

Potentiel des données massives pour la connaissance des flux de déplacements

Présenté par :

Julien Harache, Cerema Ouest

Mathieu Jacquot, Cerema Est

Avec les contributions de :

Aurélie Bousquet, Cerema Territoires et Ville,

Alice Charpe, Cerema Ouest,

Maxime Le Corre, Cerema Est,

Maria Tébar, Cerema Nord-Picardie.

Intro - « Données massives », de quoi parle-t-on ?

1. Caractéristiques des sources de données

2. Domaines de pertinence des données massives pour la connaissance de la mobilité

Conclusion

- **Définitions retenues**

- **Enquêtes**

- **par sondage aléatoire**

- Recueil d'informations via un questionnaire.
 - Échantillonnage représentatif :
 - Taux de sondage variables : 1 % pour les EMC², 20 % pour une enquête OD, etc. ;
 - Propriétés asymptotiques des estimateurs applicables (loi des grands nombres...).

- **Données massives**

- Collectées passivement (sans intervention d'un enquêteur ou de l'utilisateur) :
 - Elles peuvent être partielles :
 - Observations nombreuses ;
 - MAIS pas de maîtrise qualitative de l'échantillon ;
 - Ou quasi-exhaustives :
 - Concernent tous les véhicules qui passent en un point du réseau ou toute la population d'une zone (comptages, billettique, etc.).

« Connaissance de la mobilité », de quoi parle-t-on ?

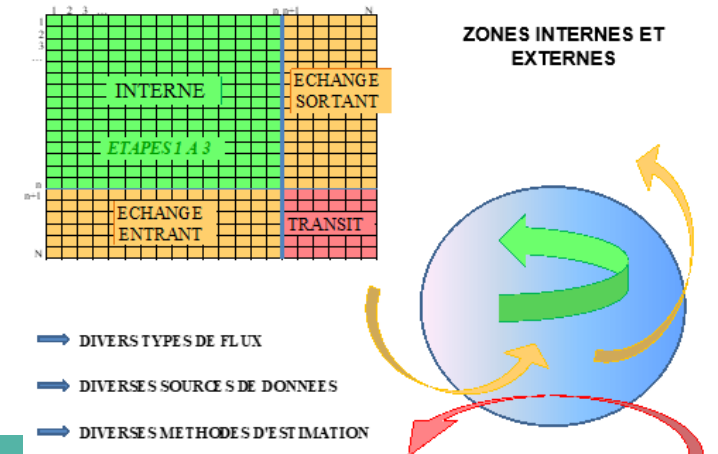
- Connaître les pratiques de déplacement et les flux qui traversent un territoire

> Pour permettre l'analyse de la mobilité :

- Qui se déplace ? Comment ? Pourquoi ?
- Quels sont les générateurs de la mobilité ? Les liens qui peuvent exister entre territoires ?

> Pour alimenter des modèles de déplacements :

- Estimer la répartition de l'ensemble flux de déplacements sur un territoire aujourd'hui et dans le futur
- Evaluer l'impact des politiques publiques et des projets de transport





1. Caractéristiques des sources

- **Fiche d'identité**
 - **Données de localisation** des possesseurs de téléphone mobile issues de l'exploitation des réseaux de téléphonie. Données liées aux événements recensés par les **antennes de téléphonie mobile**.
- **Comment découpe-t-on les déplacements ?**
 - Un **temps de stationnarité** au-delà duquel on considère que le possesseur de téléphone est à l'arrêt est défini. A **minima 15 min**.
- **Quelle est la précision spatiale de représentation des OD ?**
 - **Flux à grandes échelles**, à minima d'une ville.
 - **Dépend de la localisation des antennes**
- **Peut-on connaître les motifs et modes de déplacements ?**
 - **Motifs** : éventuellement l'ensemble des **motifs pendulaires** sans distinction et domicile
 - **Modes** : éventuellement distinction **mode ferroviaire/routier sur de grandes distances**, en fonction de la proximité des réseaux de transport

