Les données de téléphonie mobile peuvent-elles améliorer la mesure du tourisme international en France ?

Can Mobile Phone Data Improve the Measurement of International Tourism in France?

Guillaume Cousin* et Fabrice Hillaireau**

Résumé – Depuis juillet 2015, la Banque de France et le ministère de l'Économie et des Finances expérimentent l'utilisation de données de téléphonie mobile pour l'estimation du nombre et des nuitées des visiteurs étrangers en France. Cette expérience est destinée à évaluer la capacité de ces données à remplacer à terme, en tout ou partie, les données de trafic par mode de transport actuellement utilisées pour asseoir la représentativité de l'*Enquête auprès des visiteurs venant de l'étranger* (EVE). À ce jour, les données de téléphonie mobile ne sont pas intégrées en production dans le dispositif de dénombrement des visiteurs. Les estimations issues des données de téléphonie mobile présentent toutefois plusieurs intérêts en termes de délai d'obtention, de détail temporel et géographique et de suivi conjoncturel. L'expérience, encore en cours, illustre les difficultés d'exploitation de données originales de type Big Data et démontre la nécessité de l'usage complémentaire de données d'enquêtes classiques pour améliorer la qualité des estimations.

Abstract – *Since July 2015, the* Banque de France *and the French Ministry for the Economy and Finance have been experimenting with the use of mobile phone data to estimate the number and overnight stays of foreign visitors in France. The purpose of the experiment is to assess the ability of such data to eventually replace, in part or in whole, the traffic data by mode of transport currently used to establish the representativeness of foreign visitor surveys (*Enquête auprès des visiteurs venant de l'étranger *or EVE). Mobile phone data have yet to be incorporated into the method used to count tourists. However, estimates based on mobile phone data have a number of benefits in terms of the time required to obtain data, the level of temporal and geographical detail and short-term trend monitoring. This ongoing trial illustrates the difficulty of exploiting original Big Data and demonstrates the importance of drawing on traditional survey data to improve the quality of estimates.*

Codes JEL / JEL Classification : Z3, Z39
Mots-clés : données massives, Big Data, statistiques de tourisme, données de téléphonie mobile
Keywords: Big Data, tourism statistics, mobile phone data

** Banque de France (guillaume.cousin@banque-france.fr)*
*** Ministère de l'Économie et des Finances (fabrice.hillaireau@finances.gouv.fr)*

Nous remercions sincèrement François Mouriaux, Bertrand Collès, François-Pierre Gitton et Yann Wicky pour leur précieuse contribution à la conception et à la rédaction de cet article ainsi que l'équipe d'Orange avec laquelle nous avons mené cette expérimentation, notamment Sylvain Bourgeois, Jean-Michel Contet et Kahina Mokrani.

Reçu le 9 août 2017, accepté après révisions le 14 janvier 2019

Pour citer cet article : Cousin, G. & Hillaireau, F. (2018). Can Mobile Phone Data Improve the Measurement of International Tourism in France? *Economie et Statistique / Economics and Statistics*, 505-506, 89–107. https://doi.org/10.24187/ecostat.2018.505d.1967

Depuis juillet 2015, la Direction Générale des Statistiques de la Banque de France et la Direction Générale des Entreprises (DGE) du ministère de l'Économie expérimentent l'utilisation de données de téléphonie mobile pour l'estimation du nombre de visiteurs étrangers en France et de leurs nuitées, dans le cadre d'un partenariat d'enquêtes pour l'élaboration du compte satellite du tourisme et de la balance des paiements[1].

Le dénombrement des visiteurs étrangers et de leurs nuitées est nécessaire à l'élaboration des statistiques de tourisme et à l'estimation en balance des paiements des recettes de « services de voyages » de la France en provenance de l'étranger. Ces statistiques revêtent un enjeu particulier en raison de l'importance du tourisme dans l'économie française. En 2015, le poids de consommation touristique dans le PIB était de 7.27 %[2] dont 32 % du fait des visiteurs non-résidents. La France a accueilli cette même année 84.5 millions de touristes étrangers (DGE, 2016a) qui ont généré 52.6 milliards d'euros de recettes de services de voyages enregistrées en balance des paiements[3].

Le dénombrement de ces visiteurs repose actuellement sur des données de trafic par mode de transport combinées à des vacations de comptages et d'enquêtes. Les vacations permettent de ventiler les franchissements de frontière par pays de provenance des visiteurs tandis que les données de trafic par mode de transport servent de base au calcul des coefficients d'extrapolation. Les estimations actuelles du nombre de visiteurs nécessitent des redressements qui peuvent s'avérer complexes pour tenir compte de mutations rapides telles que le développement en mode *hub* des grands aéroports – qui multiplie les nationalités présentes sur un même vol – et les difficultés pour exploiter les comptages routiers aux frontières dans un pays aux multiples points d'entrées (par exemple, il existe environ trois cents points de passage entre la France et la Belgique). L'expérience d'utilisation des données de téléphonie mobile est destinée à évaluer leur capacité à remplacer à terme les données de trafic par mode de transport dans le dénombrement des visiteurs étrangers en France.

Les données de téléphonie mobile, déjà utilisées pour le suivi du tourisme en Estonie et par certains comités départementaux et régionaux du tourisme, notamment Bouches-du-Rhône Tourisme[4], pour des mesures de fréquentation ou de flux, semblaient susceptibles de remédier aux limites nouvelles affectant, ou susceptibles d'affecter, les sources actuelles.

En effet, elles permettent d'accéder à de riches informations sur les localisations et les déplacements des individus. Gonzales *et al*. (2008) ont ainsi été parmi les premiers à utiliser cette source de données pour construire un modèle des déplacements humains, à partir d'un échantillon de 100 000 téléphones mobiles suivis durant une période de 6 mois. Ce type de données a depuis été mobilisé pour repérer les déplacements (Calabrese *et al*., 2011, 2013), notamment des mouvements pendulaires (Aguilera *et al.*, 2014). Widhalm *et al.* (2015) ont cherché à construire une typologie des activités urbaines en fonction des durées, des fréquences et des lieux des déplacements. De nombreuses autres utilisations ont pu être cherchées dans des domaines variés (ONS, 2016).

La statistique publique a repéré le potentiel de ces données massives, pour le recensement de la population (Vanhoof *et al.*, 2018 ; Givord *et al.*, ce numéro), mais également pour la mesure du tourisme. L'Estonie a été pionnière dans ce domaine avec une expérience relatée dans deux principaux articles : Ahas *et al.* (2008), qui montre une forte corrélation entre les données de téléphonie et les statistiques d'hébergement, et Kroon (2012), qui expose l'expérimentation conduite par la banque centrale d'Estonie en matière d'utilisation des données de téléphonie comme intrant possible dans l'estimation des échanges de services de voyages. Eurostat (2014) a depuis produit une étude de faisabilité de ces données pour le suivi du tourisme.

Pour autant, l'analyse des données de téléphonie mobile pour mesurer le tourisme reste rare, et notre expérimentation présente l'avantage d'étudier le cas d'un pays relativement vaste (douze fois plus grand que l'Estonie) accueillant un grand nombre de touristes (28 fois plus qu'en Estonie). De plus, cet article expose le point de vue de producteurs de

---

1. Ce partenariat porte sur les enquêtes SDT « suivi de la demande touristique » et EVE « enquête auprès des visiteurs étrangers ». La première enquête collecte des données sur la demande touristique des Français. C'est une enquête sur un panel représentatif des ménages français. La seconde collecte porte sur la demande touristique des non-résidents visitant la France. C'est une enquête sur les flux de type « enquête aux frontières ». Ce partenariat permet une production intégrée de données de référence pour les statistiques officielles dont chacune des institutions a la charge.

2. Voir : DGE, Compte satellite du tourisme *(base 2010) ; Insee,* Comptes nationaux *(base 2010).*

3. Voir : Webstat, Banque de France.

4. Voir : https://www.myprovence.pro/bouches-du-rhone/projets-majeurs/projet-flux-vision-tourisme.

statistiques publiques et offre ainsi une perspective différente de celle de la plupart des autres articles, visant une utilisation opérationnelle et régulière des données massive pour la construction d'indicateurs statistiques. Dans ce cadre, il est important de tester la qualité des indicateurs en les comparants à des données alternatives, en l'occurrence l'enquête de référence sur le tourisme international en France (EVE) ainsi qu'aux données de paiements par cartes bancaires.

À ce jour, les données de téléphonie mobile ne sont pas intégrées en production dans le dispositif de dénombrement qui utilise données de trafic, comptages et enquête. La période de test a mis en évidence de nombreuses spécificités d'utilisation de ces données de type Big Data : accès aux données, contraintes d'anonymisation et contraintes techniques, qualité des données et des indicateurs construits sur la base de ces données. Elle a aussi permis de développer des méthodes de traitement pour les rendre mieux exploitables.

Le besoin initial : consolider le système actuel d'estimation de la fréquentation touristique étrangère

L'estimation de la fréquentation touristique étrangère repose sur des comptages et une enquête aux frontières

Le système actuel d'estimation de la fréquentation touristique étrangère a été construit en s'appuyant sur l'expérience héritée de l'enquête aux frontières menée entre 1963 et 2001, pour s'intégrer dans un contexte de libre circulation des capitaux, de la mise en place de la zone euro, et d'une zone de libre circulation des personnes (espace Schengen). C'est un dispositif appelé *Enquête auprès des visiteurs venant de l'étranger* (EVE) qui combine des données de trafic, des vacations de comptage et une enquête (Banque de France, 2015). La difficulté du suivi du tourisme international est qu'il n'existe pas de base de sondage à partir de laquelle il serait possible de mener une enquête classique (comme c'est le cas pour le « tourisme sortant » par exemple[5]). Le dispositif EVE repose donc tout d'abord sur un recensement du trafic aux points de sortie du territoire (ports, aéroports, gares offrant des lignes internationales, frontières routières). Les flux de passagers aériens, maritimes ou ferroviaires sont recueillis auprès des différents transporteurs tandis que les flux routiers sont estimés par le Cerema[6] à partir d'automates fixes ou mobiles répartis sur toutes les frontières (plus de cent cinquante points de comptage en tout). La seconde étape consiste à qualifier ce flux total, c'est-à-dire à le décomposer en un flux de résidents et un flux de non-résidents. Cela implique des opérations de comptage par des enquêteurs en différents points de sortie du territoire. Dans les aéroports, le comptage des non-résidents est effectué en salle d'embarquement. Il permet d'estimer la répartition entre résidents et non-résidents sur un échantillon de vols et de l'extrapoler. Pour le mode routier, les comptages ont lieu aux frontières. La répartition du trafic sortant par pays ou zone de résidence est ensuite précisée par les réponses aux questionnaires de l'enquête EVE. L'enquête EVE repose donc sur la combinaison des données de trafic (dites exogènes), des opérations de comptages spécifiques (plus d'un million de véhicules observés aux frontières, plus de 120 000 passagers aériens) et de l'enquête proprement dite, administrée à plus de 80 000 visiteurs en 2015 comme en 2016, le questionnaire étant disponible en douze langues.

