
Diagnostic qualité et apurement des données de mobilité quotidienne issues de l'enquête mixte et longitudinale Mobi'Kids

S. Duroudier¹, S. Chardonnel¹, B. Mericskay², I. Andre-Poyaud¹,
O. Bedel³, S. Depeau², T. Devogele⁴, L. Etienne⁴, A. Lepetit²,
C. Moreau⁴, N. Pelletier², E. Ployon¹, K. Tabaka¹

1. Université Grenoble Alpes, CNRS, Science Po Grenoble, PACTE, 14 av.
Marie Reynoard, 38100 Grenoble, France

sylvestre.duroudier ; sonia.chardonnel ; iandre ; estelle.ployon ;
kamila.tabaka@univ-grenoble-alpes.fr

2. UMR ESO, CNRS/Université Rennes 2, Place du recteur Henri Le Moal,
35043 Rennes Cedex, France

boris.mericskay ; sandrine.depeau ; arnaud.lepetit ;
nicolas.pelletier@univ-rennes2.fr

3. Alkante, 4 rue Alain Colas, 35530 Noyal sur Vilaine
olivier.bedel@alkante.fr

4. Laboratoire d'informatique LIFAT, 3 place Jean Jaurès, 41000 Blois
thomas.devogele@univ-tours.fr ; laurent.etienne@isen-ouest.yncrea.fr ;
clement.moreau@univ-tours.fr

RÉSUMÉ. 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 partie 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 partie établit un diagnostic qualité selon l'origine (GPS, algorithme, enquête) et la nature des erreurs (complétude, précision, cohérence). Ces typologies permettent, dans une troisième partie, de définir 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 cleansing data 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 to 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, données hétérogènes, méthodes mixtes, qualité des données, pré-traitements et nettoyage des données.

KEYWORDS: daily mobility, semantic GPS data, heterogenous data, mixed methods, data quality, pre-processing and data cleansing.

DOI: [10.3166/rig.2020.00105](https://doi.org/10.3166/rig.2020.00105) © 2020 Lavoisier

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 (Drevon *et al.*, 2014 ; Shoval *et al.*, 2014 ; Stopher, 2009). 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 » qui permettent des collectes plus massives (Chen *et al.*, 2016).

C'est dans cette perspective que le programme Mobi'Kids¹ (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 déplacements. Cette enquête suit une approche mixte en combinant différents dispositifs de collecte : suivi GPS, enquêtes mobilité, entretiens semi-directifs. Ces dispositifs 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 et conceptuel (Lenormand *et al.*, 2014 ; Lord *et al.*, 2018).

La plus-value de ces méthodes mixtes réside dans le fait que le corpus est enrichi progressivement par différentes techniques de collecte. Mais en corollaire, cette diversité des méthodes mobilisées implique un travail spécifique de mise en cohérence des données et soulève avec d'autant plus d'acuité la question de leur qualité prise au sens large. Préalable à l'analyse des données et requérant un temps qui peut être important, cette étape est souvent peu ou non évoquée dans la littérature, alors même

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).

qu'elle constitue un gage scientifique d'autant plus crucial que les corpus sont complexes et multi-sources. Aussi, l'évaluation et l'amélioration de la qualité interne et externe des données collectées sont indispensables et, au regard de l'hétérogénéité des données, elles nécessitent de formaliser une approche reproductible et transposable à d'autres corpus de données multi-sources de mobilité quotidienne.

Cet article a pour objectif de discuter cette étape de la recherche en présentant l'approche conceptuelle et méthodologique de diagnostic et les préconisations d'apurement² (c'est-à-dire de nettoyage et pré-traitement) du corpus de données hétérogènes collectées au sein du programme Mobi'Kids. Il correspond à une version revue et allongée de celui présenté au colloque SAGEO 2019 (Duroudier *et al.*, 2019).

Une première section (2) revient sur la démarche suivie dans le cadre du programme et les données collectées, amenant à pointer l'enjeu de l'évaluation de la qualité de données hétérogènes issues d'une méthode mixte et longitudinale de collecte. Une deuxième section (3) 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 (4), la définition d'une chaîne de traitements reproductible visant à améliorer la qualité interne et externe des données.

2. De la démarche de Mobi'Kids à l'enjeu de l'apurement

2.1. *Mobi'Kids : position de recherche*

Dans un contexte de transformation des modes de vie liés aux dynamiques de métropolisation, la mobilité quotidienne constitue un axe capital pour la compréhension du développement et de l'autonomie des enfants (Depeau, 2003). Les recherches ont notamment montré un recul des déplacements en autonomie comparativement aux années 1970 (Fyhri *et al.*, 2011 ; Hillman *et al.*, 1990 ; Kytä, 1997), soulignant l'impact de ce déclin sur la santé et le développement physique, social et psychologique des enfants (Shaw *et al.*, 2015). S'intéressant à cet enjeu, les études expérimentent une diversité de méthodes de recueil et d'analyse : des approches qualitatives permettent d'appréhender les expériences subjectives et sensibles de la mobilité (Christensen et O'Brien, 2003), et d'autres plus quantitatives par questionnaires ou par capteurs contribuent à mesurer l'intensité de l'activité physique des enfants (Schoeppe *et al.*, 2013). Enfin, des approches géomatiques analysent les facteurs géographiques et matériels des environnements dans lesquels évoluent les enfants (Depeau et Quesseveur, 2014 ; Kytä *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).

