Conseil national
de l'information statistique

Insee
Mesurer pour comprendre
Direction générale

LA RÉUTILISATION PAR LE SYSTEME STATISTIQUE PUBLIC DES INFORMATIONS DES ENTREPRISES

Stéphane GREGOIR

Françoise DUPONT

Rapport du groupe de travail INSEE-CNIS

Mars 2016

Sommaire

Synthèse

Tous les Instituts Nationaux de Statistiques, les Banques Centrales, s'interrogent sur les nouvelles sources de données mobilisables pour la statistique officielle dans le cadre des réflexions sur l'impact des Mégadonnées (plus couramment appelé Big Data au niveau international). Des réflexions et de nombreux échanges ont lieu depuis 2013 sous l'égide de l'ONU, d'Eurostat et de l'OCDE. Un certain nombre de sources ont été identifiées comme étant des pistes prometteuses (données issues de la gestion des opérateurs de la téléphonie mobile, données agrégées issues de la collecte des prix et des quantités élémentaires des produits passant en caisse quotidiennement dans les points de vente de la grande distribution, données issues de la gestion des transactions par cartes bancaires, données des réseaux sociaux, données issues des sites d'offres d'emploi,..). Selon les sources et selon les pays, l'instruction sur la faisabilité est plus ou moins avancée.

Six pays diffusent d'ores et déjà des données de prix basées sur les données de caisse et quatre pays s'apprêtent à le faire. Pour les données de la téléphonie mobile, seul un pays diffuse des statistiques officielles basées sur ces données : il s'agit de la banque centrale d'Estonie qui diffuse des données sur le tourisme depuis 2009. Tous les instituts nationaux de statistique européens souhaitent toutefois accéder à ces données pour mener à bien des expérimentations. Concernant les transactions par cartes bancaires, les données sont connues et utilisées par les banques centrales mais à un niveau plutôt agrégé. Aux Etats-Unis, et en Europe, il existe une offre privée d'indicateurs conjoncturels de consommation à des niveaux détaillés qui est portée par les opérateurs de cartes bancaires.

Pour analyser les conditions dans lesquelles ces nouvelles données pourraient être mobilisées par le Service Statistique Public (SSP), un groupe de travail a été constitué sous l'égide du Cnis, associant des représentants du SSP et des entreprises. Il a été chargé d'élaborer un livre blanc d'analyses et de propositions opérationnelles partagées, en vue du développement du nouveau mode de relations entre le Service Statistique Public et les entreprises, qui autorise l'utilisation des données privées pour la production de statistiques publiques. Ce travail devait contribuer à préparer une modification d'ordre législatif. Le groupe a réuni, outre son président Michel Bon, le Directeur général de l'Insee, un représentant du Cnis, le responsable des affaires juridiques de l'Insee, le rapporteur et la co-rapportrice du groupe Cnis, des experts de l'Insee en fonction des différents sujets abordés et de même, des représentants de la Banque de France pour les données de transactions par cartes bancaires (voir annexe 2).

Les objectifs du groupe étaient les suivants :

- **Pour le Service Statistique Public (SSP)**, il s'agissait d'améliorer et d'enrichir sa production de statistiques en accédant à de nouvelles données, dans le respect de ses principes d'indépendance de son code de bonnes pratiques, en restant maître des méthodes employées et en garantissant la confidentialité des données mobilisées.

- **Pour les entreprises** : conformément au principe de la liberté du commerce et de l'artisanat, il s'agissait d'envisager la fourniture de certaines de leurs données tout en veillant à ce qu'aucune atteinte ne risque d'être portée à leur valeur économique. L'usage de leurs données ne pouvait être envisagé que strictement limité à la production de statistiques publiques dûment identifiées, à l'exclusion de toute finalité lucrative ou de contrôle.

Dans ce cadre, le groupe de travail s'est intéressé à trois types particuliers de données, qui sont à des stades différents d'avancement en matière d'expérimentation par la statistique publique française.

L’utilisation des données de caisse, en premier lieu, constitue un dossier ancien. Depuis 2009, l’Insee mène des discussions avec la profession sur la faisabilité, les modalités pratiques et travaille déjà sur des expérimentations à partir de transmission de données. Dans ce projet, l'Insee s'apprête à mettre en place un dispositif de production en 2019. Pour la téléphonie mobile, les discussions avec les professionnels ont démarré en 2014. Il s’agit toutefois d’un sujet complexe et ces discussions s’inscrivent sur un temps long. Enfin, le cas d’utilisation des cartes bancaires est plus récent, puisque son instruction a démarré avec le groupe à l’automne 2015.

Le groupe s'est réuni en 2015 et au début de 2016. Le projet de loi numérique a quant à lui été rendu public le 26 septembre 2015 et un de ses articles vise effectivement à modifier la loi de 1951 pour autoriser la statistique publique à utiliser des sources privées pour l'élaboration de ses statistiques tout en encadrant cette utilisation. Dans sa version actuelle en discussion au Parlement (version du 26 janvier 2016), le texte spécifie qu’il s’agit de rendre possible l’utilisation, à certaines conditions. Il conditionne cette utilisation à la présentation d’une étude de faisabilité et d’opportunité permettant d’engager un dialogue sur le caractère à la fois utile et raisonnable de l’utilisation d’une source, ainsi qu'à un avis du Cnis sur la base de cette étude. Cet exercice doit être réalisé pour chaque source. L'article de loi ne donne qu’un cadrage général, tout comme la version actuelle de le loi de1951 pour la réalisation d’enquêtes et l'utilisation de données administratives. Il précise que chaque utilisation d’une nouvelle source fera l’objet d’un texte particulier et d’une discussion spécifique qui descendra à un niveau plus technique sur l’opportunité, la faisabilité, les coûts engendrés et les modalités pratiques de mise à disposition par chaque détenteur de données. Un dispositif d’amende est également prévu si la source ne peut être obtenue dans le cadre de ce texte particulier.

Les données de caisse ont fait l’objet de discussions qui ont abouti à un accord de principe d'une majorité d'enseignes pour une transmission des données, un accord sur l’intérêt général de cette transmission ainsi qu’une demande d’encadrement de cette transmission de façon à ce que l’obligation s’applique à toutes les enseignes. Les données transmises ne doivent servir qu’à l’élaboration de l’indice des prix. Les modalités techniques de transmission ont été largement expérimentées depuis 2009 avec la profession. Le texte de loi a suscité des discussions et des demandes de la profession. Les échanges vont se poursuivre en dehors du groupe afin que les écrits qui encadreront la mise en place de la production, en particulier le décret d'application, expriment les engagements pris par l'Insee vis à vis des entreprises de façon satisfaisante pour les professionnels. Quelques enseignes contactées n'ont pas souhaité se joindre à la concertation.

Les données des opérateurs de téléphonie mobile ont donné lieu à des discussions en septembre et octobre. Potentiellement, ces données peuvent être utilisées pour produire des statistiques sur le tourisme, la mobilité, ou la présence sur un territoire. Les discussions avec les opérateurs ont porté sur la réalisation de statistiques de population présente sur un territoire. À ce stade, la discussion n’a pas pu être très approfondie avec les opérateurs. En particulier, il n’a pas été possible d’obtenir d’éléments des opérateurs sur les charges engendrées par cette demande (et donc les coûts associés) dans le cadre du travail du groupe. L'Insee continue de discuter de la faisabilité technique en dehors de ce groupe de travail.

Enfin, concernant les cartes bancaires, l’instruction a démarré en fin d’année 2015. Le groupe s'est rapproché des opérateurs bancaires et des groupements de cartes bancaires pour instruire l’élaboration de statistiques qui permettraient d'améliorer, pour l’Insee,le suivi de la consommation en services en comptabilité nationale, et, pour la Banque de France, la production de la balance des paiements et des indicateurs conjoncturels. Les discussions menées en partenariat avec la Banque de France ont permis de rappeler les limites de la source dans un domaine par ailleurs en forte évolution. Elles ont également pointé les accords nécessaires des banques qui sont les détenteurs des données. Les discussions se poursuivront en partenariat avec la Banque de France

Comme l'ont montré par le passé la mise en place de recueil de données basé sur des partenariats avec des fournisseurs privés[1] ou même des administrations[2], de telles mises en place prennent du temps et ne

[1] Mise en place d'un indice Insee/notaire avec Perval, mise en place de l'expérimentation sur les données de caisse mise en place de nouveaux indices de prix à la production dans les branches des biens d'équipements, des biens de consommation, des services.

[2] Discussions sur l'utilisation des données administratives fiscales et sociales dans le cadre du recensement, DADS, utilisation des données du foncier,

s'inscrivent pas sur une seule année de discussion. Il s'agit d'un cheminement long qui va du premier contact en passant par une expérimentation pour déboucher sur la mise en place d'une solution pérenne avec une production régulière de données qui donne lieu à une diffusion officielle. Ce chemin est par ailleurs tributaire d'événements extérieurs indépendants de la volonté de l'institut statistique ou du fournisseur de données (modifications législatives, évolution du contexte concurrentiel...). Ainsi, les discussions sur les données de téléphonie mobile ont été impactées par des discussions concomitantes sur d'autres sujets entre les opérateurs et la Fédération Française des Télécommunications.

Les trois exemples traités par le groupe sont à des stades différents dans ce cheminement, ce qui explique en partie les conclusions différentes présentées dans ce rapport. A ce stade on peut tirer le bilan suivant :

Les garanties demandées par les entreprises

Les fournisseurs ont tous soulevé la question du respect de la vie privée ou du secret des affaires et les risques d'image encourus par eux en cas de manquement à ces obligations. Ils ont également souligné la sensibilité de ces données d'un point de vue commercial.

L'Insee de son côté, de par la nature même de son activité et du contexte légal dans lequel cette activité est réalisée, doit respecter ses obligations en matière de secret statistique. Les données d'origine du producteur sont souvent très détaillées et sont particulièrement sensibles du point de vue de la protection de la vie privée.

La solution qui semble la plus raisonnable consiste à demander dans la mesure du possible des données pré-agrégées par le fournisseur qui pourront servir de matériaux de base aux travaux d'agrégations et de redressements de l'institut statistique. Selon les cas la connaissance des données est suffisante du côté de l'institut statistique pour spécifier le niveau d'agrégation qui permet de travailler, et dans d'autres cas où la matière première est plus complexe, un travail d'instruction plus fouillé est encore nécessaire avec les fournisseurs de données pour mettre au point la maille commune d'agrégation entre les données des différents fournisseurs qui permettra de procéder aux redressements nécessaires pour corriger des biais.

L'Insee peut jouer un rôle de tiers de confiance vis à vis des fournisseurs pour élaborer des statistiques agrégées qui peuvent être diffusées sans mettre en danger la vie privée des individus, ni le secret des affaires. La mutualisation par le service statistique public des données transmises permet d'obtenir des données représentatives de l'ensemble du champ analysé.