Le dispositif EVE est confronté au besoin accru de redresser les données de trafic et de comptage

Pour les dénombrements (ou « comptages »), le principal enjeu de redressement statistique du dispositif EVE porte sur le mode routier. Pour ce mode, les difficultés proviennent à la fois des données de trafic exogènes, qui s'appuient sur des points de mesure moins nombreux, et des vacations de comptages et d'enquêtes, dont certaines sont peu productives. Afin de disposer de réponses spontanées au moment de la fin de leur séjour en France, les voyageurs sont interrogés notamment sur les aires d'autoroute proches des frontières, dont le taux de fréquentation peut varier inopinément selon la gestion d'itinéraire des autocaristes par exemple. Il s'ensuit que les résultats extrapolés de répartition du flux sortant de visiteurs par zone géographique peuvent présenter une fluctuation aberrante sur certaines provenances, nécessitant d'envisager des corrections spécifiques. La

*5. Il s'agit des Français se rendant à l'étranger. Pour ceux-ci il a été possible de composer un panel représentatif, dans le cadre de l'enquête SDT (*Suivi de la demande touristique*).*

6. Le Cerema (Centre d'étude et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement) est un établissement public à caractère administratif sous tutelle conjointe des ministères en charge de l'urbanisme et du développement durable.

même question se pose, à un degré moindre, pour répartir le trafic aérien sortant selon le pays d'origine du visiteur en tenant compte de l'importance du transit dans les aéroports parisiens et du fonctionnement en *hub* des grands aéroports. L'enquête EVE est menée avec des objectifs en termes de questionnaires par zone de provenance. Pour les aéroports, le plan d'enquête repose sur une sélection de vols échantillonnés afin de collecter des questionnaires de visiteurs de diverses provenances. Cependant, le lien entre la destination du vol et le pays de provenance des visiteurs se distend puisque les passagers peuvent rallier une destination intermédiaire avant de regagner leur pays d'origine. Par exemple, des touristes asiatiques sont susceptibles d'emprunter un vol Paris-Francfort pour quitter la France, ce qui complique le ciblage de l'enquête.

La téléphonie mobile est une source potentielle pour le dénombrement des visiteurs

La décision d'engager l'expérimentation d'utilisation de données de téléphonie a été largement motivée par le potentiel de ces données pour répondre à la problématique de suivi du tourisme international, mis en avant par plusieurs expériences. En France, certaines collectivités locales les utilisent pour la mesure de la fréquentation touristique. En Europe, l'Estonie utilise les données de téléphonie comme source principale pour la mesure du tourisme international entrant et sortant depuis 2008-2010 et d'autres pays sont intéressés par ces données dont les possibilités d'utilisation ont en outre été mises en avant par une étude approfondie (Eurostat, 2014). L'expérimentation menée par la Banque de France et la DGE est cependant singulière par son ampleur : la population d'intérêt correspond aux touristes internationaux en France qui représentent 85 millions de personnes par an. De même, le territoire sur lequel a lieu le comptage, celui de la France métropolitaine, est d'une superficie de 552 mille kilomètres carrés. En comparaison, le nombre d'arrivées touristiques en Estonie est d'environ trois millions par an, sur un territoire douze fois plus petit.

D'un point de vue technique, l'utilisation des données de téléphonie mobile est rendue possible par le fait que les opérateurs disposent de la liste des connexions entre antennes réseau et téléphones mobiles, que ce soit pour les signaux émis de façon passive ou pour l'activité téléphonique (appels, messages, réception de données par internet, etc.). Le pays de résidence de l'opérateur émetteur de la carte SIM[7] des téléphones qui se connectent au réseau français est aussi connu des opérateurs établis en France et permet de constituer une base de données massives des signaux émis par les téléphones portables utilisés en itinérance. Les données de téléphonie mobile comprennent donc des variables d'intérêt pour la production de statistiques de tourisme.

Le cadre et les modalités de l'expérience

La mise en place d'un contrat de prestation correspond au financement d'une démarche coopérative de recherche et développement

L'expérimentation des données issues de la téléphonie mobile à des fins de comptage des touristes étrangers en France relève certes d'une démarche « Big Data », mais dans un contexte qui n'est en revanche pas « open data ». Les données sont en effet détenues par les différents opérateurs disposant d'un réseau mobile sur le territoire de France métropolitaine. Afin d'y accéder sous forme de statistiques, la Banque de France et la DGE ont lancé un appel d'offres au printemps 2015 et ont reçu deux propositions, ce qui reflète d'une part l'intérêt des opérateurs pour collaborer avec les acteurs publics en vue d'évaluer un nouveau champ d'application de leurs données massives ; d'autre part, la nécessité d'un financement public reflétant un partage des frais de recherche-développement et des frais spécifiques de mise à disposition de l'information dans le cadre de l'expérimentation. Lors des auditions des candidats, les attentes en termes de transparence des méthodes de collecte et de traitement de l'information ont constitué un élément du dialogue. L'expérimentation a pu être structurée autour d'un distinguo entre le détail des algorithmes d'agrégation et d'anonymisation qui relèvent de la protection de la propriété intellectuelle du prestataire, les parts de marché de l'opérateur choisi pour les différentes populations et territoires observés relevant du secret des affaires, et les variables déterminantes pour

7. La carte SIM (de l'anglais Subscriber Identity Module*) est une puce utilisée en téléphonie mobile pour stocker les informations spécifiques à l'abonné d'un réseau mobile, en particulier pour les réseaux GSM, UMTS et LTE. Elle permet également de stocker des données et des applications de l'utilisateur, de son opérateur ou dans certains cas de tierces parties. La carte SIM contient un numéro IMSI, constitué du code pays (MCC), de l'identifiant de l'opérateur (MNC), et de l'identifiant de l'abonné (MSIN).*

évaluer la qualité statistique des données, les choix de redressements, l'ensemble des options de méthodologie, devant être maîtrisées par la Banque de France et la DGE. Ce distinguo est détaillé plus loin.

À l'issue de la procédure, le marché a été attribué à Orange Business Services. L'expérimentation porte sur les données de début juillet 2015 à fin juin 2017. Le dernier point disponible au moment de la rédaction de cet article est mars 2017.

L'offre retenue présente trois caractéristiques : préexistence d'un module de mise en forme de Big Data déjà commercialisé, équilibre entre le respect de la propriété intellectuelle et la transparence méthodologique, processus « par étapes » de la méthodologie de construction des indicateurs.

Le prestataire avait développé un module de traitement de ses données de masse, adapté notamment aux besoins d'utilisateurs à la recherche de données de fréquentation spatio-temporelles (quantifier les regroupements de personnes sur un lieu donné, par exemple un événement culturel ou sportif). En revanche, cette méthode n'avait jamais été mise en œuvre pour répondre au besoin d'observation du tourisme international sur l'ensemble du territoire français.

Le prestataire s'est engagé à fournir aux deux partenaires un niveau d'information suffisant pour qu'ils soient en mesure de s'assurer que la méthode utilisée leur permet d'être conformes aux normes de qualité statistique établies notamment pour les institutions internationales et européennes et d'être compréhensibles par les différents publics des statistiques du tourisme. La connaissance de la méthodologie utilisée garantit une capacité autonome d'interprétation des résultats et une capacité, le cas échéant, à effectuer les révisions pertinentes. Parallèlement, il convenait que le prestataire soit assuré des garanties de confidentialité nécessaires sur l'algorithme qu'il utilise pour passer de la donnée élémentaire du signal transmis par une carte SIM à une ou plusieurs antennes, à une donnée brute constituant un proxy de personne physique anonyme. Le degré de détail de l'information méthodologique partageable varie donc selon qu'on se situe aux étapes 2, 3, ou 4 et 5, telles que détaillées ci-après.

Le module préexistant de l'opérateur de téléphonie n'enregistre pas le détail des déplacements des cartes SIM pour ensuite traiter ces déplacements en agrégeant les données suivant la demande de l'étude. Le mode de traitement validé par la Commission nationale de l'Informatique et des libertés (CNIL) exige en effet que les comportements étudiés soient prédéfinis. Seuls ces comportements prédéfinis font l'objet de compteurs incrémentés en temps réel par les connexions aux réseaux du prestataire, sans que soient conservées des données à caractère personnel. La méthode de comptage comporte cinq étapes (schéma) :

- étape 1 : les critères du comptage sont définis en amont avec la Banque de France et la DGE et correspondent aux comportements touristiques d'intérêt (arrivée, nuitée) ;

- étape 2 : un algorithme s'alimente des données de connexions téléphoniques en temps réel. Ce sont des données de signalisation, qui comprennent tous les échanges entre téléphones et antennes. Les signaux enregistrés sont ceux émis de façon passive par les téléphones pour se signaler à une antenne en fonction de leur position ainsi que les données qui transitent par les antennes lors d'une activité téléphonique (appels, SMS, fonctionnement des applications mobiles). Ces données proviennent des seuls téléphones dont la carte SIM émane d'un opérateur non-résident et qui sont en itinérance sur le réseau Orange. Elles ne sont pas sauvegardées. L'algorithme traite et rend anonymes ces données au fur et à mesure de leur collecte, à la manière d'un compteur. L'algorithme construit des estimations de fréquentation en nombre de mobiles, par zone de provenance ;

- étape 3 : le prestataire procède à un redressement dit « spatio-temporel », en agrégeant les données mesurées sur les différents réseaux (« 2G », « 3G », « 4G ») et en corrigeant les effets liés à l'évolution permanente de ces réseaux (mise en service de nouvelles antennes, indisponibilité ponctuelle ou durable) ;

- étape 4 : le passage des données de connexion à un nombre estimé de mobiles étrangers présents sur le territoire de France métropolitaine se fait au moyen d'un redressement sur les parts de marché du prestataire sur la clientèle en itinérance, par couple pays de provenance - opérateur d'origine. La part de marché par pays-opérateur est mesurée à partir de la répartition entre les différents opérateurs des SMS envoyés depuis des mobiles en itinérance, connue en temps réel par le prestataire ;

- étape 5 : dans un dernier temps, l'estimation du nombre de visiteurs par zone de provenance

est effectuée grâce à un redressement statistique classique relatif aux taux d'équipement et d'utilisation des mobiles. Ce dernier redressement étant opéré *a posteriori* sur les résultats agrégés par pays (ou zones plus larges) de résidence supposée, il se prête à l'exécution de plusieurs scénarios.

