Un Modèle de Trajectoire Multidimensionnel dans le Contexte de la Collecte Participative par Micro-Capteurs

Hafsa El Hafyani, Karine Zeitouni, Yehia Taher, Laurent Yeh, Ahmad Ktaish

DAVID Lab, UVSQ - Université Paris-Saclay
45 Avenue des Etats-Unis 78000 Versailles, France
surname.name@uvsq.fr

Résumé. La qualité de l'air est l'un des principaux facteurs de risque sanitaire. Sa dégradation continue au cours des dernières décennies, en particulier en milieux urbain et péri-urbain, est désormais une préoccupation centrale dans nos sociétés. Les micro-capteurs connectés offrent la possibilité de mesurer l'exposition individuelle à la pollution atmosphérique partout et de manière continue, tout en permettant la contribution à un réseau d'observation. Le concept de collecte participative ou citoyenne est principalement basé sur cette technologie. À ce jour, les données qui en résultent sont sous-exploitées en raison de l'écart important entre la donnée brute et l'information intelligible. L'objectif de cet article est de proposer une approche basée sur l'OLAP dans le but de réduire cet écart. On introduit une méthodologie et un modèle de données multidimensionnel conçus pour analyser différentes dimensions des trajectoires individuelles enrichies de données de pollution, et ce à différents niveaux de granularité. Le modèle de données est suffisamment générique pour servir de référence dans d'autres scénarios d'analyse.

1 Introduction

Avec les progrès de l'Internet des objets, ainsi que l'utilisation généralisée du GPS et des capteurs intégrés, des applications ont vu le jour pour collecter des séries de mesures de capteurs géoréférencées. Une classe d'applications produisant ce type de données est connue sous le nouveau paradigme appelé Mobile Crowd Sensing (MCS), qui permet à des volontaires de collecter des données acquises par un boîtier multi-capteurs et un appareil mobile. Plusieurs applications basés sur le MCS existent [1], certaines concernent le captage du bruit (Ambiciti, 2021), du rayonnement (OpenRadiation, 2021), ou de la pollution de l'air comme dans notre contexte du projet Polluscope [2].

1. Le groupe de travail CASPA sur les capteurs et les sciences participatives maintient une liste de projets sur le site https ://caspa.fr/projets/
2. Projet ANR Polluscope : http ://polluscope.uvsq.fr

1.1 Contexte et Motivations

Dans le cadre du projet Polluscope, chaque participant est équipé d'un kit de capteurs et d'un appareil mobile qui permettent l'acquisition et la transmission des mesures ainsi que les coordonnées GPS associées sous forme de séries de données géoréférencées. Les participants collectent, grâce au boitier multi-capteur, des mesures de la qualité de l'air dont les particules de différents diamètres (PM10, PM2.5 et PM1), le dioxyde d'azote NO_2, le noir de carbone ou Black Carbon (BC), ainsi que des données climatiques comprenant la température et l'humidité relative. L'appareil mobile est également utilisé pour collecter les positions GPS. Par ailleurs, une application mobile est fournie aux participants afin qu'ils puissent renseigner le contexte des mesures. Ainsi, il leur est demandé d'indiquer le type d'endroit (appelé micro-environnement) à chaque fois qu'ils en changent. Ils renseignent également des évènements particuliers qui ont un impact sur les concentrations de polluants et donc sur leur exposition. Ce type d'information est appelé communément budget espace-temps. Ces annotations sont très importantes en MCS. Elles permettent d'interpréter les mesures observées car celles-ci sont largement dépendantes du type d'environnement (intérieur, extérieur ou dans les transports). Sans ces informations, les mesures collectées ne peuvent pas être interprétées correctement. En outre, elles donnent un aperçu à un niveau d'abstraction plus élevé le long des trajectoires des participants.

La combinaison de la localisation spatiale avec des mesures continues et des annotations se traduit par des trajectoires enrichies sémantiquement. En plus des mesures de l'air ambiant telles que la température et les polluants atmosphériques, ces trajectoires sont enrichies d'informations contextuelles telles que les micro-environnements des participants (exemple : maison, bureau, restaurant, etc.) et les événements liés à la pollution de l'air (exemple : fumer).

Une analyse multidimensionnelle sur des données aussi complexes est hautement souhaitable, car elle permettrait l'exploration des données suivant plusieurs perspectives. Ces données de trajectoires complexes décrivent en effet plusieurs facettes d'analyse et d'exploration, qui peuvent être regroupées en quatre perspectives : (1) La perspective longitudinale qui permet spécifiquement l'évaluation de l'exposition à la pollution sur la dimension individuelle (i.e., les participants). (2) La perspective spatiale qui donne une vue spatiale du phénomène mesuré (i.e., par région) avec une cartographie de ce phénomène et la comparaison de sa variation selon les régions. (3) La perspective temporelle qui est une facette importante de l'analyse multidimensionnelle consistant à analyser la variabilité des mesures à différents moments de la journée ou entre différentes saisons. (4) Les vues multi-échelles sont la dernière facette. Elles permettent l'exploration des données de MCS à différents niveaux de granularité en se basant sur le concept de l'escalade hiérarchique qui sera développé en détail dans la Section 3.

