DONNÉES DE MOBILITÉ POUR LA MODÉLISATION DES DÉPLACEMENTS
Fiche chapeau

Les ressources

FICHE TECHNIQUE

COLLECTE ET UTILISATION DE DONNÉES DE MOBILITÉ POUR LA MODÉLISATION DES DÉPLACEMENTS

Des enquêtes ménages-déplacements aux données massives

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cerema éditions

PRÉSENTATION DE LA SÉRIE

Pilotée et rédigée par le Cerema la série de fiches « ***Données de mobilité pour la modélisation des déplacements*** » vise à fournir un panorama des données disponibles pour la modélisation. Ces fiches donnent des critères d'analyse de la pertinence d'une source de données en illustrant son mode de collecte, les informations sur la source de données et comment sont traitées ces données (biais connu, intégration dans un modèle...).

SOMMAIRE

INTRODUCTION

La fin de la décennie 2010 a vu apparaître de nouvelles sources de données pour mesurer et analyser les mobilités, en lien avec l'utilisation de plus en plus massive d'objets connectés par les usagers des transports (téléphone, dispositifs d'aide à la conduite…).

Ces données tracent les déplacements dans l'espace de façon moins contraignante pour les usagers que les dispositifs d'enquête, en raison du caractère passif de leur collecte. Elles permettent de plus de tracer en détail l'itinéraire réalisé par un usager. C'est pourquoi elles ont suscité ces dernières années un vif intérêt dans la communauté des praticiens de la connaissance et de la modélisation des mobilités.

Elles posent toutefois des questions nouvelles. Quelle est leur représentativité des usagers des transports ? Permettent-elles vraiment de mesurer des déplacements en identifiant correctement leurs origines et leurs destinations ? Peuvent-elles se prêter à des comparaisons spatio-temporelles ? Alors que les enquêtes auprès des ménages ou des voyageurs nous renseignent sur

certaines caractéristiques des individus, des ménages et des véhicules, quelles informations individuelles peuvent nous fournir ces nouvelles sources ? L'enjeu est de savoir quel rôle elles peuvent jouer dans le paysage des études de mobilité. Peuvent-elles remplacer partiellement ou totalement certaines collectes ? Offrent-elles des visions complémentaires des approches d'observation plus classiques que sont les enquêtes par sondage et les comptages de passagers et de véhicules ?

Cette analyse du domaine de pertinence de ces nouvelles sources ne peut se faire qu'en lien avec les méthodes de modélisation utilisées. C'est pourquoi la première partie de cet ouvrage est consacrée à une présentation générale des principes de la modélisation des déplacements. Nous nous plaçons dans une optique de construction d'outils d'aide à la planification des politiques publiques de mobilité, donc sur des horizons de temps allant de quelques années à plusieurs décennies, sur des territoires allant de l'échelle du quartier à la région, voire plus large. Les phénomènes de mobilité et de trafic y sont généralement représentés de façon statique, sans considération pour l'écoulement des flux. Une analyse de la pertinence de ces sources pour des outils de prévision des trafics à court terme ou de modélisation dynamique serait nécessairement différente.

Dans une deuxième partie, nous proposons une typologie des données de mobilité en fonction de leurs caractéristiques principales pour la modélisation des déplacements. La série de critères proposée dans la troisième partie permet enfin de rentrer plus en détail dans les points forts et points faibles des sources et dans la manière dont on peut les utiliser pour mesurer ou modéliser les mobilités.

Ce document s'inscrit dans une série thématique à laquelle sont associées des fiches techniques qui balayent, selon la grille d'analyse ici proposée, les différentes sources de données de mobilité utilisées dans les modèles de déplacements. Ces fiches peuvent toutefois être consultées indépendamment de cette fiche chapeau.

Cette série de fiches s'adresse aux techniciens de bureaux d'études, de collectivités et de services de l'État ayant à commander, manipuler ou interpréter des données de mobilité, en particulier dans le cadre de travaux de modélisation des transports.

Nota : *Cette série thématique contient, outre ce document introductif, des fiches techniques faisant le point des caractéristiques et des connaissances sur les différentes sources de données de mobilité. Ces documents font ressortir les forces et faiblesses de chaque source et permettent ainsi de disposer d'une vision d'ensemble des complémentarités possibles entre les différentes sources, pour répondre au mieux aux questions posées à la modélisation.*

Rôle des encadrés

Dans la suite du document, des encadrés permettent d'approfondir certains points techniques. Le lecteur désireux d'avoir une lecture rapide du document peut les omettre.

1 • POURQUOI MODÉLISER LES DÉPLACEMENTS ?

Nous rappelons ici les grands principes de la modélisation des déplacements, en vue de situer dans son contexte le cadre d'analyse des sources de données de mobilité que nous proposons par la suite. Les modèles de déplacements sont des outils d'aide à la décision qui ont trois grandes fonctions : connaître, comprendre et prévoir les mobilités.

1.1 Connaître les mobilités

La construction d'un modèle de déplacements est l'occasion de collecter des données sur :

- l'offre de transport (réseaux, capacités routières, politique d'arrêt et horaires de transport collectif, etc.) ;
- les générateurs de déplacements (population, emplois, services, etc.) ;
- les mobilités (trafics ponctuels, enquêtes origine-destination (OD), enquêtes ménages-déplacements, etc.) ;

et de constituer une base de données territoriale structurée à partir de ces informations, initialement disparates.

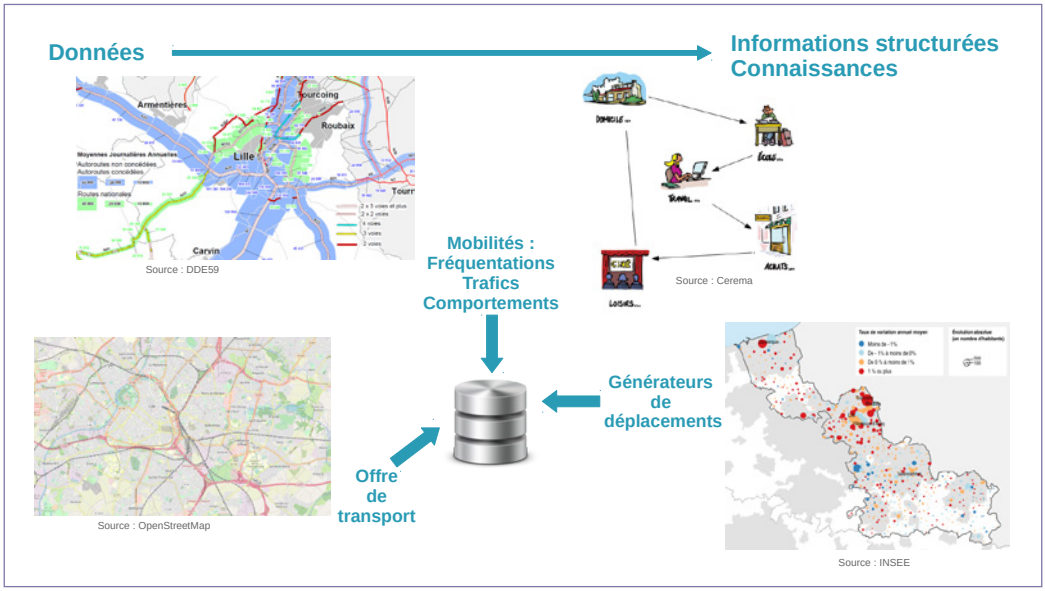


Illustration 1 – Construction et exploitation conjointe de bases de données sur les mobilités, les générateurs et l'offre de transport.

Cette base de données est en elle-même un outil de connaissance du territoire ciblé. Elle peut notamment alimenter des démarches de diagnostics territoriaux sur la mobilité. Elle permet par exemple de croiser les données d'offre et de demande par superposition (cf. Illustration 2) ou de réaliser des comparaisons multimodales sur les usages des réseaux (cf. Illustration 3).

À partir d'une base de données de ce type, il est aussi possible de réaliser des analyses d'accessibilité des territoires qui mettent en regard l'offre de transport et les opportunités d'activités offertes par le territoire (cf. bibliographie [1]). Contrairement à un modèle de déplacements, ces méthodes ne permettent pas d'identifier les niveaux de congestion des réseaux, ni de mesurer les effets de rétroaction sur la demande d'une modification du niveau de service. En revanche, elles offrent une première approche éclairante de l'adéquation entre l'offre de transport et l'organisation territoriale (cf. Illustration 4). Elles peuvent être réalisées dans un contexte mono ou multimodal.

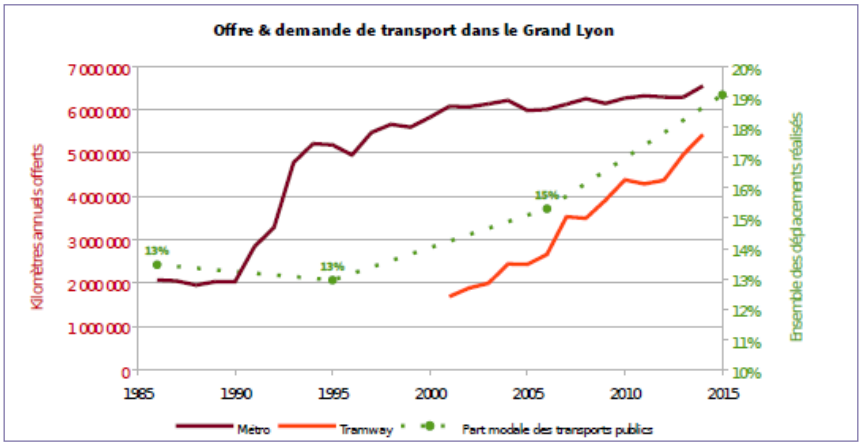


Illustration 2 – Évolution conjointe de l'offre et de la part modale des transports collectifs (sources : enquêtes ménages-déplacements de l'aire métropolitaine lyonnaise et enquêtes annuelles sur les transports collectifs urbains)

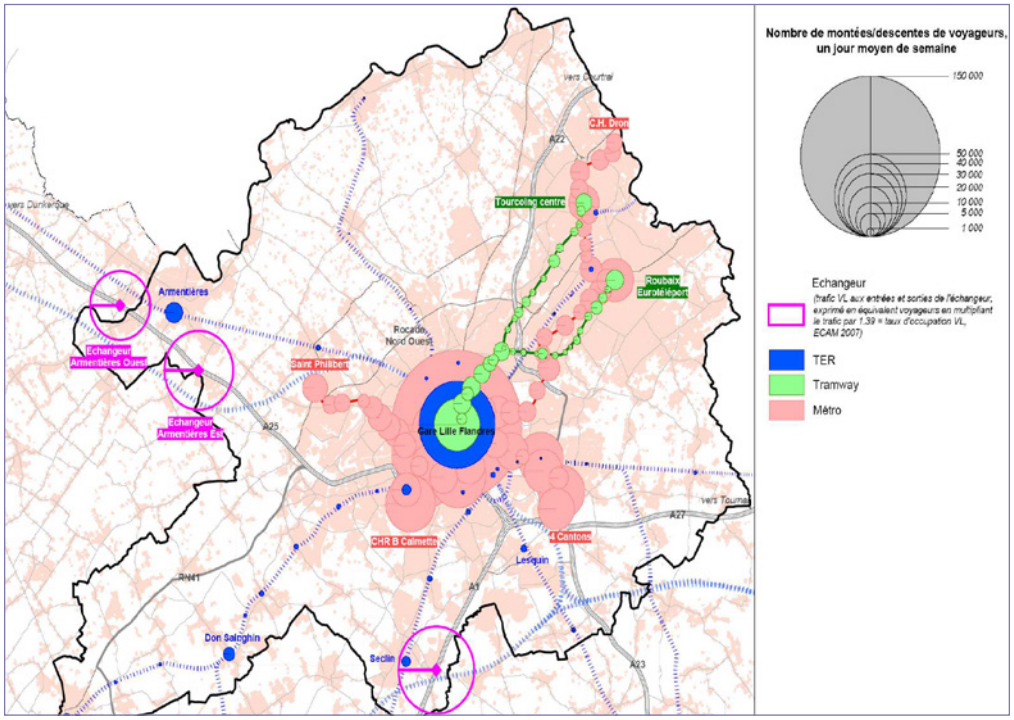


Illustration 3 – Comparaison des montées-descentes aux arrêts des modes lourds de transport collectif et des entrées-sorties au niveau des échangeurs autoroutiers (réalisation DREAL Hauts-de-France)