Une mise en place est nécessaire avant la période d'observation

La mise en place consiste à définir conjointement avec le prestataire les différents comportements d'intérêt afin qu'ils puissent être mesurés. Dans le cas de l'expérience présente, il a fallu transcrire les définitions d'arrivées et de nuitées touristiques en critères relatifs à la présence de téléphones mobiles sur un réseau. La nuitée a ainsi été définie comme présence du téléphone entre minuit et six heures du matin. Une arrivée est comptée lorsqu'une première nuitée est constatée après une absence la nuit précédente. Des hypothèses de ce type permettent d'interpréter les données de téléphonie en termes de comportements. Certains travaux ont utilisé des hypothèses semblables pour l'analyse des mobilités en supposant par exemple la présence au domicile entre minuit et huit heures (Akin & Sisiopiku, 2002). Les critères initiaux ont peu à peu été complétés afin de corriger certaines aberrations de mesure.

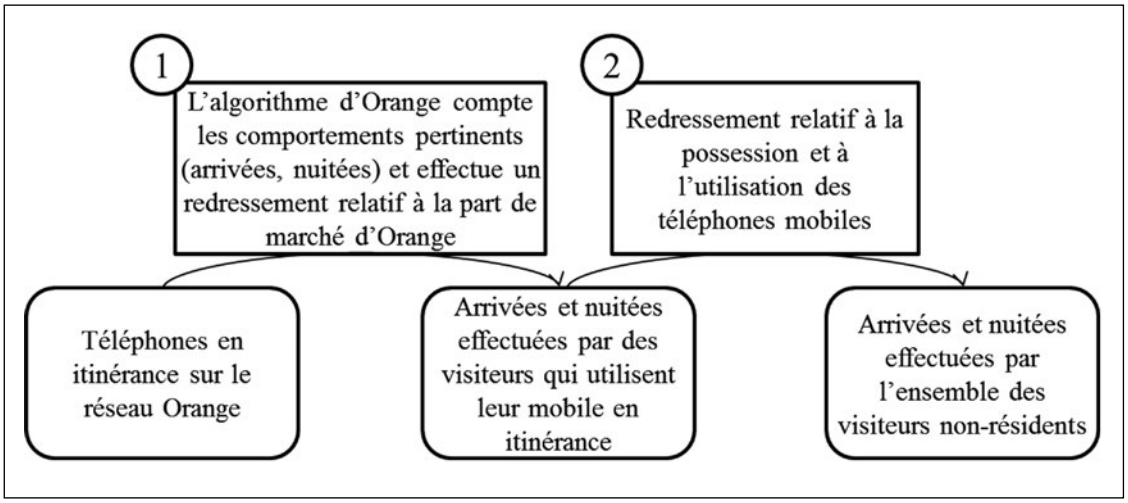
La mise en place est aussi l'occasion de définir le territoire d'intérêt. Il faut ainsi sélectionner les antennes participant aux comptages. Il a été acté de ne pas prendre en compte les flux captés par les antennes situées sur le territoire français mais à proximité immédiate de la frontière, ce point ayant été jugé indispensable dès l'analyse des premiers résultats issus du module préexistant de l'opérateur. L'efficacité des antennes n'étant arrêtée par aucune limite administrative, il s'agit d'éviter le comptage des résidents étrangers en dehors de l'Hexagone. Pour les données agrégées à l'ensemble du territoire métropolitain, qui sont le premier objectif de la Banque de France et la DGE, on peut donc s'attendre à une légère sous-estimation de la fréquentation étrangère. Celle-ci ne peut pas être mesurée car, sur la période d'expérimentation, le bruit global ne permet pas de procéder à des mesures de biais d'un tel niveau de finesse. Pour des données ventilées à un niveau régional, le problème de la sélection des antennes se pose aussi à chacune des frontières administratives, la carte des antennes et leur zone d'influence n'étant évidemment pas alignée sur la carte des régions et départements. Ceci implique un travail particulier d'affectation des antennes en fonction des regroupements spatiaux recherchés.

Les séries obtenues et leur utilité

Les séries reçues

L'opérateur transmet à la Banque de France et à la DGE les estimations d'arrivées et de nuitées

Schéma
Méthodologie simplifiée de construction des indicateurs



des touristes internationaux. Ces estimations sont communiquées mensuellement avec un délai théorique d'un mois. La fréquence des indicateurs communiqués est journalière. Les estimations sont communiquées pour 29 zones géographiques de provenance des visiteurs. Les données reçues sont au format CSV. Un fichier comporte les arrivées et un autre les nuitées. Chaque fichier comporte trois colonnes : zone de provenance, jour et nombre de nuitées ou d'arrivées pour le croisement provenance × date. Les fichiers reçus comportent ainsi environ 900 lignes (29 zones géographiques plus une ligne de total fois environ 30 jours). Il est ainsi possible de connaître, pour une date donnée et un pays de provenance donné, le nombre de touristes qui sont arrivés et le nombre de touristes qui ont passé la nuit en France.

Il faut noter que les données de téléphonie mobile avaient été initialement retenues pour étalonner la mesure des flux touristiques qui empruntent le mode routier. La prestation devait donc inclure une répartition des arrivées touristiques par frontière et mode de transport. La finesse du captage et l'expertise de l'opérateur devaient en effet permettre de discerner les modes de transport, le ferroviaire étant par exemple caractérisé par un nombre élevé de cartes SIM évoluant à la même vitesse et sur un parcours identifié. Dans la réalité, les écarts importants observés entre les données du module préexistant de l'opérateur et les données de cadrage (enquête EVE) ont été tels que la distinction du mode de transport a très tôt été abandonnée. L'approche retenue *in fine* est donc centrée sur le besoin prioritaire de dénombrement des visiteurs étrangers, agrégé pour tous modes de transport et frontières.

D'abord décevante, la qualité des données a progressé

La comparaison entre les indicateurs issus des données de téléphonie et les données d'enquêtes permet d'inférer leur fiabilité et d'étudier les éventuelles sources d'écarts. Plusieurs travaux consacrés à l'analyse de la mobilité et à la construction de matrices origine-destination ont effectué ce type de comparaisons. Les résultats obtenus avec la téléphonie mobile sont dans certains cas proches des résultats d'enquêtes mais d'un niveau plus élevé (Calabrese *et al.*, 2013). Dans le domaine de la mobilité, une étude plus récente (Bonnet *et al*., 2015) a comparé les résultats de l'enquête globale transport avec des estimations issues des données passives de téléphonie fournies par Orange. Les auteurs trouvent une forte corrélation entre ces deux types d'estimations et parviennent à des estimations proches en termes de nombre total de déplacements en Île-de-France (la différence est de 9 %). Leur étude porte cependant sur une période courte (douze jours).

Dans le cadre de l'expérimentation menée par la Banque de France et la DGE, les premières livraisons d'estimations au troisième trimestre 2015 présentaient des écarts très importants avec les estimations de l'enquête EVE, seule source disponible pour avoir une estimation des arrivées et nuitées touristiques en France métropolitaine. L'indicateur issu de la téléphonie mobile indiquait plus de cent millions d'arrivées touristiques au seul troisième trimestre alors que l'ordre de grandeur de la fréquentation touristique est de 85 millions d'arrivées par an.

Il s'est rapidement avéré que les indicateurs construits sur un captage unique sont inutilisables pour différentes raisons détaillées plus loin. Les arrivées en France de touristes et d'excursionnistes (visiteurs ne passant pas la nuit sur le territoire), se sont ainsi avérées de qualité très insuffisante. Il faut donc travailler sur les indicateurs « consolidés »[8], par exemple un nombre de nuitées, chaque nuitée étant définie par une présence confirmée plusieurs fois au même endroit sur une plage horaire définie. Les nuitées ne sont alors pas comptabilisées pour des arrivants préalablement définis par le système de mesure, c'est au contraire le nombre d'arrivées qui est déduit du constat des nuitées. Ceci respecte tout à fait l'esprit et la lettre de la définition internationale du touriste : visiteur dont la visite comprend au moins une nuit sur un territoire qui n'est pas celui de sa résidence habituelle[9]. Enfin, l'objectif de dénombrement des excursions (visites ne comportant pas de nuitée), déjà rendu difficile par l'exclusion des zones frontalières où les antennes sont susceptibles de couvrir une portion de territoire étranger, est rapidement abandonné, ne pouvant être défini avec fiabilité ni directement ni comme un solde. Les travaux d'amélioration se sont donc fondés principalement sur l'analyse des nuitées.