1.2 État-de-l'art

Suite à l'énorme génération de données spatio-temporelles, il est communément admis que les techniques d'entrepôts de données non spatiales sont insuffisantes pour exploiter pleinement la dimension spatiale des données géolocalisées (Rivest et al. (2005) et Jensen et al. (2017)). Il est devenu nécessaire de repenser les concepts OLAP traditionnels. Les solutions proposées telles que Spatial OLAP (SOLAP) combinent les fonctionnalités d'OLAP aux fonctionnalités des Systèmes d'Information Géographiques (SIG) pour l'analyse de données géolocalisées. Elles permettent d'effectuer une exploration multidimensionnelle des données qui

peuvent être présentées sous des formes détaillées ou agrégées (Bimonte et al. (2007)). Cependant, dans un modèle SOLAP, l'attribut spatial est représenté comme un objet cartographique (i.e., des points, des lignes et des polygones). Le modèle SOLAP soulève le problème d'établissement de la hiérarchie de la dimension spatiale. Typiquement, la hiérarchie spatiale est représentée par les relations topologiques (i.e., inclusion, intersection) entre les membres de niveaux spatiaux identiques et/ou différents. Ceci affecte l'exactitude du processus d'agrégation. De plus, le fait que les mesures soient des séquences d'événements horodatés fait appel à des traitements particuliers. L'OLAP séquentiel a été proposé par Lo et al. (2008) pour les opérations OLAP sur les séquences (S-OLAP). Un événement dans un système S-OLAP se compose d'une ou plusieurs **dimensions** et de **mesures**. Chaque dimension peut être associée à une **hiérarchie de concepts**. S'il existe un ordre logique parmi un ensemble d'événements, ces événements peuvent former une séquence. Un ordre logique peut être basé sur un autre attribut (exemple l'attribut *time*). Pourtant, dans le contexte du MCS, les événements ne sont pas aussi denses et réguliers que les mesures, et ils n'indiquent pas nécessairement un ordre logique.

Dans le contexte des objets mobiles, un modèle de données sur mesure avait été proposé dans Wan et Zeitouni (2006) où les concepts de dimensions continues et de faits continus permettent de capturer le fait spatio-temporel de mobilité dans un réseau pré-défini. Une méthode d'indexation adaptée permet de répondre efficacement aux requêtes agrégats spatio-temporelles. L'intérêt de ce modèle est de permettre des requêtes spatiales et temporelles à la volée, sans se limiter à un découpage préalable de l'espace ou du temps. L'inconvénient réside dans la difficulté d'implémenter ce modèle. Définir une granularité et un référentiel spatial et temporel pour les dimensions est une solution souvent adoptée dans les faits. C'est le cas dans le modèle proposé pour l'analyse des activités spatio-temporelles dans Savary et al. (2004). Ces travaux sont les plus similaires à notre contexte, mais ils étaient limités à des trajectoires sans mesures associées. Le travail de Kang et al. (2018) est aussi similaire à notre contexte. Cependant, contrairement à ce papier, nos capteurs ne sont pas fixes. De plus, notre travail ne vise pas la prédiction de la QA mais l'exploration des données selon différentes dimension. On intègre bien les méthodes d'apprentissage pour identifier les micro-environnement et à l'avenir dans la désagrégation pour compléter les valeurs manquantes. Dans ce travail, nous proposons un modèle des données multidimensionnel pour la modélisation et l'analyse des données des trajectoires riches tout en prenant en considération la particularité des données de MCS, ainsi que la nature et la sémantique spécifique du budget espace-temps déclaré par les participants (i.e., micro-environnements et événements).

1.3 Objectifs et Contributions

Dans cet article, nous étudions les capacités des systèmes de traitements analytiques en ligne (OLAP) à gérer les données de MCS et identifions leurs limites. Nous proposons un modèle multidimensionnel pour l'analyse et l'exploration des données de MCS. Ce modèle capture toutes les facettes des données et produit des résultats d'analyse suivant différentes perspectives. Pour ce faire, nous adoptons la méthode par discrétisation des dimensions spatiale et temporelle en fixant une granularité minimale. Pour la dimension spatiale, la zone d'étude

est découpée en pixels de taille prédéfinie [3]. Le temps est également discrétisé (dans notre application, la granularité de la minute a été choisie). Les trajectoires (dont les coordonnées à l'origine sont des valeurs continues et le temps continu) sont converties en références aux pixels (i.e., les membres de dimension spatiale) et à l'unité temporelle. En effet, le modèle raster convient à l'exploration à différentes échelles de manière hiérarchique. Nous adoptons un index spatial pour optimiser l'accès et les requêtes sur la dimension spatiale. Les résultats préliminaires de la mise en œuvre de notre modèle montrent l'efficacité de notre approche en utilisant des données réelles collectées dans le contexte du projet Polluscope.

Le reste de l'article est organisé comme suit. La section suivante présente les principales caractéristiques et les besoins pour la modélisation multidimensionnelle de trajectoires riches. La section 3 présente notre modèle de données conçu pour l'analyse et l'exploration de données dans le contexte de la collecte participatives de données. La section 4 présente l'implémentation de notre modèle de données. Une expérimentation est menée sur des données environnementales réelles issues de campagnes MCS. Enfin, la dernière section présente nos perspectives et résume nos principales contributions.