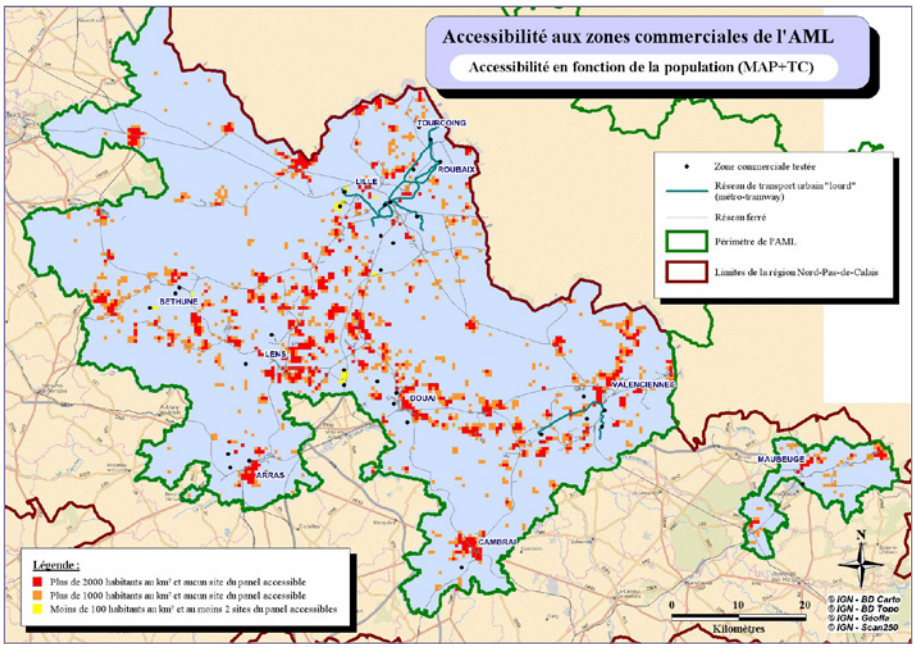


Illustration 4 – Exemple d'analyse d'accessibilité en marche + transport collectif croisant l'offre de transport, la population et les opportunités du territoire, ici les zones commerciales (réalisation Cerema)

1.2 Comprendre les mobilités

Le processus de construction d'un modèle permet ensuite de **dégager de la base de données des éléments d'explication des phénomènes observés**, par confrontation de l'offre et de la demande.

Pour cela, le modélisateur choisit les variables d'entrée adaptées et la bonne forme des équations ou algorithmes (spécification) pour expliquer le phénomène ciblé, puis ajuste les paramètres du modèle (calibrage) par confrontation de ses résultats aux données de calage. La situation de calage combine toutes les données disponibles à un horizon temporel donné (cf. Illustration 5).

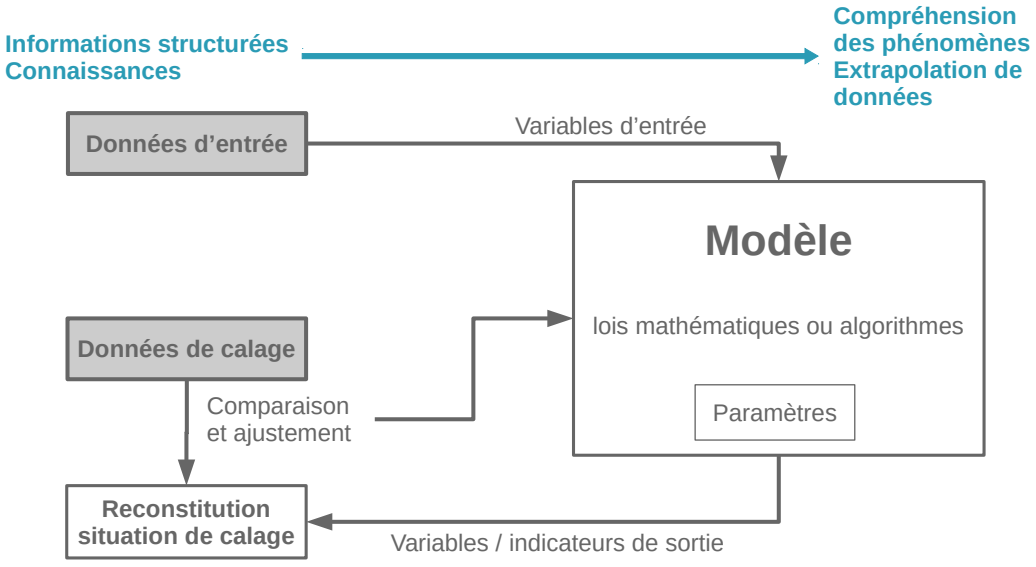


Illustration 5 – Calage du modèle à partir de données d'entrée sur l'offre et la demande, confrontées aux données de calage

Le modèle met en lumière des liens de causalité entre variables décrivant les individus, les ménages et le territoire et variables décrivant les comportements de mobilité (cf. Illustration 6). La forme du modèle retenu (types d'équations ou algorithmes, relation linéaire, quadratique, exponentielle…) ainsi que les variations des paramètres (selon les motifs par exemple) sont des éléments explicatifs des phénomènes de mobilité modélisés. Le modélisateur acquiert donc en construisant le modèle une compréhension des comportements et du fonctionnement du territoire modélisé et de ses spécificités, qu'il pourra remettre à profit lors de la phase d'analyse des résultats de la modélisation.

ACTIFS	Composition du ménage	Célibataire	Couple avec au moins un enfant de moins de 18 ans	Couple sans enfant *ou* Ménages à 2 adultes et plus	Seul avec enfant(s) de moins de 18 ans	
	Travail à temps plein ou partiel	Temps plein	Temps partiel			
	Sexe	Femme	Homme			
	Localisation	<450h/km2	Entre 450 et 1500	>1500h/km2		
	Age	18-24	25-39	40-49	50-59	>60 ans
	Motorisation	Oui	Non	Compétition		

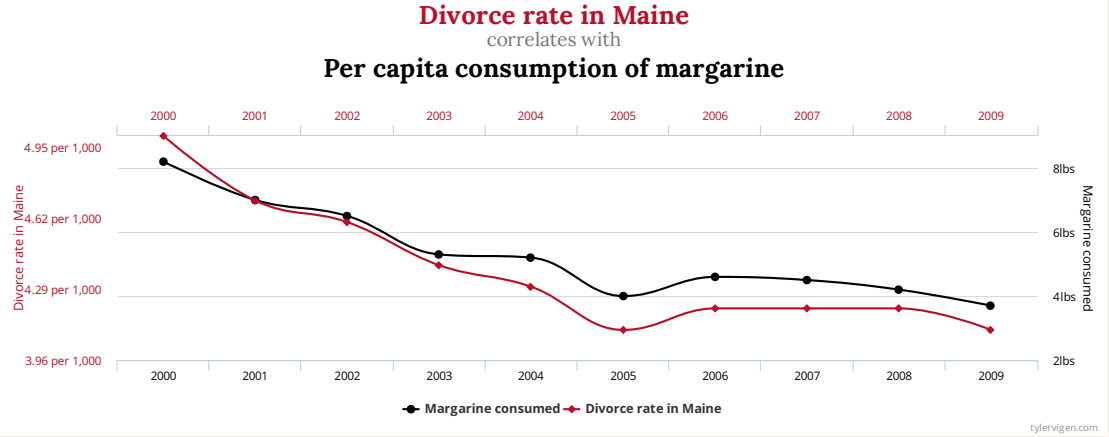
Illustration 6 – Extraction des variables discriminantes de la nature des chaînes de déplacements réalisées par les actifs (source : enquête déplacements régionale Rhône-Alpes et modèle multimodal régional rhônalpin, réalisation : Explain)

Bien choisir les variables explicatives d'un modèle : de la détection de corrélations à l'hypothèse de causalité

Intéressons-nous à une caractéristique socio-démographique du ménage que l'on cherche à lier à un comportement de mobilité de ses membres. Une analyse de corrélation peut par exemple faire ressortir qu'un haut niveau de motorisation du ménage est corrélé avec des déplacements à plus longue distance de ses membres.

C'est ensuite la connaissance que nous avons des phénomènes qui nous permet d'interpréter cette corrélation, qui n'est pas forcément synonyme d'une relation de causalité entre les deux phénomènes : est-ce le haut niveau de motorisation qui explique les plus longues distances parcourues ou bien le fait de résider dans un territoire périurbain éloigné qui explique à la fois la forte motorisation et les longues distances parcourues ?

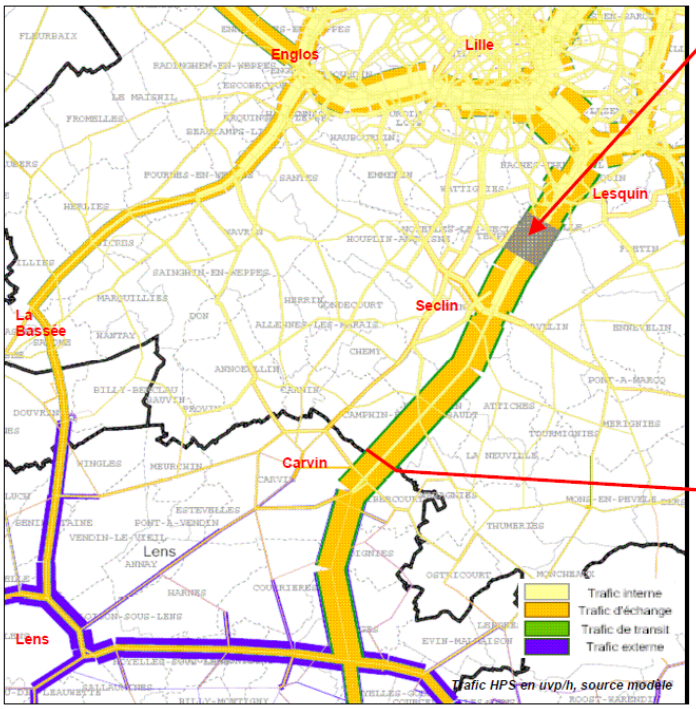
Si on sort du domaine des transports, le graphique ci-dessous reliant l'évolution temporelle du taux de divorces dans l'État du Maine et la consommation par individu de margarine, présente un exemple, certes un peu caricatural mais pédagogique, du risque lié à l'assimilation entre corrélation et causalité.



Source : https://www.tylervigen.com/spurious-correlations

En complément de cette fonction de compréhension des liens entre les variables, le modèle calé permet une **interpolation**, notamment spatiale, des **données de calage** (ex : débits, vitesses, etc.).

Le modèle permet ainsi la compréhension du fonctionnement des flux (cf. Illustration 7) sur l'ensemble du territoire modélisé.