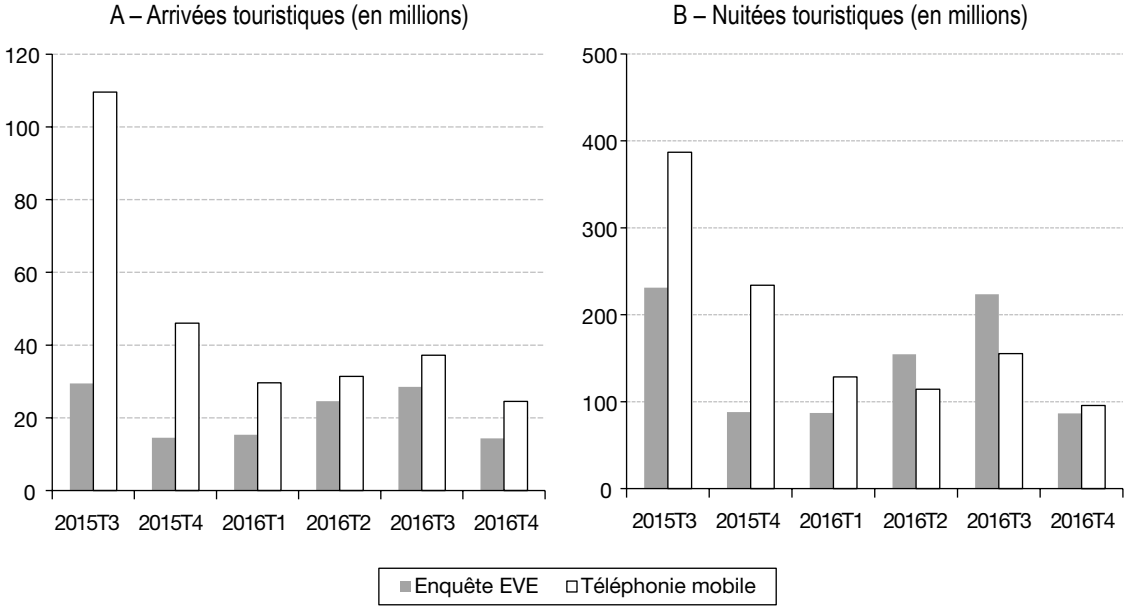
8. Plus généralement, le suivi des observations dans le temps permet de neutraliser une grande partie des défauts du système de mesure, mais les consignes de la CNIL limitent ce suivi à 3 mois consécutifs.
9. Voir le site de l'Organisation mondiale du tourisme : http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base.

Afin d'améliorer la qualité, diverses corrections ont été apportées au cours de la période d'expérimentation et sont décrites plus loin. Elles ont eu pour conséquence de rapprocher les estimations d'arrivées et de nuitées totales assises sur des données de téléphonie mobile de niveaux plus crédibles en comparaison de l'enquête EVE (figure I). L'écart entre les estimations de nuitées totales est ainsi passé de 67 % lors du premier trimestre livré (troisième trimestre 2015) à 10 % pour le dernier trimestre disponible actuellement (quatrième trimestre 2016).

Cependant, la qualité des estimations n'apparaît pas encore suffisante pour trois raisons. En premier lieu, il subsiste d'importantes différences entre les deux sources au niveau des pays ou zones de provenance. Certaines zones proches apparaissent surestimées tandis que les zones lointaines sont sous-estimées. Au 4e trimestre 2016, pour les nuitées touristiques, l'écart global de 10 % entre les données de téléphonie et l'estimation issue de l'enquête EVE résulte de la compensation d'écarts très importants au niveau des pays. Les estimations de nuitées issues de la téléphonie mobile sont ainsi supérieures de 78 % à celles de l'enquête EVE pour l'Allemagne mais inférieures d'environ 80 % pour les États-Unis, le Canada et le Brésil par exemple. Ces écarts proviennent vraisemblablement des valeurs de taux d'utilisation des téléphones portables retenues pour redresser les données ; pour les touristes en provenance de pays lointains, les facteurs de redressement reflètent mal les comportements des visiteurs. Cette limite est inhérente aux données de téléphonie : la qualité des estimations dépend de la connaissance du taux de pénétration de l'opérateur par segment de population. Cette limite a été mentionnée dans plusieurs autres études, y compris lorsque la population d'intérêt est la population résidente (Bonnet *et al*., 2015). Toutefois, contrairement aux définitions des comportements touristiques, les facteurs de redressement qui concernent l'utilisation des téléphones peuvent être modifiés *a posteriori*. Il sera donc possible d'améliorer la qualité des données en progressant dans la connaissance des comportements.

En second lieu, l'estimation des arrivées touristiques est moins robuste que celle des nuitées. Cela est dû au problème des séjours interrompus (voir partie suivante) qui a un effet relativement plus important sur les arrivées que sur les nuitées. Enfin, pour certaines provenances, les estimations issues de la téléphonie affichent une saisonnalité qui est très différente de celle de l'enquête et qui apparaît peu crédible. Ainsi, pour l'Espagne, les nuitées touristiques augmentent de 80 % entre le deuxième et le troisième trimestre d'après l'enquête EVE, ce qui correspond à la saisonnalité observée par les données des professions concernées (trafic

Figure I
Comparaison des mesures par l'enquête EVE et la téléphonie mobile



Champ : arrivées trimestrielles de touristes non-résidents.
Source : Banque de France, DGE.

aérien vers les destinations de loisirs, activité hôtelière, etc.) alors que les données de téléphonie indiquent une hausse beaucoup plus faible de 13 %. Il s'ensuit que la qualité des estimations issues des données de téléphonie mobile est encore insuffisante pour remplacer les données de trafic actuellement utilisées.

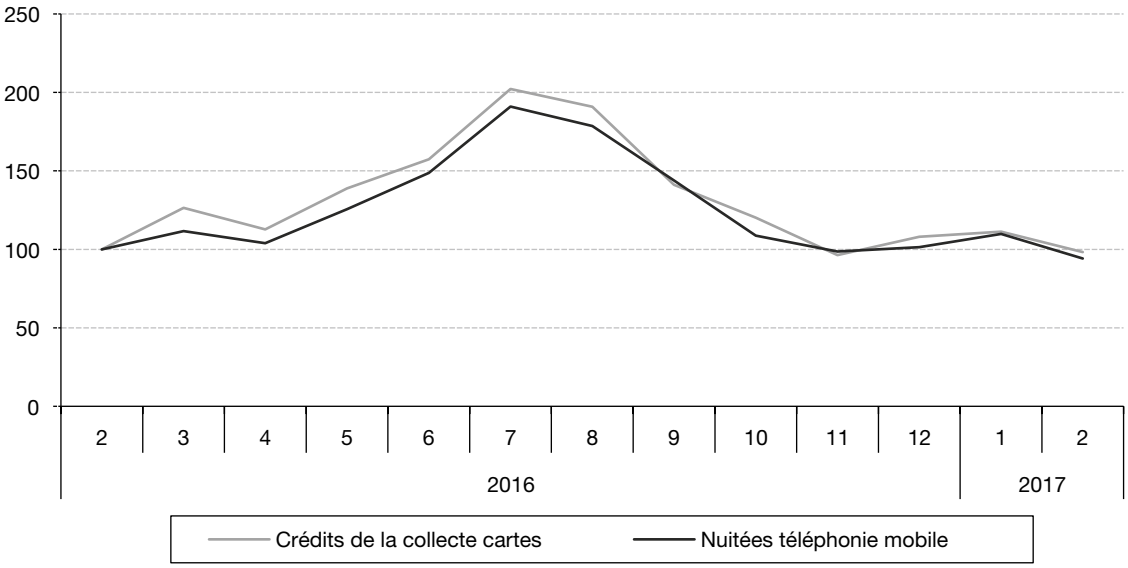
Les données sont potentiellement adaptées au suivi conjoncturel

Comme exposé *supra*, les données de téléphonie mobile ne sont pas encore fiabilisées en niveau. Cela provient principalement des redressements relatifs à l'utilisation des mobiles pour les visiteurs en provenance de zones lointaines et des séjours interrompus artificiellement.

L'analyse de ces données en évolution présente davantage d'intérêt, d'autant plus que ces données sont normalement disponibles avant celles des enquêtes classiques et possèdent un détail chronologique plus fin puisque les estimations journalières sont disponibles. Le potentiel des données de téléphonie pour le suivi de la conjoncture est mis en avant par la comparaison avec les données de cartes de paiements collectées mensuellement par la Banque de France, qui portent sur les paiements et retraits d'espèces effectués en France au moyen d'une carte non-résidente et sont agrégées par pays de contrepartie. Ces dépenses ne correspondent pas exactement aux recettes de tourisme (utilisation par des résidents de cartes étrangères, dépenses réglées en espèces retirées dans le pays d'origine ou prépayées par virement bancaire simple) mais les recoupent en grande partie, notamment pour les ressortissants de pays où l'usage du paiement par carte est prépondérant, voire exclusif. Les données sur les cartes de paiement présentent également l'avantage d'être disponibles en fréquence mensuelle avec une ventilation géographique fine, ce qui permet donc une comparaison avec les données de téléphonie mobile. Sur la période février 2016 - février 2017[10], les recettes issues des cartes de paiements et les nuitées touristiques telles que mesurées grâce aux données de téléphonie mobile sont très corrélées : le coefficient de corrélation entre les deux séries est de 0.986 (figure II). En outre, la corrélation élevée entre les recettes issues de la collecte des données de cartes bancaires et les nuitées touristiques estimées grâce à la téléphonie mobile est aussi observée au niveau des différents pays, malgré quelques exceptions. Ainsi, alors que les estimations de nuitées en niveaux connaissent d'importants écarts avec les résultats de l'enquête EVE pour certains

10. Le choix de commencer la comparaison en février 2016 est motivé par le fait que le dernier changement méthodologique important a provoqué une rupture de série entre janvier et février 2016.

Figure II
Nuitées touristiques estimées avec les données de téléphonie et recettes issues des transactions par cartes de paiements (base 100 en février 2016)



Champ : dépenses en France au moyen de cartes non-résidentes (hors internet) et nuitées des touristes non-résidents.
Source : Banque de France, DGE.

pays (États-Unis, Canada, Brésil par exemple), la corrélation entre nuitées et recettes cartes est élevée pour tous les pays hors Brésil (très faible utilisation des cartes de paiements), Maroc et Luxembourg (en raison de la forte proportion de cartes émises au Luxembourg mais utilisées par les résidents d'autres pays). Pour les autres pays, le coefficient de corrélation entre nuitées estimées par la téléphonie mobile et recettes issues de la collecte cartes varie entre 0.66 et 0.97. En outre, certains pays pour lesquels le niveau des nuitées est très sous-estimé par la téléphonie mobile voient leur évolution conjoncturelle plutôt bien estimée (Canada, États-Unis). La téléphonie mobile fournit donc des estimations crédibles d'évolution, et son utilisation pour des estimations conjoncturelles pourrait être envisageable en attendant le calibrage des estimations en niveaux.