2. Issu du vocabulaire employé dans les enquêtes statistiques sur les populations, le terme apurement renvoie à la fois aux actions de contrôle, de tests de cohérence et de correction en vue d'améliorer la qualité des données de l'enquête.

Dans ce champ scientifique analysant l'autonomie des enfants, l'ambition de MK est de caractériser les pratiques quotidiennes (les routines, les expériences informelles) des enfants dans leur contexte social (la famille, les camarades de classe) et dans leurs contextes spatiaux (de l'échelle métropolitaine à l'échelle du quartier). L'objectif du programme est d'analyser l'évolution des pratiques de mobilité 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. Il s'agit également de chercher des formes de « cultures éducatives urbaines » pour comprendre leur rôle dans les formes d'apprentissage de l'autonomie de déplacement des enfants à l'échelle du quotidien des familles et des enfants.

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 (figure 1). 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, etc.). 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.

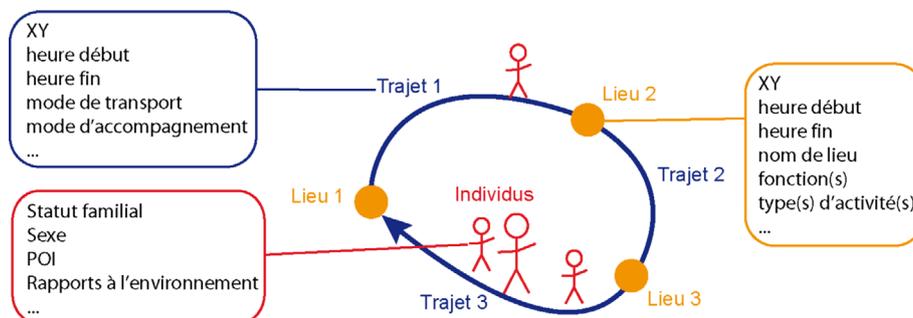


Figure 1. Les données attendues dans le corpus Mobi'Kids

Dans cette perspective, une enquête a été menée auprès d'une cohorte de 88 familles (soit 182 personnes à raison d'un enfant et un parent par famille³) au cours de deux périodes (enfant en école primaire et enfant au collège). Le recrutement s'est fait par

3. Exceptionnellement, dans certaines familles présentant des cas de jumeaux ou de parents séparés, deux enfants ou deux parents ont été suivis.

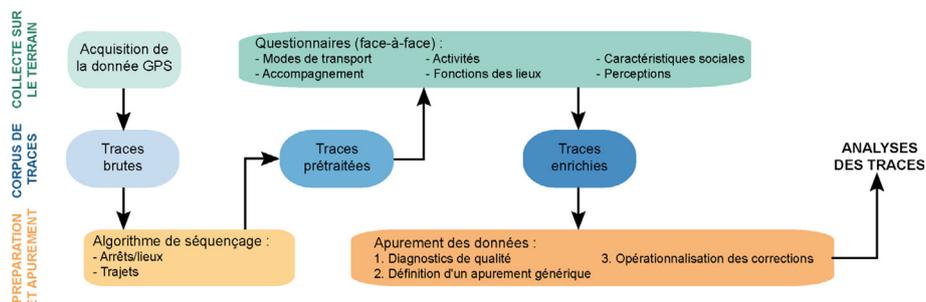


Figure 2. Saisir les traces de la mobilité quotidienne dans Mobi'Kids : entre collecte sur le terrain, préparation et apurement des données

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. Le terrain se compose de deux sites d'étude localisés dans la métropole de Rennes : le premier est dans le centre de Rennes (concernant 3 écoles et 5 collèges), le second correspond à une zone périurbaine couvrant Orgères et Saint-Erblon (2 écoles et 4 collèges).

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 afin d'enregistrer leurs traces. 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), ceci afin de les renseigner sémantiquement (mode de déplacement et d'accompagnement, activités, etc.). Enfin, les individus enquêtés sont interrogés sur leurs rapports aux mobilités enfantines et divers éléments relatifs aux attitudes éducatives.

Le phasage de ce protocole (figure 2) 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 en 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 séquençage s'appuie sur les règles suivantes : un « lieu » est un ensemble de positions regroupées dans un rayon de moins de 50 m de distance sur un temps d'au moins 300 sec ou à moins de 300 sec d'une mise en veille du

*datalogger*⁴ ; un « trajet » correspond à un ensemble de positions consécutives entre deux lieux.

Appliqué au corpus traité dans cet article, ce séquençage permet d'identifier, à partir des 28,3 millions de positions brutes enregistrées pour les 182 individus enquêtés, un ensemble de traces prétraitées constituées de 7 050 trajets et 4 450 lieux correspondant à des arrêts prolongés.

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 et 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

À 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 Morrison, 1995). Celui-ci se pose de manière singulière dans Mobi'Kids en raison du caractère inédit des données hétérogènes recueillies et des objectifs analytiques du projet de recherche. Il est donc nécessaire de définir un protocole général qui permette à la fois de diagnostiquer la qualité des données, et dans le même temps, de mettre en place des actions correctives en vue d'améliorer cette qualité pour les analyses ultérieures.