2 Challenges

Les données mesurées par des capteurs mobiles peuvent être représentées par des séries temporelles multivariées qui se caractérisent par la présence d'une dimension spatiale formant des trajectoires. De manière équivalente, on peut voir ces données comme des trajectoires spatio-temporelles enrichies par des mesures supplémentaires tout au long de la période de collecte. Ce type de données présente un certain nombre de caractéristiques et de défis :

— **Autocorrélation Spatiale et Temporelle.** Du point de vue de la modélisation, un aspect distinctif de ces séries de données est l'autocorrélation spatiale, bien connue comme la loi de Tobler Miller (2004) qui signifie que les objets proches ont tendance à être plus similaires que les objets distants. Il en va de même pour les observations consécutives d'un même capteur qui correspond à des phénomènes physiques dont la variation est généralement continue. Ainsi, contrairement aux tuples du modèle relationnel, les données collectées ne sont pas indépendantes les unes des autres. De plus, leur accès et leur analyse sont souvent ciblés sur région spatiale et intervalle de temps. Par conséquent, les dimensions spatiales et temporelles doivent être organisées et indexées de manière à faciliter leur accès et leur manipulation.

— **Imperfection des Données.** Les données collectées dans le contexte du MCS sont souvent imparfaites, en raison des limites de précision et de justesse des capteurs. De plus, dans un même kit, les capteurs peuvent être hétérogènes en terme de cycle temporel de mesure, donc de granularité temporelle variées. Par exemple, l'un peut générer des mesures à la minute tandis qu'un autre peut mesurer toutes les cinq minutes. Non seulement les mesures des capteurs, mais aussi les annotations des participants ne sont pas garanties d'être exactes. Elles sont sujettes à des erreurs, peuvent comporter du bruit ou des valeurs manquantes. Le renseignement du budget espace-temps n'est pas toujours suivi

3. Le découpage choisi suit le même découpage que l'Association de Surveillance de la Qualité de l' Air en Ile de France AirParif. Il est fixé à 12,5x12,5 m2 à Paris, 25x25m2 en petite couronne et 50x50m2 ailleurs dans le reste de la région.

par les participants, ce qui se traduit par un manque de fiabilité de ces données. Il n'est pas évident d'appréhender le modèle en présence de ces imperfections.

— **Hétérogénéité des Données.** Une caractéristique notable est l'hétérogénéité dans l'espace et dans le temps. Dans le MCS, nous avons un large spectre de capteurs avec des caractéristiques différentes pour la sensibilité, la fréquence d'échantillonnage et l'immunité au bruit. Les données collectées à partir de tous les capteurs doivent être fusionnées, ce qui pourrait conduire à des mesures à des intervalles de temps irréguliers et à des problèmes de données manquantes. On pourrait observer des horodatages très dense ou, au contraire, clairsemés dans différents cas. Certains capteurs peuvent être hors ligne pendant des heures. Ils peuvent passer en mode veille lorsque l'appareil est statique, puis passer en mode "burst" en mobilité. Certains sont configurés pour réduire la transmission des données lorsque la variation est inférieure à un seuil prédéfini. Ces sources de données hétérogènes doivent être prises en compte dans le modèle, et une vision harmonisée des données est hautement souhaitable afin de faciliter leur traitement et leur analyse. Un choix judicieux du niveau de granularité le plus bas dans l'espace et dans le temps est nécessaire afin de fournir un bon compromis entre précision et coût du stockage.

— **Multi-Granularité.** Par ailleurs, l'une des caractéristiques les plus fondamentales des données des capteurs mobiles est la diversité de leur granularité pour les dimensions temporelle et spatiale. Le domaine temporel est généralement représenté à des granularités temporelles différentes. L'entité spatiale peut être représentée à l'aide d'une représentation hiérarchique qui décrit la subdivision du domaine spatial en différentes régions ou cellules. La combinaison de plusieurs ensembles de données avec plusieurs granularités ou la modification de la granularité d'un ensemble de données sont des tâches d'analyse importantes que nous devons traiter. Nous devons également définir un cadre qui permet le passage d'une granularité à une autre.

— **Volume de Données.** D'énormes quantités de données sont collectées en continu à partir d'appareils mobiles omniprésents dotés de capteurs dans différentes zones géographiques. Cela conduit à adopter des techniques de traitement du *Big Data* afin de permettre une analyse interactive efficace des données.

— **Vélocité des Données.** Comme les données sont ajoutées en continu au fil du temps, il devient difficile de maintenir des agrégats précalculés intégrant les données les plus fraîches. Selon l'ampleur du projet de MCS en terme de nombre de participants, de nombre et de fréquences des mesures et de durée de la période de collecte, il peut être nécessaire d'adopter des techniques avancées de gestion de données massives combinant les données historiques et les flux temps réel.

3 Modèle de Données Multidimensionnel

Cette section présente le modèle multidimensionnel pour les données MCS ainsi que les dimensions pertinentes. Notre modèle sert à l'analyse et à l'exploration des données dans le contexte des capteurs mobiles. Il permet de visualiser les données du MCS à différentes échelles selon certaines hiérarchies de ses dimensions et sous différentes perspectives. La figure 1 le schéma du modèle multidimensionnel. La définition même des dimensions est un défi, du fait des particularités précitées des données du MCS. Ainsi, le modèle doit tenir compte de

l'autocorrélation spatiale et temporelle, de l'hétérogénéité des mesures, de la représentation de trajectoires riches, et enfin de la sémantique d'intervalles du budget espace-temps. Nous discuterons de la définition des tables de faits et des dimensions dans les sections suivantes.

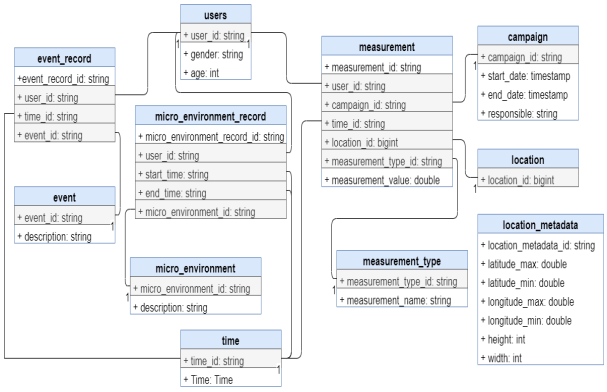

FIG. 1 – Modèle multidimensionnel proposé.

