Données hétérogènes de mobilités quotidiennes : protocole de diagnostic qualité et d'apurement à partir de la base MOBI'KIDS

Sylvestre Duroudier, Sonia Chardonnel, Boris Mericskay, Isabelle I. André-Poyaud, Olivier Bedel, Sandrine Depeau, Thomas Devogele, Laurent Etienne, Arnaud Lepetit, Clément Moreau, et al.

Données hétérogènes de mobilités quotidiennes : protocole de diagnostic qualité et d'apurement à partir de la base MOBI'KIDS

Duroudier S.[1], Chardonnel S.[1], Mericskay B.[2], Andre-Poyaud I.[1], Bedel O.[3], Depeau S.[2], Devogele T.[4], Etienne L.[4], Lepetit A.[2], Moreau C.[4], Pelletier N.[2], Ployon E.[1], Tabaka K.[1]

1. Univ. Grenoble Alpes, CNRS, Science Po Grenoble, PACTE 38000 Grenoble, France, sylvestre.duroudier@univ-grenoble-alpes.fr
2. UMR ESO, CNRS/Université Rennes 2, Place du recteur Henri Le Moal, 35043 Rennes Cedex, France
3. Alkante, 4 rue Alain Colas, 35530 Noyal sur Vilaine
4. Laboratoire d'informatique LIFAT, 3 place Jean Jaurès, 41000 Blois

RESUME. Cet article a pour objectif de présenter la méthodologie de diagnostic qualité et d'apurement des données expérimentée à partir d'une enquête de mobilité individuelle (programme MOBI'KIDS). Une première section revient sur la démarche suivie et pointe l'enjeu de l'évaluation de la qualité de données hétérogènes issues d'une méthode mixte et longitudinale de collecte (suivis GPS, enquêtes, observations). Une deuxième section établit un diagnostic qualité selon l'origine (GPS, algorithme, enquête) et la nature des erreurs (complétude, précision, cohérence). Ces typologies permettent, dans une troisième section, la définition d'une chaîne de traitements reproductible visant à améliorer la qualité interne et externe des données.

ABSTRACT. This paper aims at proposing a data quality diagnosis and clearance methodology experimented on an individual mobility survey (MOBI'KIDS program). The first section presents the theoretical approach to highlight the issue of a data quality diagnosis applied on heterogeneous data collected from mixed methods (GPS tracks, surveys, observations). Secondly, two typologies of major errors are discussed according to their origin (GPS, algorithm, survey) and their nature (completeness, accuracy, consistency). A processing chain is thirdly defined to improve both internal and external data quality in order to the perspective of a replicable methodology.

Mots-clés : mobilité quotidienne, traces GPS enrichies, méthode mixte, qualité des données

KEYWORDS: Daily mobility, semantic GPS data, mixed method, data quality

1. Introduction

Avec le développement de dispositifs mobiles capables de collecter des données de haute résolution spatiale et temporelle sur de longues périodes, l'analyse des données individuelles de mobilité quotidienne a largement évolué (Stopher, 2009 ; Shoval et al., 2014, Drevon et al., 2014). L'un des principaux enjeux méthodologiques actuels consiste à hybrider les méthodes traditionnelles des « small data » apportant des informations sémantiques riches, avec des méthodes relevant des « big data » permettant des collectes plus massives et plus longues (Chen et al., 2016).

C'est dans cette perspective que le programme MOBI'KIDS[1] (MK) analyse la mobilité des enfants et de leurs parents afin de comprendre dans quelles conditions l'autonomie de déplacement se met en place. Une enquête réalisée auprès d'une cohorte d'enfants et de parents permet de reconstituer les trajectoires quotidiennes décrivant spatialement, temporellement et sémantiquement la succession de leurs activités et leurs déplacements. L'approche mixte de cette enquête combine différents dispositifs de collecte (suivi GPS, enquêtes mobilité, entretiens semi-directifs), qui permettent la constitution d'un corpus de données originales et riches, mais dont l'hybridation pose un ensemble de défis et d'enjeux tant sur le plan technique que méthodologique (Lenormand *et al*., 2014 ; Lord *et al*., 2018).

La plus-value de ces méthodes mixtes est que le corpus est enrichi progressivement par différentes techniques de collecte. Cette diversité des méthodes mobilisées implique un travail spécifique de mise en cohérence des données soulevant avec d'autant plus d'acuité la question de leur qualité. Aussi, l'évaluation de la qualité interne et externe des données collectées apparaît indispensable et nécessite de formaliser une approche reproductible et transposable au-delà du programme MK. Cette étape, préalable à l'analyse même des données, est souvent peu ou non évoquée dans la littérature, alors qu'elle est un gage scientifique d'autant plus crucial que les corpus sont complexes et multi-sources. S'insérant dans le programme MK en cours, cet article a pour objectif de discuter cette étape de la recherche en présentant la méthodologie de diagnostic et les préconisations d'apurement des données expérimentée sur le corpus MK.

