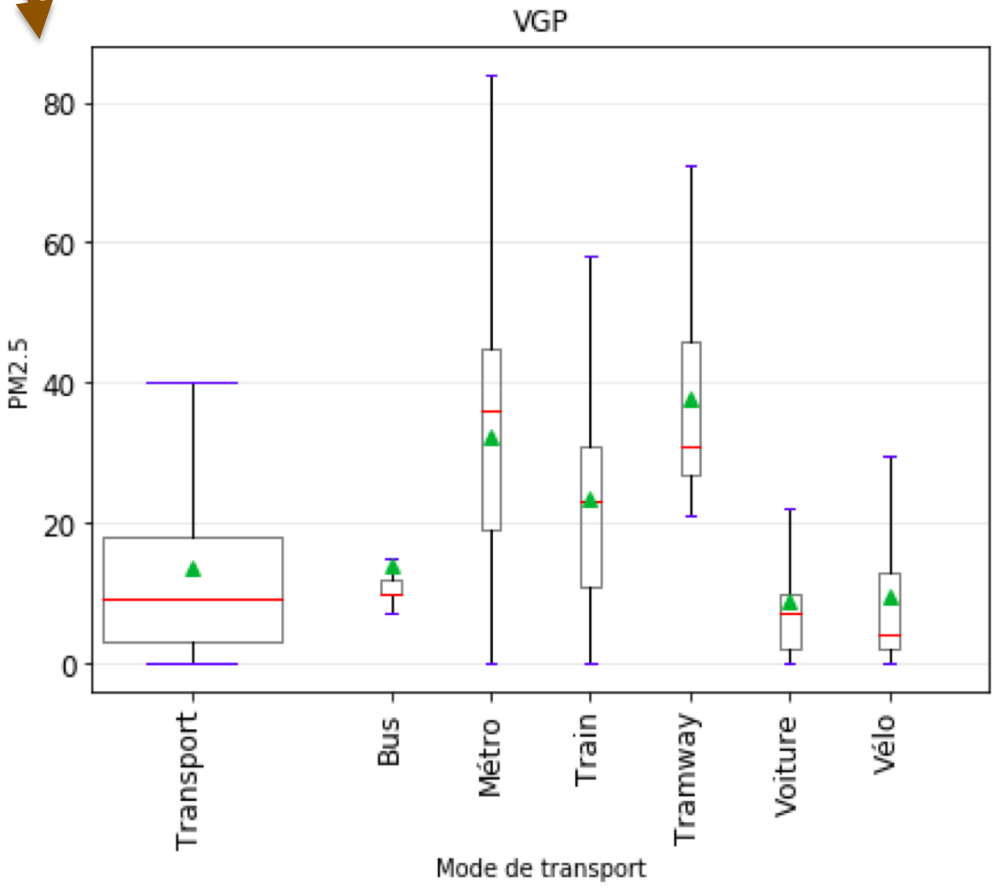


Impact sur l'analyse statistique

Analyse par microenvironnement

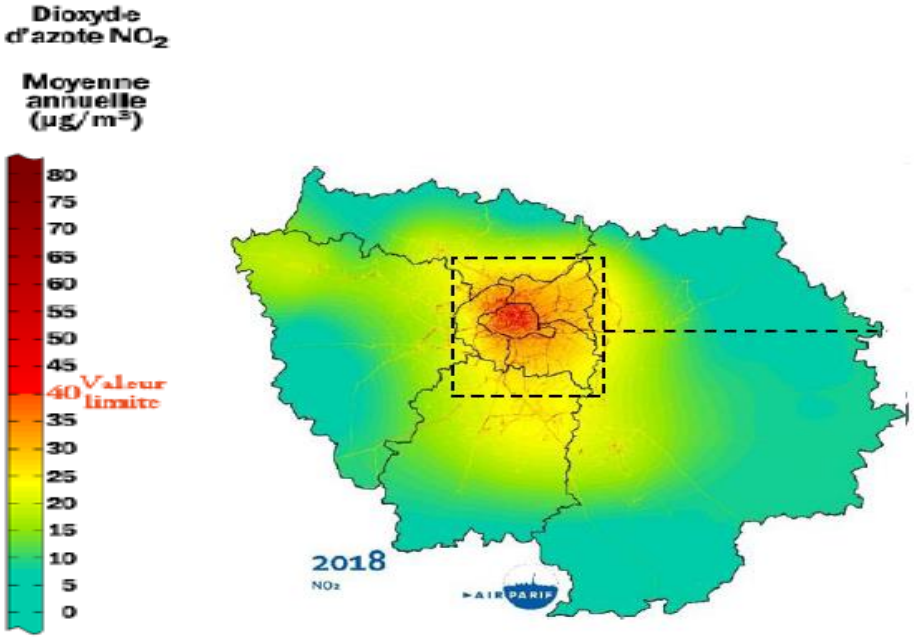
Avant / Après prétraitement