3.1 Vue générale du modèle

Une particularité des données collectées dans le cadre du MCS est la combinaison de la géolocalisation avec des observations et des mesures dans le temps, aboutissant à des "trajectoires riches". Comme exemple d'application en cours d'exécution du modèle proposé, nous considérons une base de données obtenue à partir du projet Polluscope. Une cohorte de volontaires a été équipée de capteurs individuels collectant plusieurs mesures de la qualité de l'air combinées aux données GPS. Dans le projet Polluscope, trois campagnes de collecte de données ont été menées. Chaque campagne est caractérisée par une date de début, une date de fin et un responsable. Chaque campagne s'était étalées sur 12 semaines avec une collecte réalisée en général une semaine sur deux (afin de vérifier et requalifier les capteurs). Plus de 103 personnes y ont participé. Ils ont été équipés d'un kit contenant des capteurs de pollution atmosphérique et une tablette avec GPS. Les capteurs collectent des mesures annotées dans le temps des matières particulaires (PM1.0, PM10, PM2.5), du NO2, du carbone noir (BC), de la température et de l'humidité relative. La tablette a servi à géolocaliser les participants et à renseigner leur budget espace-temps via une appli mobile développée à cet effet. Les activités durent un certain temps et représentent des micro-environnements pouvant être des milieux intérieurs (domicile, bureau, restaurant, etc.), des milieux extérieurs (parc, rue, etc.) ou encore des modes de transport (voiture, bus, métro, etc.). En outre, le participant renseigne les événements, désignant des actions temporaires sur une période brève, liées à la pollution de l'air (par exemple, ouvrir une fenêtre, cuisiner, fumer, allumer la cheminée, etc.).

L'organisation des données doit tenir compte de la propriété d'auto-corrélation des données (en maintenant la localité des données spatialement proches) et de multi-granularité (par une représentation spatiale hiérarchique). La solution proposée passe par l'indexation spatiale et est décrite dans la section suivante.

3.2 Indexation Spatiale

Le modèle multidimensionnel est principalement destiné à l'analyse et l'agrégation flexible des données. Les membres de dimensions ont des valeurs finies et généralement connues d'avance, de manière à servir au regroupement et à l'agrégation des mesures de la table des faits. C'est pourquoi les dimensions dans les entrepôts de données sont discrètes. Or à l'origine, les données spatiales et temporelles varient dans un domaine continu. Afin de permettre la représentation de la dimension spatiale, nous transformons les coordonnées exactes des positions reportées (comme la latitude et la longitude) en valeurs discrètes référençant un numéro de pixel d'une grille rectangulaire avec une résolution spatiale (ici de 50 m). De même, les données temporelles sont ramenées à la minute. De cette façon, les dimensions spatiales et temporelles peuvent être supportées par les systèmes OLAP, contrairement à la représentation originale de l'espace et du temps infinis. Reste la question du maintien de localité dans l'organisation de la dimension spatiale. Pour cela, nous adoptons l'indexation spatiale par indice de Hilbert, lequel définit une courbe de remplissage de même nom (ici d'un espace 2D). L'intérêt des indices de Hilbert est la localité (Moon et al. (2001)), autrement dit, les cellules voisines sont susceptibles d'être affectées à un indice de Hilbert proche. De plus, la courbe de remplissage de Hilbert montre une propriété fractale qui facilite l'exploration à différents niveaux de granularité spatiale, permettant le *Roll-Up* et *Drill-Down* dans la dimension spatiale à la manière d'un zoom dans les images. La figure 2 montre la représentation de données spatiales et de son indexation à l'aide de la courbe de remplissage de Hilbert. L'étendue spatiale est définie pour couvrir la zone d'étude (i.e., la région de l'Île-de-France). Notons que nous ne gardons que les cellules correspondant aux emplacements ayant des données GPS, ce qui est largement plus compact que des solutions comme les géo-cubes (OGC (2021)).

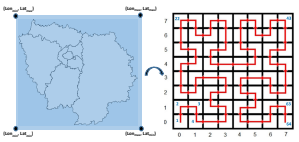

FIG. 2 – Représentation de la dimension spatiale.

L'approche de la rastérisation est avantageuse de différentes manières. Elle permet de dériver les zones d'activité statique (souvent en intérieur) des participants. Par exemple, nous pouvons découvrir les endroits où un participant passe le plus de temps en fonction des densités des cellules, ce qui peut être calculé simplement en comptant le nombre de mesures associés à la cellule (au pixel) par participant dans la période. Cette approche permet également de détecter les valeurs spatiales aberrantes et le bruit spatial. De plus, elle fournit une pixellisation à surface égale qui facilite le partage de la dimension spatiale avec les formats raster existants (geoTIFF, NetCDF, ...) permettant par exemple d'intégrer des sources externes comme les données d'Airparif (2021).