Une première section revient sur la démarche suivie et pointe l'enjeu de l'évaluation de la qualité de données hétérogènes issues d'une méthode mixte et longitudinale de collecte. Une deuxième section établit un diagnostic qualité selon l'origine (GPS, algorithme, enquête) et la nature des erreurs (complétude, précision, cohérence). Ces typologies permettent, dans une troisième section, la définition d'une chaîne de traitements reproductible visant à améliorer la qualité interne et externe des données.

[1] Les travaux présentés dans cet article sont financés par le projet MOBI'KIDS « Le rôle des cultures éducatives urbaines (CEU) dans l'évolution des mobilités quotidiennes et des contextes de vie des enfants. Collecte et analyse de traces géolocalisées et enrichies sémantiquement » (ANR-16-CE22-0009).

2. De la démarche de MOBI'KIDS à l'enjeu de l'apurement

2.1. MOBI'KIDS : position de recherche

L'analyse de la mobilité et plus spécifiquement de l'autonomie des enfants est aujourd'hui envisagée selon différentes approches. D'un côté, des recherches qualitatives interrogent les enfants sur leurs expériences subjectives et sensibles de mobilité ; de l'autre, des recherches quantitatives (questionnaire, accéléromètre) mesurent l'intensité de l'activité physique des enfants. Enfin, des approches géomatiques analysent les facteurs géographiques et matériels des environnements dans lesquels évoluent les enfants (Depeau et Quesseveur, 2014 ; Kyttä *et al*., 2018). Encore peu de travaux envisagent les dimensions de l'apprentissage de la mobilité chez les enfants par des approches hybrides (Christensen et al., 2011 ; Depeau *et al*., 2017). L'ambition de MK est de caractériser les pratiques quotidiennes des enfants dans leurs contextes socio-spatiaux. L'objectif est d'analyser l'évolution des pratiques en suivant les mêmes enfants à deux périodes, à la fin de l'école primaire et au début du collège, selon une approche à la fois descriptive des pratiques (avec une fine résolution spatiale et temporelle) et compréhensive des expériences et des représentations de l'enfant.

2.2 Les données attendues

Répondre à cet objectif scientifique requiert un corpus de positions spatio-temporelles relatives aux déplacements et aux activités des enquêtés modélisées sous la forme d'une suite ordonnée de trajets et de lieux enrichis sémantiquement. D'une part, les lieux correspondent aux arrêts prolongés, formalisés par des points, et caractérisés par des heures de début et de fin ainsi que par les activités qui y sont réalisées (domicile, école, travail, course, loisirs…). D'autre part, les trajets correspondent aux déplacements entre deux lieux, modélisés par des lignes, et caractérisés par leurs bornes temporelles, le mode de transport utilisé et les modalités d'accompagnement.

Dans cette perspective, une enquête a été menée auprès d'une cohorte de 89 familles (soit 182 personnes à raison d'un enfant et un parent par famille[2]). Le recrutement s'est fait par l'intermédiaire des écoles primaires (classes de CM1-CM2), dans le but de suivre tous les enfants d'une même classe ayant potentiellement des interactions sociales y compris en dehors de l'école. Situés dans la métropole de Rennes, les terrains d'étude se composent de 3 écoles du centre de Rennes et 2 écoles de la commune périurbaine d'Orgères. Si la cohorte a déjà été enquêtée pendant les deux périodes (primaire et collège), les analyses présentées dans cet article portent uniquement sur les traces d'une partie de la cohorte (150 individus) lors de la première période.

[2] Exceptionnellement, dans certaines familles, deux enfants ont été suivis.

2.3 Un protocole hybride et intégré de collecte de données hétérogènes

Dans la pratique, le protocole de collecte consiste pour chaque personne à porter un enregistreur GPS pendant cinq jours consécutifs. En complément, plusieurs questionnaires en présentiel sont menés auprès de l'ensemble des individus de la cohorte. Le premier vise à renseigner le profil général des familles enquêtées. Le second, passé immédiatement après le suivi GPS, consiste en un examen exhaustif du séquençage des traces collectées des personnes enquêtées *via* une tablette (Depeau *et al*., 2019) afin de les renseigner sémantiquement (mode de déplacement et d'accompagnement, activités…). Enfin, les individus enquêtés sont interrogés sur leurs perceptions des mobilités enfantines et divers éléments attenants aux attitudes éducatives.