Suivant plusieurs définitions (Devillers, 2004 ; Shi *et al.*, 2002), 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 celles collectées (figure 1). Plus précisément, l'enjeu est 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 : David et Fasquel, 1997 ; Guptill et Morrison, 1995). 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 ;

4. Le *datalogger* a été développé par le partenaire industriel RF-TRACK sur la base d'un cahier des charges répondant en priorité à un port par des enfants (Depeau *et al.*, 2019). Le boîtier enregistre les mesures brutes issues des capteurs suivants, à des fréquences prédéfinies : accéléromètre (25 Hz), magnétomètre (5 Hz) et GPS (1 Hz). Les données sont encodées dans la mémoire interne sous la forme d'un fichier binaire, au format ad hoc. Le composant GPS a été amélioré dans la dernière version pour optimiser les performances GPS, notamment le temps de démarrage, mais il n'est pas basé sur la technique A-GPS. Le capteur enregistre un point GPS par seconde (jusqu'à 600 000 informations par semaine par capteur). Cette fréquence est affinée par un accéléromètre qui limite les données collectées lors des longues périodes d'inactivité de manière à optimiser l'autonomie et atteindre 5 jours d'utilisation sans recharge.

Dalumpines et Scott, 2017 ; Lin et Hsu, 2014), 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 (Deville et Jeansoulin, 2005). Il s'agit alors d'opérer des actions et des transformations des données (agrégation, suppression, recodage, étiquetage, etc.) qui diffèrent selon les objectifs analytiques (par exemple dans MK l'autonomie des enfants ou les espaces-temps de co-présence).

Le présent article traite plus spécifiquement du diagnostic et de l'amélioration de la qualité interne des données, mais la question de la qualité externe demeure sous-jacente puisqu'elle détermine grandement la formalisation des données. Et, plus largement, ces évaluations des qualités interne et externe ouvrent des pistes de définition d'une méthode générique d'apurement de la base de données qui tient 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 d'enquête

Une première manière d'interroger la qualité interne des données est d'envisager les erreurs selon leur origine dans le protocole de collecte des données. Dans le cadre de la base de données MK⁵, les erreurs (spatiales, temporelles et sémantiques), peuvent provenir de la collecte des traces GPS, de l'algorithme de séquençage ou alors des biais inhérents à l'enquête.

3.1.1. Erreurs liées à la collecte de données GPS

Une première cause d'erreurs relève des dispositifs d'enregistrement des positions géographiques qui sont collectées au cours de l'enquête. On distingue d'abord les contraintes techniques des *dataloggers* (Morris et Conner, 2017 ; Neatt *et al.*, 2016) : le temps de chauffe (*Time To First Fix*, 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 effectif d'une mobilité (où le *datalogger* est en veille en raison d'une immobilité prolongée) et le début de 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 (figure 3) : 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

5. La structuration de la base de données relationnelle MK s'organise autour de 50 tables (personnes, lieux, trajets, notes, *point of interest* – POI, etc.) reliées entre elles par une série de clefs primaires (identifiants des enquêtés, des trajets, des lieux, codes des familles, etc.). L'administration et la gestion de cette base de données relationnelle s'opèrent via le système PostgreSQL/PostGIS.

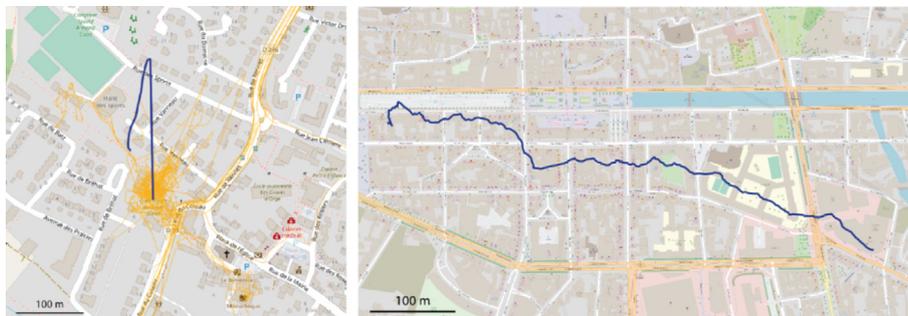


Figure 3. Effet de la perte momentanée du signal satellite (à gauche), effet de canyon urbain et décalage de la trace (à droite)

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.

Différentes analyses peuvent être envisagées pour identifier ces erreurs dans les traces brutes. 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](#) ; [Morris et Conner, 2017](#)), permet une première estimation de la qualité des données. Ainsi, 1,4 % des positions 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 GPS successives. Sur les 14,1 millions de couples de positions, 35 719 (0,25 %) présentent un écart temporel supérieur à la période d'échantillonnage (1 s). De plus, 8 887 couples se distinguent par une vitesse supérieure à 180 km/h, tandis que 14 778 cas correspondent à une distance spatiale supérieure à 50 m. Selon ces critères, près de 44 567 couples de positions distinctes (soit 0,31 % du total) présentent une aberration 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 générées par les dispositifs d'enregistrement, l'algorithme de séquençage est de son côté susceptible d'induire 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 périurbain 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 de mobilité contribue ainsi à générer des faux positifs ou négatifs ([figure 4](#)) : par exemple



Figure 4. Faux lieux sur l'autoroute (à gauche) et faux trajet à l'école (à droite)

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).

3.1.3. Les biais de l'enquête

Troisièmement, la phase d'enquête (figure 2), comporte des biais susceptibles d'introduire de nouvelles erreurs (Beaud et Weber, 1998 ; Rabaud, 2015). Dans la pratique, les conditions de passation des enquêtes sont inégales, parfois contraintes ou peu adéquates (lieux publics, disponibilité et attention variables des enquêtés, etc.). À 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. Les erreurs qui peuvent en résulter sont assez peu fréquentes et correspondent à deux principaux cas : soit des erreurs directes du protocole d'enquête (erreurs de saisie dans les questions ouvertes, utilisation fréquente des notes correctives), soit des erreurs indirectes passant par la détection d'incohérences (caractérisations incohérentes des activités ou de certains modes des trajets ; figure 5).