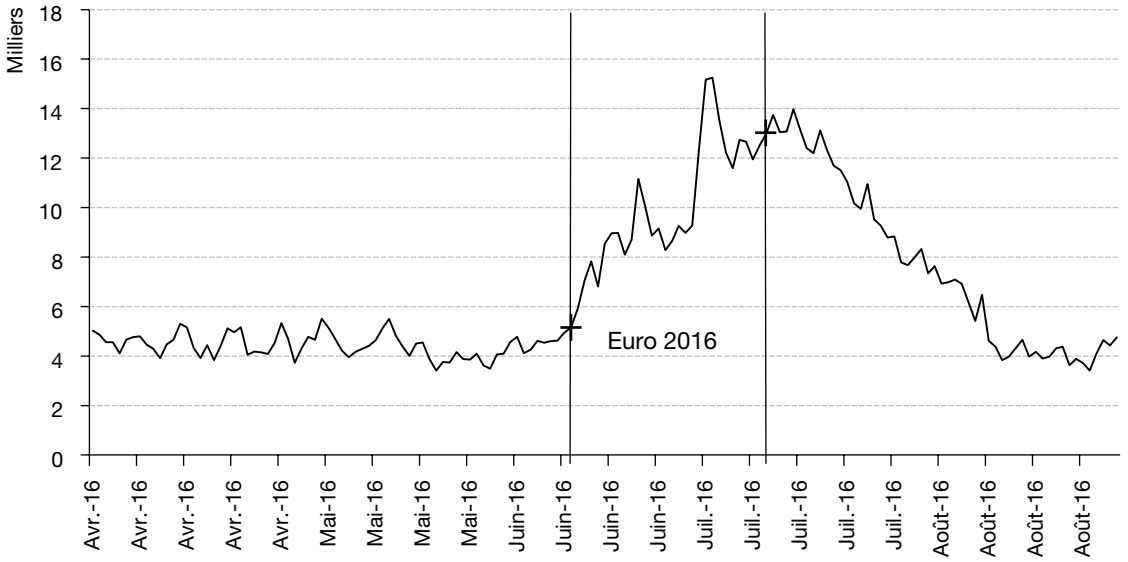
Les données de téléphonie rendent aussi compte des chocs qui affectent la fréquentation touristique. Ceux-ci ne sont pas aussi rapidement identifiables avec l'enquête EVE qui produit des résultats trimestriels. En outre, les données de téléphonie couvrent mieux certaines zones de provenance plutôt rares. L'exemple de la compétition de football Euro 2016 permet d'illustrer cette précision de mesure des données de téléphonie : le bon parcours de l'équipe islandaise se traduit par une hausse graduelle des visiteurs en provenance de ce pays (figure III).

Les sources de biais d'estimation des arrivées et nuitées

Un problème typique de mesure : les connexions sporadiques

La première cause de surestimation des volumes relève des « connexions sporadiques ». Sont ainsi désignées les connexions de téléphones portables qui peuvent apparaître en itinérance sur le réseau Orange alors qu'ils n'utilisent pas ce réseau de façon privilégiée. La présence de ces connexions sur le réseau ne correspond pas à une hypothèse de redressement utilisée dans l'algorithme préexistant du prestataire. En effet, l'opérateur observe d'abord les mobiles en itinérance présents sur son réseau, qu'ils soient actifs ou non, puis effectue un redressement relatif à sa part de marché pour en déduire le nombre total de mobiles en itinérance en France, quel que soit le réseau. La clef de ce redressement est une part de marché mesurée sur le volume de SMS des mobiles en itinérance, par couple pays-opérateur. Ce redressement est pertinent si la répartition des mobiles en termes de présence sur le réseau est égale à celle des volumes de SMS. Or ce n'est pas forcément le cas, notamment en raison des accords préférentiels qui peuvent lier des opérateurs nationaux à des opérateurs étrangers. Le téléphone mobile d'un touriste étranger (abonné dans son pays P1 auprès de l'opérateur

Figure III
Nuitées quotidiennes des touristes islandais et norvégiens



Champ : nuitées touristiques quotidiennes des touristes résidant en Norvège et en Islande en voyage en France.
Source: Banque de France, DGE.

E1P1) en France peut ainsi être prioritairement capté par une antenne de l'opérateur français F1 ayant des accords préférentiels avec E1P1, si l'état du réseau le permet. Mais il peut aussi être capté par une antenne d'un autre opérateur français F2 si l'état du réseau privilégié est insuffisant. Les connexions sporadiques sont ainsi celle d'un touriste non-résident (probablement abonné d'un opérateur ayant un accord préférentiel avec un autre opérateur français F1) capté de manière sporadique sur le réseau F2, par exemple pendant la traversée d'une zone « blanche » du réseau F1. Si ce touriste n'utilise pas son mobile activement sur cette courte période, il n'est pas représenté dans les volumes qui permettent de calculer les parts de marché. La clef de redressement est donc sous-estimée et l'extrapolation est effectuée avec un coefficient trop fort, ce qui implique une estimation trop élevée pour le nombre de cartes SIM du pays concerné sur la période et le territoire considérés. Ce problème a rendu nécessaires plusieurs changements des critères de mesure. Ils sont détaillés dans la partie suivante.

Les interruptions de captage, facteur de sous-estimation de la durée moyenne de séjour et de surestimation des arrivées

Une seconde difficulté de mesure est en partie liée à la première. Il s'agit des interruptions artificielles de séjours. Les comptages disponibles au printemps 2016 indiquent des arrivées nettement excessives et des nuitées d'un ordre de grandeur compatible avec les données de cadrage, impliquant une durée de séjour moyenne trop faible. Afin d'améliorer les mesures des durées de séjour et du nombre des arrivées, la Banque de France et la DGE ont demandé que soit étudiée l'intuition suivante : les séjours seraient artificiellement abrégés par des interruptions de captage. Ces interruptions peuvent avoir de multiples causes : téléphone portable déchargé ou volontairement éteint, passage dans une zone non couverte par le réseau Orange, etc. La surestimation induite des arrivées touristiques est mécanique : à la reprise du captage, si l'interruption a inclus une nuitée, la carte SIM est considérée comme nouvelle arrivante. Le phénomène entraîne également une sous-estimation des nuitées mais de moindre ampleur. Par exemple, lors d'un séjour touristique classique d'une semaine sur le territoire français, une carte SIM peut être captée régulièrement pendant trois jours, ne plus être captée pendant deux jours et être à nouveau captée pendant deux jours avant de quitter le territoire. Le dispositif préexistant de l'opérateur considérera alors qu'il y a eu deux arrivées touristiques, la première correspondant à un séjour de trois nuitées et la suivante à un séjour de deux nuitées.

Afin d'évaluer cette hypothèse, un premier test a été réalisé en début d'année 2016, sur un périmètre et une période réduits et propices à l'analyse : une station de montagne. L'avantage de ce périmètre est que le comportement touristique y est relativement bien connu : clientèle étrangère, part importante des séjours démarrant un samedi et s'achevant le samedi suivant, zone de captage définie sans ambigüité du fait des barrières naturelles. Il est ressorti de cette observation que la part des séjours touristiques concernés par les interruptions artificielles pouvait être très élevée. Ce résultat est conforté par une autre observation à l'échelon national, sur deux périodes d'observation : mars 2016 et septembre-novembre 2016. Pour ces périodes, il a été possible d'étudier le volume des arrivées en fonction de la contrainte d'absence qui doit précéder une arrivée. Cette contrainte est normalement fixée à un jour mais le but de l'observation est justement de déterminer si les arrivées enregistrées par l'algorithme sont de réelles arrivées ou des arrivées artificielles de personnes déjà présentes sur le territoire. Sur le mois de mars, le durcissement de la contrainte d'absence (deux jours d'absence avant une arrivée) fait diminuer les arrivées totales de 13 %. Si la contrainte est portée à six jours d'absence, le volume d'arrivées diminue de 37 %.

La sous-estimation de mesure est donc confirmée sans qu'il soit possible de la corriger, faute de pouvoir distinguer les séjours artificiellement interrompus des véritables courts séjours récurrents, alors que cette distinction est nécessaire pour reconstituer le volume des nuitées. Or, le durcissement de la condition d'absence ne va pas de soi et conduirait à s'éloigner de la définition statistique du tourisme. Au-delà de l'impact sur les agrégats, qui est important si l'on exclut par exemple toute arrivée moins de trois jours avant la précédente, un problème plus essentiel est posé : faut-il se satisfaire du raisonnement probabiliste consistant à corriger une mesure insatisfaisante en adaptant des définitions ? Une intervention directe sur les données source, un repérage physique des anomalies et une correction préalable à l'élaboration des agrégats seraient préférables mais ne sont pas réalisables pour le moment. En effet, si en termes de comportement le repérage des cas

problématiques semble sans ambiguïté (sortie brutale du réseau le plus souvent à distance d'une frontière ou sur une trajectoire incompatible avec une sortie effective du territoire, retour tout aussi brutal), ce repérage n'est pas compatible avec le système de mesure préexistant de l'opérateur. La question des séjours interrompus est mentionnée dans l'étude faite par l'Estonie sur le suivi du tourisme international (Kroon, 2012). Les auteurs pallient la difficulté en adoptant des hypothèses. Ils considèrent que le mobile était présent si l'inactivité du mobile est d'une durée inférieure à 7 jours et qu'il y a eu un départ si l'inactivité est d'une durée supérieure. Une telle solution n'est pas optimale dans le cas de la France parce que le phénomène de transit y est important.

Pays d'émission de la carte SIM et pays de résidence peuvent différer

Une troisième difficulté de mesure est liée à des comportements qui affaiblissent l'hypothèse selon laquelle le pays de la carte SIM est identique à celui de résidence du détenteur du téléphone portable. Identifiée dès le début de l'expérimentation par analogie avec les travaux menés sur les touristes français, cette difficulté est en partie corrigée.