Le phasage de ce protocole (Fig.1) oblige à séquencer *a priori* les traces brutes issues de la collecte GPS afin de mener rapidement un entretien en face à face avec les enquêtés. Le but est d'enrichir les données sur la base d'une trace prétraitée, c'est-à-dire structurée sous la forme d'une séquence ordonnée de positions spatiales correspondant à des situations d'immobilité (des arrêts dans des lieux) et à des situations de mobilité (des trajets). Cela impose de choisir à l'amont de l'enrichissement des paramètres de séquençage déterminant les valeurs qui définissent l'emprise spatio-temporelle d'un arrêt et d'un trajet. Répondant donc à des hypothèses théoriques et à des arbitrages en matière d'agrégation, l'algorithme de séquence s'appuie sur les règles suivantes : un arrêt est un ensemble de positions regroupées dans un rayon de moins de 50m de distance sur un temps d'au moins 300s ou à moins de 300s d'une mise en veille du *datalogger*[3] ; un trajet correspond à un ensemble de positions consécutives entre 2 arrêts.

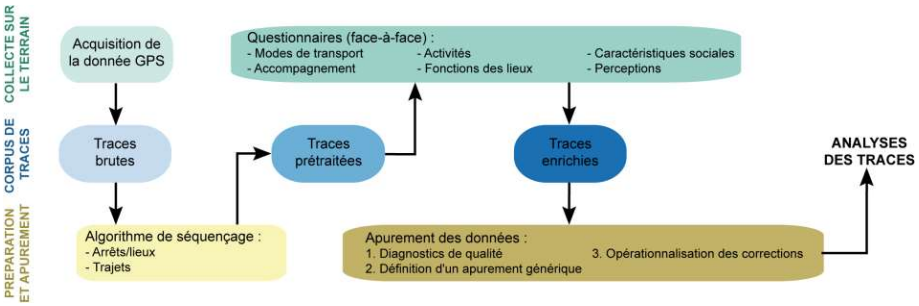

FIGURE 1. Saisir les traces de la mobilité quotidienne dans MOBI'KIDS : entre collecte sur le terrain, préparation et apurement des données.

[3] Le *datalogger* utilisé dans le projet enregistre la position géographique des individus toutes les secondes. Impliquant de très gros volumes de données (potentiellement jusqu'à 600 000 informations par semaine par capteur), cette fréquence d'enregistrement est affinée par un accéléromètre qui limite les données collectées lors des longues périodes d'inactivité.

Appliqué au corpus traité dans cet article, ce séquençage permet d'identifier à partir des 7,6 millions de positions brutes enregistrées pour les 150 individus enquêtés, un ensemble de traces prétraitées constituées de 3794 trajets et 2381 arrêts.

Par ailleurs, conscients des limites induites par un séquençage paramétré *a priori* et de manière homogène pour l'ensemble des contextes d'enregistrement, les enquêteurs pouvaient annoter les écarts (notamment les imprécisions spatiales, temporelles) relevés par les personnes enquêtées entre les arrêts et trajets détectés et ceux effectivement réalisés.

2.4. Enjeux de la qualité des données de l'enquête

A ce stade du programme, bien avant l'analyse du corpus, le principal enjeu relève du problème bien connu de l'évaluation de la qualité des données (Devillers et Jeansoulin, 2005 ; Guptill et Morisson, 2013). Celui-ci se pose de manière singulière dans MK en raison du caractère inédit des données recueillies et des objectifs analytiques du projet de recherche. Il est donc nécessaire de définir un protocole général de traitement des données qui permette d'en diagnostiquer la qualité tout en préparant la base en vue d'analyses ultérieures.