Les fournisseurs avec lesquels les discussions commencent expriment des réticences concernant la transmission au service statistique public de données sensibles du point de vue commercial. L'Insee a déjà rencontré ce type de réticences au début des discussions avec des partenaires privés lors de la mise en place d'une nouvelle enquête statistique auprès des entreprises. Le cadre légal national et européen dans lequel opère la statistique publique apporte une protection des fournisseurs par rapport à ce type de risque. Des précautions supplémentaires ont été demandées dans le cas de l'accès direct à des données privées concernant la non communication aux chercheurs de ces données.

Lorsqu'une valorisation commerciale a été mise en place par le fournisseur de données comme c'est le cas pour les données de la téléphonie mobile, la discussion a porté sur le périmètre de diffusion respectif de l'Insee et du fournisseur afin de délimiter ce qui relève de la mission d'intérêt général de la statistique publique et ce qui relève du marché concurrentiel. L'Insee a porté le point de vue qu'il est possible de délimiter, pour chacun des acteurs, des périmètres complémentaires qui permettent aux deux structures de remplir leurs missions respectives en bonne intelligence. La discussion n'a pas permis de convaincre l'ensemble des opérateurs de la téléphonie mobile à ce stade.

Les contreparties demandées par les entreprises

Les fournisseurs demandent à ce que la fourniture de données soit réalisée avec un coût de mise à disposition négligeable ou que le coût de mise à disposition soit couvert. C'est bien le cas pour les données

de caisse pour lesquelles les fichiers sont déjà constitués pour les besoins propres des distributeurs, en revanche, il n'a pas été possible d'obtenir d'éléments suffisants pour chiffrer, même très grossièrement, le coût de la mise à disposition d'informations de la part des opérateurs de téléphonie mobile.

De son côté l'Insee a indiqué que ce point ferait partie pour chaque source de l'étude de faisabilité et que la demande d'informations serait limitée au strict minimum et proportionnée à l'objectif visé par l'institut statistique en matière de diffusion.

Les fournisseurs de la téléphonie mobile ont posé la question des bénéfices reçus en retour pour les fournisseurs des données. La proposition de la statistique publique est de mettre à disposition des fournisseurs des statistiques de cadrage plus robustes pour les fournisseurs. La valeur résultant de la mutualisation des données de différentes entreprises concurrentes peut être ainsi profitable aux utilisateurs des données de l'institut statistique mais aussi aux fournisseurs des données. Ce point a été discuté avec les fournisseurs des données de caisse qui bénéficient déjà de retour d'information sur leurs données et ne souhaitent pas en l'état de retour d'information particulier. Pour les opérateurs de la téléphonie mobile, cet apport n'est pas apparu à tous comme suffisant.

La mise au point d'un dispositif cible d'accès aux données nécessite du temps et un cadre légal approprié

Les discussions au sein du groupe Insee-Cnis ont permis de prolonger ou d'initier des échanges préalables à la mise en place de partenariats potentiels fournisseur-statistique publique. Ces partenariats supposent toutefois d'établir dans la durée une relation de confiance et des échanges réguliers pour mettre au point le dispositif cible. Une phase d'expérimentation préalable est indispensable pour mettre au point les modalités concrètes de la méthodologie et de livraison de données.

Les transmissions de données sur une base volontaire ne suffisent pas à mettre en place une production statistique pérenne pour la statistique publique. Des textes juridiques sont nécessaires ; ils permettent également de mettre toutes les entreprises sur un pied d'égalité par rapport à la contribution d'intérêt général demandée.

La loi en cours de discussion sur le numérique apportera, comme pour les enquêtes de la statistique publique, un cadrage général qu'il faudra ensuite décliner pour chaque source, en élaborant des textes plus précis encadrant chaque opération une fois l'opportunité et la faisabilité attestées. La loi seule ne rend pas possible l'accès à une source privée. Une étude de faisabilité doit être réalisée et discutée avec les fournisseurs de donnée, être rendue publique et recevoir un avis du Cnis. C'est seulement à l'issue de ce travail qui garantit la bonne prise en compte des contraintes de la profession et qui lui permet de faire valoir ses objections, qu'une demande de données régulières pour la production sera possible.

L'élargissement des discussions au niveau européen est en cours

Les même questions se posent dans les autres pays européens et même plus largement au niveau international. Les fournisseurs des données sont également des acteurs qui peuvent intervenir sur un marché plus large que le marché français et dont le positionnement n'est pas seulement français. Les aspects techniques et juridiques qui interviennent dans l'instruction sont également largement européens ou internationaux. En dépit des spécificités nationales, la convergence des législations qui encadrent le travail des instituts statistiques et les réflexions communes du groupe des G29 (institutions juridiques) conduisent à réfléchir et à discuter de ces évolutions de la production du système statistique plus largement dans la sphère européenne. Les instructions à venir devraient bénéficier d'une vision étendue à l'Europe en relation avec les travaux coordonnés par Eurostat et à plus largement au niveau international avec des organismes comme l'OCDE et l'ONU. Des réflexions ont en effet été lancées sur les modèles économiques permettant au système statistique public de bénéficier des informations d'intérêt général d'origine privée.

1. Introduction

Tous les instituts nationaux de statistiques en Europe et au-delà s'interrogent sur l'impact du phénomène des Mégadonnées (Big Data dans les échanges internationaux). Les réflexions menées essentiellement depuis 2013 deviennent de plus en plus concrètes sur les méthodes de traitement informatiques et statistiques des données massives. De nouvelles sources de données apparaissent désormais pouvoir faire l'objet de traitements statistiques.

Assez naturellement, les instituts nationaux statistiques se demandent si ces données peuvent donner lieu à des statistiques nouvelles ou peuvent permettre d'améliorer les statistiques actuelles, d'en réduire le coût ou d'en augmenter la qualité. Cette réflexion assez large est menée dans les instances internationales des Nations Unies et d'Eurostat, et des échanges se nouent à des niveaux techniques sous la forme de groupes de travail. Ainsi Eurostat a lancé en février 2016 un groupe de travail[3] sur les Big Data, couvrant différentes sources et dont la France est partie prenante.

Parmi les sources potentielles figurent notamment les « données de caisse » issues des facturations de la grande distribution, une source connue et investiguée depuis longtemps déjà, ou les données de la téléphonie mobile, qui font l'objet de réflexions depuis plus de dix ans en Estonie. Des réflexions plus récentes sont également menées sur les sites d'offres d'emploi, les données de *Google Trends* ou les compteurs intelligents. Le travail des INS s'appuie dans certains cas sur de premiers travaux initiés par des chercheurs. L'Insee participe également à la Task Force « Big Data for Official Statistics » d'Eurostat et aux travaux techniques menés par l'Unece sur ces sujets (groupe « Sandbox », plateforme qui permet de tester des outils et des méthodes sur des jeux de données de type « Big Data »).

L'Insee assure depuis 2014 un rôle de coordination sur le sujet des données massives pour le Service Statistique Public. Il contribue ainsi aux réflexions sur les potentialités de ces données pour la statistique publique à moyen et long terme. l'Insee a mené plusieurs investissements sur les techniques statistiques adaptées au traitement de ces données dont notamment une étude portant sur l'utilisation des données de *Google Trends* pour améliorer la prévision économique. L'institut a également exploré les algorithmes d'apprentissage statistique et a étudié la mise en œuvre de méthodes économétriques en utilisation de technologies « Big Data » (traitements parallélisés), de méthodes de *text mining* et *webmining*, etc.

Dans ce contexte, l'Insee a lancé une concertation sous l'égide du Cnis de façon notamment à aider à concevoir un cadre juridique pour l'utilisation de ces données par la statistique publique, question que tous les pays se posent. Lancée fin 2014, cette concertation s'est terminée début 2016.Ce groupe Cnis-Insee avait pour objectif de mener une concertation avec les entreprises les plus concernées par la réutilisation par la statistique publique de données privées sans porter atteinte à la valeur économique de ces données, en respectant le secret des affaires et la vie privée des individus, ainsi que l'ensemble des principes du code européen des statistiques européennes. La mission du groupe consistait à élaborer un livre blanc de propositions pour permettre de concevoir un cadre juridique à cette réutilisation et combler ainsi un vide juridique, puisqu'il n'existait pas de cadre solide pour pouvoir utiliser ces données privées. Dans le même temps, le projet de loi numérique a été rendu public le 26 septembre 2015. Un article vise à modifier la loi de 1951 pour autoriser l'utilisation de sources privées et encadrer cette utilisation. Dans sa version actuelle en discussion au parlement (version du 26 janvier 2016), le texte spécifie qu'il s'agit de rendre possible

[3]ESSnet Big Data : réseau d'INS européens collaborant sur un sujet donné pour fournir des résultats utiles à l'ensemble du système statistique européen sous l'égide d'EUROSTAT

l'utilisation, à certaines conditions.

Le groupe de travail s'est intéressé à trois cas particuliers de données, qui sont à des stades différents d'avancement en matière d'expérimentation par la statistique publique française. L'utilisation des données de caisse, en premier lieu, constitue un dossier ancien. Depuis 2009, l'Insee mène des discussions avec la profession sur la faisabilité, les modalités pratiques et travaille déjà sur des expérimentations à partir de transmission de données. Dans ce projet, l'Insee s'apprête à mettre en place un dispositif de production en 2019. Pour la téléphonie mobile, les discussions avec les professionnels ont démarré en 2014. Il s'agit toutefois d'un sujet complexe et ces discussions s'inscrivent sur un temps long. Enfin, le cas d'utilisation des cartes bancaires est plus récent, puisque son instruction a démarré avec le groupe à l'automne 2015.

2 . Les données de caisse

2.1 le contexte international

Le premier domaine abordé par le groupe de travail est le sujet le plus avancé dans tous les pays. Quelques pays européens utilisent déjà les données de caisse pour le calcul de leur indice des prix à la consommation et la plupart développent des projets en ce sens. Dès 2001, la Norvège a été un des premiers pays à utiliser ce type de données dans son indice. Elle a ensuite été suivie par les Pays-Bas (2002), la Suisse (2008) ou plus récemment par la Suède (2012). Et puis tout récemment, en 2016, la Belgique et le Danemark. Deux autres pays devraient y recourir d'ici 2017, le Luxembourg et la Pologne. De nombreux échanges ont lieu au niveau européen sous l'égide d'Eurostat pour faire converger les méthodologies. Bien que la méthodologie du calcul de l'indice ne soit pas pour l'heure harmonisée, l'utilisation des données de caisse dans les indices européens est d'ores et déjà autorisée. Des groupes de travail ont été depuis quelques années l'occasion d'échanger sur les aspects méthodologiques. Eurostat a souhaité faire le tour des projets des différents pays. Dans ce cadre, l'Insee a présenté ses travaux aux équipes d'Eurostat fin novembre 2015. Les discussions en cours au sein d'Eurostat pourraient conduire à de premières recommandations sur l'utilisation des données de caisses dans les indices de prix européens dès cette année.