3.3 Modèle de Données

Dans cette section, nous présentons formellement notre modèle multidimensionnel typique du contexte de MCS. La figure 1 présente notre modèle conceptuel. Elle propose trois tables de faits. La première table `mesure` stocke les données des capteurs. La table de faits `mesure` relie les valeurs de la qualité de l'air, représentées par l'attribut `measurement_value`, à cinq dimensions :

1. `users` qui attribue à chaque participant ses données démographiques
2. `campaign` affecte à chaque campagne un `campaign_id` et donne des informations sur sa date de début (i.e., `start_date`), sa date fin (i.e.,`end_date`) et la personne en charge (i.e., responsable) (i.e., `responsible`).
3. `location` est la dimension spatiale qui donne des informations sur les indices de Hilbert attribués à chaque cellule de la grille.
4. `time` est la dimension temporelle.
5. `measurement_type` représente le dictionnaire des types de mesures collectées par les participants qui sont dans notre cas : `PM2.5`, `PM1.0`, `PM10`, `NO2`, `Black Carbon`, `Temperature` et `Humidity`. Il peut s'agir de n'importe quelle observation ou mesure, comme par exemple le bruit, l'ozone, le pollen, etc. Le schéma est donc générique et applicable à n'importe quel contexte d'application de collecte participative par capteurs.

Par la suite, notre modèle définit deux tables de faits supplémentaires : `micro_environment_record` et `event_record`. La table de faits `micro_environment_record` relie les informations des participants à leur budget espace-temps. Elle décrit les micro-environnements avec une heure de début et de fin de présence. La dimension `micro_environment` contient la description des micro-environnements intérieurs et extérieurs ainsi que la liste exhaustive des moyens de transport utilisés par les participants. De la même manière, la table de faits `event_record` est similaire au `micro_environment_record`, sauf que les événements sont caractérisés par un horodatage temporaire, car ils sont brefs. La dimension `event` présente le dictionnaire exhaustif des événements liés à la pollution atmosphérique. Nous soulignons que notre modèle multidimensionnel aborde la particularité des trajectoires riches collectées dans le contexte du MCS en tant que séries de données géolocalisées. Mais, en plus de cela, notre modèle multidimensionnel peut intégrer des informations externes en tant que dimensions. Nous étendons l'entrepôt de données traditionnel pour prendre en charge des informations géographiques et temporelles externes provenant d'autres sources. Par exemple, nous pouvons enrichir notre entrepôt de données avec des sources géographiques externes telles que des couches cartographiques (telles que des routes ou des limites administratives) et des points d'intérêt (PoI). Les sources temporelles externes peuvent être, par exemple, des événements temporaires liés aux phénomènes observés par ailleurs, comme un incendie qui émet des polluants ou encore un confinement entraînant une baisse importante du trafic qui est source de pollution.

3.4 Scénarios d'utilisation

Le modèle multidimensionnel ainsi proposé permet d'analyser des données à différentes échelles et hiérarchies. En outre, il permet d'examiner et de modéliser les données suivant

différentes vues, plus généralement à partir de différentes facettes de dimensions, et plus précisément à différents endroits et périodes de temps. Nous soulignons son utilité dans l'analyse et l'exploration de trajectoires riches et annotées, spécialement dans le contexte du MCS, en introduisant quelques cas d'utilisation :

— **Analyse Longitudinale** qui fait référence à l'analyse et à l'évaluation de l'exposition individuelle au fil du temps. Il permet de suivre l'évolution de l'exposition individuelle à la pollution tout en détectant les périodes de fortes et faibles niveaux de pollution. On peut agréger les données sur des périodes de temps, comme les heures de pointe, le week-end, la semaine, etc. L'analyse peut également se décomposer en périodes passées par micro-environnement, ce qui est précieux pour comprendre et comparrer les contextes d'expositions.

— **Analyse Spatiale** consiste à détecter les endroits présentant des phénomènes de pollution de haut niveau. Il permet de mettre en évidence les niveaux de pollution dans différents lieux. Pour un participant, l'analyse spatiale permet de suivre le niveau de pollution tout au long de sa trajectoire. De même, pour tous participants confondus, nous pouvons identifier les lieux présentant des phénomènes de pollution de haut niveau. L'analyse spatiale peut se généraliser aux types de micro-environnements tels que rapportés par les participants, ce qui ouvre la voie à la traçabilité du micro-environnement avec l'exposition la plus élevée et la plus faible pour chaque participant ou, pour tous participants confondus.

— **Analyse Temporelle** fait référence à l'analyse des mesures au fil du temps. En plus de l'analyse longitudinale précitée qui se concentre sur la dimension individuelle, nous nous intéressons à d'autres analyses temporelles qui combinent les données de plusieurs participants. Un exemple consiste à analyser l'ensemble des mesures rapportées pour différentes périodes de temps (e.g., heures de pointe, jour de la semaine, week-end, mois, saison, année). Un autre consiste à se concentrer sur un micro-environnement ou un domaine spécifique pour évaluer l'impact de certaines politiques sur le niveau de pollution au fil du temps.

4 Implémentation et Expérimentation

4.1 Contexte et Outils

Notre modèle est implémenté sous *Spark* 2.1.2, *Python* 3.6.1, *Hadoop* 2.8.3 et *Hive* 2.1.0. *Hive* est utilisé pour gérer le schéma de *Spark* tandis que ce dernier est utilisé pour le calcul. *Hive* a été adopté pour l'interrogation, l'analyse et l'entreposage des données. Il est construit sur *Hadoop* et fournit une interface de type SQL pour l'interrogation et le traitement des données. *Hive* utilise le HDFS de *Hadoop* pour le stockage des données. Grâce à son modèle relationnel étendu, *Hive* peut répondre à des requêtes simples et fournit un opérateur OLAP optimisé. Cela permet de créer des cubes OLAP basés sur les données gérées par *Hive*.