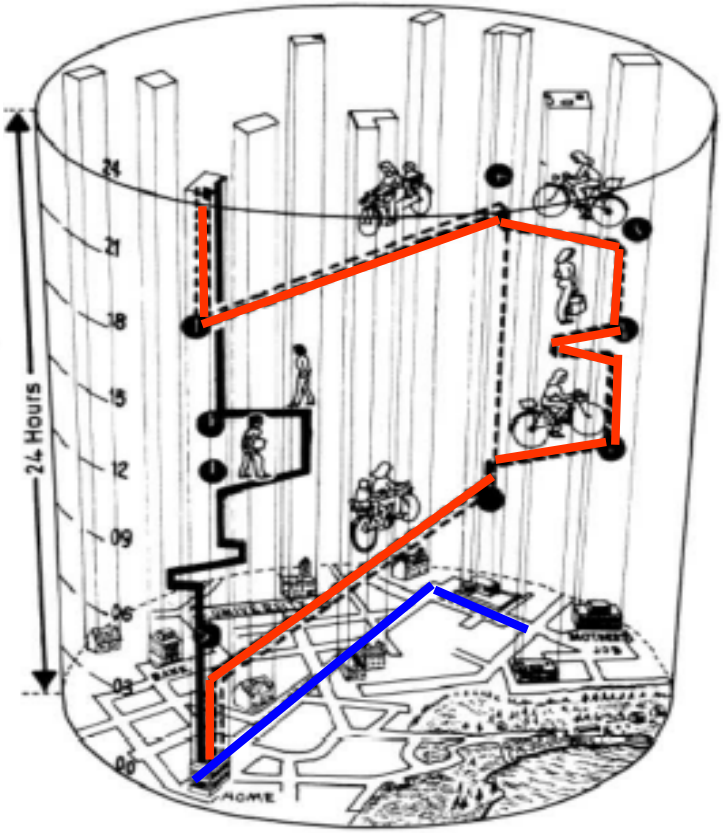


Comparaison
Micro-capteurs - Modèle de QA

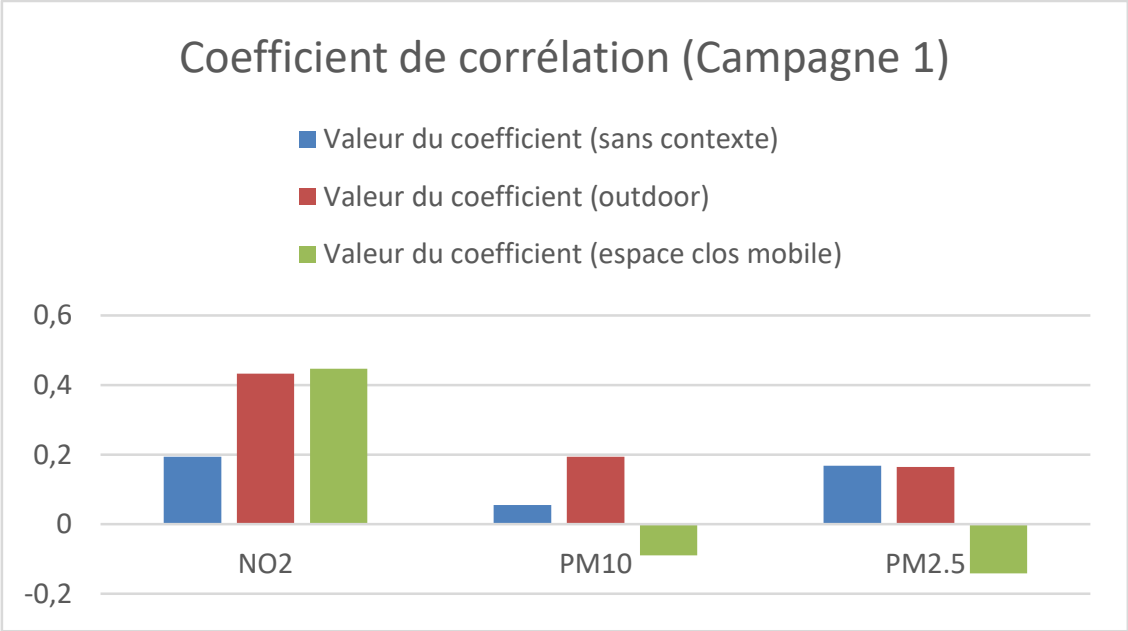
Données Airparif sur la qualité de l'air - Résultat de modèles