2.2 le projet français d'utilisation des données de caisse pour l'indice de prix à la consommation

Des besoins français et européen

L'indice des prix à la consommation (IPC) est une opération phare de l'Insee. Un réseau de 200 enquêteurs, animé par neuf sites prix en régions (environ 75 personnes), un pôle Prix à Bordeaux et la division des prix à la direction générale (20 personnes), réalise chaque mois environ 180 000 relevés de prix dans 27 000

points de vente situés dans 104 agglomérations de plus de 2 000 habitants de la métropole et des DOM.

Les discussions sur la qualité et l'intérêt de l'IPC français au moment de l'introduction de l'euro ont permis de réaffirmer sa pertinence pour le suivi de l'érosion monétaire. Elles ont également montré le besoin de le compléter par de nouveaux indicateurs (commission Quinet de 2008 sur la mesure du pouvoir d'achat). D'autres besoins sont également apparus, notamment celui des comparaisons des niveaux de prix entre territoires à des niveaux fins (agglomérations, régions) ou celui du suivi de marchés très spécifiques (produits éco-labellisés notamment).

Ces nouveaux besoins d'une part, et l'intérêt manifesté par la Commission européenne (Eurostat) pour la nouvelle source de données que représentent les données de caisses des enseignes de la distribution pour les statistiques de prix d'autre part, ont décidé l'Insee à lancer, à la fin de l'année 2009, un projet d'exploitation de ces données. L'accès aux données de caisse est aujourd'hui un moyen d'améliorer la qualité de l'indice produit. Il doit à terme permettre le calcul d'indice directement à partir des données fournies par les enseignes. Ce projet s'inscrit dans une modernisation plus large de l'IPC qui inclut notamment la refonte de l'échantillon d'agglomérations, un renforcement du suivi de la vente par internet et une évolution des processus de collecte (notamment du matériel de collecte).

Deux phases expérimentales pour mettre au point le protocole de transmission et valider la méthodologie

Le projet vise à calculer des indices de prix en reprenant la méthodologie actuelle de l'IPC (indice de Laspeyres chaîné annuellement). Ces indices seraient fondés sur un échantillon de produits (EAN[4]x magasin), chaque produit pesant dans l'indice d'ensemble en proportion de son poids dans le chiffre d'affaires total observé sur l'ensemble de l'année précédente. Les données de caisse exploitées seront celles relatives aux biens de consommation courante vendus dans les super et hypermarchés, c'est-à-dire principalement les produits alimentaires (hors produits frais), les produits d'hygiène et les produits d'entretien courant de la maison. Leur champ fait actuellement l'objet de 35 000 relevés par mois (soit 20 % du champ total). A l'avenir, les données des supérettes pourraient faire l'objet d'une même exploitation.

Les travaux méthodologiques menés depuis 2012 démontrent la faisabilité statistique du projet :

- deux tests de comparaison des données de caisse avec la collecte actuelle par enquêteurs ont permis de valider la qualité des données de prix et de documentation des articles ;

- les travaux de simulations d'indices sont concluants pour la précision des indices issus des données de caisses et leur comparaison avec les indices IPC actuels. Les indices de prix issus des données de caisse sont sensiblement plus précis que ceux de l'IPC actuel, du fait qu'ils sont fondés sur un nombre de séries plus important (au minimum de 10 à 20 fois plus grand). Ils sont concordants avec les indices actuels, ces derniers étant inclus dans les intervalles de confiance à 95 % des indices issus des données de caisse.

Ces travaux ont été réalisés dans un premier temps à partir des ventes des années 2007 à 2009 relatives à 10 familles d'articles et 1 000 points de vente de six enseignes. Les chiffres d'affaires et les quantités vendues étaient agrégés par mois. Les prix utilisés pour le calcul des indices étaient donc des prix mensuels moyens. Ces données de caisse étaient enrichies par des informations issues des référentiels d'articles et de points de vente de la société IRI. Ces travaux préliminaires ont pu débuter grâce à la mise en place d'un groupe de travail entre l'Insee et la distribution réunissant la Fédération des entreprises du commerce et de la distribution et des représentants de 5 enseignes.

[4] L'EAN (European Article Number) est un identifiant à 13 chiffres pour les produits manufacturés. Il est géré mondialement par un organisme, GS1, et permet d'identifier le produit de manière unique.

Dans un second temps, les travaux ont été menés sur les données de caisse quotidiennes transmises par 4 enseignes depuis novembre 2012. Ces données sont transmises à l'Insee à titre gracieux dans le cadre de conventions triennales conclues avec les enseignes volontaires pour ce deuxième temps. La réception par l'Insee de données de caisses dans cette phase expérimentale s'est inscrit dans le programme des enquêtes statistiques non obligatoires depuis décembre 2011 (au sens de la loi de 1951).

Les formats et procédures de transmission sont les mêmes que ceux utilisés pour la transmission des données de caisse aux sociétés Nielsen et IRI. Les données de ventes sont agrégées quotidiennement au niveau de chaque produit par point de vente. Il ne s'agit pas de données personnelles et il n'est pas possible de réidentifier un individu à partir de ces données. Ces données sont en revanche extrêmement sensibles du point de vue du secret des affaires. Les transmissions de données sont réalisées sous forme cryptées selon les bonnes pratiques en vigueur au moment de la transmission. Toutes les informations annexes utiles pour les utiliser que l'Insee peut être amené à connaître relèvent du secret professionnel.

Les conventions stipulent que :

- les données ne sont utilisées que dans le cadre de l'élaboration de l'indice des prix à la consommation, à l'exclusion de tout autre usage

- seuls les agents de la division des prix à la consommation et ceux des services informatiques dédiés à la maintenance de l'IPC ont le droit d'utiliser ces données, à l'exclusion de tout autre agent

- les données de caisse transmises à la division des prix sont détruites sous un délai de 24 mois. Seules sont conservées les données de prix relatives à l'échantillon à partir duquel a été effectivement calculé l'IPC et les éléments de pondération généraux utilisés.

Les premières études méthodologiques sont terminées. Elles ont permis à l'Insee d'affiner la méthodologie statistique de calcul d'indice à partir des données quotidiennes, de construire un processus de traitement des données et de dessiner l'infrastructure informatique nécessaire à sa réalisation en faisant appel en partie à des technologies Big Data.

Dans le cadre de ces tests, l'Insee a conclu par ailleurs un marché public avec IRI pour l'achat de ses référentiels d'articles et de points de vente sur la période 2012-2016. Ceci a permis à l'Insee de valider les conclusions du premier test sur un champ plus large d'articles et de points de vente mais aussi d'acquérir une expérience opérationnelle dans la réception de ces données et leur exploitation.

Une nécessaire sécurisation juridique pour diffuser un indice des prix utilisant les données de caisse

D'un point de vue pratique, l'Insee souhaite adosser le calcul de l'IPC sur les données de caisses, pour le champ de la consommation couvert par ces données. Ce champ correspond aujourd'hui à 20% de la collecte terrain des prix à la consommation pour l'IPC. Il n'est pas envisageable, étant donné le degré de sensibilité de l'IPC, d'asseoir la production de cet indice sur des données dont la transmission pourrait s'interrompre à tout moment, comme cela pourrait être le cas si elles étaient obtenues, comme aujourd'hui, sur une base volontaire. Sécuriser la réception de ces données est donc une condition essentielle de leur utilisation pour l'IPC. La voie privilégiée aujourd'hui pour y parvenir reste la mise en place d'un cadre juridique adapté permettant l'accès aux données privées à des fins de production de statistiques reconnues d'intérêt public.

L'Insee ayant également réfléchi collectivement au statut des informations numériques du secteur privé, la réflexion sur la sécurisation juridique des données de caisse dépassant le seul champ du projet, un groupe de travail créé au sein du Cnis et de l'Insee a permis de mener une concertation sur ce sujet avec les acteurs du domaine.

2.3 la concertation

Deux réunions ont été organisées dans le cadre du groupe de travail Insee-Cnis. La première s'est tenue le 13 janvier 2015 et a été préparée avec la Fédération du Commerce et de la Distribution.

Dans le cadre de la préparation du projet de loi sur le numérique, une consultation a été organisée sur la base d'une version mise en ligne le 26 septembre. Le caractère très général de l'article 7 (voir annexe) a suscité de nouveaux échanges et des craintes de la profession sur le champ des données qui seraient demandées. Une deuxième réunion de discussion dans le cadre du groupe Cnis-Insee a été organisée pour finaliser les échanges avec les professionnels de la distribution et dissiper leurs interrogations sur l'impact de cette modification législative.

Des contacts ont été pris avec la FCA pour leur présenter le projet et son orientation d'intérêt général entre les deux réunions. Certains représentants d'enseignes (membres de la FCD ou pas), sollicités pour participer au groupe, n'ont pas donné suite.

Dans ces réunions, Michel Bon explique que la réforme de l'Etat nécessite une aide de la société publique pour permettre des gains en efficacité. L'Insee souhaite dans cet esprit utiliser des données de caisse dans l'indice des prix. La mission du groupe Cnis-Insee est rappelée : il s'agit de discuter des conditions de réalisation de cette transmission de données, comme celles d'autres données disponibles dans le domaine de la téléphonie mobile et des cartes bancaires.

Les projets en matière d'utilisation des données de caisse ont été présentés aux participants aux réunions. Il s'agit d'un des chantiers important de modernisation de l'Insee, avec d'autres chantiers comme le développement de la collecte par internet en particulier dans le recensement. Ce chantier a déjà fait l'objet d'une phase de test qui a permis d'évaluer la capacité de l'Insee à recevoir les flux importants de données brutes et à les traiter. Ceci a été réalisé dans le cadre d'un partenariat avec quatre enseignes et une contractualisation avec un tiers de transmission (IRI). L'Insee indique que la pertinence des calculs a été vérifiée et que les données de caisse permettent d'envisager des calculs de prix plus détaillés au niveau géographique.

l'Insee après avoir exposé ses projets en matière de données de caisse a indiqué que l'enjeu de la concertation était de garantir la disponibilité des données sous des conditions qui permettent de réaliser des économies et d'assurer la participation de toutes les enseignes à la transmission des données. L'Insee indique qu'il souhaite en effet s'assurer de la participation pérenne des enseignes avant de publier des résultats.