Suivant plusieurs définitions (Aalders, 2002 ; Devillers, 2004), on distingue ici la qualité interne et la qualité externe des données. Concernant la qualité interne, il s'agit d'identifier les différences entre les données attendues et réelles, c'est-à-dire de quantifier et de qualifier les erreurs sémantiques, géographiques et temporelles selon leur origine (modes de collecte et compilation ; Servigne *et al*., 2005) et leur nature (complétude et exhaustivité, cohérence, précision et indexation ; Guptill et Morrison, 1995 ; David et Fasquel, 1997). Sur ce plan, le programme MK réalise un grand écart entre d'une part les méthodes de nettoyage et de séquençage de données GPS (Biljecki *et al*. 2013 ; Lin et Hsu, 2014 ; Dalumpines *et al*., 2017), et d'autre part les procédures d'apurement de données d'enquêtes classiques (statistiques publiques, enquêtes ménages-déplacements ; Certu, 2008 ; INED, 2019). De manière concomitante, la qualité externe caractérise l'adéquation entre les données et les besoins de ses utilisateurs. Il s'agit alors d'opérer des actions (agrégation, suppression, recodage, étiquetage, etc.) qui diffèrent selon les objectifs analytiques (autonomie des enfants, co-présence…). Finalement, ces évaluations des qualités interne et externe ouvrent des pistes de définition d'une méthode générique d'apurement de la base de données qui tienne également compte des spécificités des analyses envisagées.

3. Typologie et diagnostic de la qualité interne des données de l'enquête

3.1. Typologie des erreurs selon l'étape du protocole

3.1.1. Erreurs liées à la collecte de données GPS

Une première cause d'erreurs relève des dispositifs d'enregistrement des positions géographiques collectés au cours de l'enquête. On distingue d'abord les contraintes techniques des *dataloggers* (Neatt et al., 2016) : le temps de chauffe

(TTFF), le déchargement de la batterie, la mise en veille ou la défaillance de l'accéléromètre. Ainsi, le TTFF peut conduire à la perte d'informations entre le début d'une mobilité (où le *datalogger* est en veille due à une immobilité prolongée) et l'enregistrement des positions une fois le signal GPS actif. Un autre ensemble d'erreurs relève des limites techniques du dispositif selon les environnements pratiqués par les enquêtés (Fig.2) : perte de signal à l'intérieur des bâtiments ou dans le métro, nombre insuffisant de satellites captés, réverbération des signaux lors « d'effets de canyon urbain ». Ce type de cas génère des positions GPS aberrantes qui se répercutent sur les trajets : traces décalées ou hors du réseau routier, vibrations et sinuosités, formes rectilinéaires, etc.

FIGURE *2. Effet de la perte momentanée du signal satellite (à gauche), effet de canyon urbain et décalage de la trace (à droite).*

Une analyse simple des positions GPS collectées, reposant sur le nombre de satellites captés ou l'indicateur DOP (*dilution of precision* ; Kim et Park, 2017), permet une première estimation de la qualité des données. Ainsi, 0,86% des positions réparties sur 25% des trajets ont capté strictement moins de 4 satellites.

Une seconde analyse s'inspirant des règles de filtrage de Neatt et al. (2016) a porté sur les écarts de distance, de temps, et les vitesses moyennes associées, pour chaque couple de positions successives. Ces mesures permettent d'identifier 18901 couples de positions qui ont un écart temporel supérieur à la période d'échantillonnage (1s). De plus, 394 couples se distinguent par une vitesse supérieure à 180km/h, tandis que 3424 cas correspondent à une distance spatiale supérieure à 50m. Selon ces critères, près de 19 279 couples de positions distinctes présentent une aberration, soit 0,25% du total, et peuvent faire l'objet d'un filtrage.

3.1.2. Forces et limites de l'algorithme de séquençage des traces brutes

Deuxièmement, et indépendamment des erreurs liées aux dispositifs d'enregistrement, l'algorithme de séquençage est susceptible de générer d'autres erreurs. En effet, cet algorithme se base sur un ensemble de règles spatiales et temporelles (cf. 2.3) parfois inadaptées devant l'hétérogénéité des conditions d'enquêtes : par exemple entre un milieu urbain dense et un milieu ouvert, ou

lorsque l'amplitude spatiale des déplacements est réduite (domicile proche de l'école ou au-dessus d'un centre commercial). La multitude de situations contribue ainsi à générer des faux positifs ou négatifs (Fig.3) : par exemple des arrêts détectés sur des autoroutes en raison de bouchons ou encore des trajets au sein de lieux (domiciles, écoles) voire de « grands lieux » (centre commerciaux, parcs, cimetières).

FIGURE *3. Faux lieux sur l'autoroute (à gauche) et faux trajet à l'école (à droite).*

3.1.3. Les biais de l'enquête

Troisièmement, la phase d'enquête (Fig.1), comporte des biais susceptibles d'introduire de nouvelles erreurs (Beaud et Weber, 1998). Dans la pratique, les conditions de passation des enquêtes sont inégales, parfois contraintes ou peu adéquates (lieux publics bruyants, disponibilité et attention variables des enquêtés...). A cela s'ajoute l'hétérogénéité d'expérience des enquêteurs mobilisés, qui implique différentes sensibilités aux informations importantes à retenir et à saisir dans l'examen des traces. En outre, l'entretien peut être long (plus d'une heure) et altérer la précision des réponses de certaines personnes enquêtées. En résumé, ces erreurs sont de différentes natures : caractérisations incohérentes des activités ou de certains modes des trajets, erreurs de saisie dans les questions ouvertes, utilisation fréquente des notes correctives, etc.