Données Polluscope sur l'exposition individuelle



Comparaison
Micro-capteurs - Modèle de QA

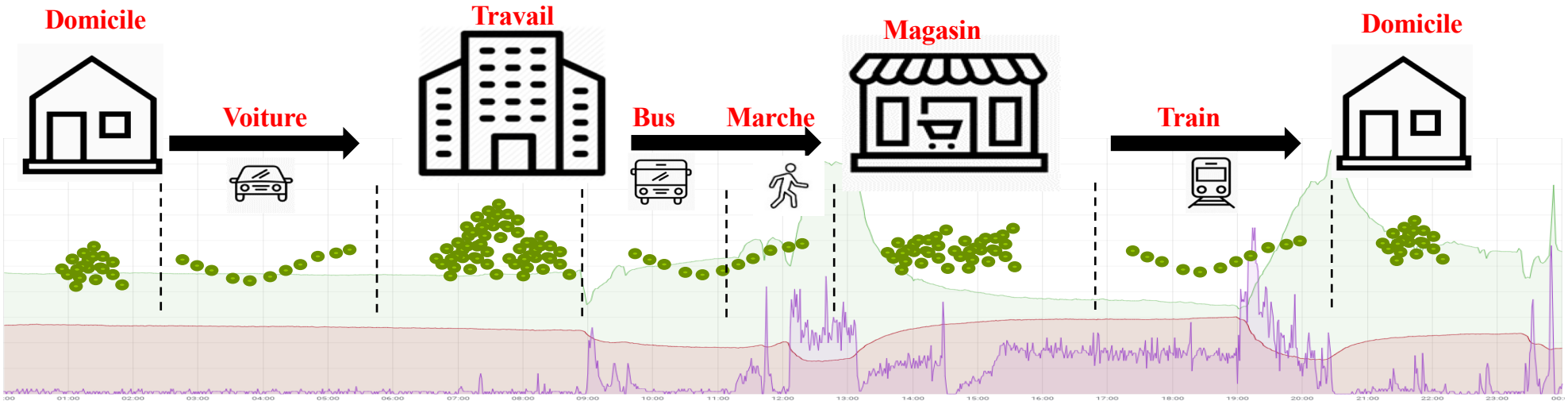


- Corrélation toute relative

Résultat préliminaire à confirmer après pré-traitement

Problème de contextualisation

- La qualité de l'air dépend fortement du micro-environnement
  - Tenir un journal de ses activités précis est difficile ☹