La loi de 1951 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques permet à l'Insee de recueillir des données à travers des enquêtes et de rendre ce recueil obligatoire. La transmission des données de caisse a été dans la phase expérimentale placée sous l'égide de cette loi au travers d'un décret qui a permis à l'Insee de décrire les conditions de protection des données sur lesquelles il s'engageait vis-à-vis des enseignes volontaires. Or, à ce jour, un acteur important du secteur ne souhaite pas participer à cette collecte d'information sur le mode du volontariat. Pour affirmer la nécessité de cette réponse et mettre en place un système d'amende adapté aux enjeux, il est nécessaire de modifier la loi actuelle (loi de 1951 du 7 juin 1951 sur

l’obligation, la coordination et le secret en matière de statistiques)

Pour les entreprises de la grande distribution, l'enjeu de cette transmission est de garantir la confidentialité les données pour protéger la vie privée et le secret des affaires, d'assurer un coût minimal pour cette transmission, d'examiner les bénéfices en retour possibles de cette transmission pour les distributeurs.

Compte tenu de l'antériorité des discussions avec la profession qui ont permis d'expérimenter des modalités de transmission des données sur 2009-2012, de réaliser des exploitations probantes sur la faisabilité méthodologique et de valider la valeur ajoutée qu'apportent ces données, les discussions ont pu déboucher assez rapidement.

Les participants de la première réunion du 13 janvier 2015 ont ainsi indiqué leur accord sur le fait que la transmission de ces données de caisse à titre gracieux représente une participation raisonnable des enseignes à une mission d'intérêt général, à la condition que cette participation s'applique à toutes les enseignes. Tous les participants ont donc conclu que le volontariat avait ses limites et qu'une obligation juridique assortie d'une amende permettrait de clarifier le statut de cette participation citoyenne des enseignes.

Il n'est pas apparu de demande de retour d'information de la part des enseignes dans la mesure où elles disposent d'un droit de tirage d'études en contrepartie de leur transmission de données. Les entreprises conviennent qu'elles peuvent tirer des bénéfices réputationnels de cette participation à une mission relevant de l'intérêt général. En terme d'organisation pérenne et efficace, les enseignes ont exprimé une préférence pour que celles-ci soient responsables juridiquement de la transmission, tout en délégant, si elles le souhaitent la réalisation de cette transmission au concentrateur IRI, comme c'est le cas actuellement. Les conditions pratiques de transfert des données de l'expérimentation ont paru satisfaisantes et n'ont pas soulevé de remarques particulières.

La deuxième réunion a eu lieu après la parution du projet de loi numérique mis en consultation le 26 septembre avec un article 7 qui matérialise l'évolution de la loi de 1951 pour y inclure le recours aux données privées. La réunion a eu lieu juste avant que le texte de loi ne soit proposé en conseil des ministres puis proposé à l’Assemblée Nationale.

L'Insee a rappelé qu'il ne maîtrise pas totalement la rédaction du texte, mais que l'article 7 a fait l’objet d’échanges avec le ministère, puis d’une discussion en Conseil d’État en présence de l'Insee. Dans ces discussions avec les juristes, l’Insee a cherché à clarifier autant que possible ses intentions et le mode de fonctionnement souhaité qui est la concertation. Tout ceci sera plus visible dans les textes du décret et des conventions qui seront élaborés pour décliner l'application du texte de loi dans chaque cas.

La mise à disposition des données ne serait pas rémunérée, l’institut couvrirait le cas échéant les frais supplémentaires générés par la mise à disposition des données. Dans le cas présent, la transmission à l'Insee ne crée pas de surcoûts significatifs. La loi prévoit bien sûr également des garanties quant aux usages possibles et conditions d’usage des données pour les détenteurs de données.

Le texte actuel intègre plus clairement le fait qu’une statistique ne pourra être mise en place qu’après une étude de faisabilité rendue publique, une concertation de la profession qui passera par le Cnis, que les données collectées ne peuvent être utilisées que pour l’élaboration de statistiques et ne peuvent être cédées en l’état pour d’autres usages. Elle intègre également le fait que les modalités concrètes sont définies dans un décret.

Par ailleurs, les opérations qui précèdent la mise en place d’une production statistique intègrent toujours des

discussions préalables avec la profession, mais aussi des mises au point lourdes pour l'Insee. Les productions qui résulteront de la loi ne pourront donc de ce fait être nombreuses. En outre, l'expérience montre qu'il n'est pas possible de réaliser un recueil des données sans la coopération de l'entreprise concernée, ne serait-ce que pour des questions de format des données, de documentation. Il est donc également de l'intérêt de l'Insee de mettre en place une relation de confiance pérenne.

Les professionnels ont été invités à lister leurs craintes afin que l'Insee y apporte des réponses dans le cadre du dialogue.

Les représentants ont demandé des garanties complémentaires :

- que le texte de loi utilise le terme exclusif pour renforcer le fait que l'usage qui sera fait des données ne pourra être que statistique
- que le caractère agrégé de la diffusion soit également renforcé
- que le fait que l'Insee prenne toutes les précautions pour assurer la sécurité des données soit plus explicitement mentionné

L'Insee a rappelé que la rédaction actuelle permet d'ores et déjà de garantir que l'usage ne sera que statistique, ce que le règlement européen n°223/2009 révisé garantit également.

Suite à des craintes exprimées sur la possibilité de contester les amendes, il a été rappelé que les amendes sont examinées par un comité du contentieux qui dépend du Cnis et qui comprend des représentants des organisations professionnelles. Une contestation est possible auprès du tribunal administratif. Il a été rappelé que le montant des amendes proposé dans le texte de loi a diminué au cours des discussions, en particulier sous l'effet de la demande des professionnels.

Les professionnels se sont interrogés sur le contact utilisé pour cette dernière demande compte tenu de leur organisation interne. L'Insee précise que les entreprises désigneront un contact suite à une demande officielle du Directeur général de l'Insee. La désignation des points de contact en particulier pour les franchisés devra être un point de vigilance pour les entreprises et l'Insee.

L'Insee a proposé que les professionnels transmettent leurs remarques sur le projet de loi dans le cadre de la discussion qui a lieu à l'assemblée nationale.

L'Insee a proposé de travailler en amont avec les professionnels sur le texte du décret à venir et sur les conventions afin de bien intégrer les remarques qui portent sur les modalités plus concrètes. L'Insee a poursuivi ces échanges avec la profession comme prévu.

3. Les données de la téléphonie mobile

3.1 le contexte international

Le potentiel statistique des données de la téléphonie mobile font l'objet d'un vif intérêt au niveau international. De nombreux articles de recherche ont été publiés depuis une dizaines d'années en particulier par des géographes afin de démontrer leur intérêt pour analyser les problématiques de transport (au sein d'un pays, d'une ville) et de tourisme.

Le challenge D4D[5] d'Orange au Sénégal a contribué à augmenter récemment encore davantage l'attention sur le potentiel de ces données pour éclairer les politiques publiques. Les thèmes ciblés sont plus larges et concernent la santé, l'agriculture, l'énergie, le transport, l'urbanisme, les statistiques nationales.

Dans le contexte européen, l'Estonie fait figure de pionnier. Les géographes de l'université de Tartu s'intéressent depuis plus de dix ans au potentiel de ces données. Un partenariat entre l'université et les opérateurs de téléphonie s'est créé progressivement et permet à l'entreprise Positium de rassembler les données des opérateurs de téléphonie[6]. Ce travail de longue haleine a permis de déboucher sur une diffusion de statistiques officielles Ainsi, suite à des restrictions de crédits, la banque centrale Estonienne diffuse mensuellement depuis 2009 des statistiques concernant le tourisme de non résidents dans le pays de référence et le tourisme des résidents à l'étranger (nombre de voyages, durée de séjour). Les données sont rassemblées par l'entreprise Positium qui les agrège et les redresse sur la base d'une méthodologie validée par la banque centrale estonienne. Positium élabore également des données de populations présente de jour et de nuit. Elles sont utilisées par l'État pour calibrer des équipements liés à des situations d'urgence.

D'autres pays ont également cherché à diffuser des statistiques officielles basées sur ces données. L'office du Tourisme tchèque diffuse depuis 2012 des statistiques sur le nombre de visiteurs sur 45 sites. L'office statistique des Pays Bas a mené une étude expérimentale sur la base de données CDR recueillies sur deux semaines en 2010 pour construire des statistiques sur le tourisme et les déplacements. Plusieurs offices statistiques nationaux ont cherché à acquérir les données : l'Irlande, la Slovénie, le Monténégro et la Finlande. D'autres pays se sont intéressés au sujet Portugal, Autriche, Suisse, Royaume uni.

Une étude de faisabilité commanditée par Eurostat[7] et rendue publique en avril 2014, analyse le potentiel des données de téléphonie mobile en matière de statistiques de tourisme à partir des nombreuses recherches déjà menées en Europe, en particulier en Estonie. Elle évoque également le potentiel de cette source pour les analyses de mobilité. Elle détaille les différents aspects à prendre en compte pour une utilisation de ces données. L'accès aux données est examiné sous l'angle de la faisabilité technique, des coûts, des aspects légaux, des problèmes de confidentialité et de perception des utilisateurs, des risques de changement de technologie pouvant affecter la source. Le rapport analyse également la méthodologie de traitement, la qualité des résultats, les coûts pour les opérateurs et les utilisateurs. Il s'intéresse également à d'autres utilisations. Elle conclut que le principal obstacle à l'utilisation à ce jour est le problème de l'accès à ces données sur le plan légal, en raison des barrières que mettent les opérateurs sur les aspects financiers et sur le secret des affaires, et en raison de la perception que peut avoir l'opinion publique sur cette utilisation.

Cette étude conclut que ces données ne peuvent pas se substituer entièrement aux statistiques actuelles sur le tourisme mais qu'elles peuvent en alléger les coûts et apporter des éclairages complémentaires. La rapidité d'obtention, la fréquence de recueil, le détail géographique, le faible coût potentiel en regard des coûts d'une enquête font partie de leurs atouts. Ces sources présentent néanmoins quelques faiblesses : problèmes de sous-représentation et sur-représentation[8], manque d'information sur le motif du voyage, sur le budget, difficulté à reconstituer le moyen de transport.

[5] http://www.d4d.orange.com/fr/Accueil

[6] EMT, Elisa, Tele2.

[7] http://mobfs.positium.ee/index.php?id=reports

[8] Il n'y a pas d'équivalence exacte entre un portable et un individu. Par exemple, les enfants et les personnes âgées sont sous ou pas équipés de portables, lors de déplacements de touriste, il peut y avoir un portable pour une famille ou un groupe. D'un autre côté un individu peut être équipé de deux portables (professionnel et personnel). Une partie des individus peuvent également ne pas avoir de portables en voyage ou l'éteindre. On dispose de peu d'information sur les comportements d'utilisation des portables par les touristes.