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

Une seconde entrée dans la qualité interne des données revient à réaliser une typologie des erreurs selon leur nature (Guptill et Morrison, 1995 ; Servigne *et al.*, 2005). Dans le cadre de données spatiales, temporelles et sémantiques telles que celles du programme MK, trois principaux types peuvent être retenus : l'incomplétude, l'imprécision et l'incohérence.

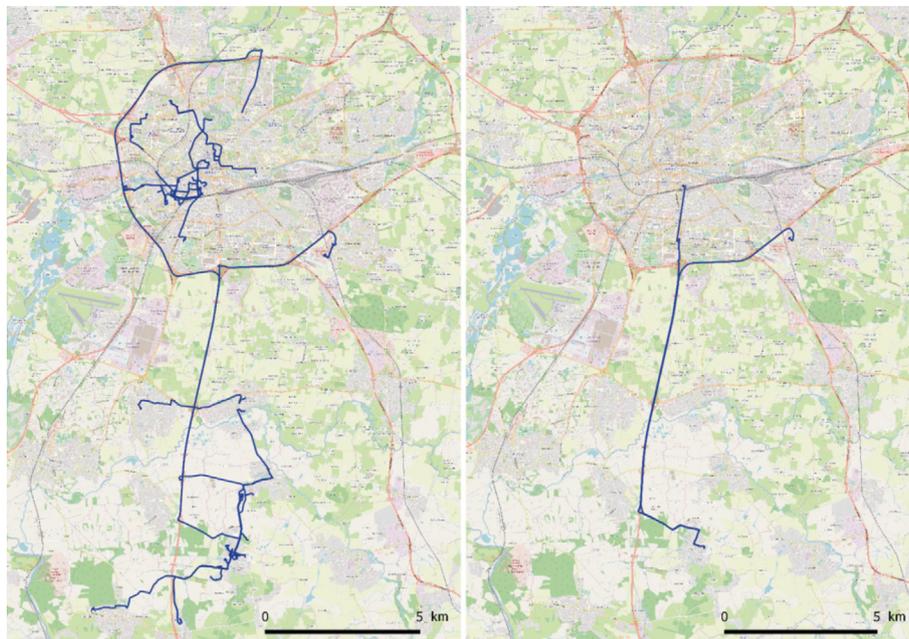


Figure 5. Incohérences dans la saisie des données : « enfants conducteurs » (gauche) et « piétons » sur l'autoroute (droite)

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 traces séquencées (lieux et trajets) et au niveau de l'homogénéité des différentes informations sémantiques associées (mode de transport ou d'accompagnement, activité, récurrence, etc.).

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 : absence de données enregistrées (figure 6), alternance stricte arrêt-trajet, adresses identiques d'arrivée et de départ de trajets successifs.

D'autre part, l'incomplétude sémantique des données peut être liée à l'absence d'informations attributaires sur les trajets et des lieux qui ont fait l'objet d'un enrichissement lors des phases d'entretiens (tableau 1). On dénombre par exemple plus de 12 % des trajets non renseignés selon les modes de transport ou d'accompagnement, quand 11 % des lieux ne sont pas caractérisés par une fréquence

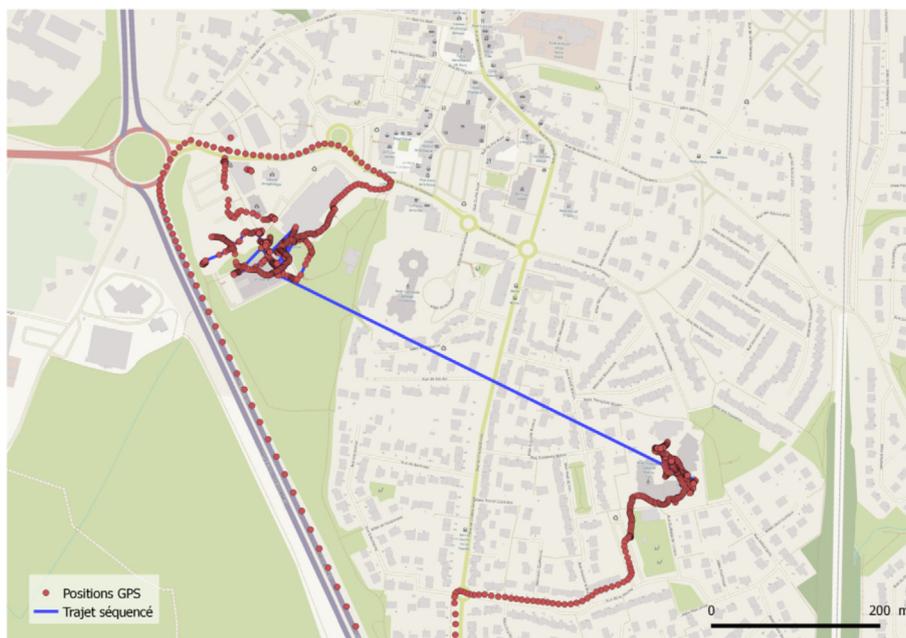


Figure 6. Exemple de non continuité des traces

Tableau 1. Exemples d'incomplétudes sémantiques

Table	Critères	Conditions	Nombre	Pourcentage (%)
Trajets	Mode de transport	Non renseigné	467	12,31
	Mode d'accompagnement		460	12,12
	Type de lieu de départ		291	7,67
	Type de lieu d'arrivée		304	8,01
Lieux	Fréquence	Non renseigné	392	16,39
	Type (domicile, travail, loisirs, courses, etc.)		289	12,08
	Nom		221	9,24
	Activité (manger, sociabilités, sport, soins, etc.)		604	25,25