Problème de contextualisation

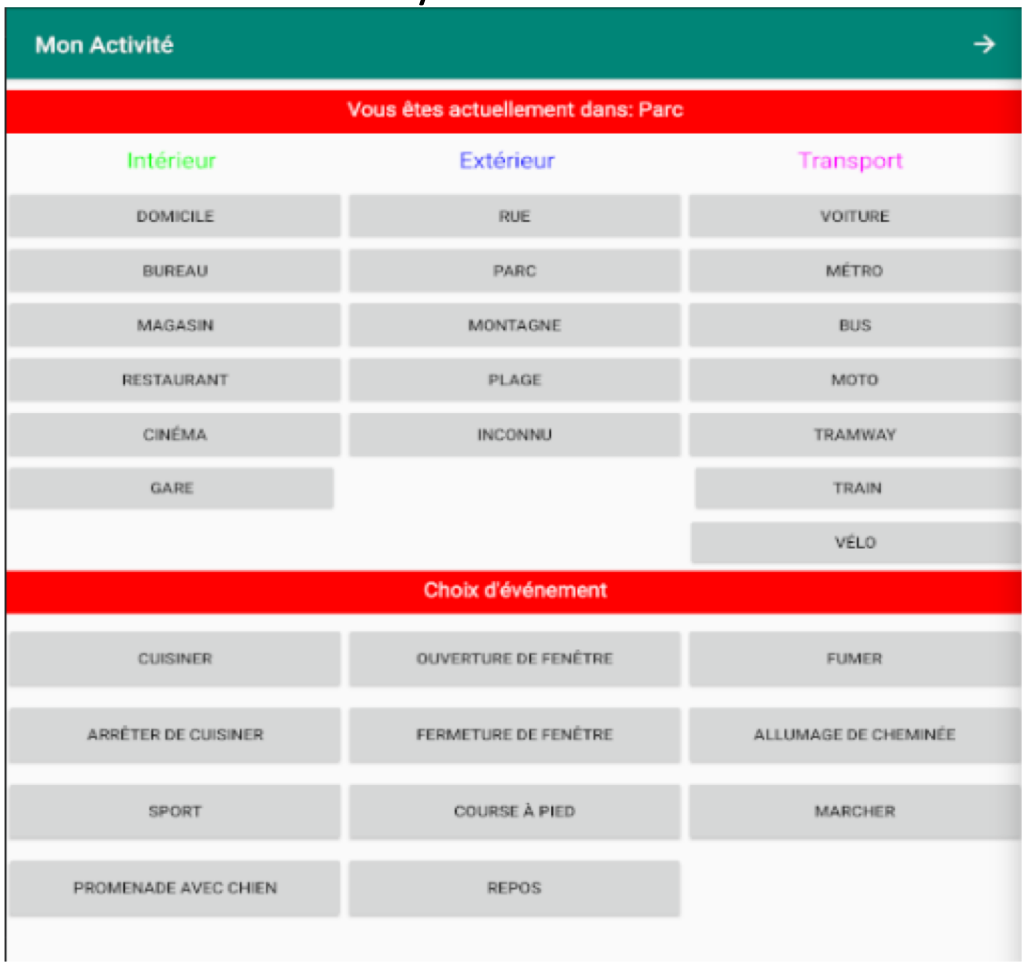
➢ Solutions :

1. Une **appli mobile** pour annoter les débuts d'activités / évènements

Malgré tout, annotations peu fiables !

Oubli de transitions

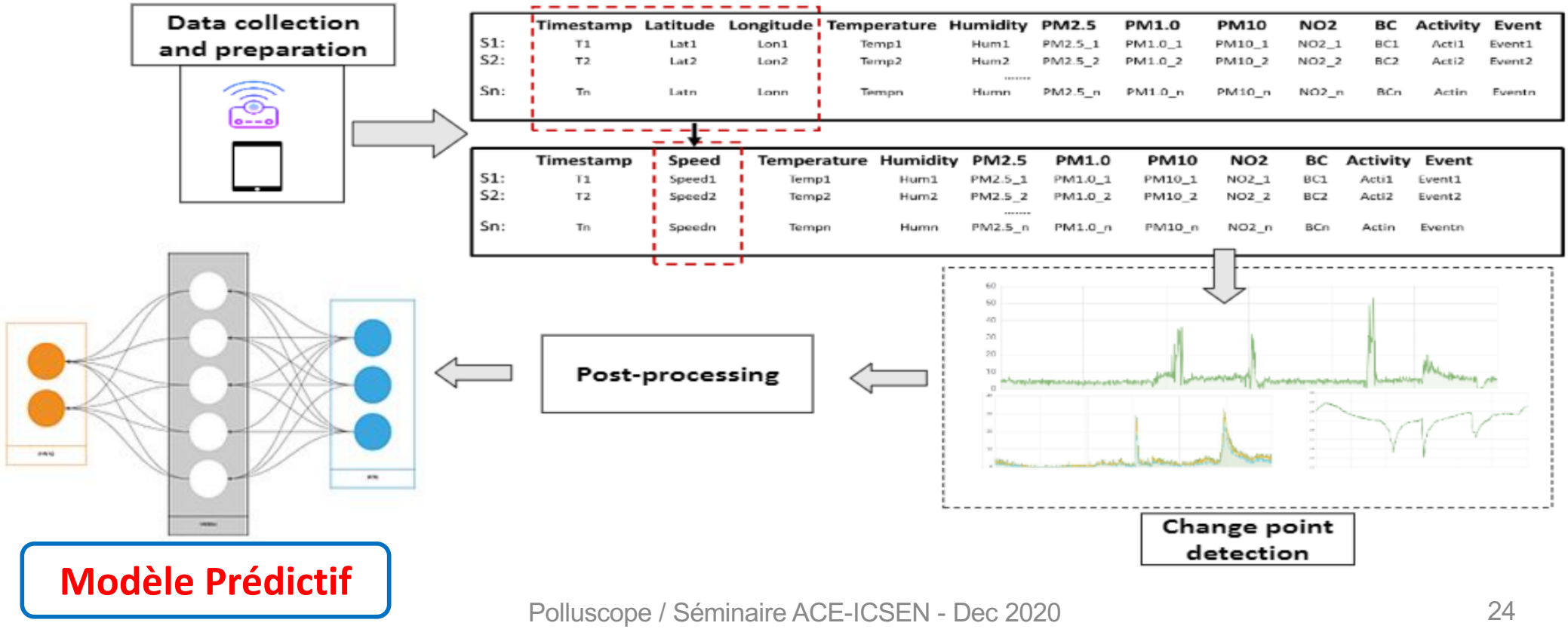
- *Anomalie d'horaire/durée*
- *Contradiction avec la mobilité*
- *Contradictions avec les évènements*