Les statistiques de tourisme

Le tourisme se réfère à l'activité de visiteurs se rendant dans une destination située en dehors de leur environnement habituel, pour une période inférieure à un an. Il peut avoir tout motif principal (notamment les affaires, les loisirs ou d'autres raisons personnelles) autre que le fait de travailler pour un résident, un ménage ou une entreprise du pays visité.

Les statistiques touristiques consistent en deux composantes principales : d'une part, les statistiques relatives à la capacité et à l'occupation des hébergements touristiques collectifs et, d'autre part, les statistiques relatives à la demande touristique. Dans la plupart des États membres, les premières proviennent d'enquêtes complétées par les établissements d'hébergement, les secondes étant principalement collectées via des enquêtes auprès des voyageurs aux frontières ou des enquêtes auprès des ménages.

Les statistiques sur la capacité des hébergements touristiques collectifs portent notamment sur le nombre d'établissements, le nombre de chambres et le nombre de places-lits. Ces statistiques sont disponibles par type d'établissement ou par région et sont publiées annuellement.

Les statistiques sur l'occupation des hébergements touristiques collectifs portent sur l'occupation des chambres (ou ou appartements ou emplacements de campings), le nombre d'arrivées et de nuitées passées par les touristes dans ces établissements), répartis par type d'établissement, zone géographique de l'hébergement et nationalité des touristes .

Les statistiques sur la demande touristique renvoient à la participation touristique, c'est-à-dire le nombre de personnes faisant un voyage d'au moins quatre nuitées au cours de la période de référence. Ces statistiques portent sur le nombre de séjours touristiques effectués (et le nombre de nuitées de ces séjours), ventilés selon :

- *le pays de destination;*
- *le mois du départ;*
- *la durée du séjour;*
- *le type d'organisation de voyage;*
- *le mode de transport;*
- *le type d'hébergement (yc dans le non marchand et les logements proposés par les particuliers)*
- *les dépenses.*

Les données peuvent également être analysées selon des variables explicatives sociodémographiques telles que l'âge et le sexe; ces statistiques sont collectées sur une base trimestrielle et annuelle. Cette source examine aussi les non-départs et également les intentions de voyages.

Des données provenant d'autres sources officielles peuvent également être utilisées pour étudier le tourisme, notamment :

- *les données sur l'emploi dans le secteur de l'hébergement touristique, tirées de l'enquête sur les forces de travail (EFT) et analysées selon le temps de travail (temps plein/temps partiel), le statut de travail, l'âge, le niveau d'éducation, le sexe, la continuité et l'ancienneté du travail auprès d'un même employeur (données annuelles et trimestrielles);*
- *les données sur les recettes et dépenses pour les voyages privés tirées de la balance des paiements ;*
- *les statistiques relatives au transport (transport aérien de passagers, par exemple);*
- *les statistiques structurelles sur les entreprises (SC), qui peuvent donner des indications supplémentaires sur les flux touristiques et la performance économique de certains secteurs liés au tourisme.*

3.2 le contexte français

La mesure de la population présente sur un territoire :

Le projet de la statistique publique discuté dans le cadre de la concertation concerne un éclairage structurel sur la population présente pour calibrer des équipements liés à une situation d'urgence. L'éclairage souhaité a peu de recouvrement avec le marché développé par les opérateurs.

La mobilité sous toutes ses formes entraîne une distorsion de plus en plus forte entre la population "résidente" et la population "présente". La population résidente est mesurée en France par le recensement de la population dans lequel chaque individu est affecté à un lieu de résidence unique. Mais avec le développement de la mobilité, la population réellement présente en un lieu donné peut être notablement différente de la population résidente. L'afflux de population est parfois très important en certains lieux et à certaines époques. Ce surcroît de population doit être géré par les pouvoirs publics. Il est donc important qu'il soit mesuré. En effet certains équipements doivent être calibrés en fonction de la population maximale pouvant être présente sur les lieux. Certains départements vont ainsi jusqu'à doubler de population à certaines périodes de l'année, tandis que d'autre ont une population présente presque toujours inférieure à leur population résidente. Les utilisations possibles sont nombreuses :

- Santé et épidémiologie : organisation de la permanence des soins, prévision du dimensionnement optimal de l'offre de soin (notamment dans les services d'urgence), gestion des risques sanitaires (épidémie, pollution de l'eau…)…
- Adaptation des moyens de secours ou de sécurité publique (par ex pompiers et matériel de secours) à des augmentations temporelles de la population
- Connaissance/anticipation des accidents routiers en fonction des variations de population
- Dimensionnement des équipements (réserve d'eau potable, collecte sélective et traitement des déchets…) et programmation plus précise des investissements à réaliser.
- Dimensionnement des infrastructures et services liés au transport
- Adaptation des infrastructures touristiques afin d'améliorer la qualité de l'offre.

Si la mesure de la population présente ouvre la voie à un autre regard sur les problématiques de gestion et d'aménagement des territoires, elle sert également de base à une nouvelle approche de l'économie territoriale, l'économie présentielle, dont le principe est qu'une consommation, et donc une activité économique, est induite par la présence de personnes à un moment donné sur ce territoire.

Une estimation a été réalisée pour la première fois en 2005 par le service statistique du ministère en charge du tourisme. Cependant l'estimation repose sur des sources anciennes et elle est limitée en termes de finesse au niveau géographique en raison des tailles des échantillons utilisés. Avec les données issues de l'usage des téléphones mobiles, on peut envisager de diffuser des estimations de population à un niveau de maille et une temporalité assez fins.

Avec les données issues de l'usage des téléphones mobiles, l'INSEE peut envisager de diffuser annuellement des estimations de population présente à la maille communale. Le détail temporel pourrait être une moyenne par mois, complétée d'un maximum.

Les sources mobilisables par la statistique officielle pour élaborer la population présente

*L**'enquête DGE-Banque de France auprès des visiteurs venant de l'étranger (EVE)** réalisée auprès de 130 000 visiteurs représente un coût important. Elle mesure le volume trimestriel des flux touristiques des non-résidents à leur sortie du territoire pour les modes de transports collectifs (train, avion et bateau) et sur les aires d'autoroutes pour le mode routier. L'enquête permet de connaître l'ensemble des dépenses réalisées par les non-résidents et contribue à élaborer le poste «recettes» de la ligne «voyages» de la balance des paiements dont la Banque de France a la charge.*

*Les **enquêtes de fréquentation des hébergements collectifs (Insee en partenariat avec la DGE et des partenaires régionaux – CRT)** sont basées sur des interrogations postales et internet réalisées par l'Insee auprès de 12 000 hôtels, 6000 campings et 2500 hébergements collectifs. Les résultats sont publiés au niveau département mais disponibles mensuellement au niveau communal.*

*Réalisée par la DGE en partenariat avec la Banque de France **l'enquête sur le Suivi de la Demande Touristique des Français (SDT**) permet de suivre l'évolution des comportements touristiques des personnes résidant en France. Elle est réalisée mensuellement auprès d'un panel de 20 000 personnes représentatif de la population française âgée de 15 ans et plus. Un volet trimestriel permet par ailleurs de décomposer les voyages en « séjours », définis par le fait d'avoir passé au moins une nuit en un même lieu, et de détailler le type d'hébergement, les raisons du séjour, les activités pratiquées, les dépenses réalisées, etc. C'est ce volet trimestriel qui a été exploité pour élaborer les tableaux relatifs à l'activité « vélo ».Les résultats les plus fins sont diffusés au niveau région et agglomération.*

*Le **Recensement de la population** permet de disposer d'une population résidente au niveau communal mise à jour chaque année.*

*Le **recensement de la population** et les **DADS** donnent des éléments sur les mobilités domicile travail au niveau communal*

Les opérateurs

En France, trois des quatre opérateurs SFR, Orange et Bouygues ont créé récemment une offre payante à partir de leurs données internes.

SFR a mis au point depuis 2013 une offre marchande en direction des acteurs du transport, des collectivités locales, mais aussi pour le secteur de la distribution qui représente aussi l'une des principales cibles du groupe. Orange de son côté a mis au point en 2013 avec les acteurs du tourisme d'une part et avec les acteurs qui s'intéressent aux statistiques du trafic routier d'autre part, une offre d'indicateurs sous le terme de Flux vision. Orange diffuse désormais sous le terme Fluxvision les résultats tous les mois au niveau départemental à de nombreux Comités départementaux du Tourisme (CDT), à des Comités régionaux du Tourisme (membres de la FNCRT) et au niveau France à Atout France qui les intègre dans son bilan, avec des données provenant d'autres sources. Bouygues a mis au point depuis un an une offre payante.

L'Insee avait déjà amorcé depuis mi 2014 des discussions techniques avec Orange et SFR pour monter des expérimentations permettant d'analyser la faisabilité de statistiques de populations présente.

Les données des opérateurs

D'après le rapport d'Eurostat de 2014[9](voir supra), les opérateurs disposent de plusieurs types de données :

- *Des données actives en particulier les CDR (Call Detail Records) qui correspondent aux informations enregistrées systématiquement chaque fois qu'il y a un appel ou un SMS et qui du point de vue de la confidentialité sont particulièrement sensibles car elles contiennent de nombreuses informations sur les appels (durée, numéro appelés.....) à l'exception de leur contenu. Le consentement des personnes est nécessaire pour y accéder*
- *Des données passives liée à la localisation des mobiles par le réseau des antennes et qui peuvent être conservées sur décision d'enregistrement des opérateurs*

Les données souhaitées pour la statistique publique sont plutôt les données dites «passives», c'est-à-dire les enregistrements des échanges entre les téléphones allumés ou en veille et l'antenne téléphonique la plus proche. En effet elles permettent de limiter au maximum la part de la population non couverte et donc les biais dans les statistiques qui en découlent. Cependant le rapport d'Eurostat souligne que les données CDR sont plus faciles, moins coûteuses à extraire pour les opérateurs et qu'elles sont en outre plus standardisées.

La juxtaposition des données pour un portable donné de la connexion aux antennes de sa zone posent des problèmes de protection de la vie privée. En effet, même si on ne conserve pas l'identifiant du portable, il est possible de reconnaître un individu dans une base de donnée à partir de quelques déplacements. La CNIL a donc encadré fortement cette collecte de données des opérateurs et a donné son accord pour une diffusion d'indicateurs agrégeant un nombre minimal d'individu d'une part et imposé des limites dans la conservation des données individuelles qui permettent de construire ces indicateurs.