A hauteur de Templemars

Fonction de l'autoroute A1 en milieu urbain entre Seclin et Fâches Thumesnil (à l'heure de pointe du soir):	Part
Interne arrondissement de Lille	27%
Relations Arrondissement de Lille - SCOT Lens-Lievin-Hénin-Carvin	22%
Relations Arrondissement de Lille - SCOT Grand Douaisis	11%
Relations Arrondissement de Lille - France Hors Région	11%
Relations Arrondissement de Lille - Arrondissement d'Arras	8%
Relations Arrondissement de Lille - reste de la Région	7%
Transit arrondissement de Lille	14%
Autres	1%
Total	100%

33%

Trafic HPS en uvp/h, source: modèle de trafic

A hauteur de Phalempin:

La moitié du trafic PL recensé en limite d'arrondissement est en transit

	VL	PL	Total
Echange	84 500	13 470	97 970
Transit	8 900	14 110	23 010
Total	93 400	27 580	120 980

Trafic mjo en véh/j, source ECAM 2006

Illustration 7 – Analyse des trafics en situation de calage sur une section d'autoroute visée par un projet d'échangeur, différenciant flux internes, flux d'échange et flux de transit par rapport à l'agglomération lilloise (réalisation : DREAL Hauts-de-France)

Selon le niveau d'agrégation du modèle, les données à mobiliser ne seront pas tout à fait de la même nature. La modélisation de la demande de déplacements peut être réalisée :

- à une échelle désagrégée sur les individus, en considérant les choix de chaque individu, avant une éventuelle phase de passage à l'échelle de la population totale ;
- à une échelle agrégée sur les individus, en considérant des groupes à comportements homogènes représentés chacun par un individu moyen.

Rôle des données de pratique de mobilité dans la chaîne de modélisation

Selon le type de modèle que l'on construit, les données de mobilité n'y seront pas intégrées de la même manière.

Les modèles de demande de déplacements ont pour principaux objectifs d'estimer les grandes variables de mobilité que sont le nombre de déplacements réalisés, l'heure ou la période de déplacement, les distances parcourues, les modes de transports choisis, etc. Dans ce type de modèles, les données décrivant les mobilités sont plutôt utilisées comme **données de calage**.

Les modèles d'affectation ont pour objectif d'estimer les flux passant sur chaque tronçon du réseau. Dans ce type de modèle, les données décrivant les mobilités sont utilisées comme **données d'entrée** (pour la construction de matrices origine-destination) et comme **données de calage** (en confrontant les résultats à des débits observés en différents points du réseau).

Il est bien entendu possible de combiner les deux approches (demande et affectation). Cela correspond à la démarche classique de modélisation à 4 étapes (génération, distribution, choix modal et affectation) .

Le lecteur désireux d'aller plus loin dans la compréhension des différentes approches de modélisation statique peut se référer aux ouvrages [**5**], [**7**] et [8].

Lorsqu'on calibre un modèle désagrégé sur les individus, on cherche des corrélations entre les caractéristiques individuelles et les comportements de mobilité. On peut donc se contenter de **données de calage non représentatives de la population** que l'on cherche à modéliser, sous réserve :

- qu'elles soient suffisamment variées pour décrire tous les types de comportements souhaités ;
- qu'elles décrivent suffisamment finement les individus eux-mêmes.

Le passage à la population totale (on parle d'« inférence statistique ») se fait alors en appliquant le modèle calé sur les individus à une population synthétique, reproduction de la population totale sur les variables utilisées par le modèle (cf. Illustration 8).

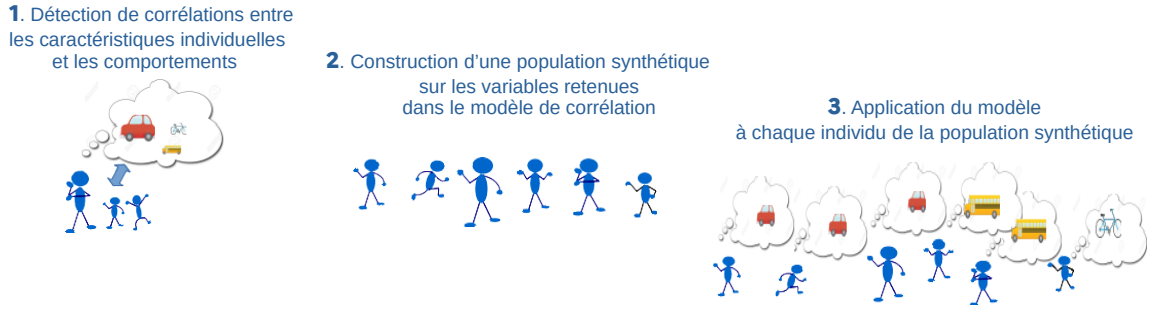


Illustration 8 – Étapes de mise en œuvre d'un modèle désagrégé (source : Cerema)

Un modèle agrégé suppose que chaque individu de la population adopte les comportements d'un individu moyen. Lorsqu'on calibre un modèle agrégé sur les individus, on a besoin de disposer de **données de calage représentatives de la population** sur laquelle s'applique le modèle, afin d'être en mesure de construire ces valeurs moyennes. Des niveaux intermédiaires d'agrégation sont couramment utilisés, en subdivisant la population globale en sous-populations, chacune constituant un groupe d'individus sur lequel les comportements sont supposés homogènes. Chaque groupe est représenté par un individu moyen. Les échantillons observés des données de calage doivent être généralisés à chacune de ces sous-populations, *via* une opération de redressement (voir encadré ci-après).

L'exigence de représentativité des données de calage est donc plus forte si le modèle se base sur une représentation agrégée des individus. Les données de calage doivent également offrir suffisamment de variables auxiliaires[1] pour permettre le redressement.

Redressement de la donnée de calage sur une population

Il s'agit de dénombrer des quantités à l'échelle d'une population (au sens statistique du terme), sans avoir mesuré ces quantités sur chaque unité de la population. L'exploitation d'une enquête OD routière passe par exemple par cette étape de redressement : il s'agit de connaître les origines et destinations des véhicules empruntant un axe à partir de celles observées sur un échantillon de véhicules empruntant l'axe. Le redressement permet de pondérer les observations de l'échantillon enquêté pour permettre de le ramener au flux total passé sur l'axe.

On parle d'« inférence statistique » pour décrire cette opération de passage d'un échantillon à une population. Celle-ci n'est possible de façon agrégée que si l'échantillon de départ est suffisamment représentatif de la population sur laquelle on cherche à construire les indicateurs. En théorie, cette représentativité ne peut être garantie que si l'échantillon résulte d'un sondage totalement aléatoire.

Les opérations de redressement cherchent des corrélations entre les variables à estimer et des variables auxiliaires. Ce sont ces corrélations qui vont permettre de généraliser les grandeurs obtenues à la population.

1 « *variable qualitative ou quantitative autre que la variable d'intérêt qu'on cherche à estimer et que les variables de gestion de l'échantillon* » (P. Ardilly, Techniques de Sondage). Autrement dit, ce sont les variables qui caractérisent les ménages, les individus, les déplacements autres que celles qui mesurent directement le déplacement, donc essentiellement des variables socio-démographiques (taille du ménage, CSP, âge, sexe, composition du ménage, etc.)

1.3 Prévoir les mobilités

À partir d'un modèle calé, il est ensuite possible de simuler les effets d'évolutions de l'offre et de la demande (cf. Illustration 9). Le modèle est donc aussi un **outil pour prévoir**. Cette étape de prévision n'est toutefois pas systématique. Lors de l'étape de prévision, l'objectif est d'évaluer **l'impact quantitatif d'un projet sur les flux de mobilité futurs** (cf. Illustration 10). Contrairement à certains modèles de prévision à court terme, il n'est pas possible dans un contexte de planification à long terme de prévoir l'état du système de transport uniquement à partir de ses états précédents. En effet, les horizons de temps considérés vont jusqu'à plusieurs décennies, rendant la seule prolongation des tendances passées inenvisageable. Il faut donc passer par des scénarios d'évolution des variables d'entrée.

Le besoin de prévision contraint donc également le choix des variables d'entrée du modèle : on se restreint aux variables qu'il est possible de **projeter en situation future *via* des scénarios** (cf. Illustration 11 page suivante). Étant donné l'effort de projection, voire de prospective, nécessaire pour projeter des variables dans le futur, elles sont forcément en nombre restreint. Selon que le modèle est construit prioritairement pour comprendre ou pour prévoir, les choix des variables à retenir ne sont donc pas toujours les mêmes.

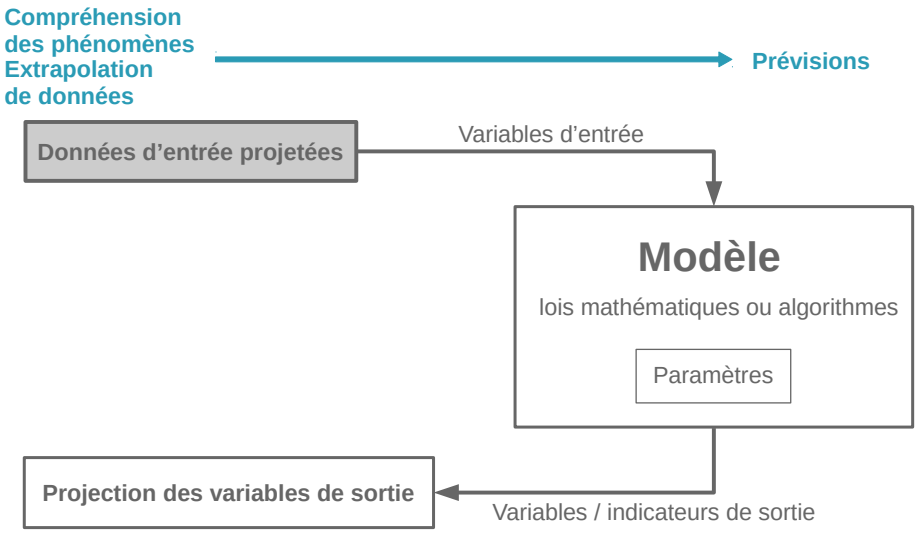


Illustration 9 – Simulation de l'impact de scénarios projetés à partir du modèle calé sur la situation actuelle

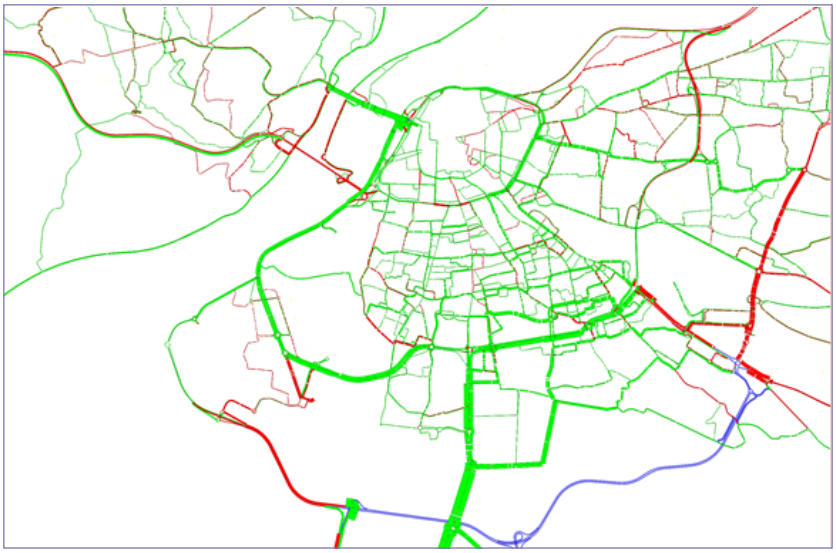


Illustration 10 – Carte de différences : simulation des volumes qui se chargent (en rouge) et qui se déchargent (en vert) suite à la mise en place d'une nouvelle infrastructure routière (source : modèle d'Avignon, réalisation Cerema)

Sur le plan strictement mathématique, la vocation principale du modèle (comprendre ou prévoir) peut également modifier le choix du type de modèle à mettre en place. Comme le soulignent notamment Breiman [**2**] ou Saporta [**3**], certains modèles issus de méthodes d'apprentissage automatique (*machine learning*) ont de très bonnes capacités prédictives, aux dépens de leurs capacités explicatives, en raison de leur fort effet « boite noire ». À l'inverse, les méthodes statistiques et économétriques ont de meilleurs potentiels d'explication des phénomènes, mais parfois de moindres performances en prévision.