d'usage. Semblant importantes, ces métriques sont toutefois à relativiser car certains trajets ou lieux séquencés par l'algorithme ne renvoient pas réellement dans les pratiques des enquêtés à des déplacements ou des activités (*cf.* 3.1.2). En ce sens, ces « faux » trajets et lieux ne sont pas qualifiés sémantiquement par les attributs classiques, mais peuvent dans certains cas être caractérisés au sein des notes correctives spécifiées au cours des entretiens par les enquêteurs (236 cas sur 645 incomplétudes sémantiques).

3.2.2. Précision géographique et temporelle

Un deuxième 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 temporellement à la seconde près. La procédure de séquençage résume l'information temporelle des traces en une suite d'objets (arrêts et trajets) définis par leur borne temporelle de début et de fin, ce qui parfois peut être préjudiciable. En effet, les seuils retenus pour le séquençage n'ont pas permis d'identifier dans les trajets des ruptures de vitesse suffisantes pour détecter, par exemple, des changements de mode pourtant effectivement réalisés : dans ce cas de figure, il n'est alors pas possible de retrouver l'horaire et le lieu où s'opèrent ce changement à partir des objets séquencés.

Par ailleurs, l'imprécision géographique est plus importante au sein des données séquencées puisque la localisation des lieux ou le tracé des trajets dépendent des positions GPS et donc de la qualité du signal ainsi que de l'algorithme de séquençage. Les localisations des trajets et des lieux peuvent donc être imprécises et se situer à distance des lieux ou des infrastructures repérables dans des référentiels comme ceux de l'IGN ou d'OpenStreetMap (*figure 7*). La distance entre des traces ayant utilisé la même route par exemple peut dépasser plusieurs centaines de mètres. De même, pour les lieux caractérisant des



Figure 7. Imprécision des trajets à l'entrée d'Orgères (à gauche) et des lieux associés à l'une des écoles (à droite)

« écoles »⁶, la distance au barycentre des points peut varier entre 84 m et 350 m. Cette imprécision s'avère problématique pour comparer l'intensité de la mobilité et l'étendue des espaces de vie des enfants à l'aide d'indicateurs de distance.

3.2.3. Cohérence logique et sémantique

Un troisième type d'erreurs correspond aux incohérences logiques et sémantiques dans, et entre, les tables enrichies. Les trajets et les arrêts de l'enquête MK sont caractérisés au sein de la base par une multitude de champs. Cette richesse d'informations, qui proviennent à la fois d'opérations automatisées et de saisies lors des entretiens, est synonyme d'incohérences diverses et pose de réels problèmes en matière d'intégrité et de validité des données (tableau 2). La détection de ce type d'erreur est d'abord passée par une démarche exploratoire de la base de données menée par plusieurs personnes en parallèle, amenant dans un second temps à l'élaboration collective d'une liste d'incohérences qui ont ensuite été détectées et quantifiées automatiquement.

Tableau 2. Exemples d'incohérences logiques et sémantiques

Table	Critères	Conditions	Nombre	Pourcentage (%)
Trajets	Durée	< 30 sec et > 3 h	225	5,93
	Vitesse	< 1 km/h et > 100 km/h	151	3,98
	Statut familial et mode de transport	Enfant conducteur de voiture	36	0,95
	Vitesse et mode de transport	Marche et > 20 km/h	63	1,66
Lieux	Durée	< 300 sec	247	10,37
	Date de fin	Ultérieure à la date de rendu	74	3,11
	Type de lieu et durée	Non domicile et > 12 h	162	6,8
	Type de lieu et fréquence	Domicile fréquenté une seule fois	111	4,86

6. Chaque point correspond à un lieu issu du séquençage, et donc chacun caractérise une occurrence de fréquentation du lieu par un unique enquêté (hormis certains points fréquentés à plusieurs reprises lorsqu'ils avaient été préalablement déclarés POI par les enquêtés).

Une partie de ces incohérences correspond aux longueurs, durées et vitesses aberrantes : par exemple près de 5 % des trajets de la base présentent soit des durées inférieures à 30 sec ou supérieures à 3 h ou bien des vitesses inférieures à 1 km/h ou supérieures à 130 km/h. De même, d'autres erreurs concernent les activités, les modes de transport et les modes d'accompagnement. Ces incohérences peuvent être mises en exergue 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 20 km/h ».

4. Améliorer la qualité des données de l'enquête MK : un apurement générique et reproductible

La qualification des types d'erreurs et leur quantification constitue un préalable à la définition d'une procédure d'apurement visant à améliorer la qualité interne des données. Dans la mesure où les sujets de l'enquête MK sont enquêtés d'abord en fin d'école primaire et plus tard au début du collège, il est nécessaire que ce protocole d'apurement soit suffisamment générique pour être reproductible sur les corpus récoltés lors des deux phases d'enquêtes. Ce protocole s'organise en 4 principales étapes : 1) le diagnostic généralisé des données ; 2) la classification des erreurs étiquetées en vue de définir des actions correctives adaptées ; 3) les tests d'implémentation de ces corrections et leur chaînage ; et 4) l'opérationnalisation définitive des corrections dans la base de données relationnelle. Cette partie présente ainsi les principales lignes directrices associées à chaque étape et quelques résultats préliminaires de mise en œuvre de ce protocole d'apurement dans la base de données MK.