- **Quels sont les biais de collecte spécifiques à cette source ?**
 - **Traitement algorithmique complexe** et évolution du réseau téléphonique
 - **Échantillonnage temporel irrégulier** et hétérogène
 - Biais lié au choix de l'opérateur, avec des données ne provenant que d'un seul opérateur à la fois
 - **Sous échantillonnage pour les déplacements courts**
- **Comment généraliser les données collectées à l'ensemble de la population ?**
 - **Redressement à partir de la part de marché de l'opérateur** : pas nécessairement sur l'âge ou les caractéristiques socio-démographiques
 - **Documentation faible sur le sujet**
- **Peut-on mesurer des évolutions ?**
 - Données conservées un an : **analyses saisonnières, semaine/week-end**
 - Évolutions rapides des algorithmes et du fonctionnement des réseaux rendant cela incertain
 - Nécessité de choisir le même opérateur
- **De quelles informations dispose-t-on sur les individus et leur foyer ?**
 - Offre hétérogène selon les opérateurs :
 - **Données socio-démographiques : incertaines** compte tenu de la source (données de facturation)
 - **Lieux d'activité et de résidence** : en fonction de l'opérateur, **plus ou moins fiable**

- **Fiche d'identité**

- **Données de localisation horodatées fournies par un système GNSS** (via un assistance d'aide à la conduite, intégré ou non au véhicule, ou une application smartphone) permettant :

- La connaissance des **vitesse**s et des **congestions**
- La connaissance des **itinéraires** et donc des **flux OD**

- **Quelle est la précision spatiale de représentation des OD ?**

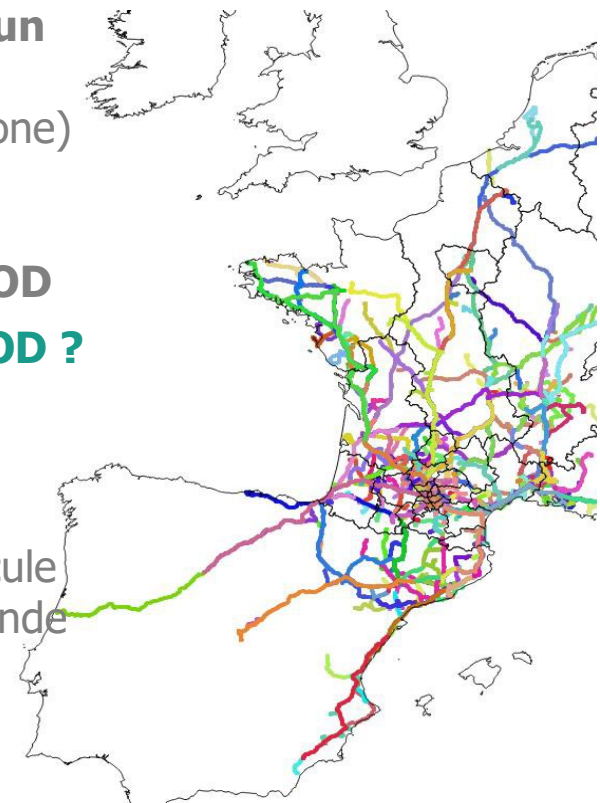
- **Précision** liée à la fiabilité des données GPS (**25 m au maximum**, mais **généralement inférieure à 10 m**)

- **Comment découpe-t-on les déplacements ?**

- **Coupure à partir d'une valeur de temps** où le véhicule est considéré à l'arrêt : nombreuses limites, car une grande partie des traces se termine sur une vitesse non nulle.
- Possibilité d'utiliser des algorithmes plus poussés (par exemple, pour filtrer les pauses pour les PL).

- **Peut-on connaître les motifs de déplacements ?**

- **Non**, sauf si historique plus long permettant par déduction de repérer le domicile et éventuellement le travail (R&D + **contraintes d'anonymisation**)



- **Quels sont les biais de collecte spécifiques à cette source ?**
 - Pour les véhicules légers
 - **Vitesse légèrement au-dessus de la moyenne** (problématique en situation fluide)
 - Motifs de déplacement et classes de distance non-représentatifs (moins de déplacements courts pour motif personnel)
 - Pour les poids-lourds :
 - Avec un seul fournisseur, les pavillons ne sont pas représentatifs
- **Comment généraliser les données collectées à l'ensemble de la population ?**
 - **Vitesse** : sans redressement, si l'échantillon est assez important.
 - **Flux OD** : un **redressement** est **indispensable, méthodologie** à définir **au cas par cas**
 - Si flux trop peu représentatifs, le redressement ne suffira pas à corriger les OD.
- **Peut-on mesurer des évolutions ?**
 - **Vitesse** : **Oui**, si la taille de l'échantillon le permet
 - **Flux OD** : **Difficilement**, le taux d'équipement évoluant vite...
- **De quelles informations dispose-t-on sur les véhicules ?**
 - FCD « classique » :
 - direction (cap) et vitesse instantanée du véhicule
 - l'identifiant unique propre au véhicule
 - xFCD : FCD « classique » + données disponibles via le bus Can du véhicule
 - ID-xFCD : xFCD + informations sur le propriétaire du véhicule