3.2. Diagnostic de l'intégrité des données selon la nature des erreurs

3.2.1. Complétude spatio-temporelle et sémantique des données

Un premier type de problème correspond à l'incomplétude des données collectées. Dans le cadre du programme MK, le problème se pose à la fois dans la nécessaire continuité spatio-temporelle des informations et dans l'ajout de données sémantiques par le processus d'enquête.

D'une part, concernant les données issues de la collecte GPS, il est essentiel de s'assurer de leur continuité spatio-temporelle dans la perspective d'analyser les agendas individuels. Pour cela, plusieurs vérifications sont mises en œuvre sur l'ensemble de la cohorte (Fig.4) : absence de données enregistrées, alternance stricte

arrêt-trajet, adresses identiques d'arrivée et de départ de trajets successifs, continuité des identifiants dans les tables.

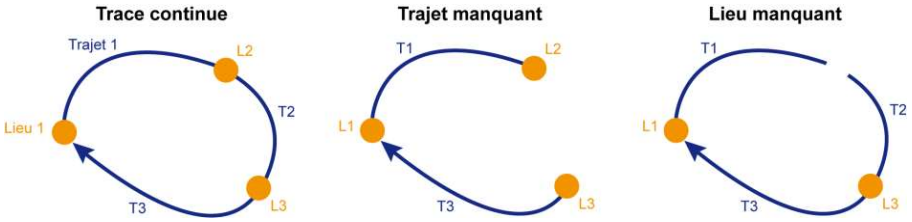

FIGURE *4. Exemples d'erreurs de non continuité des traces.*

Table	Critères	Conditions	Nombre	En %
Trajets	Mode de transport et accompagnement	non renseigné	467	12,31
	Lieu de départ	non renseigné	291	7,67
	Lieu d'arrivée	non renseigné	304	8,01
Lieux	Type de lieu	non renseigné	288	12,08
	Fréquence d'usage	non renseigné	390	16,36

TABLE 1. Exemples d'incomplétudes sémantiques

D'autre part, l'incomplétude des données est liée à l'absence de renseignements sémantiques des trajets et des lieux ayant fait l'objet d'un enrichissement par l'enquête (Tab.1). On dénombre par exemple 12,3% des trajets non renseignés selon les modes de transport et d'accompagnement, quand 16,4% des lieux ne sont pas caractérisés par une fréquence d'usage.

3.2.2. Précision géographique et temporelle

Un second type de problème relève de la précision des traces. Sur le plan temporel, les données collectées sont de bonne facture puisque chaque position est caractérisée à la seconde près. Cependant, chaque arrêt et trajet est défini seulement par ses bornes de début et de fin, ce qui implique une perte des temporalités détaillées au sein de cet intervalle.

Par contre, l'imprécision géographique est plus importante au sein des données séquencées puisque la localisation des arrêts ou le tracé des trajets dépendent des positions GPS et de l'algorithme de séquençage. Ceux-ci peuvent donc être imprécis et situés à distance des lieux ou des infrastructures (Fig.5) : la distance entre des traces ayant utilisé la même route peut dépasser plusieurs centaines de mètres. De même, la distance des points au barycentre d'une école varie entre 84m et 450m.

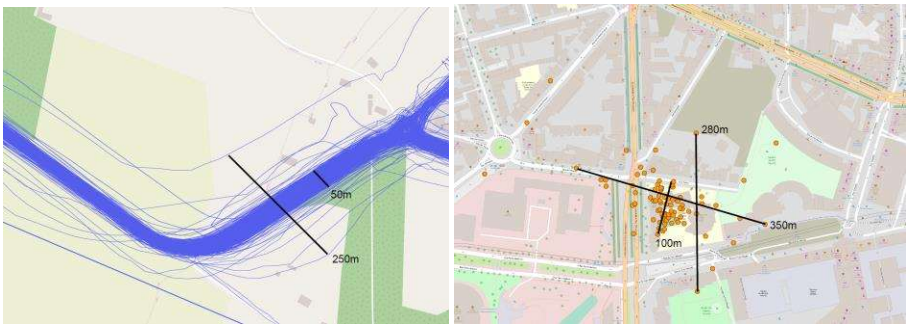

FIGURE 5. Imprécision des trajets à l'entrée d'Orgères (à gauche) et des lieux associés à l'une des écoles (à droite).