4.1. Étape 1 : un diagnostic généralisé des erreurs

La première étape réside dans l'identification semi-automatique des erreurs de complétude, de cohérence et de précision grâce à des requêtes prédéfinies collectivement (cf. 3.2). Ces requêtes sont appliquées à toutes les tables de la BDD, y compris celles ne comportant pas de géométrie spatiale (personnes, familles, logements). D'un point de vue technique, cette étape est menée en parallèle au sein de trois environnements logiciels complémentaires (requêtes SQL au sein d'un SGBD PostgreSQL/PostGIS, scripts dans l'environnement R et de manière plus manuelle dans un logiciel SIG). Ces dispositifs garantissent notamment la généralisation et la reproduction des diagnostics aux différents temps de l'enquête, quand leur vérification croisée permet la validation des résultats. Ces diagnostics permettent ainsi d'étiqueter les données aberrantes dans la base de données par l'ajout de colonnes indiquant la validité des entités (trajets, lieux, positions, personnes, etc.) pour chacun des critères d'évaluation de la qualité retenus dans le diagnostic.

La mise en œuvre de ces diagnostics sur l'ensemble de la base montre que ces procédés sont bien généralisables (tableau 3). De manière rassurante, les niveaux

Tableau 3. Exemples d'erreurs aux différents temps de l'enquête Mobi'Kids (T1 pour le primaire, T2 pour le collège)

Table	Critères	Conditions	NB T1	% T1	Nb T2	% T2	NB total	% total
	Durée	< 30 sec et > 3 h	281	6,08	84	3,45	365	5,18
	Vitesse	< 1 km/h et > 100 km/h	202	4,38	65	2,67	279	3,79
Trajets	Statut familial et mode de transport	Enfant conducteur de voiture	38	0,82	5	0,04	43	0,61
	Vitesse et mode de transport	Marche et > 20 km/h	99	2,15	30	1,23	129	1,83
Lieux	Durée	< 300 sec	307	10,58	194	12,52	501	11,26
	Type de lieu et durée	Non domicile et > 12 h	190	6,55	123	7,94	313	7,03

d’erreur sont comparables entre les différentes phases d’enquête, hormis une diminution des erreurs liées aux biais d’enquête (par exemple « enfant conducteur ») liée notamment à une sensibilisation redoublée aux erreurs de codage lors de la formation des enquêteurs. Enfin, à l’issue de ce diagnostic de la qualité des données, toutes les erreurs sont étiquetées dans les tables de la base de données.

4.2. Étape 2 : définition des actions correctives

Sur la base du diagnostic générique des erreurs et de leur étiquetage dans la base de données, l’enjeu consiste à définir des actions correctives en vue d’améliorer la qualité des données collectées. Comme le montre la partie 3, la difficulté réside notamment dans la multiplicité des erreurs rencontrées : d’une part, selon leur type et leur origine (spatial et temporel en lien avec le *datalogger* et l’algorithme de séquençage), d’autre part, selon leur nature (incomplète, imprécision, incohérence). Le croisement de ces deux typologies permet de formaliser un ensemble de 8 classes différentes d’erreurs dans la base de données (figure 8), et ce faisant, ouvre la définition d’actions plus génériques.

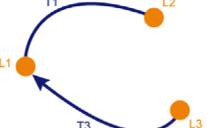
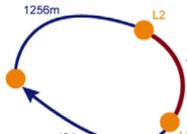
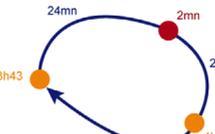
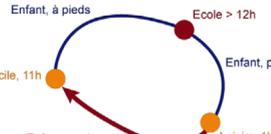
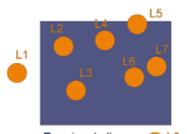
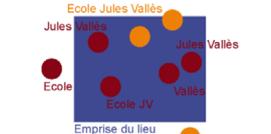
	Spatial (sources : datalogger et séquençage)	Temporel (sources : datalogger et séquençage)	Sémantique (source : enquête)
Complétude	<p>Points GPS manquants</p> 	<p>Trajet (ou lieu) manquant</p> 	<p>Traces non renseignées</p> 
Coherence	<p>Trace incohérente - distance</p> 	<p>Trace incohérente - durée</p> 	<p>Traces incohérentes</p> 
Précision	<p>Localisations multiples</p> 		<p>Traces imprécises</p> 

Figure 8. Typologie croisée des erreurs dans la base de données

Si de multiples actions plus ou moins lourdes peuvent être conduites afin d’améliorer la qualité des données, il s’agit aussi d’appliquer des correctifs efficaces qui permettent de passer au mieux et rapidement aux analyses. En ceci, les 8 types croisés des erreurs peuvent être traités selon 3 grands types d’actions (tableau 4) :

Tableau 4. Actions correctives de l'apurement générique

	Spatial (sources : datalogger et séquençage)	Temporel (sources : datalogger et séquençage)	Sémantique (source : enquête)
Complétude	<p>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.</p>	<p>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.</p>	<p>Exemple : absence de renseignement dans les tables des trajets et des lieux. Actions : étiquetage des individus, recodage si possible à l'aide des notes des lieux ou trajets, suppression si l'enquête n'a aucune information.</p>
Cohérence	<p>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.</p>	<p>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 ».</p>	<p>Exemple : enfants conducteurs, enfants à l'école le dimanche, piétons à plus de 20 km/h Actions : étiquetage des ind., recodage si possible à l'aide des notes correctives des lieux ou trajets.</p>
Précision	<p>Exemple : Localisations diffuses des trajets et des lieux. Actions: agrégation et appariement au réseau routier ou aux parcelles/adresses.</p>		<p>Exemple : diversité des noms déclarés des lieux Actions : étiquetage des ind., création de dictionnaires, recodage dans les tables.</p>

1) Une première action correspond à la création dans la BDD d'un type d'objet « inconnu » en plus des trajets et des lieux, qui regroupe les espaces-temps invalides (c'est-à-dire où le diagnostic de complétude et/ou de cohérence a révélé des erreurs).