Le passage d'un nombre de cartes SIM à un nombre de touristes n'est pas simple

Enfin, la dernière difficulté concerne la phase ultime de redressement, qui intervient *a posteriori*, indépendamment du dispositif préexistant de l'opérateur pour répartir les arrivées et les nuitées selon le bassin émetteur des touristes. Comme mentionné dans la partie relative à la qualité des données, les facteurs de redressement qui permettent d'extrapoler des volumes de touristes à partir des volumes de téléphones ne sont pas toujours adaptés. Ces données sur les taux d'équipement dans les différents pays proviennent de GSM Alliance. Mais l'utilisation de ces données sur l'équipement des populations est fortement limitée par, d'une part, la différence de représentativité entre la population totale d'un pays et la fraction de cette population visitant la France et d'autre part, par les comportements spécifiques en situation de voyage à l'étranger. Les taux d'utilisation peuvent en effet varier selon les tarifs pratiqués par les opérateurs du pays concerné et selon des facteurs socio-culturels (populations plus ou moins sensibles aux questions de sécurité et aux habitudes de connexion plus ou moins intensives).

Faute de données exogènes sur les taux d'utilisation, le prestataire combine les données des taux d'équipement à plusieurs jeux de coefficients déterminés par grands groupes de pays selon leur éloignement.

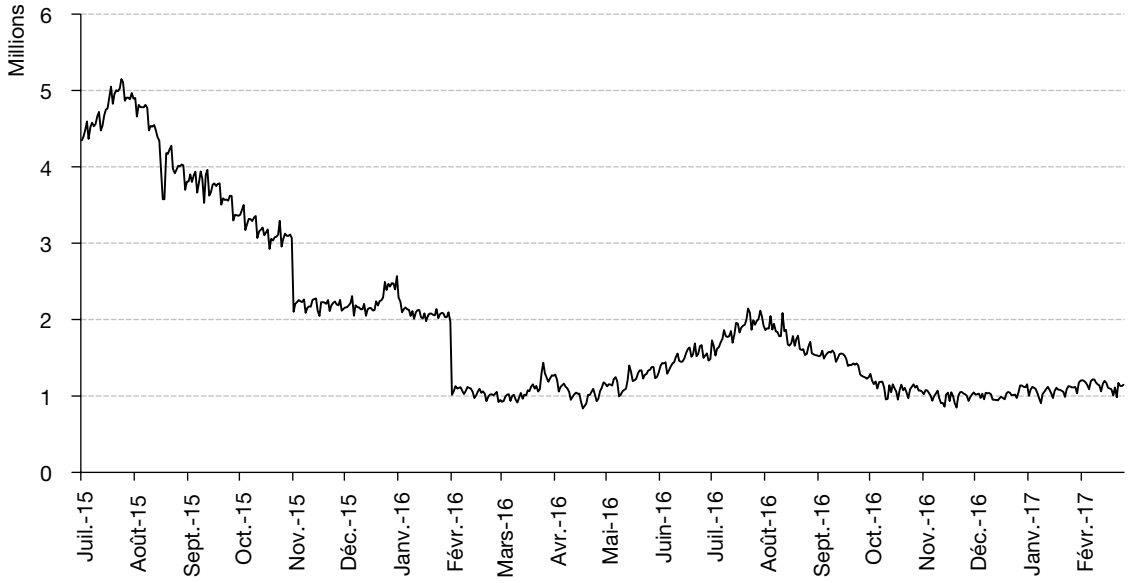
Cette difficulté de redressement, si elle n'empêche pas l'analyse en évolution[11] d'indicateurs pays par pays, rend difficile l'utilisation d'indicateurs agrégés. En admettant que l'évolution des nuitées touristiques pour les résidents du pays A soit mesurée convenablement et qu'il en soit de même pour le pays B, l'évolution des nuitées pour les résidents de l'ensemble des deux pays ne peut être calculée en l'absence de données exogènes sur les poids respectifs de ces deux pays dans la fréquentation touristique de la zone considérée. Ainsi, outre les volumes par pays qui sont dans certains cas très sous-estimés, les données en évolution sont affectées dès lors que l'on s'intéresse à un ensemble de nationalités.

Corrections apportées, bénéfices obtenus et limites

La première correction apportée a consisté à introduire une contrainte de fidélité des téléphones portables au réseau de l'opérateur afin de réduire le bruit dû aux connexions sporadiques de téléphones qui ne se connectent au réseau qu'en cas de perte de leur réseau privilégié. Comme indiqué *supra*, la comptabilisation de ces téléphones provoque une surestimation du nombre de nuitées et d'arrivées en raison du redressement que l'opérateur effectue relativement à sa part de marché sur la clientèle en itinérance. Afin de distinguer les utilisateurs réguliers et les utilisateurs occasionnels, le premier critère introduit concerne le temps de présence cumulé sur le réseau qui doit être supérieur à 9 heures sur 21 heures. Ce premier critère a été ajouté pour la livraison des données à partir du mois de novembre 2015 et a provoqué une baisse de l'estimation des nuitées touristiques d'environ 30 % (voir figure IV). Il a ensuite été renforcé par l'ajout d'une nouvelle contrainte de fidélité pour la livraison

11. Sur des périodes courtes à tout le moins : l'analyse de longue période suppose une stabilité des comportements d'utilisation des téléphones portables.

Figure IV
Nuitées touristiques estimées à partir de données de téléphonie mobile



Champ : nuitées quotidiennes des touristes non-résidents en voyage en France.
Source: Banque de France, DGE.

des données à compter de février 2016. Cette contrainte, toujours utilisée, impose la réalisation d'au moins trois évènements réseau sur les 24 heures qui précèdent ou suivent la nuitée touristique comptabilisée. Malgré ces améliorations successives, les connexions sporadiques de mobiles continuent d'introduire du bruit dans la mesure. Les travaux actuels s'orientent vers la sélection des opérateurs du pays d'origine des visiteurs en fonction de leur fidélité au réseau lors de l'itinérance en France.

La seconde correction de mesure importante concerne les résidents français qui utilisent un téléphone portable avec une carte SIM étrangère, ce qui peut notamment être le cas des travailleurs frontaliers (résidents français travaillant à l'étranger) qui seraient détenteurs d'un abonnement téléphonique auprès d'une société étrangère. Ce comportement limite quelque peu la validité de l'hypothèse selon laquelle le pays de résidence de l'utilisateur est identique à celui de sa carte SIM et conduit à surestimer les nuitées touristiques. La correction apportée s'est appuyée sur la contribution importante du groupe de travail piloté par Tourisme & Territoires[12] à propos de la segmentation de la population observée. Elle est incontournable pour les données des résidents puisque la fréquence et la durée des déplacements permettent de répartir les individus en différentes catégories. Le nombre de nuits passées sur un territoire donné permet ainsi de considérer qu'un porteur de téléphone portable réside dans ce territoire, indépendamment des caractéristiques de sa carte SIM. Appliquée de manière plus sommaire aux cartes SIM portant un code étranger, cette segmentation permet d'écarter les individus qui passent plus de la moitié de leurs nuitées en France sur une période de deux mois. Le prestataire a donc introduit une condition de non résidence dans l'algorithme de traitement des données de téléphonie mobile. Les personnes qui ont passé plus d'un mois sur le territoire au cours des deux derniers mois sont considérées comme résidentes et donc exclues de la mesure du tourisme international, ce qui couvre bien le cas des travailleurs frontaliers.

En raison de l'apprentissage nécessaire, la première livraison de données prenant en compte cette correction est celle de février 2016. Cette correction ajoutée au durcissement de la contrainte de fidélité au réseau mobile a réduit l'estimation totale des nuitées d'environ 50 %. L'impact sur les nuitées est nettement plus important que sur les arrivées. La durée moyenne d'un séjour touristique en France est en effet de 6.8 jours toutes clientèles internationales confondues (DGE, 2017), alors que les

12. Voir : http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/.

résidents passent la quasi-totalité des nuitées en France. La correction est cependant imparfaite puisqu'elle conduit à exclure de la mesure les touristes qui restent en France pour un séjour d'une durée supérieure à un mois alors que la définition statistique comprend les séjours d'une durée allant jusqu'à une année.

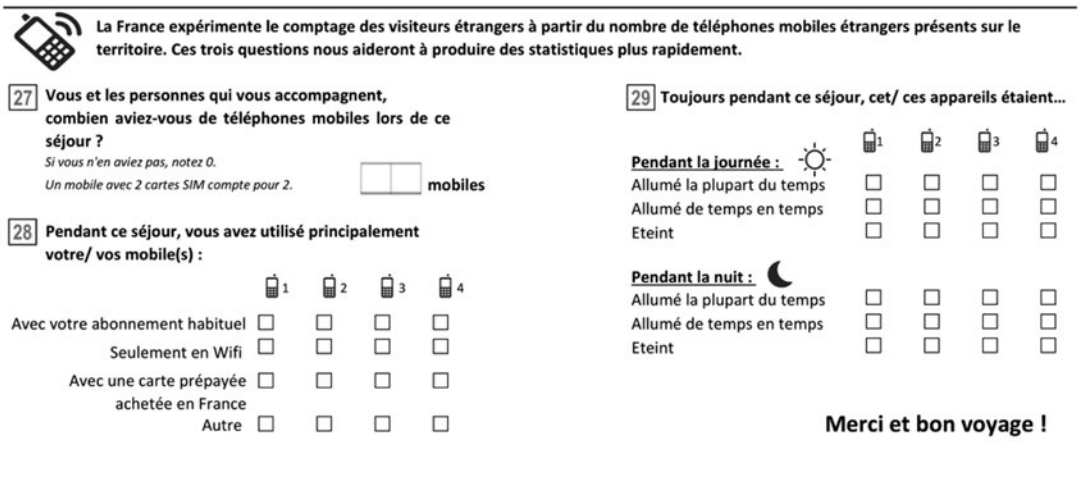
La contrainte d'anonymisation empêchant de sauvegarder les données individuelles de connexions entre mobiles et antennes, la modification des critères de définition des comportements pertinents pour le suivi du tourisme ne peut pas être rétropolée. L'effet des corrections décrites peut donc être observé sur la figure IV sous la forme des ruptures de série de novembre 2015 et février 2016.