• Fiche d'identité

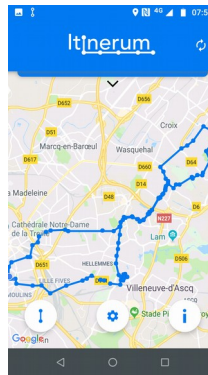
- Selon l'appli : **localisations obtenues à l'aide du GPS du téléphone, des antennes Wifi** ou des antennes du réseau de téléphonie mobile

• Applis d'enquêtes

- Recrutement des individus dans le cadre d'une enquête : ils téléchargent une application et connectent leur GPS.
- Selon l'appli, interactions avec l'utilisateur possibles : pour chaque déplacement, le corriger et le valider + renseigner des informations (motif, etc.)
- Anonymisation à faire par le développeur de l'appli

• Applis collectant la position de l'utilisateur

- Collecte des traces sans action spécifique de l'utilisateur, pas de recrutement ;
- Anonymisation par le développeur de l'appli ou l'opérateur téléphonique.



Crédit photo : PhonAndroid



• Comment découpe-t-on les déplacements ?

- Continuité d'une trace à **croiser avec un fichier « arrêts »** (applis d'enquête)
- Fichier déplacements **directement saisi** ou inféré par un algorithme (applis d'enquête)
- Trace continue : **algorithme à développer** a posteriori (applis d'enquête + guidage)

• Quelle est la précision spatiale de représentation des OD ?

- Si **GPS** : **Itinéraire et OD précis**, si **antennes wifi** : précision **comparable aux FMD**

• Peut-on connaître les motifs et modes de déplacements ?

- Selon l'appli : **possibilité d'avoir une interprétation du mode / du motif**

- **Quels sont les biais de collecte spécifiques à cette source ?**
 - **Tout le monde ne dispose pas d'un smartphone** / d'un forfait « data », ne sait pas l'utiliser ;
 - **Perte de données possibles** : panne de batterie, désactivation du GPS, volontaire ou non ;
 - Différences de tracking selon le système d'exploitation du téléphone, sa version, la marque et le modèle ;
 - Les applications doivent être mises à jour régulièrement ;
 - **Biais spécifiques aux applis d'enquêtes**
 - **Acceptabilité** : côté intrusif du tracking ;
 - **Recrutement complexe** : pas de méthode éprouvée à ce jour
 - **Biais spécifiques aux applis collectant la position de l'utilisateur**
 - Souvent spécifiques à un certain type d'utilisateurs, **pas représentatif de la population**, possibilité de doubles comptes en croisant 2 jeux de données, **fréquence de collecte dépendante de l'appli** ;
- **Comment généraliser les données collectées à l'ensemble de la population ?**
 - **Pas de méthode de redressement validée à ce jour** : besoin d'autres données
- **Peut-on mesurer des évolutions ?**
 - Oui, sur le champ de population concerné
- **De quelles informations dispose-t-on sur les individus et leur foyer ?**
 - **Applis d'enquêtes** (selon l'appli) : **possibilité d'avoir un questionnaire complémentaire** pour connaître le profil de la personne

- **Fiche d'identité**

- **Données de validation / paiement** sur un réseau donné
- Soit en un **point de passage**, soit en **entrée/sortie** selon les dispositifs et les réseaux

- **Peut-on connaître les motifs de déplacements ?**

- **Non**

- **Peut-on mesurer des évolutions ?**

- **Oui, données fortement dynamiques**, possibilité de mesurer des évolutions de jour en jour, entre saisons différentes, d'une année sur l'autre
- Possibilité théorique de suivre un véhicule ou un usager sur plusieurs jours, mais généralement, les contraintes d'anonymisation rendent cette information indisponible.

- **De quelles informations dispose-t-on sur les individus et leur foyer ?**

- Titre de transport utilisé. **Aucune autre information** en général (contraintes d'anonymisation).