Comme mentionné ci-dessus, plus de 103 volontaires ont participé à la phase de collecte de données qui dure une semaine pour chaque participant. Les données GPS sont collectées à une fréquence de 1 à 30 secondes, tandis que les mesures des polluants sont collectées toutes les minutes, engendrant plus de 10 millions de tuples de mesures dans le temps, plus quelques

annotations de données par le type de micro-environnement et des événements liés à la pollution.

La dimension spatiale est définie par les pixels (ensemble fini et plus facile à comparer) plutôt que par la position exacte. Elle est organisée selon l'ordre de l'indice de Hilbert. Cette organisation spatiale est avantageuse pour l'analyse dimensionnelle. En effet, en se référant à l'analyse longitudinale, un individu peut découvrir par exemple les pixels où la pollution élevée, la durée passée et générer ainsi des cartes de chaleur (heat map) de son exposition. De même, cette comparaison peut être effectuée pour différents participants combinés au même endroit (au niveau du pixel fin ou à tout niveau de la hiérarchie spatiale), afin d'identifier les participants avec l'exposition la plus élevée. Comme nous pouvons le voir, il existe de nombreuses facettes différentes pour explorer les données. Toutes les combinaisons possibles doivent être accessibles. Dans la section suivante, nous présentons et discutons certaines combinaisons possibles des dimensions, i.e., la localisation, le temps, les types de polluants, etc.

4.2 Analyse Longitudinale

L'analyse longitudinale consiste à analyser l'exposition par participant au fil du temps. Elle capture la vue d'exposition individuelle. Dans notre contexte, nous avons l'intention de comparer l'exposition individuelle des participants à un profil d'exposition moyen fictif. Pour ce faire, nous calculons différents agrégats de polluants par participant et pour tous participants confondus afin de constituer le profil d'exposition moyen fictif. Nous tirons profit de l'opérateur `ROLLUP` pour naviguer dans la hiérarchie de dimension et explorer toutes les facettes possibles. Cet opérateur permet de répondre à des requêtes telles que les requêtes présentées dans l'exemple suivant.

Example 4.1. *Quelle est l'exposition individuelle au polluant PM2.5 ? Et quelles sont les périodes d'expositions maximales ?*

```
1 SELECT M.user_id,  day(T.time) AS Day,hour(T.time) AS Hour, avg(M.
      measurement_value)*count(*) AS Exposure, max(M.measurement_value) AS
      Peak_value
2 FROM measurement M, measurement_type MT, time T
3 WHERE MT.measurement_type_id = M.measurement_type_id AND T.time_id = M.
      time_id AND MT.measurement_name = 'pm2.5'
4 GROUP BY M.user_id, day(T.time),  hour(T.time) WITH ROLLUP
```

La première requête de l'exemple 4.1 retourne une vue sur l'exposition individuelle au PM2.5 ainsi que l'exposition globale pour tous participants confondus. Cela permet aux participants de comparer leurs expositions à celle des autres participants pour savoir s'il est plus ou moins exposé. La Fig 3 montre un extrait de la sortie de la première requête 4.1. Cette sortie exprime l'exposition au PM2.5 et sa valeur maximale par participant avec un roll up aux tous participants confondus (désigné par `null` dans l'attribut `user_id`). D'autre part, La Fig 4 montre un extrait de la sortie de la deuxième requête de l'exemple 4.1. La requête renvoie l'exposition individuelle aux PM2.5 à partir de la hiérarchie des participants (`user_id`) en descendant dans la hiérarchie temporelle (`Houret Day`). Les agrégats au niveau du participant sont désignés par `null` dans l'attribut `Hour`, et les agrégats pour tous participants confondus sont désignés par `null` dans l'attribut `user_id`. L'analyse longitudinale peut être envisagée pour couvrir l'exposition individuelle par micro-environnement au fil du temps (i.e., pour chaque

user_id	Exposure	peak_value
null	134457148	5010.0
99999M	73571633	5010.0
99999E	81236	202.0
99999D	60492	59.0
99999C	837	37.0
9999997	359	22.0
9999994	41455	109.0
9999993	1214	38.0
9999992	994677	23.0
9999991	2790727	350.0

FIG. 3 – Analyse longitudinale.

user_id	Day	Hour	Exposure	Peak_value
null	null	null	5948.0	5.0
9999921	null	null	5948.0	5.0
9999921	28	null	5948.0	5.0
9999921	28	18	2476.0	5.0
9999921	28	17	0.0	0.0
9999921	28	16	82.0	1.0
9999921	28	15	619.0	1.0
9999921	28	14	1074.0	1.0
9999921	28	13	1697.0	2.0

FIG. 4 – Analyse longitudinale à différentes granularités temporelles.

user_id	micro_environment	time	Exposure	Peak_value
null	null	null	5216138.0	2577.0
9999950	null	null	5216138.0	2577.0
9999950	Voiture	null	105960.0	90.0
9999950	Voiture	2019-11-03 12:26:21	157.0	1.0
9999950	Voiture	2019-11-02 18:51:10	7849.0	9.0
9999950	Voiture	2019-11-02 17:05:44	6011.0	17.0
9999950	Voiture	2019-11-02 15:29:51	1922.0	3.0
9999950	Voiture	2019-11-02 14:21:56	34326.0	90.0
9999950	Voiture	2019-11-02 10:10:52	3600.0	16.0
9999950	Voiture	2019-11-01 00:26:46	16131.0	44.0
9999950	Voiture	2019-10-31 20:03:36	27616.0	49.0
9999950	Voiture	2019-10-31 17:15:00	8348.0	18.0
9999950	Rue	null	2627.0	30.0
9999950	Rue	2019-11-02 14:20:00	277.0	3.0
9999950	Rue	2019-11-02 12:30:00	520.0	3.0

FIG. 5 – Analyse longitudinale par micro-environnement.