Si le choix de modèles « boites noires » peut se révéler intéressant pour la prévision sur le plan mathématique, il n'est généralement pas privilégié en modélisation des déplacements, pour deux raisons principales :

- cette discipline apporte une aide à la décision publique, avec un engagement généralement assez important de fonds publics sur les projets évalués et, en retour, une attente forte des citoyens et des associations de transparence de cette décision. Les politiques publiques doivent pouvoir être justifiées par des critères compréhensibles et ne peuvent uniquement résulter d'une décision prise par un algorithme ;
- les modèles du *machine learning* ont tendance à utiliser de très nombreuses variables en entrée, variables qui comme on l'a vu précédemment seront pour certaines difficiles à projeter dans le futur.

Pour ces raisons, les méthodes statistiques et d'économétrie sont plus fréquemment employées en modélisation des déplacements que les méthodes d'apprentissage automatique.

Nota *: la section 1. ci-avant présente de façon volontairement succincte les principes de la modélisation des déplacements. Le lecteur souhaitant approfondir ce sujet peut se référer aux documents [4], [5], [6], [7] et [8].*

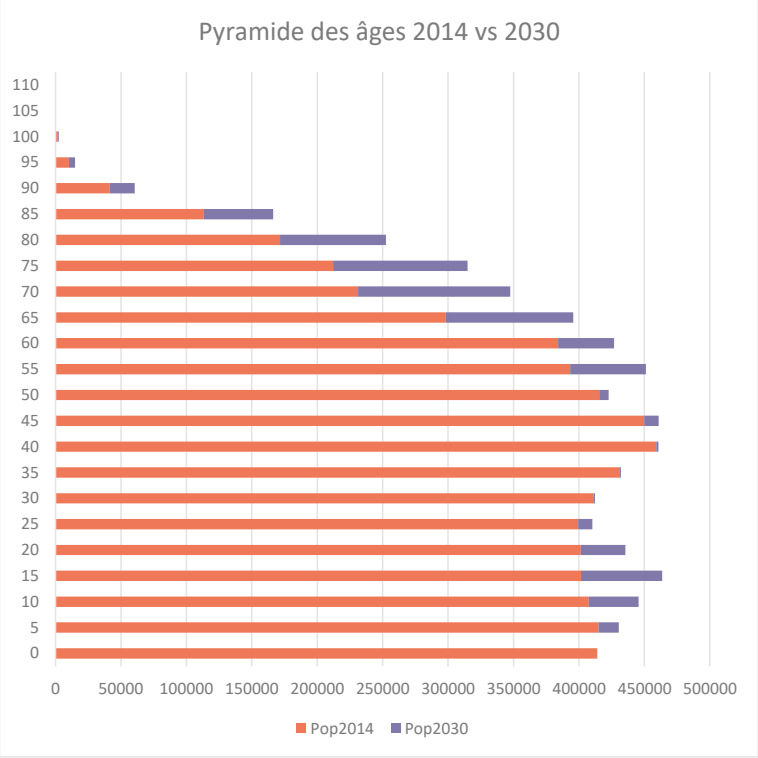


Illustration 11 – Exemple de projection des populations par tranche d'âges d'après le modèle Omphale de l'Insee (source : d'après modèle multimodal régional rhônalpin - Explain)

2 • TYPOLOGIE DES SOURCES DE DONNÉES DE MOBILITÉ

Les grands principes de la modélisation des déplacements étant rappelés, nous nous intéressons maintenant à l'implication que les choix de modélisation vont avoir sur les données de calage à réunir.

Pour qualifier les données de mobilité selon la manière dont elles s'intègrent dans la chaîne de modélisation, nous avons retenu deux principaux axes discriminants :

- le niveau de maîtrise du processus d'échantillonnage ;
- le caractère actif ou passif de la collecte.

Ces deux axes sont détaillés ci-après, ainsi que les quatre types de sources de données qui en résultent.

2.1 L'échantillonnage est-il maîtrisé ? [2]

La première question concerne les possibilités, au moment de la collecte, de maîtriser la représentativité de l'échantillon par rapport à une population-cible.

Un plan de sondage maîtrisé et basé sur un tirage aléatoire permet de limiter et d'estimer les biais liés à l'échantillonnage. Il est alors possible de mettre en œuvre des méthodes de redressement respectant la théorie statistique, basées sur la loi des grands nombres (voir encadré page suivante).

Population-cible, inférence statistique et représentativité

Dans le cadre d'une approche quantitative des phénomènes de mobilité, on cherche à estimer les déplacements réalisés par une population-cible, qu'il est nécessaire de bien définir. S'agit-il des usagers des transports résidents d'un territoire précis, d'une sous-population de ces résidents (étudiants, actifs, etc.), de ceux empruntant une infrastructure précise, d'une sous-population de ces usagers (poids lourds, usagers en transit par rapport à une agglomération, etc.) ?

Les données de mobilité collectées sont rarement exhaustives par rapport à cette population-cible. Une fois la population-cible définie, on peut choisir les données dont l'échantillon et le mieux à même de la représenter dans le cadre d'une procédure d'inférence statistique.

On dit des données qui permettent de reconstituer correctement la population-cible au travers de la procédure d'inférence qu'elles possèdent une bonne représentativité statistique. La représentativité d'un échantillon s'évalue toujours relativement à la grandeur que l'on veut mesurer (indicateur à construire).

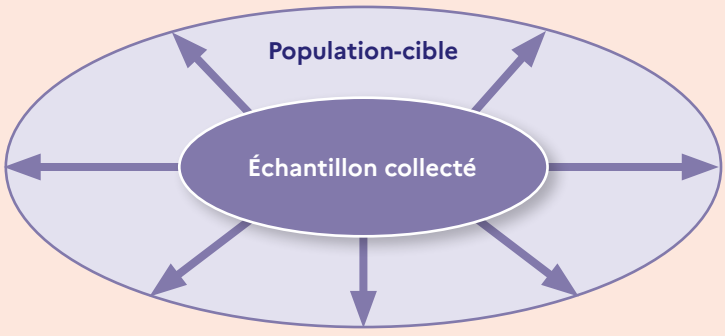


Illustration 12 – Principes de la procédure d'inférence statistique

2 P. Ardilly : http://www.editionstechnip.com/en/catalogue-detail/113/techniques-de-sondage-les.html

Des difficultés peuvent malgré tout se présenter, notamment lorsque la population-cible ne correspond pas exactement au champ de l'enquête. Par exemple, dans les enquêtes auprès des ménages, la population-cible souhaitée correspond à la totalité des résidents d'un territoire. En pratique, les bases de sondage dans lesquelles les échantillons sont tirés n'incluent généralement pas les résidences collectives (résidences universitaires, casernes, maisons de retraite, etc.). Il faut donc en tenir compte si on souhaite étudier plus en détail la population étudiante, dont une part significative est exclue du champ de l'enquête. Une connaissance détaillée du plan de sondage est souhaitable pour qualifier avec précision les usages qui pourront être faits de la source de données et pour pouvoir calculer les précisions des indicateurs. Le tirage est-il aléatoire ? Le plan de sondage est-il stratifié, en grappes, etc. ? Un exemple est donné par l'illustration 13 pour les enquêtes de mobilité certifiées Cerema.

Si on ne maîtrise pas l'échantillonnage, l'échantillon contient généralement les individus qu'on a le plus de facilité à enquêter. Il n'est plus aléatoire. Il faut alors se poser la question suivante : « *La manière dont on collecte les données a-t-elle un impact sur la grandeur que l'on veut mesurer ?* ». Si c'est le cas, l'échantillon présente un biais systématique et on ne pourra plus appliquer la loi des grands nombres.

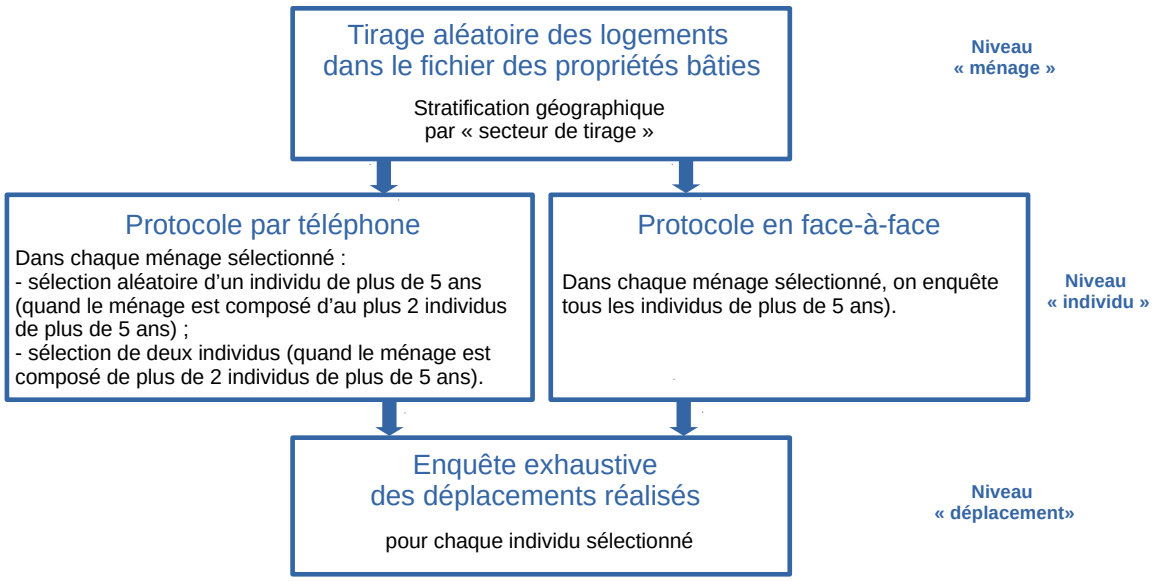


Illustration 13 – Description du plan de sondage des enquêtes de mobilité certifiées Cerema

Loi des grands nombres

La loi des grands nombres indique que la moyenne empirique des valeurs d'une variable sur un échantillon converge vers l'espérance mathématique (moyenne théorique de la loi de probabilité sous-jacente), lorsque la taille de l'échantillon tend vers l'infini (on parle de « propriétés asymptotiques »).

Outre cette propriété de convergence vers l'espérance, de très nombreuses propriétés et formules statistiques découlent de la loi des grands nombres, notamment celles des intervalles de confiance, qui offrent des estimations de la précision des indicateurs calculés. De façon formelle, elle n'est valable que si l'échantillon collecté a été tiré de façon aléatoire.

Les quotas : une méthode pour garantir la variabilité des échantillons, mais pas leur représentativité

La méthode dite « des quotas » assure une variabilité de l'échantillon, en s'assurant *a posteriori* qu'un minimum d'individus de chaque catégorie de population a bien été enquêté. Toutefois, cette méthode ne garantit pas le caractère aléatoire de l'échantillon. En assurant par exemple une répartition par classes d'âges dans l'échantillon conforme aux données Insee du territoire, elle n'empêche pas que d'autres biais apparaissent : par exemple, que les personnes les moins mobiles restent plus faciles à joindre que les autres dans le cadre d'une enquête par téléphone.