Crédit photo : Le blog auto



Crédit photo : Actu Maint

• Fiche d'identité

- **Lecture Automatisée de Plaques d'Immatriculation :**
 - Connaissance des volumes en un point donné (comptage)
 - Connaissance des trafics d'échange et de transit
 - Par rapport à un périmètre donné => **pas d'OD**
- Exhaustives sauf masquage de plaque
 - Redressement sur des comptages

• Quels sont les biais de collecte spécifiques à cette source ?

- Conditions de mauvaise visibilité (nuit, fortes intempéries)
- Erreurs de lecture entièrement imputées aux flux d'échange

• Peut-on connaître les motifs de déplacements ?

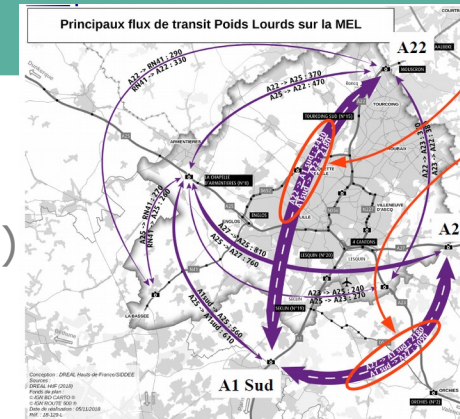
- **Non**, connaissance impossible

• Peut-on mesurer des évolutions ?

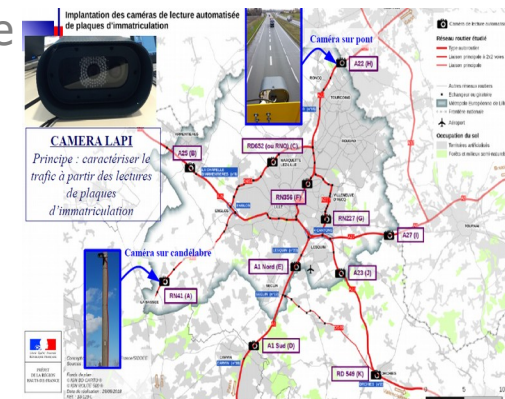
- **Oui**, si on situe la caméra au même endroit

• De quelles informations dispose-t-on sur les véhicules ?

- **Pays d'immatriculation** du véhicule
- Type de véhicule : **grand PL, petit PL, VUL, VP, moto**
- Avec dispositif spécifique (RGPD et autorisation CNIL) : possibilité d'apparier avec des données issues du fichier SIV (vignette, motorisation, etc)



Source : Dreal HF



Source : Dreal HF



• Fiche d'identité

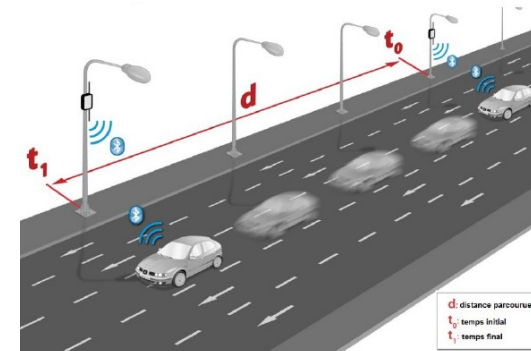
- **Données d'adresses MAC horodatées** : la norme bluetooth permet de connaître la catégorie de l'équipement (smartphone, oreillette, casque audio, etc.)
- **Dispositif « bord de route »**
 - antenne directionnelle ou omnidirectionnelle
 - rayon de capture paramétrable
 - paramétrage complexe

• Comment découpe-t-on les déplacements ?

- **Déplacements point à point** entre capteurs, dépendant de la répartition des capteurs dans le cadre d'une campagne de collecte => **pas d'OD**

• Peut-on connaître les motifs et les modes de déplacements ?