L'enjeu spécifique du redressement comportemental rend nécessaire une collecte exogène

La connaissance des comportements d'utilisation des téléphones mobiles par les visiteurs étant insuffisante pour permettre de redresser de façon satisfaisante le comptage des cartes SIM, la Banque de France et la DGE ont décidé d'intégrer au questionnaire de l'enquête EVE un bloc de questions relatif à l'utilisation des téléphones mobiles. Ces questions (encadré) ont été ajoutées en janvier 2017 et les premiers résultats pourront être analysés à la fin de l'année. Il sera alors possible de déterminer un coefficient pour chacun des principaux pays de résidence et d'améliorer ainsi le redressement par pays de provenance. Cependant, si les comportements d'utilisation s'avèrent d'une grande variabilité temporelle et spatiale, l'utilisation des données de téléphonie mobile sera plus coûteuse en raison de la nécessité d'une collecte dédiée au redressement. Cela pèserait donc sur le bilan coût-avantage qui devra être effectué en fin d'expérience, pour prendre la décision d'intégrer ou non ces données dans le processus courant de production.

Parmi les difficultés de redressement liées au taux d'équipement et d'utilisation, la composition familiale des groupes de touristes peut probablement jouer un rôle important : il ne s'agit pas seulement d'estimer le nombre de porteurs de mobiles mais aussi le nombre d'accompagnants. L'impact de la composition du groupe n'est d'ailleurs pas une difficulté propre à la mesure de la fréquentation touristique étrangère. Des données de cadrage sont disponibles pour les touristes français : selon une étude adossée au dispositif permanent de suivi de la demande touristique (SDT), réalisée par la DGE et la Banque de France au printemps 2015, si le taux de possession du mobile est assez uniforme pour les résidents âgés de 15 ans et plus, il varie très fortement jusqu'à l'âge de quinze ans. Le nombre de touristes de moins de quinze ans accompagnant un touriste de plus de quinze ans dépend en outre fortement de la période (vacances scolaires ou non), du type d'hébergement (hôtel/camping/location) et donc de la zone. La prise en compte de cet impact sera également une étape importante

ENCADRÉ – **Questions dédiées à l'utilisation du téléphone mobile, collecte EVE 2017**



La France expérimente le comptage des visiteurs étrangers à partir du nombre de téléphones mobiles étrangers présents sur le territoire. Ces trois questions nous aideront à produire des statistiques plus rapidement.

27 **Vous et les personnes qui vous accompagnent, combien aviez-vous de téléphones mobiles lors de ce séjour ?**
Si vous n'en aviez pas, notez 0.
Un mobile avec 2 cartes SIM compte pour 2.
☐☐ **mobiles**

28 **Pendant ce séjour, vous avez utilisé principalement votre/ vos mobile(s) :**

	1	2	3	4
Avec votre abonnement habituel	☐	☐	☐	☐
Seulement en Wifi	☐	☐	☐	☐
Avec une carte prépayée achetée en France	☐	☐	☐	☐
Autre	☐	☐	☐	☐

29 **Toujours pendant ce séjour, cet/ ces appareils étaient…**

	1	2	3	4
Pendant la journée :				
Allumé la plupart du temps	☐	☐	☐	☐
Allumé de temps en temps	☐	☐	☐	☐
Eteint	☐	☐	☐	☐
Pendant la nuit :				
Allumé la plupart du temps	☐	☐	☐	☐
Allumé de temps en temps	☐	☐	☐	☐
Eteint	☐	☐	☐	☐

Merci et bon voyage !

dans l'amélioration de la mesure de la fréquentation française. Le ratio touristes/cartes SIM est nécessairement plus élevé dans une zone de camping à la fréquentation familiale au cœur de la saison estivale que, hors vacances scolaires, dans une zone où prédomine le tourisme professionnel et ses représentants dotés de plusieurs mobiles, voire de mobiles dotés de plusieurs cartes SIM ; la difficulté est alors de disposer de facteurs de redressement adaptés et suffisamment fins.

Les évolutions futures et les incertitudes

La fin des frais d'itinérance dans l'Union européenne

La suppression de la refacturation aux abonnés des coûts d'itinérance au sein de l'Union européenne est effective depuis juin 2017. Les opérateurs avaient commencé il y a plusieurs années à limiter cette refacturation sur certaines destinations. En conséquence, les comportements des touristes devraient s'homogénéiser au sein de l'Union européenne, chacun étant supposé à terme utiliser son téléphone portable comme s'il était dans son pays de résidence habituelle. Cette situation améliorerait la précision des estimations pour les pays européens mais la phase de transition entre les habitudes actuelles et l'homogénéité escomptée risque d'induire du bruit dans les estimations, la hausse attendue de l'utilisation du téléphone portable en itinérance à l'étranger risquant d'entraîner une hausse artificielle de fréquentation mesurée.

Ce risque doit cependant être relativisé puisque le système ne repose pas seulement sur l'utilisation effective du téléphone portable mais aussi sur des connections passives au réseau, qui ne font pas l'objet de facturation ; l'impact ne sera donc pas forcément élevé. De plus, certains opérateurs ayant déjà commencé à ne plus facturer les frais d'itinérance à leurs abonnés sur les destinations internes à l'Union européenne, la phase de transition est en fait amorcée depuis de nombreux mois et son impact sera lissé. Les données collectées auprès des touristes dans le cadre d'EVE devraient permettre de mesurer ces changements de comportement.

Si la clientèle européenne assure environ 79 % des arrivées touristiques étrangères en France (DGE, 2016b), les clientèles lointaines constituent cependant une part non négligeable et croissante de la fréquentation. Pour ces clientèles, les frais d'itinérance sont susceptibles de rester élevés et de continuer à dissuader une grande partie des touristes d'utiliser leurs cartes SIM d'origine.

La connexion en Wifi

Des pratiques spécifiques de connexion peuvent limiter la portée du dispositif de mesure basé sur les réseaux de téléphonie mobile. Ainsi, certains touristes, par exemple nord-américains et chinois, privilégieraient d'ores et déjà la recherche de spots Wifi et la connexion à un réseau internet, afin d'utiliser des applications spécifiques de communications vocales sans solliciter le réseau de téléphonie mobile. Le recours à ces applications, populaires chez les jeunes voyageurs et les technophiles, échappe à ce jour à la mesure. De plus, la mise à disposition d'une connexion Wifi est identifiée par les sites touristiques comme un élément d'attractivité et les solutions techniques fleurissent, par exemple sur certaines zones littorales. Les questions dédiées ajoutées à l'enquête EVE début 2017 devraient permettre de mieux mesurer l'ampleur du phénomène.

Les abonnements supranationaux

Autre évolution affectant non plus l'exhaustivité mais la précision du dispositif, le développement des abonnements supranationaux est susceptible de distendre le lien entre le pays de la carte SIM captée et celui de résidence de son utilisateur. Ce développement pourrait être encouragé par la suppression de la facturation de l'itinérance au sein de l'Union européenne (cf. *supra*), quoique celle-ci soit accompagnée de restrictions. Ces restrictions visent à dissuader les usages extrêmes (caractérisés par une consommation minoritairement située dans le pays d'émission du forfait). La difficulté de déduire la résidence de l'utilisateur de la nationalité de la carte SIM existe déjà. En attestent les comptages très élevés des nuitées supposément luxembourgeoises en France : ce constat, qui fut l'un des premiers de l'expérimentation menée par la Banque de France et la DGE, a résisté aux améliorations successives de la méthode. La seule explication réside donc dans la commercialisation par des sociétés luxembourgeoises de forfaits utilisés par des résidents d'autres pays, ce qui est sans doute à rapprocher du nombre de travailleurs frontaliers au Luxembourg, qui peuvent résider en France mais aussi en Belgique ou en Allemagne.

Bilan de l'expérimentation et pistes d'amélioration pour l'avenir

Le type de partenariat mis en place et la démarche de travail apparaissent adaptés aux spécificités de ces données massives

L'expérimentation suggère quelques clés pour un partenariat réussi entre une entreprise privée dépositaire de données massives et des institutions responsables de l'élaboration de statistiques officielles. Pour l'expérience en cours, les travaux se sont construits à partir d'un module de mise en forme des données déjà éprouvé, mais pour servir des besoins différents de ceux exprimés ici (analyse d'événements sur un territoire limité – ville, département, etc. – *versus* évaluation en niveau des flux de visites et des durées de séjour sur l'ensemble du territoire métropolitain). L'avantage est que cela a permis de disposer de jeux de données en tout début de partenariat, ce qui a favorisé une démarche empirique et une confrontation avec les données de référence dont disposent la Banque de France et la DGE. L'inconvénient réside dans la faible significativité des premiers résultats, dans un contexte où les traitements en amont (phases 2 à 4) relevaient de la seule expertise du fournisseur. Cet obstacle a pu être surmonté par la mise en place d'une démarche de co-développement. Ceci implique un engagement de moyens proportionnés et équilibrés entre les parties, d'où l'importance d'une structure de pilotage du partenariat qui mobilise à la fois les experts mais aussi le niveau décisionnel approprié. Dans ce contexte, la capacité des deux parties à introduire des adaptations rapidement (méthode agile) s'est avérée essentielle. Par exemple, la Banque de France et la DGE ont décidé d'intégrer dans le questionnaire d'enquête 2017 un module permettant de collecter des variables sur le comportement d'utilisation du téléphone mobile qui, si elles s'avèrent robustes, amélioreront les possibilités d'exploitation des données de téléphonie à des fins statistiques.

Approche en co-développement et méthode agile apparaissent en outre adaptées aux spécificités des Big Data : l'observation d'événements très fréquents sur l'ensemble du territoire métropolitain mobilise des capacités de calcul plus élevées que celles requises pour les traitements actuels d'où l'importance de pouvoir disposer d'une forte réactivité pour adapter les capacités de calcul. Certaines adaptations de définitions passent par une phase d'examen de variantes, inhérente à ce type d'expérimentation. La stabilisation des définitions, nécessaire à la construction de séries et à leur confrontation avec les sources existantes, ne peut donc être atteinte d'emblée et les chercheurs doivent accepter d'interpréter des résultats successifs intégrant des ruptures de méthode, ce qui nécessite une forte interaction entre eux. Un test par échantillonnage ou par la restriction de l'expérimentation à un territoire limité aurait permis d'atténuer cette difficulté mais n'aurait pas répondu à l'objectif de l'exhaustivité de mesure sur l'ensemble du territoire qui constitue un des principaux apports attendus de ces données.