Chaque opérateur a développé son propre système d'agrégation et son propre système de redressement pour redresser les biais des données initiales. Pour disposer d'indicateurs fiables, il faut en effet extrapoler, la population non couverte en raison de la part de marché de l'opérateur, des limites techniques …

3.3 La concertation

Des contacts réguliers ont été pris avec la Fédération Française de la Téléphonie en février 2015. L'année 2015 a été jalonnée d'annonces et d'évènements et dans le monde de la téléphonie. Aussi, Il a été difficile de trouver le moment propice et les modalités adéquates de la concertation. A la demande de la Fédération et des opérateurs des rencontres ont finalement été réalisées en bilatéral pour préserver la confidentialité de l'offre commerciale (méthodologie et usages) qui interviennent dans les discussions. Les réunions en bilatéral ont été suivies d'échanges techniques qui n'ont pas permis de converger vers la définition de modalités concrètes de transmission de données, ni la production d'estimation des charges de mise à disposition des données.

9 http://mobfs.positium.ee/index.php?id=reports

Le sujet est complexe sur le plan de la méthodologique statistique et chaque opérateur a mené des investissements importants pour mettre au point son offre. Les méthodes des opérateurs diffèrent dès le premier niveau de traitement des données individuelles, en lien avec la nécessaire protection de la vie privée de leurs clients. Les opérateurs ne souhaitent pas dévoiler trop d'éléments relatifs à leur méthodologie en raison des risques commerciaux. Dans ce contexte, il a été difficile d'avancer substantiellement.

L'Insee a présenté sa demande de production de statistique de population présente ainsi que les motivations associées. Il s'agit d'éclairer une demande structurelle pour calibrer des besoins en termes d'équipements ou de moyens humains de différents services publics (sécurité sanitaire, force de l'ordre, sécurité incendie…). L'Insee rassemblerait les données des différents opérateurs pour limiter les sources de biais.

Il ne s'agit pas de répondre à un besoin de connaissance conjoncturelle. La réponse à ce besoin est donc différente en termes de niveaux de détail géographique et temporel de celle proposée par les opérateurs à leurs clients dans leur offre privée. L'indicateur serait diffusé annuellement, avec un détail mensuel et communal pour permettre de recomposer les zones de présence des utilisateurs. Il porterait plutôt dans un premier temps sur la population présente la nuit par souci de simplicité. Il serait accessible gratuitement sur le site de l'Insee pour des questions d'égalité de traitement des demandes ; une description synthétique de la méthodologie accompagnerait cette diffusion pour répondre aux obligations de transparence qui régissent la statistique publique.

Le niveau de détail des données transmises à l'Insee sur lesquelles travailler serait bien sûr supérieur au niveau diffusé pour permettre de réaliser les redressements nécessaires. Le niveau de détail des données transmises est à mettre au point avec les opérateurs dans le respect de la vie privée (en particulier en plein accord avec la Cnil), dans le respect du secret des affaires (pour les données comme pour les informations complémentaires sur la méthodologie ou tout autre aspect technique que l'Insee serait amené à connaître), avec des coûts appropriés à l'objectif. Il peut s'agir de données individuelles sous réserve d'acceptation par la Cnil ou de données pré- agrégées.

L'Insee n'envisage pas de financer les opérateurs. Il s'agit de demander une extraction de données dont le coût serait marginal et donc assumable par les opérateurs. La mise en œuvre concrète envisagée s'inspirerait de la mise en place de l'utilisation des données de caisse pour les indices de prix, en mettant au point avec la profession concepts, méthode et modalités de transmission dans une phase expérimentale, avant de passer à la mise en œuvre d'une production statistique régulière.

Dans la discussion, l'Insee a fait valoir que la mutualisation au niveau de l'Institut dans le respect du secret commercial permettait de fournir en retour aux opérateurs des totaux de population qui leur permettraient de recaler leurs propres productions en bénéficiant de la consolidation des données réalisée par l'Insee.

Tous les opérateurs ne sont pas insensibles à l'idée d'une utilisation des données ayant un objectif d'intérêt général sur un champ qui ne concurrence pas directement leur offre. Une condition nécessaire est que la charge qui en résulte soit gérable par l'opérateur, dans un contexte d'un nombre croissant d'obligations de la part de l'État. Il faut pour cela estimer plus en détail les coûts induits et donc poursuivre l'instruction technique.

La question de l'évolution de la demande de l'Insee dans le temps et des utilisateurs visés a été posée. L'Insee ne devrait pas être concurrent de la diffusion de l'opérateur, en particulier sur le tourisme.

A la proposition d'une acquisition payante, l'Insee a rappelé que cela ne lui paraissait pas possible en l'état dans la mesure où l'institut est tenu d'expliciter ses méthodes pour tous ses utilisateurs et qu'il doit donc être en capacité d'en maitriser le contenu.

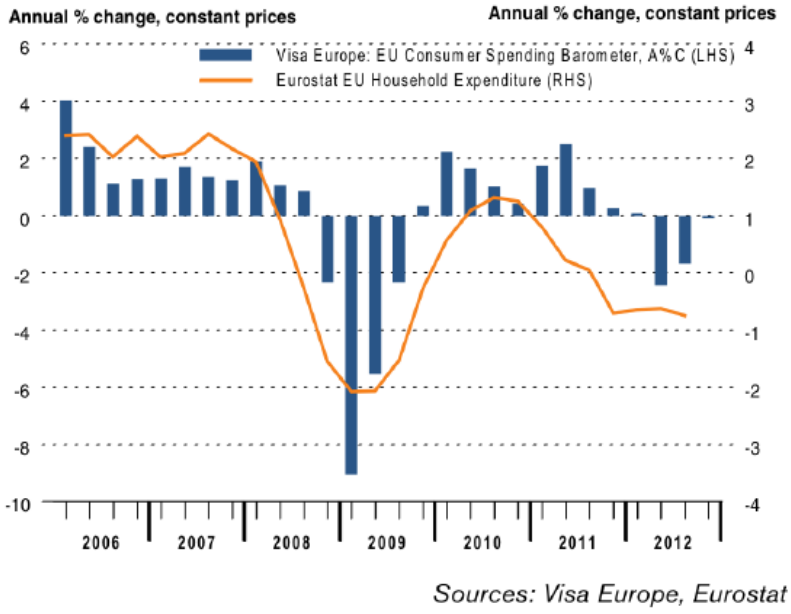
Les discussions ont permis d'échanger les points de vue. Les discussions sont amenées à se poursuivre à l'avenir en dehors du groupe de travail.

4. Les utilisations possibles des données de cartes bancaires

4.1 le contexte international

Le dernier domaine traité par le groupe de travail était celui des cartes bancaires. Dans le cadre des réflexions générales qui sont menées au niveau international sous l'égide d'Eurostat pour analyser les nouvelles sources de données susceptibles d'être utilisées par la statistique officielle, une analyse prospective réalisée pour Eurostat datée d'avril 2014[10] évoque les données de Visa Europe. Elles sont utilisées pour construire un indicateur sur la consommation des ménages au niveau européen dont les résultats sont assez convergents avec les indicateurs diffusés par Eurostat (voir infra).

Figure 2. Year-on-year relative change of Visa Europe's EU Consumer Spending Barometer (left-hand-side) and EU Household Expenditure (right-hand side).



Source: Visa Quarterly report on EU Consumer Spending Barometer (March 2013)[26]

Dans le colloque international sur le Big Data pour les statistiques officielles sous l'égide de l'ONU qui s'est tenu en octobre 2015 à Abu Dhabi[11], le Bureau of economics analysis (BEA) indiquait qu'il travaillait sur des expérimentations (Mastercard, PayPal) pour améliorer les estimations de dépenses de consommation et développer des estimations locales (agglomération et county)

[10] http://www.cros-portal.eu/content/analysis-methodologies-using-internet-collection-information-society-and-other-statistics-1
[11] http://unstats.un.org/unsd/trade/events/2015/abudhabi/default.asp

Une offre privée s'est développée. Ainsi des indicateurs trimestriels ont été construits par la société de services **Markit** spécialisée dans l'information financière pour les niveaux Europe, UK et Suède, basée sur des données fournies par **Visa Europe**. Les indices UK et Suède sont désagrégés en huit postes correspondant à des regroupements de la nomenclature COCOIP en douze postes:

01 + 02 produits alimentaires et boissons (alcoolisées ou non) tabac
03 habillement et chaussures
04 + 05 logement, eau, gaz, électricité et autres combustibles ameublement, équipement ménager et entretien de la maison
06 + 10 santé et éducation
07 + 08 transports et communications
09 loisirs et culture
11 hotellerie, cafés restauration
12 autres biens et services

Pour mettre au point ces indices, de nombreux traitements sont effectués à partir des données journalières qui comprennent les informations suivantes : numéro de carte, identifiant du client, date heure transaction, type de transaction, type de dépense, monnaie, taux de change, montant, pays de la transaction, TVA, type de carte -prépayée, débit, crédit- le service fournisseur. Un apurement est réalisé pour délimiter le champ de la consommation, apurer les remboursements, le cashback. Des redressements sont effectués pour corriger les défauts de représentativité. Ils utilisent le taux de change, les évolutions d'usage des cartes, l'inflation, puis un traitement est réalisé pour éliminer les variations saisonnières. La dernière parution pour l'indice européen date d'octobre 2014[12] pour des raisons inconnues. Les comparaisons entre les indices publiés par Markit et ceux publiés par Eurostat sont plutôt convergents.

Les données de **Mastercard** sont également utilisées pour bâtir une offre commerciale d'indicateurs http://www.mastercardadvisors.com/spendingpulse.html. Les indicateurs de ventes sont détaillés par secteurs fins en fonction du besoin du client, par semaine et sont mobilisables rapidement (quelques jours après la période de référence)e[13]. Ils proposent également des prévisions. Les indicateurs sont utilisés pour le Royaume-Uni, le Brésil, les USA, au Canada , Hong Kong.

Au niveau des discussions entre Banques Centrales, ces données sont également identifiées comme un bon point d'appui pour la construction de la balance des paiements. Elles sont déjà mobilisées, mais avec difficulté et à un niveau très agrégé.

4.2 le projet français et la concertation

Après avoir conduit une concertation avec les professionnels de la distribution, puis les professionnels de la téléphonie mobile, le groupe de travail a engagé une concertation avec les acteurs des données issues des cartes bancaires pour construire des statistiques d'intérêt général.

12 http://www.visaeurope.com/newsroom/all-reports

13 "Reports deliver authoritative industry sales data far earlier than government or industry sources"

Ce volet du travail a été mené en association avec la Banque de France qui utilise les données issues des cartes bancaires à la fois dans le cadre de l'Observatoire de la sécurité des cartes de paiement[14] et pour élaborer des statistiques de la balance des paiements. Pour des contraintes de calendrier, et parce qu'elle n'a pas bénéficié d'une instruction préalable, celle-ci n'a pu être qu'amorcée dans le cadre du groupe de travail. Elle sera poursuivie conjointement par l'Insee et la Banque de France ultérieurement et sera présentée au Cnis dans le cadre de ses activités régulières.