Problème de contextualisation

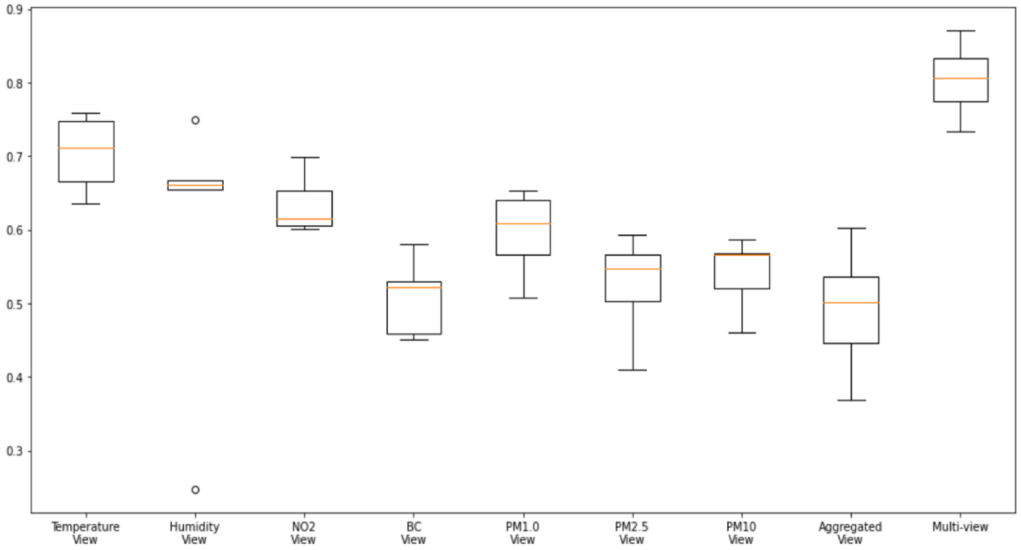
➢ Solutions :

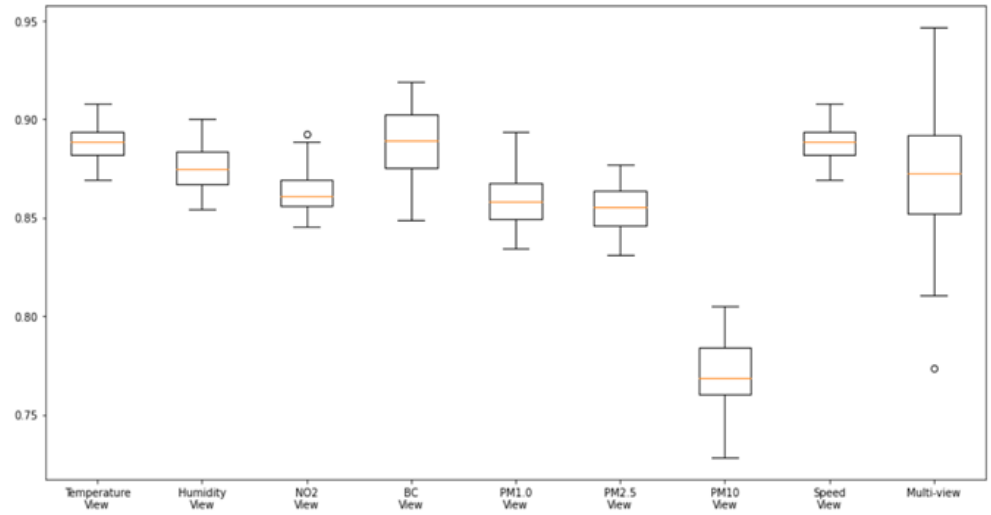
2. Entrainer un **modèle d'apprentissage** automatique pour **segmenter** les **longues séries temporelles** pour **détecter** **les micro-environnements**