Les avancées obtenues en cours d'expérimentation ouvrent des possibilités d'utilisation pour le suivi conjoncturel de court terme et la « régionalisation » des données nationales de tourisme

La mise en place des traitements des données apporte des résultats adaptés au suivi conjoncturel de la fréquentation touristique dans le court terme et aux mesures de chocs dans le cas d'événements ponctuels (manifestation sportive, festival, attentat, comparaison entre des populations en période haute et période basse). Hors période de mise en place, la rapidité de mise à disposition des données issues de la téléphonie (moins de trente jours avant la fin du mois observé) est un atout incontestable face aux modes de collectes par enquête tout en étant comparable à ceux des données de carte bancaire. Ces utilisations nécessitent certaines précautions, notamment le recours à des indicateurs d'évolution plutôt qu'à des indicateurs de volumes. Second domaine d'intérêt, les traitements statistiques issus de l'expérimentation qui devraient fournir, pour chacun des principaux pays de résidence, une répartition satisfaisante des nuitées selon les treize régions métropolitaines.

Une utilisation pérenne suppose toutefois d'autres améliorations

Afin de permettre une diffusion par les différents utilisateurs, les travaux futurs doivent permettre des améliorations notables dans deux domaines. Le premier concerne la réduction du bruit global de la mesure. Cela relève du cœur du système préexistant de l'opérateur, qui a servi de base de départ à l'expérimentation et implique des changements sur les algorithmes

de base. Le contournement d'une qualité insuffisante des données brutes par des adaptations de définitions des comportements prédéfinis ne peut être jugé comme une solution convenable. Le second domaine relève de la connaissance des comportements d'utilisation des téléphones mobiles. La création d'une base de données exogène sur les taux d'utilisation du téléphone portable des visiteurs étrangers en France, segmentée selon les principaux pays de provenance, est indispensable. Le coût de collecte de données exogènes de bonne qualité constitue dans ces conditions un des paramètres entrant dans l'évaluation de l'intérêt du recours aux données de téléphonie mobile. Ces dernières devant initialement alléger, ou se substituer, à la collecte de données exogènes relatives au trafic global des différents modes de transport[13], il serait peu efficient de déployer un dispositif de collecte important pour caler les données recueillies sur les données dont elles devaient être les alternatives.

À court terme, et pour s'inscrire dans la démarche adoptée visant à obtenir des progrès visibles selon des jalons relativement proches, les parties prenantes de l'expérimentation ont entrepris de tester une méthode promettant de concilier une meilleure maîtrise de la qualité de l'information brute et la conservation des algorithmes de base : pour limiter le bruit lié aux connections sporadiques et aux trop courts séjours, il s'agit de mesurer un taux de sporadicité pour chacun des opérateurs étrangers, pour ne plus conserver que les opérateurs les plus fidèles au réseau Orange. Point fort de cette sélection, elle ne sera pas définie *a priori* sur la base des accords d'itinérance préférentiels, mais sera mesurée sur le terrain. Elle sera aussi évolutive, la liste des opérateurs intégrant les comptages devant être régulièrement mise à jour. Se posera la question de la représentativité des différents opérateurs, le profil de la clientèle pouvant être plus ou moins marqué selon que l'opérateur est *low cost*, historique, ciblé sur les technophiles, etc. Séduisante sur le principe, cette nouvelle version n'a pas encore pu être évaluée.

Atteindre les objectifs initiaux de l'expérimentation conduit à développer des traitements qui distendent le lien entre les données massives et la série statistique produite

Les données de téléphonie mobile ne permettent pas à ce jour de consolider les données de trafic sortant du territoire de France métropolitaine. Elles ne peuvent se substituer aux données de trafic et le dispositif EVE doit donc être maintenu dans son architecture actuelle.

Les solutions envisagées pour améliorer la qualité des estimations privilégient des stratégies d'échantillonnage. La sélection des opérateurs étrangers aux connexions les moins sporadiques relève de ce domaine. Le suivi d'individus volontaires afin de mieux connaître les comportements serait aussi une solution envisageable. Les opérateurs de téléphonie mobile maîtrisent le suivi d'un échantillon d'utilisateurs volontaires et commercialisent plus fréquemment ce genre d'étude que les études portant sur des populations entières. Ces études échappent à une partie des inconvénients rencontrés dans les tentatives de mesures exhaustives, notamment le volume des données et la rigidité des algorithmes d'anonymisation. Dans le cas du tourisme, une telle méthode permettrait d'obtenir des résultats détaillés sur les comportements de mobilité, en particulier la fréquence, la durée et la destination des voyages. Pour les opérateurs disposant d'un réseau sur un ou plusieurs pays frontaliers, un suivi de part et d'autre de la frontière serait même envisageable. Au-delà des aspects statistiques (extrapolation à la population totale des comportements observés sur un échantillon de personnes volontaires), l'adoption d'une telle méthode nécessite un cadre juridique approprié au traitement de données sur des personnes physiques.

L'idée d'envisager une approche par échantillonnage constitue, d'une certaine manière, un aboutissement paradoxal de l'expérience, dont la motivation de départ était d'exploiter une source de données exhaustives de manière simple. Aller dans cette direction suppose d'évaluer la pérennité d'une telle approche, et de sa transparence, compte tenu de la rapidité de l'évolution des technologies et des comportements associés. Cela pourrait générer des coûts élevés pour la maintenance de la base de sondage, dans un contexte où la statistique publique est redevable à ses utilisateurs d'une information claire sur ses méthodes et sur les évolutions de celles-ci.

---

13. Ces données et notamment celles de l'enquête Cerema pour le mode routier, fournissent une référence pour déterminer le plan de sondage et permettent de calibrer les traitements statistiques assurant la représentativité des données du questionnaire. À noter que le questionnement des visiteurs reste, lui, indispensable à la connaissance de leurs comportements touristiques : dépenses par nature, type d'hébergement et d'activités.

Conclure à la nécessité de mettre en œuvre une stratégie de traitement par échantillonnage de données massives revient à renoncer à une de leurs potentialités présumées, à savoir l'obtention rapide à partir de sources brutes et exhaustives de résultats très représentatifs et facilement explicables. C'est, d'une certaine manière, les dénaturer pour les transformer en données classiques, c'est-à-dire dont l'utilisation va de pair avec des coûts d'acquisition et de retraitement.

* *
*

L'expérimentation menée par la Banque de France et la DGE conduit donc à considérer les données de téléphonie mobile comme une source complémentaire d'informations et non comme une source susceptible de remplacer les collectes existantes pour le moment. Cette conclusion était également celle d'Eurostat dans son rapport de 2014 sur les données de téléphonie. Les utilisations les plus pertinentes de ces données dans le contexte du suivi du tourisme international en France sont l'analyse conjoncturelle et la régionalisation des données de l'enquête EVE. Le suivi du tourisme international en France représente cependant un contexte d'étude bien particulier en raison, d'une part, de la taille de la population d'intérêt et de son hétérogénéité (modes de transport, pays de provenance, comportements relatifs à la téléphonie mobile) et, d'autre part, des caractéristiques du territoire (frontières, superficie, phénomènes de transit et de travail frontalier). L'utilisation des données de téléphonie pour la production des statistiques de tourisme en niveau reste envisageable. Elle est subordonnée à une amélioration des algorithmes et à une meilleure connaissance des comportements des visiteurs en matière d'utilisation des mobiles. □

BIBLIOGRAPHIE

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Million, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C.: Emerging Technologies,* 43(2), 198–211.
http://dx.doi.org/10.1016%2Fj.trc.2013.11.007

Ahas, R., Aasa, A., Roose, A., Mark, Ü. & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486.
https://doi.org/10.1016/j.tourman.2007.05.014

Akin, D. & Sisiopiku, V. P. (2002). *Estimating Origin-Destination Matrices Using Location Information from Cellular Phones.* Puerto Rico, USA: Proc. NARSC RSAI.
https://s3.amazonaws.com/academia.edu.documents/7109318/PuertoRicopaper_finall.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires-=1549286310&Signature=h7qw7eNszCA7KWGLEd4lvSaLzWw=&response-content-disposition=inline;filename=Estimating_origin_destination_matrices_u.pdf

Banque de France (2015). *Méthodologie – La balance des paiements et la position extérieure de la France*.
https://www.banque-france.fr/sites/default/files/media/2016/11/16/bdp-methodologie_072015.pdf

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M. & Smoreda, Z. (2015). Passive Mobile Phone Dataset to Construct Origin-Destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, 11, 381–398.
https://doi.org/10.1016/j.trpro.2015.12.032

Calabrese, M., Di Lorenzo, Liu, L. & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing,* 10(4), 36-44.
http://dx.doi.org/10.1109/mprv.2011.41

Calabrese, M., Di Lorenzo, G., Ferreira Jr., J. & Ratti, C. (2013). Understanding Individual Mobility Patterns From Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies*, 26, 301–313.
https://doi.org/10.1016/j.trc.2012.09.009

DGE (2016a). *Chiffres clés du tourisme.* Édition 2016.
https://www.entreprises.gouv.fr/files/files/directions_services/etudes-et-statistiques/stats-tourisme/chiffres-cles/2016-Chiffres-cles-tourisme-FR.pdf

DGE (2016b). *Le 4 pages de la DGE*, N° 60.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-60-touristes-etrangers-france-2015

DGE (2017). *Le 4 pages de la DGE*, N° 71.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-71-touristes-etrangers-france-2016

Eurostat (2014). Feasibility Study on the Use of Mobile Phone Positioning Data for Tourism Statistics. *Consolidated Report Eurostat Contract* N° 30501.2012.001-2012.452
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Gonzales, M. C., Hidalgo, C. A. & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature,* 453(7196), 779–782.
https://doi.org/10.1038/nature06958

Kroon, J. (2012). Mobile Positioning as a Possible Data Source for International Travel Service Statistics. United Nations, Economic Commission for Europe, Geneva, Switzerland, 31 October-2 November 2012, *Seminar on New Frontiers for Statistical Data Collection*.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

ONS (2016). Statistical uses for mobile phone data: literature review. *Methodology working paper series* N° 8.
https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno8statisticalusesformobilephonedataliteraturereview

Organisation mondiale du tourisme. *Comprendre le tourisme : glossaire de base*.
http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base

Tourisme & Territoires. *Zoom sur le projet Flux Vision Tourisme*.
http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/

Widhalm, P., Yang, Y., Ulm, M. Athavale, S. & Gonzales, M. C. (2015). Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, 42(4), 597–623.
https://doi.org/10.1007/s11116-015-9598-x