Pour la Banque de France, cette source pourrait être mobilisée davantage à un niveau plus fin pour alimenter la balance des paiements et pour permettre de compléter ses indicateurs conjoncturels. En effet, pour suivre la conjoncture, la Banque de France interroge chaque mois des milliers d'entreprises : les données des cartes bancaires allégeraient ce dispositif en remplaçant une partie des données collectées par des données similaires, notamment auprès des secteurs ayant pour clients les particuliers. Les banques centrales mènent également des réflexions au niveau international sur les sources de données en particulier les cartes bancaires et les données de la téléphonie mobile.

Côté Insee, il s'agit d'instruire la faisabilité de leur utilisation pour mieux suivre la consommation en services dans les comptes nationaux.

Des contacts ont été pris avec la Fédération Bancaire Française en septembre 2015. Des échanges d'informations ont eu lieu avec la Banque de France pour faire le point sur l'utilisation actuelle des données de cartes bancaires et les informations utiles pour cette démarche.

Une réunion du groupe Insee-Cnis s'est déroulée le 13 janvier 2016 à la Banque de France. Une fois rappelés le contexte général des nouvelles sources de données privées pour l'Insee, le contexte de l'article 12 de la loi numérique et les objectifs du groupe de travail, les travaux déjà menés dans ce cadre, la discussion s'est portée sur le potentiel des données des Cartes Bancaires et les conditions d'un éventuel accès.
Il est ressorti de l'entretien que l'utilisation par l'Insee ou par la Banque de France de ces données, seraient soumises aux mêmes restrictions que celles qui s'appliquent aux diffusions des données réalisées par le groupement cartes bancaires :

- toute utilisation des données est soumise à l'accord des banques qui sont propriétaires des données
- la diffusion doit respecter le secret bancaire
- la diffusion de données ne doit pas créer de distorsion de concurrence

Il y a donc trois aspects juridiques à, prendre en compte : le droit de la propriété, le secret bancaire, le droit de la concurrence.

Les aspects de sécurité sont cruciaux. Les normes de protection des données sont drastiques en matière de cartes bancaires[15]. Il est donc indispensable que tout traitement sur les données individuelles présente le

[14] https://observatoire.banque-france.fr/accueil.html

[15] *Le standard PCI DSS (Payment Card Industry Data Security Standard) a été développé pour sécuriser les données bancaires. En effet, malgré une augmentation constante des fraudes, les systèmes d'informations restent encore très vulnérables. Que ce soit au niveau du réseau, des serveurs ou des outils de stockage, les failles sont nombreuses. Mis en place par le comité PCI SSC (PCI Security Standards Council), ce standard est composé d'un ensemble de points de contrôle pour agir au niveau technique et organisationnel. Le GIE CB, qui représente les établissements de paiement, recommande fortement cette accréditation à l'ensemble des acteurs qui ont à gérer des flux bancaires dans leur système d'information.Toutes les entreprises, qui transportent, stockent ou traitent à un moment donné ces informations sensibles, devraient respecter cette norme pour réduire la fraude électronique et empêcher les violations de sécurité.*

même niveau de sécurité. La question de la possible distorsion de concurrence s'est déjà posée lors de la diffusion de données locales.

Il a été rappelé que le groupement cartes bancaires transmet des données au titre des centralisations statistiques de l'Eurosystème.

Il est ressorti de l'entretien que pour cette source comme pour les autres sources qui font l'objet d'instruction dans le cadre du Big data pour la statistique publique, il est utile de faire apparaître un tiers de confiance qui prend en charge les premiers niveaux d'agrégation des données. Cela permet ensuite de travailler sur des données moins sensibles du point de vue de la protection de la vie privée voire du secret des affaires en rassemblant des données d'entreprises concurrentes. Dans le contexte actuel où la Banque de France recueille les données cartes bancaires directement auprès des banques – solution qui s'est imposée pour assurer une couverture satisfaisante et une centralisation rapide – la Banque de France pourrait jouer le rôle de tiers de confiance compte tenu de son habilitation. La Banque de France propose donc d'être le tiers de confiance qui permettrait de résoudre les problèmes de confidentialité au titre de ses missions.

Du point de vue du potentiel statistique de ces données, il est à noter que les moyens de paiements connaissent de fortes évolutions qui sont susceptibles d'affecter la pérennité d'outils statistiques basés sur les données de cartes bancaires.Le paysage général est celui d'une concurrence qui s'exacerbe entre les émetteurs de cartes. Une partie des flux mesurés actuellement par le groupement cartes bancaires pourrait dans ce mouvement de concurrence être réduit. Les évolutions des cartes bancaires font l'objet de réflexion dans l'espace européen : nouveaux systèmes de paiements et en particulier paiements instantanés pourraient fortement modifier le système des cartes bancaires.

Au niveau Européen se produit un mouvement de convergence qui s'est déjà traduit par des normes (SEPA) et qui se poursuit par un règlement européen du 29 avril 2015 relatif aux commissions d'interchanges qui plafonne les commissions interbancaires liées aux paiements par carte à 0,3 % par opération pour les cartes de crédit (cartes à débit différé ou cartes avec réserve de crédit) et à 0,2 % pour les cartes de débit (cartes à débit immédiat).

D'autres limites ont été évoquées dans la discussion sur l'utilisation des données de cartes bancaires pour les projets envisagés : le caractère volatile des acteurs du secteur du commerce électronique et plus généralement les précautions d'utilisation de l'information établissement pour une utilisation économique. Par ailleurs, par construction de la source, l'enregistrement est rattaché à l'établissement et ne permet pas de distinguer certains biens de nature différentes achetés dans la même enseigne.

Les discussions sont amenées à se poursuivre à l'avenir avec des représentants du secteur bancaire en dehors du groupe de travail, en concertation entre l'Insee et la Banque de France pour instruire la faisabilité de statistiques officielles basées sur cette source.

5. Conclusion

La statistique publique se trouve à un tournant de son évolution, comme celui qu'elle a connu dans les années 70 avec l'apparition de nombreux fichiers administratifs. Ceux-ci constituaient en effet grâce au progrès de l'informatisation une nouvelle ressource immédiatement disponible sans resaisie. Ils permettaient d'éviter des collectes par enquêtes - et de limiter la charge d'interrogation qui pèse sur les ménages et les entreprises- de diminuer les coûts de collecte et d'améliorer la qualité des statistiques. En effet l'exhaustivité des informations augmentait la précision des statistiques : par ailleurs les informations administratives pouvaient apporter dans certains cas une information plus homogène et plus précise que la déclaration directe des enquêtés.

La mise en place de l'exploitation de ces données amorcée dès les années 50 pour l'exploitation des données sur les salaires, dans un contexte légal compatible, s'est installée avec le temps et n'a pas manqué de soulever dans les années 70 la question de la préservation de la vie privée des individus et des précautions à prendre pour éviter toute dérive d'utilisation. Elle a été à l'origine de la création de la CNIL et de la loi de 1978 relative à l'informatique aux fichiers et aux libertés. Depuis, les fichiers administratifs sont devenus une ressource incontournable de la statistique officielle comme en témoigne l'article 17 bis de la loi statistique européenne qui assure un accès gratuit et immédiat à l'ensemble des fichiers administratifs (voir annexe) afin de réduire la charge pesant sur les répondants.

L'évolution actuelle a laquelle est confrontée la statistique publique est en de nombreux points comparable à ce changement. Elle présente néanmoins une différence majeure dans la nature du fournisseur des données et pose à nouveau la question de l'évolution juridique de la loi de 1951 pour permettre de franchir une nouvelle étape dans l'évolution de l'offre de la statistique publique. Cette évolution se joue dans un contexte général de montée en puissance de l'utilisation des données, sous l'impulsion conjuguée de l'open data et du Big data)

Les trois exemples étudiés dans ce groupe de travail permettent malgré leur stade d'avancement différent de tirer des conclusions confortées par l'analyse des discussions qui sont menées au niveau international et tout particulièrement européen sur les mêmes sujets.

En effet, dans les trois cas d'utilisation analysés dans le groupe de travail, les fournisseurs potentiels demandent des garanties strictes concernant la préservation du secret des affaires et la non distortion de concurrence ainsi que la protection de la vie privée de leurs clients et par là même la préservation du capital confiance de l'entreprise vis à vis de ces clients. Ces demandes ne sont pas nouvelles pour la statistique publique qui doit répondre aux mêmes exigences vis-à-vis des entreprises mais aussi des ménages dans le cadre de la collecte d'information actuelle. Un encadrement par la loi de 1951 offre déjà des garanties confortées par la loi statistique européenne récemment révisée.

Le caractère général du texte du projet de loi numérique rendu public à mi-parcours a soulevé des inquiétudes. Il a conduit certaines entreprises à penser que toute donnée de leur entreprise pouvait être demandée dans des délais courts. Les discussions ont conduit à souligner qu'il n'en était rien et que, comme pour la mise en place d'une nouvelle enquête, des préalables prouvant l'opportunité et la faisabilité, discutées avec la profession concernée seraient bien sûr nécessaires avant toute mise en place d'une production régulière basée sur une donnée privée. Dans tous les cas, il est apparu que seule une discussion basée sur une expérimentation pouvait permettre de mettre au point des modalités d'utilisations des données et de transfert qui répondent aux exigences des entreprises et qui installent un partenariat de production pérenne dans la confiance réciproque.

Selon l'usage déjà réalisé par l'entreprise à partir de ces données de gestion, la demande de la statistique publique de valoriser ces données dans l'intérêt général pour le bénéfice de la collectivité nationale au service de l'efficacité de l'État reçoit un accueil différent à ce stade de la concertation.

Lorsque préexiste une valorisation marchande des données qui concerne les utilisateurs naturels de la statistique publique se pose le problème d'une ligne de partage entre l'offre gratuite de la statistique publique et l'offre payante du détenteur initial des données. C'est le cas pour l'utilisation par la statistique publique des données de la téléphonie mobile en France.

En effet, le Big Data a créé une prise de conscience collective du côté des entreprises de la valeur qui peut être produite à partir de leurs données internes de gestion. De nombreux changements internes s'opèrent au sein des grandes entreprises pour utiliser au mieux cette nouvelle ressource identifiée. La valorisation de ces données donne ainsi lieu à des optimisations d'organisation interne, des modifications de l'offre client, mais aussi dans certains cas à une offre marchande sur le marché de l'information. Ce phénomène est relativement récent en Europe et les offres marchandes des entreprises sont également assez récentes et pour certaines encore en cours de développement.