3.2.3. Cohérence logique et sémantique

Un troisième type d'erreurs correspond aux incohérences logiques et sémantiques dans – et entre – les tables enrichies. Mises en exergue notamment par des requêtes sur les variables qui caractérisent les trajets et les arrêts, ces incohérences sont diverses (Tab.2). Une partie correspond aux longueurs, durées et vitesses aberrantes : par exemple près de 6% des trajets présentent des durées inférieures à 30s ou supérieures à 3h. De même, d'autres erreurs concernent les activités, les modes de transport et les modes d'accompagnement. Ces incohérences peuvent également être révélées par la combinaison de plusieurs variables : on observe par exemple des « enfants conducteurs d'une voiture », des « enfants à l'école le dimanche », ou encore des « piétons se déplaçant à plus de 20km/h ».

TABLE 2. Exemples d'incohérences logiques et sémantiques

Table	Critères	Conditions	Nombre	En %
Trajets	Durée	< 30sc et > 3h	225	5,93
	Vitesse	< 1km/h et > 100km/h	151	3,98
	Statut familial et mode de transport	Enfant et conducteur	36	0,95
	Mode de transport et vitesse	Marche et > 20km/h	63	1,66
Lieux	Durée	< 300sc	247	10,37
	Date de fin	Supérieur à la date de rendu des GPS	74	3,11
	Type de lieu et durée	Non-domicile et > 12h	162	6,80

4. Mise en application de l'apurement et des corrections des données

Cette classification associée à une quantification des erreurs constitue un préalable à la définition d'une procédure d'apurement visant à améliorer la qualité générale des données. Dans la mesure où l'enquête MK se déroule en deux temps, il est nécessaire que le protocole d'apurement soit à la fois reproductible et qu'il tienne également compte de la diversité des besoins d'analyse. Cette partie présente en ce

sens les principales lignes directrices et quelques résultats préliminaires du travail en cours de définition et de mise en œuvre de ce protocole d'apurement.

La clé de voute du protocole consiste en un apurement générique, appliqué à l'ensemble de la base de données relationnelle et qui vise à améliorer la qualité interne. Cet apurement s'organise en plusieurs étapes :

- Étape 1 : Identification semi-automatique des erreurs de complétude, de cohérence et de précision grâce à des requêtes prédéfinies collectivement (cf. 3.2 ; Tab.1 et 2). Ces requêtes (menées en SQL, dans R et dans des SIG) sont appliquées à toutes les tables de la BDD, y compris celles ne comportant pas de géométrie spatiale.
- Étape 2 : Définition des actions correctives (Tab.3). Après un étiquetage des erreurs dans les tables, d'autres actions sont envisagées tel que la création dans la BDD d'un type d'objet « inconnu » en plus des trajets et des lieux, qui regroupe les temps invalides selon la complétude ou la cohérence. En outre, plusieurs actions géométriques visent à améliorer la précision spatiale des traces, notamment leur appariement au réseau routier à l'aide d'un SIG. Enfin, d'autres actions portent sur la sémantique : soit la modification si des notes correctives permettent de rectifier les informations, soit le recodage des modalités (notamment pour les réponses aux questions ouvertes), soit la désignation de l'individu comme invalide s'il n'a aucune sémantique.
- Étape 3 : Essais d'implémentation et chaînage des actions correctives. Il s'agit d'abord de tester la mise en œuvre technique et les effets des actions correctives envisagées : en les implémentant dans la BDD, puis en réitérant l'étape 1 de diagnostic. Par exemple, la suppression des positions GPS erronées selon le nombre de satellites permettra, sans re-segmenter, d'améliorer la précision et la cohérence des lieux et des trajets. Ces essais indépendants permettent ensuite de chaîner les corrections en tenant compte des relations entre les différentes tables de la BDD. Par exemple, les « temps inconnus », pour lesquels les données sont incomplètes ou insuffisamment fiables, sont d'abord capturés à partir de l'absence de traces brutes, puis complétées par les mini-trajets (moins de 10m), et enfin terminés par les trajets et les lieux ayant une sémantique lacunaire.
- Étape 4 : Opérationnalisation de la chaîne de corrections dans la BDD. Menée dans un environnement Postgresql, cela doit aboutir à la création d'une version apurée qui servira de base aux différentes analyses. Cependant, la mise en œuvre de ce protocole pourra être renouvelée après chaque nouvelle phase du suivi longitudinal (enfants scolarisés au collège).