2) Une seconde action d'ordre géométrique vise à améliorer la précision spatiale des traces. À l'aide de méthodes d'analyse et de transformation des données spatiales, il s'agit notamment d'apparier les objets séquencés à des référentiels externes et connus : par exemple certains trajets à des réseaux routiers, cyclables ou pédestres ou bien des lieux à des emprises de bâtiments ou de parcelles cadastrales.

3) Une troisième action porte uniquement sur la sémantique, quel que soit le type d'erreur. Il s'agit soit d'une modification si des notes correctives permettent de rectifier les informations, soit d'un recodage des modalités (notamment pour les réponses aux questions ouvertes), soit de désigner l'objet comme invalide s'il ne possède aucune sémantique.

4.3. Étape 3 : chaînage des actions correctives

Une fois ces actions correctives définies, il s'agit ensuite de mener des essais d'opérationnalisation dans la BDD pour vérifier leur mise en œuvre technique et évaluer les effets de ces actions. En raison de la multiplicité des erreurs et des données à améliorer, la difficulté de cette étape tient à la définition d'un ordre pertinent dans lequel ces actions sont exécutées.

Ainsi, des essais sont d'abord menés de manière indépendante pour chaque type et chaque origine d'erreur. Ils sont systématiquement suivis d'une répétition de l'étape 1 qui permet éventuellement d'ajuster l'action définie en étape 2. Cette première phase précède des essais croisés, qui se déroulent de la même manière, et dont le but est de tenir compte des relations entre les différentes tables de la BDD. Par exemple, la suppression des positions GPS erronées selon le nombre de satellites améliore, sans re-segmenter, la précision et la cohérence des lieux et des trajets. L'itération de ces essais amène enfin à définir un ordre pertinent dans lequel les actions correctives sont appliquées. Et par exemple, les « espaces-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 trajets anormalement courts (moins de 10 m), et enfin terminés par les trajets et les lieux ayant une sémantique lacunaire.

4.4. Étape 4 : opérationnalisation dans la base de données

Finalement, la dernière étape de l'apurement correspond à l'opérationnalisation de la chaîne de corrections dans la BDD. Menée dans un environnement PostgreSQL/PostGIS, elle doit aboutir à la création d'une version apurée de la base de données qui servira aux différentes analyses. La mise en œuvre de ce protocole pourra être

renouvelée après chaque nouvelle phase du suivi longitudinal (enfants scolarisés au collège).

En outre, pour vérifier si cette opérationnalisation a abouti, la quatrième étape se termine et ouvre une phase de pré-analyse du corpus apuré, qui consiste à valider la possibilité de créer des « indicateurs » pour les analyses thématiques envisagées. En effet, 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.

5. Conclusion : Vers un apurement adapté et reproductible

Suivant l'objectif d'analyser l'apprentissage de l'autonomie chez les enfants, 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 enregistrées par suivi GPS, des lieux et des trajets séquencés par un algorithme, des renseignements sur les mobilités et lieux fréquentés acquis à partir de questionnaires sur tablettes. Si cette chaîne de méthodes mobilisées permet d'atteindre une grande richesse informationnelle, elle constitue également un défi important en générant des problèmes dans la cohérence, la précision et la complétude des données collectées.

Dans un objectif de disposer de traces spatio-temporelles séquencées et enrichies qui soient précises, complètes et cohérentes, le temps comme les moyens dédiés à la phase de diagnostic et d'amélioration de la qualité des données ne doivent pas être minorés. Au contraire, cette étape nécessaire et préalable aux analyses forme la clé de voute de l'utilisabilité comme de la validité scientifique des données. Dans le cadre de MK, cet enjeu est d'autant plus présent que l'approche mixte nécessite à la fois de pouvoir formaliser et structurer, au sein d'une base de données, des données qualitatives et quantitatives et surtout de pouvoir les faire dialoguer entre elles.

Fondé sur les expériences acquises dans le cadre du programme Mobi'Kids, cet article suivait l'objectif de discuter cette étape d'apurement des données, par ailleurs souvent éludée dans l'explicitation des travaux scientifiques. Dans la continuité de quelques travaux sur la notion de qualité des données (Devilleers, 2004 ; Devillers et Jeansoulin, 2005 ; Guptill et Morrison, 2013 ; Servigne *et al.*, 2005 ; Shi *et al.*, 2002), 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 de diagnostic ouvre la voie à la définition et la mise en œuvre d'un protocole d'apurement de la base de données relationnelle. Appliqué aux données de MK, ce protocole est avant tout confronté à des enjeux d'objectivité, de généralité et de reproductibilité aux différentes phases de l'enquête longitudinale.