La loi des grands nombres ne s'applique pas en général à des données obtenues par une méthode d'échantillonnage par quotas.

2.2 La collecte est-elle active ou passive ?

Si la collecte est réalisée via un questionnaire, on considère qu'elle est active, que l'enquête soit auto-administrée ou administrée par un enquêteur. Une collecte passive est quant à elle réalisée après accord (plus ou moins explicite) de l'enquêté, mais avec peu ou pas d'intervention de sa part (suivi GPS, données de communication avec les antennes de téléphonie mobile, etc.). Il est bien sûr possible de combiner les deux approches (par ex. Enquêtes de mobilité certifiées Cerema, avec dispositif optionnel de suivi GPS sur une semaine).

Le caractère actif de la collecte offre, via son questionnaire, des possibilités de collecter des variables sur les individus (et les véhicules) et une précision dans les réponses que n'offrent pas les collectes passives. En revanche, il introduit une contrainte qui peut être forte sur l'usager, en particulier lorsque la collecte est réalisée en cours de déplacement. Les temps de passation des questionnaires (donc le nombre et la précision des questions) doivent souvent être réduits pour améliorer leur acceptabilité.

Les collectes passives offrent, par leur caractère discret pour l'usager, la possibilité de le suivre pendant des périodes de temps plus longues. En revanche, l'enrichissement de la donnée de traces de mobilité (par des informations sur l'individu, sur son foyer, sur les modes de transport utilisés, sur les motifs de déplacement…) est faible voire nul, soit par manque de disponibilité de l'information, soit par nécessité de respect de la vie privée. En raison du caractère continu de la collecte, malgré leur faible enrichissement, les collectes passives peuvent permettre d'identifier des personnes et sont donc alors soumises aux contraintes de confidentialité relatives aux données personnelles (RGPD)[3], tout comme les collectes actives.

2.3 Quatre types de sources de données

Du croisement de ces deux axes discriminants découle une typologie en quatre groupes, décrite dans l'illustration 14. Le groupe de rattachement d'une source de données donne une première indication sur la manière dont elle pourra être intégrée dans le processus de modélisation.

	Collecte active	Collecte passive
Bonne maîtrise de l'échantillonnage	**1** Échantillons représentatifs d'une population, obtenus par sondage aléatoire. Recensements de la population Enquêtes OD EMC² Panel parc-auto	**3** Observations tendant vers l'exhaustivité du flux en un point du réseau. Comptages ponctuels Billettique Transactions péage Lecture de plaques minéralogiques
Faible maîtrise de l'échantillonnage	**2** Recherche d'une variabilité de l'échantillon, sans garantie de représentativité. Enquête préférences déclarées	**4** Observations qui peuvent être nombreuses, mais dont on ne maîtrise pas qualitativement l'échantillon. Bluetooth, wifi, FCD, FMD Données GPS d'applications smartphone

Illustration 14 – Typologie des données de mobilité

Collectes actives avec bonne maîtrise de l'échantillonnage

Ces collectes sont les plus classiquement utilisées dans les modèles de déplacements. Il s'agit d'enquêtes par sondage aléatoire, dont l'échantillonnage est maîtrisé et sur lesquelles la théorie statistique s'applique, grâce à la loi des grands nombres. Elles ont été conçues selon un processus maîtrisé, pour répondre à une ou plusieurs questions précises (voir tableau ci-après pour quelques exemples).

3 Il y a un travail particulier de la Cnil selon les données : pour les téléphones https://www.cnil.fr/en/node/24869, pour les véhicules connectés https://www.cnil.fr/fr/vehicules-connectes-un-pack-de-conformite-pour-une-utilisation-responsable-des-donnees par exemple.

Protocoles d'enquête	Principales variables de mobilité ciblées
Enquêtes de mobilité certifiées Cerema (EMC²) réalisées auprès des ménages	Taux de mobilité (nombre de déplacements par personne et par jour), parts modales, distances parcourues
Enquêtes OD bord de route	Origines et destinations des flux, taux d'occupation des véhicules légers, types de marchandises transportées par les poids lourds.
Recensement de la population	Navettes domicile-travail et domicile-études

Les variables ciblées par les protocoles d'enquêtes de mobilité les plus couramment utilisés (EMC², enquêtes OD bord de route, dans les transports collectifs, enquêtes marchandises en ville, etc.) ont été calées sur les besoins des modèles de déplacements. De plus, elles se prêtent naturellement au dénombrement (loi des grands nombres) et à la détection de corrélations grâce à la présence de variables individuelles. C'est pourquoi, malgré leur coût et leur temps de mise en œuvre relativement élevés par rapport aux autres sources, ces enquêtes restent des données de référence pour la connaissance et la modélisation des mobilités.

Les possibilités de détection de corrélations individuelles sont d'autant plus fortes que le questionnaire est riche, ce qui est le cas du recensement de la population et des EMC². L'approche d'échantillonnage est parfois zonale, portant sur les ménages (recensement, EMC²) et parfois par axe, portant sur les véhicules (enquêtes OD).

Ce type de données est principalement utilisé pour la modélisation de la demande de déplacements (génération, distribution et choix modal). On peut également rattacher à cette catégorie les données issues de panels, qui suivent l'évolution des comportements d'un ensemble d'individus au cours du temps. Actuellement, ces données sont peu utilisées en modélisation. Les mesures d'évolution des comportements qu'elles fournissent peuvent toutefois être utilisées pour vérifier la transférabilité temporelle des modèles, c'est-à-dire vérifier que le calibrage effectué a de bonnes chances de rester pertinent en situation future.

Collectes actives avec faible maîtrise de l'échantillonnage

Comme présenté au paragraphe 1.2, pour construire des modèles désagrégés sur les individus, il n'est pas nécessaire que les données de mobilité utilisées pour le calage soient représentatives de la population-cible. En revanche, il est nécessaire qu'elles disposent d'une variabilité minimale pour décrire tous les types de comportements que l'on cherche à modéliser. Pour cela, on utilise fréquemment des **méthodes d'échantillonnage par quotas, non aléatoires**. Ceci permet de réduire les coûts liés à la construction d'une base de sondage exhaustive et aux relances des non-répondants. Parmi ces collectes, on trouve fréquemment des « enquêtes de préférences déclarées » (voir encadré ci-après).

Modèles désagrégés, préférences révélées et préférences déclarées

Les modèles désagrégés peuvent être alimentés par des enquêtes dites de « préférences révélées » (les enquêtes OD en bord de route sont les plus utilisées pour les modèles de trafic interurbains) qui révèlent *a posteriori* les choix effectivement réalisés par l'usager, lorsqu'il est confronté à plusieurs alternatives d'offre de transport. En complément, les modèles utilisent parfois des enquêtes dites de « préférences déclarées », qui révèlent *a priori* les comportements des usagers des transports face à des options hypothétiques d'offre de transport.

Les enquêtes de préférences déclarées et de préférences révélées renseignent sur l'alternative choisie, mais également sur les alternatives laissées de côté par l'usager.

En pratique, on réalise rarement des « enquêtes de préférences révélées », les enquêtes de mobilité auprès des ménages (EMC²) étant utilisées en substitution (bien qu'elles ne présentent pas les alternatives laissées de côté par l'usager).

En revanche, on réalise fréquemment des « enquêtes de préférences déclarées » pour tester des alternatives non disponibles en situation actuelle, en particulier des alternatives de choix modal. Étant utilisées uniquement dans le cadre de modélisations désagrégées, elles sont généralement conçues selon un échantillonnage non aléatoire, par une méthode de quotas. Elles rentrent donc dans cette deuxième catégorie de sources.

L'étape d'inférence est réalisée dans un second temps, par application du modèle désagrégé à une population simulée, dite « **population synthétique** ». C'est la population synthétique qui garantit la représentativité finale.

Collectes passives avec une bonne maîtrise de l'échantillonnage

En pratique, les collectes passives avec une bonne représentativité sont aujourd'hui généralement des **collectes exhaustives** (comptages routiers, transactions aux péages) ou **quasi exhaustives** (billettique, qui est entachée par la fraude, lecture automatique de plaques d'immatriculation, qui présente des échecs de lecture). Dans le cas des collectes quasi exhaustives, il convient, comme pour les sources échantillonnées, de bien connaître le défaut de couverture et d'en tenir compte dans les redressements qui pourront être réalisés des données (cf. Illustration 15, qui montre la nécessité de prendre en considération l'âge dans le redressement des données billettiques).

Il est toutefois envisageable de produire des données échantillonnées qui rentreraient dans cette catégorie, par exemple grâce à un protocole d'enquête ménages-déplacements avec un suivi GPS systématique de tout l'échantillon (sous réserve de l'acceptabilité par tous les enquêtés). Lorsqu'elles ne sont pas couplées à une collecte active, les collectes passives sont généralement **pauvres en informations individuelles**, d'une part parce que le caractère automatique de la collecte ne permet pas l'enrichissement et d'autre part pour respecter la vie privée des usagers.

Ces données sont en revanche **particulièrement pertinentes pour le dénombrement de flux, de personnes ou de véhicules** et constituent, à ce titre, des données de choix pour le calage des modèles de déplacements, en particulier pour l'étape d'affectation, sous réserve que l'on sache évaluer correctement les défauts de couverture (fraude, échecs de lecture, etc.).

Collectes passives avec une faible maîtrise de l'échantillonnage

Dans cette catégorie, les volumes de données et le nombre d'unités observées peuvent être élevés, mais leur répartition dans la population n'est jamais garantie, en raison du protocole qui ne permet pas de toucher des personnes qui auraient été ciblées à l'avance par un échantillonnage aléatoire.

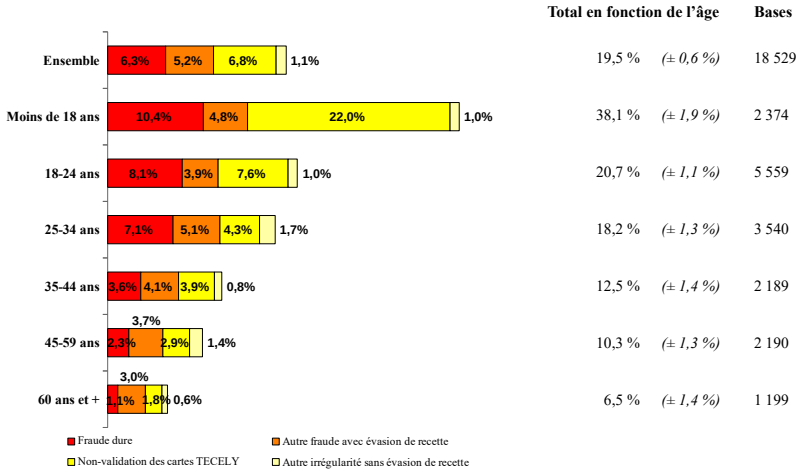


Illustration 15 – Taux de fraude par tranche d'âges sur le réseau des Transports en commun lyonnais (TCL), (source enquête fraude et [14])

Données massives ?

Bien qu'il n'existe pas de définition totalement partagée des « **données massives** », la plupart des auteurs (cf. Wikipedia) font référence à cette expression pour désigner une source de données **volumineuses, variées et avec une mise à jour rapide** (donc une collecte passive). Dans le domaine de la mobilité, il s'agit principalement de données **générées par des services commerciaux numériques** (donc sans maîtrise de l'échantillon). Les « données massives de mobilité » appartiennent donc bien à cette 4e catégorie de sources pour laquelle la collecte est passive et l'échantillonnage n'est pas maîtrisé.