- **Motifs** : **pas d'informations** sur les origines et destinations des déplacements, donc pas possible d'inférer des motifs
- **Modes** : **pas de détermination possible** sur la base des temps de parcours ; **localisation** (route, gare, etc.) et paramétrage des dispositifs, **éventuellement**





- **Quels sont les biais de collecte spécifiques à cette source ?**
 - **Taux d'équipement** (difficulté de faire le lien appareils → personnes → véhicules) et **nécessité que le bluetooth soit activé**
 - **Nouvelle norme bluetooth pas forcément détectable**, selon la génération des équipements (sur tests récents, beaucoup d'équipements de type oreillettes et très peu de véhicules, par exemple)
 - A venir : adressage MAC dynamique => ne permet pas de suivre un appareil dans le temps
- **Comment généraliser les données collectées à l'ensemble de la population ?**
 - **Vitesse : oui**
 - **Flux OD :**
 - **Redressement sur comptages automatiques**, en parallèle
 - Taux de pénétration variable autour de 10 % tous équipements (attention aux doubles comptes)
 - possible de redresser une campagne « cordon » (entrée/sortie)
- **Peut-on mesurer des évolutions ?**
 - **Vitesse : oui, si on situe les équipements au même endroit**
 - Suivi dans le temps compliqué par l'évolution de l'équipement électronique de la population : variation du taux de pénétration, évolution des normes et technologies
- **De quelles informations dispose-t-on sur les individus et leur foyer ?**
 - **Pas d'enrichissement possible**



- **Fiche d'identité**
 - **Enquête des résidents d'un territoire par sondage aléatoire** en face-à-face et/ou par téléphone
 - **Photographie à l'instant « t »** de leurs déplacements sur l'ensemble de la journée (jours de semaine hors vacances + option week-end)
- **Quelle est la précision spatiale de représentation des OD ?**
 - **A minima l'IRIS**, voir jusqu'à une géolocalisation précise (coordonnée x,y) pour les générateurs de déplacements sur le territoire
- **Peut-on connaître les motifs et les modes de déplacements ?**
 - **Oui, avec énormément de détails**
- **Quels sont les biais de collecte spécifiques à cette source ?**
 - **Biais classiques liés aux enquêtes statistiques** (biais enquêteurs, non-réponse, etc.), connus et maîtrisés
- **Comment généraliser les données collectées à l'ensemble de la population ?**
 - **1 à 3 % de la population enquêtée**
 - Redressement de la non-réponse à partir de la base DGI des logements
 - **Redressement** de l'échantillon au travers des **données du recensement** de la population de **l'INSEE** par un calage sur marges



- **Peut-on mesurer des évolutions ?**
 - **Oui** (méthodologie stable), **mais sur des périodes de 10 ans**, sauf option « Fréquence + »
- **De quelles informations dispose-t-on sur les individus et leur foyer ?**
 - **Individus :**
 - **âge, genre, PCS**, statut d'occupation, niveau de scolarité,
 - **permis de conduire, abonnement TC**,
 - **adresse précise du lieu de travail ou d'études.**
 - **Ménages :**
 - **taille du ménage**, type de logement,
 - **nombre de véhicules, lieu et type de stationnement** à domicile et au travail.

- **Fiche d'identité**

- **Enquêtes réalisées en bord de route par interview** des conducteurs
- Questionnaires courts pour ne pas trop gêner l'écoulement du trafic : 30 sec à 1 min
- Dispositifs complexes à mettre en œuvre sur certaines voiries

- **Quelle est la précision spatiale de représentation des OD ?**

- **Généralement à la commune pour la France**, sur un zonage agrégé à l'étranger ;
- Parfois, **collecte infracommunale pour les zones proches** du poste d'enquête.



- **Peut-on connaître les motifs de déplacements ?**

- **Oui**, toujours collectés pour les VL, souvent nature de la marchandise pour les PL

- **Quels sont les biais de collecte spécifiques à cette source ?**

- Collecte entre 7 heures et 19 heures classiquement : on **manque les OD de nuit** ;
- Non-réponse : notamment conducteurs ne parlant pas français : + difficile pour les enquêteurs ;

- **Comment généraliser les données collectées à l'ensemble de la population ?**

- **10 à 25 % du trafic enquêté**,
- **Redressement sur des comptages** manuels sur la période d'enquête (distingue finement les types de véhicules) et automatiques sur 2 semaines (distingue uniquement VL et PL).