Impact sur la qualité de prédiction

Avant / Après prétraitement





➢ On gagne plus de 25% en précision

Conclusions

Il y a un gap entre la collecte et l'exploitation des données

➢ Principaux problèmes :

- Les micro-capteurs donnent des mesures approximatives
- Difficile de distinguer le bruit des variations réelles & Perte de données
- Non respect du protocole
- Données sur le micro-environnement manquantes ou erronées

➢ Solutions :

Analyse des problèmes 1 par 1 pour adapter la solution

- ✓ Détection algorithmique des anomalies
- ✓ Méthodes d'estimation des données manquantes
- ✓ Utilisation d'une appli « Budget espace-temps »
- ✓ Fiabilisation « manuelle » d'un échantillon & entrainement d'un modèle prédictif.

⇨ **On arrive à une bonne caractérisation par micro-environnement**

Perspectives

➢ Suite de la campagne :

- Pour les ***cohortes VGP – Patients***
- Pour la cohorte VIMA (Personnes âgés --- > femmes enceintes)
- Introduction de nouveaux capteurs envisagée
- On prévoit un séminaire de restitution aux participants actuels et potentiels
- Ajuster le protocole / motiver les volontaires à le suivre

➢ Traitement et analyse :

Travaux à consolider et à compléter

- Détection d'évènements, apprentissage semi-supervisé, règles a priori
- Comparaison entre micro-environnements
- Indicateurs d'exposition
- Amélioration de l'interface et du système
- Anonymisation en vue du partage des données …

Quelques publications

- Brahem, M., El Hafyani H., Mehanna S., Zeitouni, K., Yeh, L., Taher Y., Kedad Z., Ktaish A., Chachoua M., Ray C. (2020) Data perspective on environmental mobile crowd sensing, In Intelligent Data-Centric Systems, *Intelligent Environmental Data Monitoring for Pollution Management, Academic Press.*
- Languille, B., Gros, V., Bonnaire, N., Pommier, C., Honoré, C., Debert, C., Gauvin, L., Srairi, S., Annesi-Maesano, I., Chaix, B. and Zeitouni, K. (2020) A methodology for the characterization of portable sensors for air quality measure with the goal of deployment in citizen science. *Science of the Total Environment*, 708, p.134698.
- El Hafyani H., Zeitouni, K., Taher, Y., Leveraging Change Point Detection for Activity Transition Mining in the Context of Environmental Crowdsensing. SIGKDD International Workshop on Urban Computing, UrbComp 2020.
- Noureddine, H., Ray, C. and Claramunt, C., 2020, June. Semantic trajectory modelling in indoor and outdoor spaces. In 2020 21st IEEE International Conference on Mobile Data Management (MDM)
- Brahem, M., Zeitouni, K., Yeh, and El Hafyani H.: Prospective Data Model and Distributed Query Processing for Mobile Sensing Data Streams. ECML/PKDD workshop, MASTER 2019.
- Brahem, M., Chachoua, M., El Hafyani, H. Z. Kedad, A. Ktaish, S. Mehanna, C. Ray, Y. Taher, R. Thibaud, L. Yeh and K. Zeitouni. "Polluscope – Vers un observatoire participatif de l'exposition individuelle à la pollution de l'air et de ses effets sanitaires", SAGEO 2019, Clermont-Ferrand 13-15/11/2019.
- Mustapha, A., Zeitouni, K., Taher, Y. (2018) Towards Rich Sensor Data Representation - Functional Data Analysis Framework for Opportunistic Mobile Monitoring. GISTAM 2018: 290-295

+ 1 thèse (Baptiste Languille encadrée par LSCE-Airparif, plusieurs rapports de stages, ….

➢ **Soumission prochaine d'un article commun ACE-ICSEN - Polluscope !**

Campagne Santé

Philippe Agearther

Labo VIMA