A l'issue de ces 4 étapes, une phase de pré-analyse de ce corpus apuré consiste à valider la possibilité de créer des « indicateurs » pour les analyses thématiques envisagées. Ainsi, afin de mesurer l'évolution de l'autonomie de déplacements des enfants entre les différents groupes d'enfants et les phases d'enquêtes, il est nécessaire de créer des indicateurs synthétiques sur les modes de transport et d'accompagnement, sur les itinéraires empruntés ou les séquences d'activité.

TABLE 3. Actions correctives de l'apurement générique

	Spatial (sources : *datalogger* et séquençage)	**Temporel (sources : *datalogger* et séquençage)**	**Sémantique (source : enquête)**
Complétude	Exemple : Pas de géométrie associée à la trace du lieu ou du trajet. Action : étiquetage des individus, définition d'un temps invalide dans les séquences trajets-lieux.	Exemple : Discontinuité dans la succession des activités ou dans l'alternance des lieu-trajet Actions : balisage des "moments" concernés, définition d'un temps invalide dans les séquences trajets-lieux.	Exemple : Absence de renseignement dans les tables des trajets et des lieux. Actions : étiquetage des ind., recodage si possible à l'aide des notes des lieux ou trajets, suppression si l'enquêté n'a aucune information.
Cohérence	Exemple : Trajet avec une distance trop courte ; lieu inexistant ou non indexé à une adresse. Actions : étiquetage des ind., définition d'un temps invalide, correction des lieux par les coordonnées et/ou adresses.	Exemple : Durée trop courte ; Traces hors des périodes de collecte. Actions : étiquetage des ind., définition d'un temps invalide selon la durée, suppression des traces « hors délai ».	Exemple : Enfants conducteurs, enfants à l'école le dimanche, piétons à plus de 20km/h Actions : étiquetage des ind., recodage si possible à l'aide des notes correctives des lieux ou trajets.
Précision	Exemple : Localisations diffuses des trajets et des lieux. Actions : agrégation et appariement au réseau routier ou aux parcelles/adresses		Exemple : Diversité des noms déclarés des lieux Actions : étiquetage des ind., création de dictionnaires, recodage dans les tables

5. Conclusion : Vers un apurement adapté et reproductible

Le programme MOBI'KIDS repose sur un protocole méthodologique mixte de collecte de données de mobilités quotidiennes. Celles-ci sont à la fois hétérogènes et multidimensionnelles, spatio-temporelles et sémantiques : des traces numériques par suivi GPS, des renseignements sur les mobilités et lieux fréquentés à partir de questionnaires. Outre les erreurs qu'impliquent ces étapes de collecte, la mise en relation de ces données requiert, au préalable des analyses, la mise en œuvre d'un protocole de diagnostic qualité et d'apurement.

Dans cette perspective, cet article avait pour objectif de présenter la démarche exploratoire suivie dans MOBI'KIDS de diagnostic de la qualité des données collectées. La discussion permet d'établir un état des lieux des erreurs sémantiques,

géométriques et temporelles selon leurs origines (dispositif technique, biais d'enquête, algorithme de séquençage) ou leurs natures (imprécision, incomplétude, incohérence). Ce travail exploratoire de diagnostic ouvre la voie à la définition et la mise en œuvre d'un protocole d'apurement de la base de données relationnelle. Ce protocole est avant tout confronté à un enjeu d'objectivité et de reproductibilité. Il s'agit de construire une chaîne méthodologique qui pourra être déployée aux différents temps de l'enquête longitudinale MOBI'KIDS, puisque les enfants sont également suivis au collège. Mais plus largement, l'enjeu est de proposer une méthode reproductible à d'autres cas d'analyses de mobilités quotidiennes croisant de multiples et diverses informations.

Remerciements : Ces travaux sont financés par MOBIKIDS (ANR-16-CE22-0009)

Bibliographie

Aalders H. J. G. L. (2002), The Registration of Quality in a GIS, *Spatial Data Quality* (W. Shi, P. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, p. 186-199.

Beaud S., Weber, F. (1998). *Guide de l'enquête de terrain : produire et analyser des données ethnographiques*, La Découverte, Paris.

Biljecki F., Ledoux H., Oosterom P. van (2013). Transportation mode-based séquençage and classification of movement trajectories. *International Journal of Geographical Information Science* 27, 385–407.

Chen C., M J. ,Susilo Y.,Liu, Y.,Wang M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies, 68, 285–299.