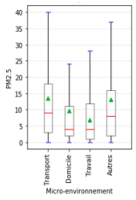

FIG. 6 – Exposition aux PM2.5 par micro-environnement.

user_id	micro_environment	Exposure	Peak_value
null	null	130189865	5010.0
99999M	null	73580029	5010.0
99999M	Voiture	6282954	5010.0
99999M	Vélo	53721	17.0
99999M	Rue	20180	41.0
99999M	Magasin	198021	455.0
99999M	Domicile	3681732	79.0
99999M	Bureau	63343421	5009.0
99999E	null	81236	202.0
99999E	Voiture	41	4.0

FIG. 7 – Exposition individuelle aux PM2.5 par micro-environnement.

intervalle de temps que l'individu a passé dans ce micro-environnement). Similaire à la requête de l'exemple 4.1, et en adaptant la requête pour acquérir `micro_environment`, notre modèle multidimensionnel peut être utilisé pour explorer l'exposition individuelle à la pollution par participant et par micro-environnement au fil du temps. La Fig 5 représente un extrait de la sortie. Nous pouvons aller plus loin dans l'analyse et illustrer la perspective de l'analyse globale individuelle indépendamment du temps et de l'espace. Nous considérons la dimension utilisateur uniquement pour calculer l'exposition à la pollution par micro-environnement. La Fig 7 représente la sortie de l'exposition individuelle aux PM2,5 par micro-environnement, et la Fig 6 représente la concentration de PM2,5 par micro-environnement.

4.3 Analyse Spatiale

L'analyse spatiale aborde le problème de la détection des endroits présentant des phénomènes riches en pollution. Elle consiste à répartir ce phénomène dans la localisation géographique et sélectionner les endroits visités par tous les participants, présentant un haut ou bas niveau de pollution. Il faut mentionner que les courbes de Hilbert est par définition hiérarchique. Cela signifie qu'en effectuant un zoom avant et/ou un zoom arrière, nous pouvons systématiquement passer au niveau suivant de la hiérarchie. Pour passer à un niveau supérieur de la hiérarchie, il faut diviser l'indice de Hilbert par 2^{2n}. Comme le montre l'exemple 4.2, cette représentation d'indexation spatiale permet de répondre à des requêtes telles que : *Quel est le niveau de pollution des endroits fréquemment visités par les participants à différents niveaux de la hiérarchie spatiale ?* `P64` et `P16` indiquent des niveaux grossiers de la hiérarchie où chaque pixel contient respectivement un groupement de 64 et 16 cellules fines.

Example 4.2. *Quel est le niveau de pollution à différents niveaux de la hiérarchie spatiale ?*

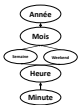

FIG. 8 – Hiérarchie temporelle.

weekend	Day	Hour	Exposure	Peak_value
null	null	null	21854.0	31.0
7.0	null	null	21854.0	31.0
7.0	3.0	null	21854.0	31.0
7.0	3.0	23.0	6158.0	14.0
7.0	3.0	22.0	3204.0	17.0
7.0	3.0	21.0	2385.0	6.0
7.0	3.0	20.0	2866.0	9.0
7.0	3.0	19.0	3708.0	31.0
7.0	3.0	18.0	428.0	1.0
7.0	3.0	17.0	1218.0	2.0

FIG. 9 – Exposition à différents niveaux de la hiérarchie temporelle.

P64	P16	Frequency	Exposure
null	null	532545	7448413.0
129017	null	36	270.0
129017	516071	28	220.0
129017	516068	8	50.0
129016	null	783	4054.0
129016	516067	27	108.0
129016	516066	52	230.0
129016	516065	45	293.0
129016	516064	659	3423.0
127807	null	542	2940.0

FIG. 10 – Exposition à des niveaux grossiers de la hiérarchie spatiale.

```
SELECT FLOOR(M.location_id/64) AS P64, FLOOR(M.location_id/16) AS P16,
    count(*) AS Frequency, avg(M.measurement_value) * count(*) AS Exposure
FROM measurement M, measurement_type MT
WHERE MT.measurement_type_id = M.measurement_type_id AND MT.
    measurement_name='pm2.5'
GROUP BY FLOOR(M.location_id/64), FLOOR(M.location_id/16) WITH ROLLUP
```

La requête de l'exemple 4.2 crée un sous-total du niveau groussier de la hiérarchie (P64) en descendant à un niveau moins grossier (P16). Cela est équivalent au calcul des agrégats pour les ensembles de regroupement suivants : (P64, P16), (P64) et () (i.e., tous). La requête renvoie l'exposition aux PM2,5 à ces deux niveaux grossiers de la hiérarchie (i.e., pixels de 16 et 64 cellules). Un extrait de la sortie de la requête est affiché dans la Fig 10 qui montre la fréquence de visite de chaque niveau grossier de la hiérarchie ainsi que l'exposition au PM2,5 sur ces niveaux. De plus, afin d'obtenir les lieux les plus visités, l'ajout d'une clause `ORDER BY Frequency` à la fin est suffisante.