Dans la suite du document, nous utiliserons donc le terme « données massives » pour désigner les données de mobilité appartenant à cette 4e catégorie.

La répartition se fait par les usages. Via une application smartphone, le jeu de données fourni est orienté vers les utilisateurs de l'application, qui sont rarement représentatifs de la population-cible. Si la population-cible correspond à la population totale, quelle application peut aujourd'hui prétendre être autant utilisée par les plus jeunes et les plus âgés, par les hommes et les femmes, par les actifs et les inactifs, par chacune des catégories socioprofessionnelles, etc. ?

Dans le cas où on souhaite réaliser une inférence statistique agrégée, il faut donc déterminer si l'imprécision est acceptable au vu de la population-cible et de la variable à estimer. Par exemple, pour estimer les flux réalisés par la population totale à partir de *Floating Mobile Data* (FMD), dont on sait qu'elles sous-estiment la part de personnes âgées, il faut estimer la part des personnes âgées dans la population totale du territoire. Si elle est importante, les chiffres issus des FMD risquent d'être faussés, car cette catégorie d'usagers est mal représentée (cf. Illustration 16). *A minima*, il faudra donc pour corriger ce biais envisager un redressement des données sur les tranches d'âges.

Les informations individuelles sur chaque unité de l'échantillon sont limitées par l'absence d'un questionnaire. On ne dispose que des données de gestion du service commercial dont les données sont issues.

Certains algorithmes peuvent permettre de reconstituer des informations *a posteriori*, comme le lieu du domicile, lorsqu'on dispose de traces sur une période couvrant plusieurs jours. Les informations utilisables sont toujours limitées par la protection de la vie privée.

En revanche, s'agissant de données collectées de façon automatique, en continu et sans gêne pour l'usager, il est beaucoup **plus facile et moins coûteux qu'avec les autres types de sources de disposer de données sur de longues périodes**, donc éventuellement d'établir des **corrélations entre les comportements de mobilité d'un même individu sur plusieurs jours**.

Cette catégorie de données présente donc des **possibilités limitées pour le dénombrement, comme pour la détection de corrélations individuelles.** Leur utilisation en modélisation des déplacements n'est donc pas naturelle et nécessite toujours un travail approfondi de compréhension de la source, de préparation des données (redressement notamment) et de comparaison à des sources plus classiques, que l'on maîtrise mieux.

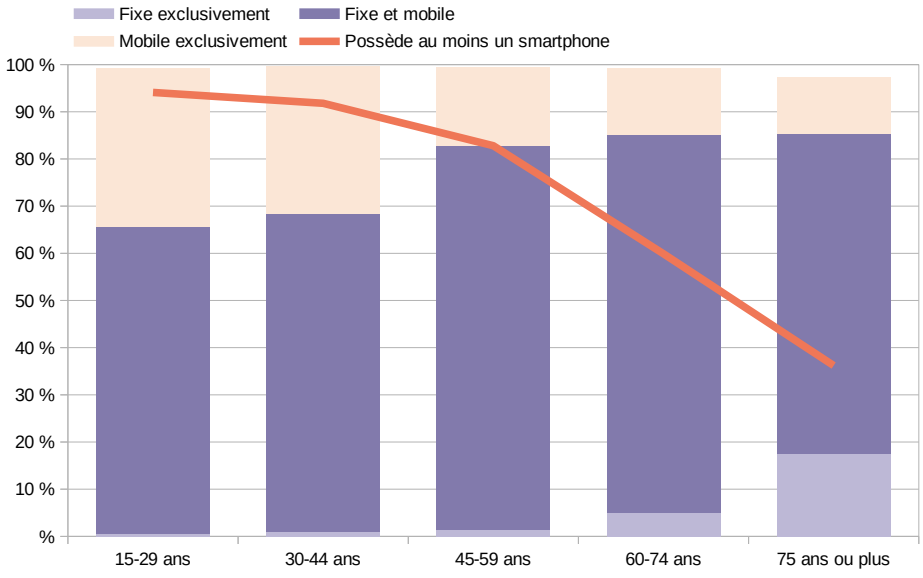


Illustration 16 – Taux d'équipement téléphonique selon l'âge en 2021. Champ : France hors Mayotte, personnes de 15 ans ou plus vivant en logement ordinaire. Source : Insee, enquête TIC ménages 2021.

3 • CRITÈRES POUR L'ANALYSE DE PERTINENCE D'UNE SOURCE DE DONNÉES DE MOBILITÉ

La typologie présentée au paragraphe 2. donne une première idée de la manière dont une source de données de mobilité va pouvoir être utilisée dans la modélisation. Toutefois, pour utiliser une source de données de façon opérationnelle et pertinente, il est nécessaire d'aller plus loin en analysant plus en détail ses caractéristiques. Nous proposons ci-dessous une grille d'analyse, sous forme d'une série de questions, du domaine de pertinence de chaque source de données qui peut être mobilisée dans le cadre de la construction d'un modèle. Elle a été déployée dans les fiches associées à cet ouvrage, qui décrivent les propriétés des différentes sources de données utilisées en modélisation. Pour la construire, nous nous sommes notamment inspirés du code de bonnes pratiques de la statistique européenne (cf. [**9**]).

Cette grille d'analyse doit ensuite être déclinée encore plus précisément pour chaque source et chaque fournisseur de données, avant d'envisager son utilisation.

La pertinence d'une source de données doit toujours être analysée également en fonction des objectifs des études à réaliser à partir du modèle, des principes de modélisation retenus, ainsi que des autres données disponibles.

3.1 Méthode de collecte

S'intéresser à la manière dont les données ont été collectées permet de comprendre quelles informations sont disponibles et quels en sont les droits de réutilisation, ainsi que d'appréhender la qualité finale de la source et sa capacité à offrir un suivi dans le temps.

Qui produit ces données ?

Le producteur de la donnée peut avoir un impact sur la stabilité spatio-temporelle des méthodes de collecte. Le producteur fait-il l'objet d'une certification sur la qualité de ses données ? Répond-il à des protocoles standardisés ? Publie-t-il sa méthode de collecte et de retraitement des données ? S'agit-il d'un organisme pérenne ou d'une structure qui peut du jour au lendemain cesser la mise à disposition de ses données ?

En fonction de ces éléments, on pourra notamment déterminer si des comparaisons sont possibles entre plusieurs jeux de données émanant de ce producteur, dans l'espace et dans le temps.

Quels sont les protocoles de collecte ?

Les protocoles de collecte influent sur le temps disponible pour interroger chaque usager, donc sur la richesse possible du questionnaire. Ils influent également sur la capacité à maîtriser l'échantillonnage. Le détail du protocole de collecte permet également d'identifier, le cas échéant, les biais systématiques de l'échantillon.

Par exemple, une enquête origine-destination à bord des trains aura tendance à sous-estimer les OD courtes, le temps de parcours de l'usager étant alors trop réduit pour la passation complète du questionnaire.

Quelles sont les unités observées ?

Quelle est l'unité de base d'observation des données : un véhicule, un ménage, un individu, une chaîne de déplacements ou un déplacement ? L'échantillonnage peut comporter plusieurs niveaux : par exemple, d'abord les ménages, ensuite les individus puis les déplacements (cf. exemple des enquêtes de mobilité certifiées Cerema à l'illustration 13).

Comment est constitué l'échantillon ?

La représentativité de l'échantillon par rapport à la population-cible ne peut être estimée que si on connaît en détail la méthode de constitution de l'échantillon. Si la donnée est échantillonnée (non exhaustive) et que des biais systématiques sont identifiés dans le processus d'échantillonnage (ex. faible taux d'équipement

des personnes âgées en téléphones mobiles pour les FMD), la généralisation à l'ensemble de la population-cible sera imprécise voire impossible.
Il faut également s'interroger sur la non-réponse. Dans le cadre des collectes actives, elle correspond à un refus explicite ou à une impossibilité de joindre la personne. Dans les collectes passives, elle peut être liée à un refus de l'usager d'être tracé ou à une difficulté de mesure (plaque d'immatriculation illisible dans le cas d'une collecte par lecture vidéo par exemple).

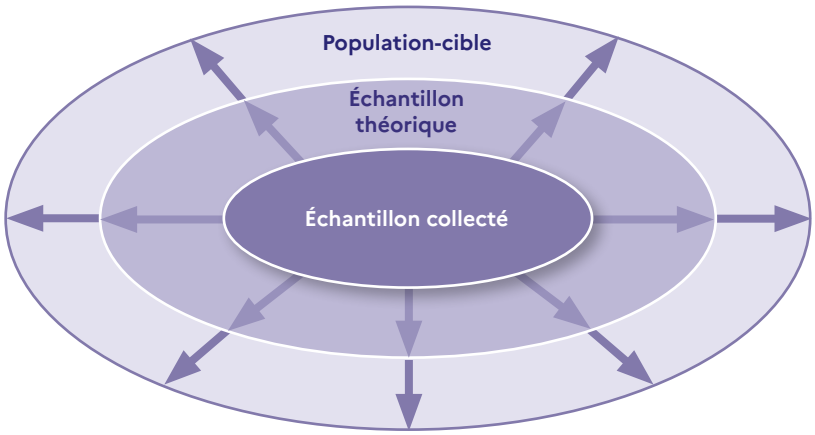


Illustration 17 – La non-réponse induit souvent une différence entre l'échantillon collecté et l'échantillon théorique.

Sur quelle période les données ont-elles été collectées ?

Quel type de période est évalué par la source de données (jours de la semaine, saison particulière) ? Les données sont-elles collectées en continu ou de manière ponctuelle et à quelle fréquence ?

Les données massives présentent un avantage sur ce plan, avec des collectes possibles en continu, sans gêne à l'usager. Elles autorisent des comparaisons de comportements entre saisons ou périodes de la journée, à échantillons comparables. Lorsque l'intervalle de temps dépasse un an, les données massives ne permettent généralement plus de garantir la stabilité des méthodes et il faut alors se tourner vers les autres types de données.

Il est également nécessaire de s'interroger sur l'actualité des données par rapport à la situation de calage du modèle et aux risques associés. Un petit écart temporel pourra parfois être corrigé par une procédure d' « actualisation » du redressement, sur des données de référence plus récentes (population, comptages). Lorsque l'écart devient trop important, le risque de divergence avec les autres sources de données devient fort.

Comment la vie privée et le secret commercial sont-ils protégés ?

Lorsqu'on est responsable de la collecte, il est nécessaire de s'interroger en amont sur les contraintes imposées sur la collecte par la Commission nationale informatique et libertés (CNIL) et le Règlement général de protection des données (RGPD). Elles limitent les caractéristiques individuelles qui peuvent être recueillies. Par exemple, les plaques minéralogiques collectées dans leur totalité constituent un identifiant personnel. Elles peuvent également nécessiter un floutage *a posteriori*, si certaines informations de mobilité permettent d'identifier un individu par recoupements.

Lors de l'achat ou de la réutilisation de sources produites par d'autres, il est également nécessaire de se renseigner sur la manière dont le producteur a anonymisé les données, car ce processus peut conditionner la qualité de la donnée finale. En complément, des contraintes relatives au respect du secret commercial et des affaires peuvent également intervenir dans certains cas.

À titre d'exemple, en France, les règles du secret statistique, qui doivent être respectées par la statistique publique, sont résumées dans le tableau ci-après (cf. [**10**] pour plus de détails).