Christensen P., Mikkelsen M.R., Nielsen T.A.S., Harder H. (2011). Children, Mobility, and Space: Using GPS and Mobile Phone Technologies in Ethnographic Research. *Journal of Mixed Methods Research* 5, 227–246.

CERTU (2008). *L'enquête ménages déplacements « standard Certu »*, Guide méthodologique.

Dalumpines R., Scott D.M. (2017). Making mode detection transferable: extracting activity and travel episodes from GPS data using the multinomial logit model and Python. *Transportation Planning and Technology* 40, 523–539.

David B., Fasquel P. (1997). Qualité d'une base de données géographique : concepts et terminologie. *Bulletin d'information de l'IGN*, n°67.

Depeau S., Bedel O., Cherel P., André-Poyaud I., Chardonnel S., Gombaud J., Jambon, F., Lepetit A., Mericskay B., Quesseveur E. (2019) MK-MOBIBACK : un dispositif hybride et intégré pour enquêter finement les mobilités quotidiennes des familles, Communication Acceptée, Conférence SAGEO 2019.

Depeau S., Chardonnel S., André-Poyaud I., Lepetit A., Jambon F., Quesseveur E., Gombaud, J., Allard T., Choquet C.-A. (2017). Routines and informal situations in children's daily lives. *Travel Behaviour and Society* 9, 70–80.

Depeau, S. & Quesseveur, E. (2014). A la recherche d'espaces invisibles de la mobilité : usages, apports et limites des techniques GPS dans l'étude des déplacements urbains à l'échelle pédestre. *Netcom*, vol 28, n°1-2, p. 35-54.

Devillers R. (2004). *Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales*. Thèse de doctorat, Université de Marne la Vallée.

Devillers R., Jeansoulin R. (2005). *Fundamentals of Spatial Data Quality*, ISTE, London, Newport Beach.

Drevon, G.,Jambon, F.,Chardonnel, S.,Christophe, S.,André-Poyaud, I.,Davoine, PA , Lutoff, Céline (2014). Évaluation comparée de l'apport de l'assistance GPS aux enquêtes de mobilité. *Netcom*, (28–1/2), 13–34.

Guptill S.C., Morrison J.L. (1995). *Elements of Spatial Data Quality*, Elsevier, Oxford.

Kim N. H., Park, C. H. (2017). Simulation Analysis of GPS/GLONASS Absolute Positioning Performance in an Urban Canyon Environment. *International Journal of Computer Theory and Engineering*, 9(1).

Kyttä M., Oliver M., Ikeda E., Ahmadi E., Omiya I., Laatikainen T. (2018). Children as urbanites: mapping the affordances and behavior settings of urban environments for Finnish and Japanese children. *Children's Geographies* 16, 319–332.

Lenormand M., Picornell M., Cantú-Ros O.G., Tugores A., Louail T., Herranz R., Barthelemy M., Frías-Martínez E., Ramasco J.J. (2014). Cross-Checking Different Sources of Mobility Information. *PLoS ONE 9*, e105184.

Lin M., Hsu W.-J. (2014). Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing*, 12, 1–16.

Lord S., Després M., Kestens Y. (2018). Approcher la complexité de la mobilité et des territoires dans la vieillesse: l'intégration de données et de méthodes mixtes dans l'analyse des relations personne-environnement. *Oser les défis des méthodes mixtes en sciences sociales et sciences de la santé*, Montréal, ACFAS p 194-207.

Neatt K., Millward H., Spinney J. (2016, April). Aggregation and spatial analysis of walking activity in an urban area: results from the Halifax space-time activity survey. In *IOP Conference Series: Earth and Environmental Scienc*e (Vol. 34, No. 1, p. 012022). IOP Publishing.

Servigne S., Lesage N., Libourel Th. (2005). Quality Components, Standards, and Metadata. *Fundamentals of Spatial Data Quality*, ISTE, London, Newport Beach, p. 179-210.

Shoval N., Kwan M.-P., Reinau K.H., Harder H. (2014). The shoemaker's son always goes barefoot: Implementations of GPS and other tracking technologies for geographic research. *Geoforum* 51, 1–5.

Stopher P. (2009). Collecting and Processing Data from Mobile Technologies, in: Bonnel, P. (Ed.), *Transport Survey Methods: Keeping up with a Changing World*, United Kingdom, Emerald, p. 361-391.

Site de l'INED : Saisie, codage, apurement, documentation, https://www.ined.fr/fr/ressources-methodes/methodologie-enquete/les-choix-methodologiques/saisie-codage-apurement-documentation/ ; consulté en avril 2019.