4.4 Analyse Temporelle

L'analyse temporelle consiste à analyser les mesures de pollution au fil du temps. Il permet d'avoir un aperçu de phénomène mesuré pendant des périodes spécifiques tout en se déplaçant dans la hiérarchie temporelle. Cette dernière peut être définie de plusieurs manières. La Fig 8 illustre un exemple de cette présentation.

Comme pour l'analyse longitudinale au fil du temps, il est possible d'obtenir les mesures de pollution tout au long de la hiérarchie temporelle sans tenir compte de la dimension individuelle, ce qui est déjà illustré dans l'exemple 4.1 en supprimant la dimension individuelle. Ainsi, nous pouvons répondre à des requêtes du genre : *Quel est le niveau de pollution le week-end ?*. La Fig 9 illustre un extrait de la sortie de cette requête. Les week-ends sont représentés par les nombres 6 (Samedi) et 7 (Dimanche).

5 Conclusion and Perspectives

Cet article aborde l'exploration et l'analyse de séries de données géoréférencées collectées dans le contexte du captage participatif des données (MCS). Plusieurs travaux ont tenté de traiter la nature complexe de ces données, mais cet article tente de combler l'écart entre les données brutes et les informations utilisables, en fournissant une vue multidimensionnelle des

données. Après avoir analysé les exigences de la modélisation de données multidimensionnelles dans le contexte du captage participatif, cet article présente un tel modèle de données multidimensionnel conçu pour traiter et interroger les différents aspects des trajectoires individuelles ainsi que les mesures de pollution sous-jacentes. L'implémentation du modèle a été basée sur l'outil NoSQL Hive bien connu de l'éco-système Hadoop pour l'analyse des données afin de prendre en compte tous les aspects des données. Le modèle de données de base et la méthodologie considérée sont appliqués aux données de mobilité urbaine et de pollution, mais est suffisamment générique pour servir de modèle de référence pour d'autres applications.

Nous prévoyons utiliser le modèle OLAP, d'une part, pour détecter les anomalies en explorant les données et en les corrigeant par l'application par exemple des méthodes de lissage statistique telles que la moyenne mobile et la moyenne mobile exponentielle. D'autre part, le modèle OLAP peut être utilisé pour l'enrichissement des données tel que la dérivation d'arrêts (i.e., les lieux de séjour) en fonction de la densité de points par cellule, ou en fonction de la rareté des lectures GPS au fil du temps. De plus, nous prévoyons de travailler sur des implémentations de désagrégation spatiale et temporelle afin d'unifier les données spatio-temporelles sur le même niveau de granularité en utilisant des données auxiliaires. Enfin, l'étude expérimentale sera étendue à une analyse approfondie des performances en utilisant des données volumineuses. En utilisation continue, le volume augmente continuellement, ce qui peut être ingérable. Comme le volume de données dans les projets expérimentaux (y compris Polluscope) n'atteint pas encore le volume de données attendu, un générateur est nécessaire pour le contexte particulier du captage participatif.

Remerciements

Ce travail a été financé par le projet Polluscope ANR-15-CE22-0018. Les auteurs tiennent à remercier tous les membres du projet ayant participé de près ou de loin.

Références

Airparif (Last accessed February 2021). `https://www.airparif.asso.fr/`.

Ambiciti (Last accessed February 2021). `http://ambiciti.io/`.

Bimonte, S., A. Tchounikine, et M. Miquel (2007). Spatial olap : Open issues and a web based prototype.

Jensen, S. K., T. B. Pedersen, et C. Thomsen (2017). Time series management systems : A survey. *IEEE TKDE 29*(11), 2581–2600.

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

Lo, E., B. Kao, W.-S. Ho, S. D. Lee, C. K. Chui, et D. W. Cheung (2008). Olap on sequence data. In *Proceedings of SIGMOD '08*, pp. 649–660.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *AAG 94*(2), 284–289.

Moon, B., H. V. Jagadish, C. Faloutsos, et J. H. Saltz (2001). Analysis of the clustering properties of the hilbert space-filling curve. *IEEE TKDE 13*(1), 124–141.

OGC (Last accessed February 2021). http://www.ogc.org/.

OpenRadiation (Last accessed February 2021). https://openradiation.org/.

Rivest, S., Y. Bédard, M.-J. Proulx, M. Nadeau, F. Hubert, et J. Pastor (2005). Solap technology : Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *ISPRS (PRS) 60*(1), 17–33.

Savary, L., T. Wan, et K. Zeitouni (2004). Spatio-temporal data warehouse design for human activity pattern analysis. In *Proceedings. DEXA, 2004.*, pp. 814–818. IEEE.

Wan, T. et K. Zeitouni (2006). Représentation et indexation d'objets mobiles dans un entrepôt de données. In D. Grigori, S. Lopes, B. Nguyen, et K. Zeitouni (Eds.), *Actes de la 2ème journée francophone EDA*, Volume B-2 of *RNTI*, pp. 139–154. Cépaduès.

Summary

Air quality is one of the major risk factors in human health. Its continued degradation in the recent decades, particularly in urban and industrialized environments, is becoming a central concern for our societies. Emerging connected micro-sensors offer the opportunity to measure each person exposure to air pollution anywhere and anytime, while contributing to the observation network. Mobile Crowd Sensing (MCS) is a trending concept based on this technology. To date, MCS data are under-utilized due to the gap between raw data and usable information. The objective of this paper is to investigate an OLAP approach in filling this gap. In this respect, this paper introduces a methodology and a multidimensional data model designed for processing and querying the different aspects of individual trajectories together with underlying pollution data at different granularity levels. The core data model is generic enough to act as a reference model for further analysis.