	Fichiers agrégés	Fichiers de données individuelles
Entreprises	Jamais moins de trois unités par case d'un tableau, jamais une case du tableau où une entreprise représente plus de 85 % du total de la case.	Pas d'accès, sauf dérogation après passage par le comité du secret statistique (organe du comité national de l'information statistique)
Individus	Il ne doit pas être possible d'identifier une personne. Pour le recensement de la population, cette règle a été déclinée de façon plus précise dans l'arrêté du 19 juillet 2007.	Il ne doit pas être possible d'identifier une personne. Pour cela, on retire les variables les plus susceptibles de mener à l'identification (commune de résidence, profession détaillée, etc.). La liste des variables à retirer est adaptée au cas par cas (cf. [11] pour des exemples de méthodes).

Illustration 18 – Les règles du secret statistique à respecter

3.2 Informations fournies par les données

Ce deuxième paragraphe s'intéresse aux données collectées et aux indicateurs qu'il est possible de construire à partir de ces données.

Quelles sont les principales variables de mobilité observées ?

La source de données peut fournir différentes informations : des OD, des débits ponctuels, des données individuelles de comportements de mobilité (chaînes de déplacements avec motifs, etc.). Chacune étant utilisée à un moment précis dans le processus de modélisation, la connaissance des variables de mobilité mesurées permettra de savoir comment la source peut être intégrée à la chaîne de modélisation.

Quelles sont les informations disponibles sur les unités observées ?

La source de données est-elle enrichie de données socio-démographiques à l'échelle du ménage ou de l'individu ? Ces informations seront nécessaires aux modèles pour faire des corrélations entre les caractéristiques de l'individu ou du ménage et les comportements de mobilité. Elles seront également utilisées pour redresser les données échantillonnées.

Les déplacements sont-ils enrichis d'informations sur les motifs et les modes de transport utilisés ? La distinction des motifs de déplacement est un élément important pour l'analyse des corrélations entre variables.

Quelle est la précision de localisation des origines et des destinations ?

La finesse spatiale de la donnée est liée à la précision du dispositif de mesure ou à la précision de localisation des OD demandée dans l'enquête, toujours contrainte par le temps de passation du questionnaire et par le respect de la vie privée.

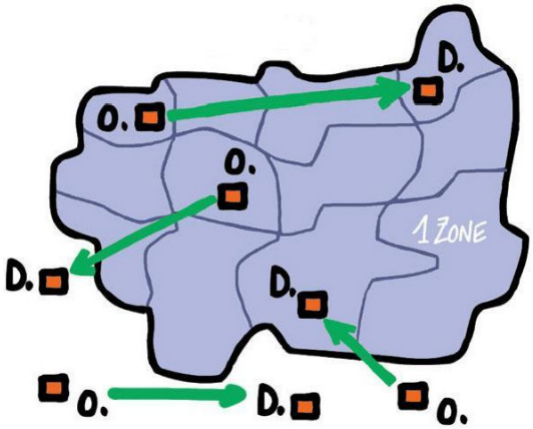


Illustration 19 – Imbrication du zonage de tirage de l'échantillon – trait noir épais – et des zones fines de repérage des origines et destinations – trait mauve fin – dans le dispositif des enquêtes de mobilité certifiées Cerema (Source : Cerema - dessin H. Baudry)

Illustration 20 – Découpage de la région Grand Est par des zones de détection de la téléphonie mobile. Les limites théoriques, basées sur un découpage communal, sont représentées en noir. Le zonage de couverture des télécommunications de l'opérateur est représenté sous la forme d'une palette de couleur (Source : Opérateur)

La précision spatiale de localisation des OD n'est pas forcément similaire au niveau d'agrégation spatiale auquel on peut se permettre d'analyser la donnée.

Par exemple, dans les enquêtes de mobilité certifiées Cerema, le niveau des zones fines est utilisé pour la localisation des OD et donc le calcul des distances parcourues, mais les analyses statistiques ne se font jamais à un niveau spatial plus fin que le secteur de tirage, qui correspond à un regroupement de zones fines (cf. Illustration 18). Ceci permet de respecter les stratifications spatiales de l'échantillon et de garantir des échantillons suffisants pour construire les indicateurs souhaités, en particulier les volumes par OD qui sont utilisés pour caler le modèle. Le modèle permet ensuite en général un affinement spatial des données de flux par OD (rôle d'extrapolation spatiale abordé au paragraphe 1.2).

On a une problématique similaire avec les données issues des antennes de la téléphonie mobile, pour lesquelles les OD sont localisées selon la densité des antennes de télécommunication (cf. Illustration 19), mais l'analyse se fait toujours à un niveau d'agrégation plus élevé pour avoir un échantillon suffisant et pour limiter les biais liés à des contraintes géographiques locales.

Quels sont les principaux biais connus ?

La connaissance des biais permet d'interpréter au mieux les exploitations des données et d'éviter de sortir de leur domaine de pertinence.

Parmi les biais que l'on peut rencontrer, figurent par exemple la sous-estimation systématique des courtes distances pour les enquêtes OD, des motifs récurrents pour les données FCD, du nombre d'hyper-mobiles dans les enquêtes ménages-déplacements par téléphone, etc.

3.3 Traitement des données

Comment généralise-t-on les données à la population-cible ?

La généralisation d'un échantillon à une population-cible correspond à l'opération de **redressement des données** (cf. Illustration 17).

Le redressement ne doit pas être une boîte noire. L'enjeu de cette question est de comprendre en détail sur quelles variables il a été effectué et comment ces variables ont été choisies, en particulier lorsque ce n'est pas le modélisateur qui effectue ce redressement. En effet, ce sont l'utilisation de variables de redressement bien corrélées à la variable à expliquer et la faible dispersion des poids de redressement qui assurent la représentativité finale des données. Il est donc utile de demander au fournisseur de la donnée de fournir des analyses de corrélations et de dispersion des poids de redressement, y compris lorsque les données sont fournies de façon agrégée.

Peut-on mesurer des évolutions à partir de cette source ?

La comparabilité temporelle des résultats issus d'une source de données est un critère de qualité pour mesurer des évolutions ou assurer une certaine stabilité des équations et des paramètres de modélisation entre deux mises à jour. Cette question est en lien avec la méthode de collecte et le type de producteur de la donnée.

Comme le soulignent notamment l'Insee et le Cnis (cf. [**12**] et [**13**]), la pérennité des sources de données massives est généralement fragile et cela limite leurs utilisations dans le cadre des processus de statistiques publiques.

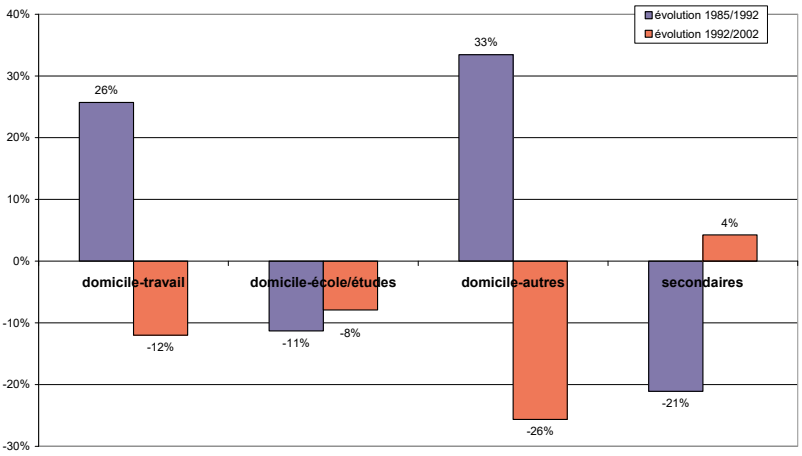


Illustration 21 – Évolution des motifs de déplacement à vélo entre les trois enquêtes ménages-déplacements de 1985, 1992 et 2002 à Grenoble, source [15].

Comment les données s'intègrent-elles dans les modèles ?

Lors du choix d'une source de données, il est intéressant de se poser quelques questions sur le processus de modélisation en lui-même et l'adéquation entre la source et le modèle. Quelle est la pertinence de la donnée par rapport à ce que le modèle cherche à estimer ? Existe-t-il des précédents d'intégration de ce type de données dans les modèles ? Si oui, comment l'intégration a-t-elle été réalisée (pour quelle étape de modélisation, avec quel retour d'expérience) ?

Les questions précédentes sont d'autant plus prégnantes lorsque l'on cherche à intégrer dans un modèle des données massives de mobilité, issues d'objets connectés. Ces données, collectées passivement et sans maîtrise de l'échantillonnage, interrogent les pratiques de modélisation. En effet, elles souffrent à la fois d'un manque de garantie sur leur représentativité et d'un faible enrichissement en variables individuelles. Les utiliser pour caler un modèle est donc toujours une opération complexe, que le modèle soit agrégé ou désagrégé (cf. paragraphe 1.2). Pourtant, leurs capacités de suivi des comportements sans gêne pour l'usager et en continu sont des atouts pour comprendre les phénomènes de mobilité.

Nous détaillons ci-après les quatre principaux cas d'usage des données massives de mobilité que nous avons recensés.

a) Mesure des comportements d'une population-cible bien couverte

Les données massives peuvent fournir des informations sur les comportements de sous-populations qui seraient correctement couvertes par l'échantillon : par exemple les étudiants d'un campus, à partir des données issues d'une application smartphone spécifique très utilisée sur ce campus. Une étude préalable doit permettre de démontrer que la population couverte par la source de données tend vers l'exhaustivité ou ne présente pas de biais trop important sur les principales variables socio-démographiques connues pour avoir une influence sur le comportement de mobilité (âge, sexe, composition du ménage, motorisation, etc.). Un redressement des données doit toujours être effectué.

b) Mesure des comportements à une échelle spatiale très agrégée afin de limiter les biais

Diminuer la précision spatiale des données peut permettre de limiter les biais par rapport à une population-cible. C'est par exemple le cas pour les données FMD, dont on constate qu'en général, plus elles sont considérées sur un zonage agrégé, moins elles sont biaisées. Ce type d'usage donne des ordres de grandeur des grands flux sur un territoire, notamment sur des territoires peu denses qui ne seraient pas équipés d'autres dispositifs d'observation.

c) Mesure d'évolutions à court et moyen termes

Ces données peuvent fournir des éléments de suivi temporel (sous réserve de la stabilité de la méthode de collecte, donc sur un an ou deux au maximum[4]). Les mesures relatives d'évolutions seront en effet moins entachées d'erreurs que des mesures absolues à un instant *t*. Les données massives peuvent offrir une fréquence de mesure supérieure aux collectes actives, d'une part en raison de la moindre gêne à l'usager qui permet de renouveler plus régulièrement la collecte, et d'autre part parce qu'on peut en général s'attendre à un coût moindre. La question du coût reste toutefois à vérifier au cas par cas, car elle est très dépendante du fournisseur, du territoire et de la nature des données demandées. Les coûts d'appropriation et de retraitement de la donnée doivent également être considérés dans l'estimation.

Les collectes actives offrent une maîtrise de l'échantillon et des informations individuelles riches. La récupération de traces de mobilité sur des périodes longues (par exemple une semaine) pour les usagers enquêtés dans le cadre d'une collecte active peut apporter un complément intéressant à l'enquête principale, en permettant d'étudier les récurrences de comportement sur plusieurs jours, avec une gêne mineure de la personne enquêtée qui améliore l'acceptabilité du dispositif. De cette manière, les traces de mobilité collectées sont représentatives et sont associées à des informations individuelles riches, issues de l'enquête.

d) Remplacement d'une collecte physiquement impossible ou dangereuse

L'usage des données massives de mobilité peut également être envisagé en complément des sources classiques, dans des cas d'impossibilité physique de collecte. Par exemple, certains postes d'enquêtes origines-destinations en bord de route peuvent présenter des risques en termes de sécurité routière et de création de congestion tels qu'on préfère leur substituer des *Floating Car*

4 En raison de l'évolution rapide des méthodes et des réseaux de collecte.

Data (FCD), dont on connaît pourtant les limites pour estimer des OD. Dans ces cas particuliers, il faudra toujours doubler les données massives par d'autres dispositifs d'observation dont on maîtrise mieux la représentativité et qui permettront de les redresser.