- **Peut-on mesurer des évolutions ?**
 - **Oui**, on actualise habituellement un poste **tous les 10 ans**
- **De quelles informations dispose-t-on sur les véhicules / les individus et leur foyer ?**
 - **VL** (selon les enquêtes), informations classiquement collectées :
 - **taux d'occupation**,
 - **âge, sexe, PCS, communes de résidence et de travail** du conducteur
 - **PL** (selon les enquêtes), informations classiquement collectées :
 - **Catégorie de marchandises, tonnage**

• Fiche d'identité

- **Enquêtes réalisées à bord des transports collectifs** (toutes les courses d'une ligne ou un échantillon)
- Questionnaires courts pour capter les OD courtes : 30 sec maxi
- Connaissance des arrêts de montée / descente / correspondance, des lignes empruntées / des modes de rabattement à l'origine et à la destination

• Quelle est la précision spatiale de représentation des OD ?

- Généralement à la **commune**, parfois, collecte **infracommunale**

• Peut-on connaître les motifs de déplacements ?

- **Oui**

• Quels sont les biais de collecte spécifiques à cette source ?

- **Sous-représentation des OD courtes** : usagers qu'on n'a pas le temps de capter ;
- **Période horaire limitée** : on enquête rarement toute la journée, pour des raisons de budget.

• Comment généraliser les données collectées à l'ensemble de la population ?

- **Taux de sondage très variable** selon les enquêtes,
- **Redressement sur des comptages** manuels sur la période d'enquête et éventuellement automatiques (billétique...), à la course et à la ligne.



Enquête à Montpellier – Crédit photo : France 3

- **Peut-on mesurer des évolutions ?**
 - **Oui**, selon les réseaux, on actualise les enquêtes **tous les 3 à 10 ans**
- **De quelles informations dispose-t-on sur les individus et leur foyer ?**
 - Selon les enquêtes : **âge, sexe, PCS, communes de résidence et de travail**, type de titre de transport, fréquence d'utilisation



Enquête à Montpellier – Crédit photo : France 3

- **Recensement de la population de l'INSEE**
 - **Enquêtes par sondage** aléatoire avec contrôle de l'échantillon
 - Actualisé tous les ans, mais **comparaison** de deux vagues nécessite un **intervalle de 5 ans** (rotation des communes enquêtées)
 - Fournit la commune de résidence et la commune de travail / étude : **pas de flux de déplacements, mais des potentiels**
 - Taux de sondage :
 - questions principales : 40 % à 100 % selon les communes
 - questions complémentaires (dont lieu de travail/d'étude et mode de transport utilisé) : 20 à 40 % selon les communes
- **Enquêtes en gares/aéroports/aux points d'arrêt**
 - **Enquêtes par sondage aléatoire** avec contrôle de l'échantillon
 - Pour **connaître les pratiques de rabattement** sur le point d'arrêt
 - Environ 40 % de taux de sondage observés sur des petites gares, très dépendant du terrain

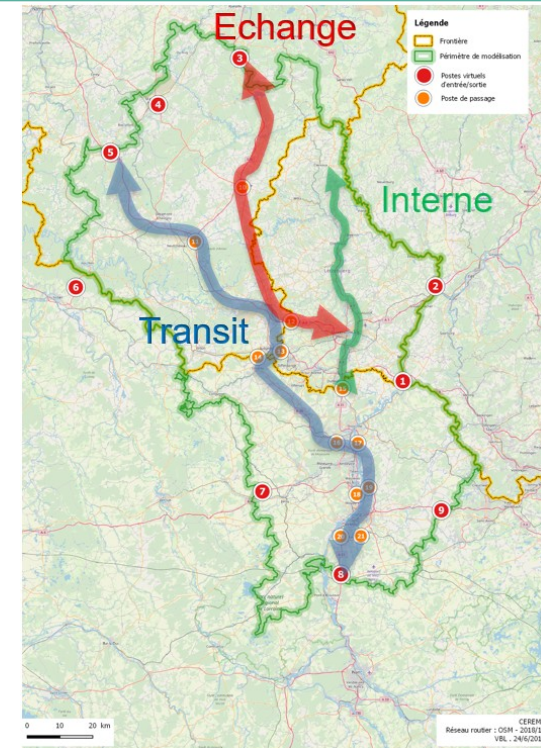
- **Enquêtes de préférences déclarées**
 - Pour connaître les **comportements des usagers** (choix de mode notamment) face à des situations hypothétiques ;
 - Vise à lier les variables individuelles (âge, PCS, etc.) aux comportements de choix ;
 - Enquête sur des **échantillons non représentatifs** mais **besoin de variabilité** ;
 - Généralisation à l'ensemble de la population via une population synthétique ou des ratios par catégorie de population :
 - **Nécessite la construction d'un modèle de choix discrets, ne peut pas être exploitée par des tris à plat.**