Plus largement, cet article propose une approche théorique et méthodologique reproductible à d'autres cas d'analyses de mobilités quotidiennes croisant de données nombreuses, hétérogènes et multidimensionnelles.

Remerciements

Ces travaux sont financés par l'ANR dans le cadre du programme Mobi'Kids (ANR-16-CE22-0009). Cet article a fait l'objet d'une présentation orale lors du colloque SAGEO 2019.

Bibliographie

- 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*, vol. 27, p. 385-407.
- Chen C., Ma 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*, vol. V 68, p. 285-299, doi: 10.1016/j.trc.2016.04.005.
- Christensen P.M., O'Brien M. (2003). *Children in the city home, neighborhood, and community*, London, New York, Routledge Falmer.
- 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*, vol. 5, p. 227-246.
- CERTU (2008) L'enquête ménages déplacements « standard Certu », collection « Références », n° 72, Lyon, France, 2008.
- 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*, vol. 40, p. 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. (2003). *L'enfant en ville : autonomie de déplacement et accessibilité environnementale*. Thèse de Doctorat en Psychologie, Université René Descartes.
- Depeau S., Bedel O., Cherel P., André-Poyaud I., Chardonnel S., Gombaudo 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, Actes Conférence SAGEO 2019, 13-15 novembre, Clermont-Ferrand, p. 238-250, Url: <https://sageo2019.inrae.fr/actes/>.
- Depeau S., Chardonnel S., André-Poyaud I., Lepetit A., Jambon F., Quesseveur E., Gombaudo J., Allard T., Choquet C.-A. (2017). Routines and informal situations in children's daily lives, *Travel Behaviour and Society*, vol. 9, p. 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 P.-A., Lutoff C. (2014). Évaluation comparée de l'apport de l'assistance GPS aux enquêtes de mobilité, *Netcom*, vol. 28, n° 1/2, p. 13-34.
- Duroudier S., Chardonnel S., Mericskay B., Andre-Poyaud I., Bedel O., Depeau S., Devogele T., Etienne L., Lepetit A., Moreau C., Pelletier N., Ployon E., Tabaka K. (2019). « Données hétérogènes de mobilités quotidiennes: diagnostic qualité et protocole d'apurement à partir de la base MOBI'KIDS », SAGEO 2019, Clermont-Ferrand, 13-15 novembre 2019. p. 58-70. Url: <https://sageo2019.inrae.fr/actes/>.
- Guptill S.C., Morrison J.L. (1995). *Elements of Spatial Data Quality*, Elsevier, Oxford.
- Hillman M., Adams J.G., Whitelegg J. (1990). *One false move: a study of children's independent mobility*, London, Policy Studies Institute.
- INED (2019) Saisie, codage, apurement, documentation, <https://www.ined.fr/fr/ressources-methodes/methodologie-enquete/les-choix-methodologiques/saisie-codage-apurement-documentation/>; consulté en avril 2019.
- 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*, vol. 9, n° 1, p. 38-42, DOI: 10.7763/IJCTE. 2017. V9. 1108.
- Fyhri A., Hjorthol R., Mackett R.L., Fotel T.N., Kyttä M. (2011). Children's active travel and independent mobility in four countries: Development, social contributing trends and measures. *Transport Policy*, vol. 18, p. 703-710. <https://doi.org/10.1016/j.tranpol.2011.01.005>
- Kyttä M. (1997). Children's independent mobility in urban, small town, and rural environments, in Camstra R. (Ed.), *Growing up in a Changing Urban Landscape*, Assen, Royal Van Gorcum, p. 41-52.
- 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*, vol. 16, p. 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*, vol. 9, n° 8.
- Lin M., Hsu W.-J. (2014). Mining GPS data for mobility patterns: A survey, *Pervasive and Mobile Computing*, vol. 12, p. 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.
- Morris G., Conner L.M. (2017). Assessment of accuracy, fix success rate, and use of estimated horizontal position error (EHPE) to filter inaccurate data collected by a common commercially available GPS logger. *PLoS one*, vol. 12, n° 11.
- Neatt K., Millward H., Spinney J. (2016). Aggregation and spatial analysis of walking activity in an urban area: results from the Halifax space-time activity survey, *IOP Conference Series: Earth and Environmental Science*, vol. 34, n° 1, p. 012-022.
- Rabaud M. (2015). L'évolution de la méthodologie des enquêtes déplacements standard Certu, entre continuité et adaptations, in Armoogum, J. (Ed.), *Mobilité en transitions: connaître, comprendre et représenter*. Cerema, Lyon.
- Servigne S., Lesage N., Libourel Th. (2005). *Quality Components, Standards, and Metadata. Fundamentals of Spatial Data Quality*, ISTE, London, Newport Beach, p. 179-210.
- Shaw B., Bicket M., Elliott B., Fagan-Watson B., Mocca E., Hillman M. (2015). *Children's Independent Mobility: an international comparison and recommendations for action*. Final Report. London: PSI.
- Shi W., Fisher P., Goodchild M. F. (2002). The Registration of Quality in a GIS, in Shi W., Fisher P., Goodchild M. F. (Eds), *Spatial Data Quality*, Taylor & Francis, Chapter 13 p. 186-199.
- Schoeppe S., Duncan M.J., Badland H.M., Oliver M., Curtis C. (2013). Associations of children's independent mobility and travel with physical activity, sedentary behavior and weight status: A systematic review. *Journal of Science and Medicine in Sport*, vol. 16, n° 4, p. 312-319.
- 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*, vol. 51, p. 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.