Pour redresser des FCD afin de remplacer un poste OD bord de route, il faudra par exemple réaliser en parallèle des collectes sur le même site (comptages routiers, relevés minéralogiques), qui fourniront des volumes et éventuellement des postes OD sur des voiries comparables, qui fourniront des distributions de distances parcourues de référence.

3.4 Coûts

Il est difficile de comparer les coûts de différentes sources de données dans l'absolu, chacune ayant des points forts et des points faibles différents. L'analyse de coûts doit donc, dans la mesure du possible, être faite à l'échelle de l'ensemble des données utilisées pour construire le modèle.

Pour les données massives, les coûts varient selon les fournisseurs. On constate que dans certains cas, les fournisseurs alignent leurs prix sur ceux de collectes terrain équivalentes : lorsqu'on achète des données FCD pour remplacer un poste d'enquête OD, on observe la plupart du temps des coûts à peu près équivalents à ceux d'un poste réalisé sur le terrain. Le prix est d'autant plus élevé que le commanditaire exige des traitements précis, notamment de redressement ou de filtrage des traces à utiliser. Les coûts de retraitement des données après livraison ne doivent pas être négligés dans l'analyse.

***Nota** : Les fiches techniques associées à cet ouvrage ne traitent pas de la question des coûts, en raison de son caractère fortement évolutif dans le temps et selon la nature des prestations demandées.*

4 • CONCLUSION

La modélisation des déplacements est un processus qui doit toujours s'adapter aux questions posées, aux données disponibles, ainsi qu'au temps et au budget que le commanditaire décide de dédier à la démarche.

Les questions posées à la modélisation évoluent sans cesse. D'outils principalement mobilisés pour évaluer des grands projets d'infrastructures routières dans les années 1970, les modèles sont aujourd'hui sollicités pour répondre à des questionnements beaucoup plus divers. Des questionnements sont apparus ces dernières années par exemple sur l'optimisation du positionnement de bornes de recharge des véhicules électriques, l'évaluation de la pertinence de nouveaux aménagements cyclables et de leur impact prévisible sur le partage modal et sur les distances parcourues, l'impact des flux touristiques sur le fonctionnement d'un réseau en période d'affluence, ou encore l'effet d'une pandémie et des différentes mesures de restriction des déplacements associées sur les mobilités et le partage modal. Les questionnements qui apparaissent ont tendance à nécessiter des approches plus individualisées des comportements (allant donc vers des modèles de plus en plus désagrégés sur les individus, sur la représentation des chaînes d'activités et sur le zonage) et des visions plus dynamiques des flux (dans la journée, dans la semaine, sur les différentes saisons de l'année, sur quelques années).

En parallèle, sont apparues des données massives décrivant les mobilités, issues des objets connectés. Comme l'a montré ce document, ces données présentent en général des limites en termes de représentativité et de richesse des informations fournies. Toutefois, elles représentent aussi une opportunité pour répondre à certaines questions pour lesquelles les sources traditionnelles ne sont pas adaptées ou sont difficiles à mettre en œuvre, en particulier :

- lorsqu'on souhaite mesurer des évolutions des mobilités à court ou moyen terme, c'est-à-dire à des horizons allant de quelques jours à quelques années ;
- lorsqu'une collecte est impossible ou dangereuse sur le terrain, en particulier pour remplacer certains postes d'enquêtes OD routières ;
- pour mesurer les comportements d'une population spécifique que l'on sait bien couverte par une source massive et moins bien par les sources traditionnelles (par exemple les étudiants d'un campus, si on peut disposer des traces remontées par une ou plusieurs applications smartphone très utilisées par cette population) ;
- pour offrir une vision à une échelle spatiale relativement agrégée (nationale et régionale notamment) des flux, en particulier des flux routiers, difficiles à capter de façon suffisamment exhaustive par des enquêtes sur site, en raison du caractère très maillé du réseau.

Sur ces sujets, les données massives apportent un réel supplément d'informations par rapport aux sources traditionnelles. Elles doivent toutefois toujours être utilisées après redressement, c'est-à-dire après pondération par un facteur correctif pour compenser leurs problèmes de représentativité. Dans certains cas, il est même nécessaire d'aller vers des procédures de fusion de données, capables d'agréger plusieurs sources massives entre elles ou des sources massives avec des sources traditionnelles, en tenant compte des incertitudes associées à chacune d'elles. Les méthodes de fusion à utiliser sont encore expérimentales voire du domaine de la recherche.

Les sources traditionnelles restent donc nécessaires pour servir de référence, pour permettre la correction et la validation des données massives, mais également pour apporter de la richesse sémantique aux données, avec des caractéristiques des individus, des véhicules, des ménages ou des déplacements (modes et motifs) que les sources massives ne savent pas ou mal mesurer aujourd'hui.

Les sources massives et les sources traditionnelles apparaissent donc comme très complémentaires. Les défis à venir de la modélisation des déplacements consistent à intégrer toutes les données, pour répondre au mieux aux nouvelles questions qui se posent aux planificateurs et aux gestionnaires de réseaux, notamment à la nécessaire prise en compte de l'incertitude sur l'évolution des comportements comme l'a montré le contexte de crise sanitaire.

Bibliographie

[1] Bousquet A., et Caubel D. (juin 2015). ***Mesurer l'accessibilité multimodale des territoires – État des lieux et analyse des pratiques***, Bron : Cerema, Connaissances, https://www.cerema.fr/fr/centre-ressources/boutique/mesurer-accessibilite-multimodale-territoires

[2] Breiman L. (2001). ***Statistical Modeling: The Two Cultures Statistical science***, vol. 16, n° 3, p. 33, http://www2.math.uu.se/~thulin/mm/breiman.pdf

[3] Saporta G. (2017). ***Expliquer ou prédire ? Les nouveaux défis***, p. 49, https://hal-cnam.archives-ouvertes.fr/hal-02507348

[4] Certu (mars 2003). ***Modélisation des déplacements urbains de voyageurs - Guide des pratiques***, Lyon : Certu, Références, https://www.cerema.fr/fr/centre-ressources/boutique/modelisation-deplacements-urbains-voyageurs

[5] Cerema (octobre 2015). ***Modélisation multimodale des déplacements de voyageurs - Concevoir un modèle de choix modal***, Bron : Cerema, Références, https://www.cerema.fr/fr/centre-ressources/boutique/concevoir-modele-choix-modal

[6] Setra (juillet 2010). ***Calage et validation des modèles de trafic : techniques appliquées à l'affectation routière interurbaine***, https://www.cerema.fr/fr/centre-ressources/boutique/calage-validation-modeles-trafic

[7] Bonnel P. (décembre 2001). ***Prévision de la demande de transport***, Laboratoire d'économie des transports - UMR ENTPE, université Lyon 2, CNRS, https://www.presses-des-ponts.fr/notre-librairie/228-prevoir-la-demande-de-transport.html

[8] Dios Ortúzar J. De, et Willumsen L. G. (2011). ***Modelling transport***, 4. ed., Chichester, Wiley.

[9] Eurostat (16 novembre 2017). ***Code de bonnes pratiques de la statistique européenne - À l'intention des autorités nationales de la statistique et d'Eurostat (autorité statistique de l'Union européenne)***, Eurostat, https://www.insee.fr/fr/statistiques/fichier/4140105/Code_Bonnes_Pratiques_Stat_Euro_nov2017.pdf

[10] Insee (2020). ***Guide du secret statistique***, Insee, https://www.insee.fr/fr/statistiques/fichier/1300624/guide-secret.pdf

[11] Insee (2016). ***La gestion de la confidentialité pour les données individuelles***, Insee, https://www.insee.fr/fr/statistiques/2535625

[12] Combes S. ***Utilisations des données massives pour les statistiques publiques***. Présentation, https://www.insee.fr/fr/information/2653115

[13] Elbaum M. (2 juillet 2018). ***Rencontres du CNIS - Les enjeux des nouvelles sources de données - Questions introductives***, https://www.cnis.fr/evenements/rencontre

[14] Egu O. (septembre 2015). ***Analyse du potentiel des données billettiques - Le cas de Lyon***, [Mémoire de Master Recherche], Laboratoire d'Économie des Transports.

[15] Treil S. (2005). ***Comprendre l'évolution de la mobilité entre deux Enquêtes Ménages Déplacements : le cas du vélo à Grenoble entre 1992 et 2002***, Altermodal.

La série de fiches « Données de mobilité pour la modélisation des déplacements »

- **Fiche chapeau - Collecte et utilisation de données de mobilité pour la modélisation des déplacements -** Des enquêtes ménages-déplacements aux données massives
- Fiche n°1 - Enquêtes origine-destination
- Fiche n°2 - Navettes : apport du recensement de la population

Fiches à paraître :

- Données issues des antennes de la téléphonie mobile
- Données issues des enquêtes spécifiques sur le transport de marchandises
- Enquêtes déplacements auprès des ménages
- Lecture de plaques d'immatriculation de véhicules
- Traces GPS de véhicules
- Traces GPS d'applications smartphone
- Enquêtes de préférences déclarées
- Comptages ponctuels, permanents ou temporaires
- Données billettiques et de péage
- Données des capteurs Bluetooth et Wifi

LE CEREMA, DES EXPERTISES AU SERVICE DES TERRITOIRES

Le Cerema est un établissement public qui apporte son expertise pour la transition écologique, l'adaptation au changement climatique et la cohésion des territoires. Grâce à ses 26 implantations partout en France, il accompagne les collectivités dans la réalisation de leurs projets. Le Cerema agit dans 6 domaines d'activité : Expertise & ingénierie territoriale, Bâtiment, Mobilités, Infrastructures de transport, Environnement & Risques, Mer & Littoral.

Téléchargez nos publications dans la rubrique « centre de ressources » sur cerema.fr

COLLECTE ET UTILISATION DE DONNÉES DE MOBILITÉ POUR LA MODÉLISATION DES DÉPLACEMENTS

Des enquêtes ménages-déplacements aux données massives

Vue du hall de la gare TGV de l'Aéroport de Roissy-Charles de Gaulle (95)

CONTRIBUTEURS

Autrices et coordinatrices du groupe de travail :
Aurélie Bousquet et Julie Tricoche (Cerema)

Membres du groupe de travail :
Alice Charpe, Barbara Christian, Julien Harache, Gaëlle Jaillet, Maria Tebar et Damien Verry (Cerema)

Relecteurs :
Bernard Allouche, Stéphane Chanut, Thierry Gouin, Joris Marrel, Luc Mathis, Nicolas Nuyttens et Patrick Palmier (Cerema)

Auteurs des fiches techniques :
Aurélie Bousquet, Alice Charpe, Barbara Christian, Julien Harache, Gaëlle Jaillet et Maria Tebar (Cerema)

CONTACTS

modelisation-deplacements@cerema.fr

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

www.cerema.fr

Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement
Siège social : Cité des mobilités - 25 avenue François Mitterrand - CS 92803 - 69674 Bron Cedex - Tél. +33 (0)4 72 14 30 30 -
Dépôt légal : novembre 2022 - ISSN : en cours / Crédit photo de couverture : A. Bouissou/Terra