2. Domaines de pertinence des données massives pour la connaissance des mobilités

Différents cas d'usage :

- **Flux de voyageurs internes à une zone d'étude :**
 - Pertinent à une échelle macro (EPCI, tâches urbaines) avec les FMD (tests en court en Vendée, région Grand Est)
 - A consolider à l'échelle d'un quartier avec les données d'applis Smartphone (pas de tests publics connus)
 - **Flux de voyageurs d'échange et de transit tous modes :**
 - avec les FMD, mais pas de reconstitution du trajet, imprécision géographique et perte du signal possible
 - A consolider avec les données d'applis Smartphone (pas de tests publics connus)
 - **Flux VL/PL d'échange et de transit :**
 - avec les FCD mais nécessité de corriger le biais sur la distance de déplacement et problématique de représentativité des extrémités du déplacement
 - Pour le mode routier avec les LAPI, mais caractérisation de l'échange peu utilisable (DREAL Hauts-de-France)
- => vigilance vis-à-vis des problématiques de réseaux pour les secteurs transfrontaliers



- **Adaptation temporelle de la connaissance des flux**

- Les **enquêtes par sondage** : **pas de mesure des variations** en fonction du jour ou de la saison ;
- Ces variations sont importantes, en particulier dans des zones touristiques ou pour certains modes (pointe du vendredi soir pour le ferroviaire liée en grande partie aux retours des étudiants).
- Les **données massives permettent de mesurer des évolutions saisonnières**, mais n'offrent généralement pas de représentativité de la population.
- **Combiner les deux sources** pourrait permettre de transposer les résultats d'une enquête à d'autres périodes que sa période de collecte => **Pas de retour d'expérience à ce jour en France**

- **Connaissance du choix d'itinéraire**

- Usage fréquent des FCD pour les **temps de parcours**, développement en cours pour la **fréquentation ponctuelle** (territoire transfrontalier du Luxembourg)
- Usage des données LAPI (études A31bis), mais sur quelques itinéraires assez longs et simples
- Fortes potentialités sur les applis d'enquête, les FCD et les applis collectant la position des usagers (mais **pas de retour d'expérience à ce jour en France**)

Conclusion

- **Les données massives apportent des éléments de caractérisation générale de la mobilité.**
 - Les **données massives de déplacement sont peu qualifiées** (pas de motif, de mode, de caractéristiques des individus...), **souffrent d'imprécisions** (localisation précise de l'origine et de la destination, biais de représentativité mal connus)

=> L'accès aux données des applis smartphone pourraient toutefois changer la donne sur certains éléments (modes de déplacements, motifs), au détriment de la représentativité à ce stade.

 - Pour connaître les mobilités actuelles et prévoir les futures, il convient de **combiner ces données avec des sources classiques** (EMC², enquêtes OD, comptages...). Les méthodes afférentes sont en cours de test et présentent encore un très fort caractère expérimental.

=> Il est d'autant plus crucial de disposer de bases de données structurées pour les données plus classiques et de continuer d'en produire régulièrement.
- **Les données massives permettent de mesurer des variabilités temporelles que les enquêtes par sondage aléatoire ont du mal à capter.**
 - **MAIS** elles doivent être utilisées en combinaison de données classiques, qui fournissent une situation de base représentative.
- **Les coûts ne sont pas nécessairement plus faibles pour des données massives que pour des collectes par sondage sur le même champ.**

Pour aller plus loin...

FCD :

Représentativité des origines - destinations des usagers utilisant un système GPS

Maxime Le Corre, Gilles Bedat, Marie Giraud – Présentation et article pour le congrès ATEC 2018

FMD :

Données de téléphonie mobile pour la connaissance de la mobilité : enseignements de trois expérimentations

Alice Charpe, Julien Harache, Maxime Le Corre, Olivier Richard, Wilfried Raballand

Présentation et article pour le congrès ATEC 2019

Utilising mobile network data for transport modelling

Catapult Transport Systems, Department for Transport of the United Kingdom

<https://www.gov.uk/government/publications/mobile-phone-data-in-transport-modelling>

Billétique TC :

Analyse du potentiel des données billétiques – Le cas de Lyon

Oscar Egu, rapport de stage LAET, sept. 2015.



Merci de votre attention