Dans le secteur de la téléphonie mobile qui cristallise l'attention de nombreux acteurs publics, les discussions sur les modalités d'utilisation de ces sources au titre de l'intérêt général sont marquées par les fortes réticences des entreprises. La période trop mouvementée dans le secteur de la téléphonie a été particulièrement peu propice aux discussions qui n'ont pas pu être approfondies sur le plan technique faute de disponibilité et/ ou d'appétence du côté des opérateurs. L'argument de l'intérêt général est reçu de façon différente par les opérateurs. La valeur ajoutée potentielle pour les opérateurs résultant d'une mise en commun de leurs données, l'Insee jouant le rôle de tiers de confiance agrégateur, dans le respect de la vie privée et du secret commercial n'est pas à ce stade un argument recevable pour les opérateurs, compte tenu du climat de compétition actuel.

Malgré une analyse de faisabilité dans d'autres pays et un existant statistique en Estonie les modalités pratiques d'utilisation de ces données doivent encore être instruites dans le détail pour permettre d'utiliser les données de différents opérateurs. En effet il faut trouver la bonne méthode statistique pour amalgamer des données de départ potentiellement différentes en raison des choix de méthodologies différentes des opérateurs. Il faut également intégrer le respect de la vie privée des personnes et limiter le coût de mise à

disposition des données. Il y a une tension entre ces différents aspects dans la mesure où pour mieux protéger la vie privée il est souhaitable d'agréger davantage ce qui demande des traitements plus lourds et donc plus coûteux pour l'opérateur. Pour intégrer ces éléments il faut mener des discussions plus approfondies basées sur des expérimentations à partir de données réelles ce qui n'a pu être réalisé dans le cadre du groupe de travail. Faute de coopération des opérateurs, les coûts de mise à disposition des informations n'ont pas pu être évalués même grossièrement.

Les discussions menées dans ce groupe ne constituent ainsi qu'une étape dans le cheminement long de discussion qui va de l'analyse de la faisabilité jusqu'à la mise en place effective d'une production adossée à une source de donnée privée. Elles devront donc se poursuivre pour espérer déboucher à moyen ou à long terme sur une production pérenne. Ainsi, les discussions sur les données de caisse ont pu aboutir assez rapidement compte tenu des longues années d'échanges et d'expérimentations. A l'autre extrémité du spectre, les discussions sur le sujet des cartes bancaires n'en sont qu'à leur début et devront être poursuivies pour confirmer le caractère prometteur de cette source pour la statistique publique. Enfin les discussions sur les données de la téléphonie mobile bénéficient déjà de nombreux éclairages techniques grâce à des chercheurs et des expériences étrangères mais ils nécessitent néanmoins des échanges plus approfondis pour aboutir à un mode opératoire qui respecte toutes les contraintes.

Des discussions de même nature se tiennent dans les différents pays européens et plus largement. En dépit des spécificités nationales, la convergence des législations qui encadrent le travail des instituts statistiques ainsi que les réflexions communes du groupe des CNIL européennes G29 (institutions juridiques) conduisent à échanger surces évolutions de la production du système statistique plus largement dans la sphère européenne. Les instructions à venir devraient en effet bénéficier d'une vision étendue à l'Europe en relation avec les travaux coordonnés par Eurostat et plus largement au niveau international avec des organismes comme l'OCDE et l'ONU.

Annexes

Annexe 1 - Mandat du groupe de travail

Conseil national
de l'information statistique

INSEE
DIRECTION GÉNÉRALE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Paris, le 8 décembre 2014 - n°161/H030

Objet : Élaboration d'un livre blanc sur l'utilisation de données privées par le service statistique public

Monsieur le Président,

Depuis de nombreuses années, les volumes d'informations collectées et produites par les entreprises connaissent un développement exponentiel. Elles pourraient constituer, pour la production de statistiques publiques, une source de données très précieuse.

Dans cette perspective, le cadre juridique dans lequel agit le service statistique public (SSP), conçu et arrêté en 1951, doit être adapté car - quoique aménagé depuis lors - il ne prévoit pas le recours à des données privées.

Conformément aux conclusions d'un groupe de travail interne à la statistique publique associant l'Insee et un service statistique ministériel concerné par ce sujet, l'accès à des données privées pour la production de certaines statistiques publiques devrait être subordonné aux conditions suivantes :

- qu'il en résulte une amélioration ou un enrichissement de la production de statistiques, dans le respect des principes d'indépendance du SSP, de son code de bonnes pratiques et de ses méthodes ;
- qu'il en résulte une réduction de la dépense publique consacrée à la production de statistiques publiques, sans diminution du volume et de la qualité de cette production. Cela impliquera d'instaurer un principe de gratuité : les données seront transmises au SSP à titre gratuit, sauf prise en charge éventuelle des coûts de transmission ;
- que, conformément au principe de la liberté du commerce et de l'artisanat, aucune atteinte ne soit portée à la valeur économique des données des entreprises, au regard des usages qu'elles en font. L'usage de leurs données par le SSP devra donc être strictement restreint à la production de certaines statistiques publiques, à l'exclusion de toute finalité lucrative ou de contrôle.

Avant de traduire ces principes dans la loi, il est souhaitable qu'un groupe de travail, associant des représentants du SSP et des entreprises, élabore un « livre blanc » d'analyses et de propositions opérationnelles partagées (techniques, organisationnelles et juridiques), en vue de développer, au bénéfice de la collectivité nationale, ce nouveau mode de relations entre les deux catégories de partenaires.

.../...

RÉPUBLIQUE FRANÇAISE
Secrétariat général du Cnis : Timbre H030 - 18, Bd. A. Pinard 75675 PARIS Cedex 14 – Tél. : 01 41 17 52 62 – Fax : 01 41 17 55 41 – secretariat-general@cnis.fr – www.cnis.fr

Ce groupe de travail est proposé conjointement par l'Insee, en charge de la coordination de la statistique publique, et le Conseil national de l'information statistique (Cnis), en tant que lieu d'échange entre producteurs et utilisateurs de la statistique publique au sein duquel les questions de charge pesant sur les entreprises répondant aux demandes de la statistique publiques sont régulièrement abordées.

Nous vous remercions d'avoir bien voulu accepter de présider ce groupe, qui sera notamment composé de représentants de fédérations d'entreprises et de représentants (directeurs des services concernés) d'un échantillon des entreprises susceptibles d'entrer dans ce nouveau mode de relation avec le SSP. Il s'agit d'entreprises à réseaux, disposant de systèmes d'information performants, et donc de volumineuses bases de données ou de mégadonnées comportant entre autres des informations sur des tiers.

Pour mener vos travaux, vous serez assisté par un rapporteur et un rapporteur adjoint (respectivement Stéphane Gregoir, directeur de la méthodologie, de la coordination statistique et internationale à l'Insee et Françoise Dupont, administratrice de l'Insee). Le secrétariat du groupe de travail sera assuré par le secrétariat du Conseil national de l'information statistique.

Nous souhaiterions que votre rapport nous soit remis à la fin du mois d'avril 2015.

Nous vous renouvelons nos remerciements et vous prions d'agréer, Monsieur le Président, l'assurance de notre haute considération.

La Présidente du Conseil national
de l'information statistique

Yannick Moreau

Yannick Moreau

Le Directeur général de l'Insee

Jean-Luc Tavernier

RÉPUBLIQUE FRANÇAISE
Secrétariat général du Cnis : Timbre H030 - 18, Bd. A. Pinard 75675 PARIS Cedex 14 – Tél. : 01 41 17 52 62 – Fax : 01 41 17 55 41 – secretariat-general@cnis.fr – www.cnis.fr

Annexe 2 - Composition du groupe de travail et liste des acteurs conviés dans la concertation

Composition du groupe de travail

Monsieur Michel BON, président du groupe de travail
Pierre AUDIBERT, secrétaire général du Cnis
Jean-Luc TAVERNIER, Directeur général, Insee
Michel ISNARD, responsable des affaires juridiques, Insee
Stéphane GREGOIR, rapporteur du Groupe de travail, Insee
Françoise DUPONT, corapportrice du Groupe de travail, Insee
Fabrice LENGLART et Patrick SILLARD Insee, pilotes du projet d'utilisation des données de caisse dans l'indice de prix à la consommation, Insee
François MOURIAUX, Directeur de la balance des paiements, Banque de France
François BRUNET, Directeur adjoint des enquêtes et statistiques sectorielles, Banque de France

Les personnes suivantes ont été conviées :

Pour les données de caisse :

Isabelle SENAND, Fédération du Commerce et de la Distribution.
Éric ADAM, groupe Carrefour
Nathalie NAMADE Directrice des Affaires publiques, groupe Carrefour
Éric DUPRE, Système U
Franck GERETZHUBER Secrétaire Général, groupe Auchan France
J-L. FECHNER Direction des Relations Extérieures, groupe Casino
Michel-Edouard LECLERC, groupe Leclerc
Jean-Pierre BARTHEL, Directeur Général d'ALDI
M. Friedrich FUCHS, Président de LIDL

Pour les données de la téléphonie mobile :

Des contacts ont été pris avec la Fédération Française des Télécoms en la personne d'Yves LE MOUËL Directeur Général de la Fédération Française des Télécoms

Pour l'opérateur Orange :

Bertrand ROJAT
Jean-Luc CHAZARAIN
Pascal CHAMBREUIL.

Pour l’opérateur SFR Numéricable :
Pierre-Emmanuel STRUYVEN
Mathieu GRAS
Marie-Georges BOULAY

Pour l’opérateur Bouygues Telecom :
Renan ABGRALL
Anthony COLOMBANI
Aude LAUNAY

Pour l'opérateur Free
Maxime LOMBARDINI, Directeur général Iliad

Pour les données des cartes bancaires :

Des contacts ont été pris avec la Fédération Bancaire Française : Valerie OHANNESSIAN, Directrice générale adjointe de la Fédération.

Martine BRIAT Directeur des affaires juridiques et bancaires, groupement des Cartes Bancaires
Antoine SAUTEREAU, Directeur des opérations, groupement des Cartes Bancaires

Annexe 3 - Présentation de l'INSEE sur le projet de calcul de l'indice des prix à la consommation sur les données de caisse

Projet de calcul de l'IPC à l'aide des données de caisses





Direction des statistiques démographiques et sociales

Insee

Mesurer pour comprendre

13/01/2015

L'indice des prix à la consommation (IPC)

L'IPC, une production phare de l'Insee:

- 200 000 observations de prix par mois réalisées par 200 enquêteurs dans toute la France + 100 000 observations de tarifs (centralisées)
- Calcul et publication mensuels de l'indice d'ensemble, de 150 sous-indices, de prix moyens et d'un indice de la Grande distribution
- Détermination de la composante française de l'indice européen IPCH: cadre fixé par des règlements européens

- Un intérêt général incontestable pour :
  - Le pilotage de la politique monétaire
  - L'indexation du SMIC, des pensions de retraites, des minimas sociaux, des obligations du Trésor OATI et OATIe, etc.
  - La construction du déflateur de la consommation des ménages dans les Comptes nationaux

LE PROJET « DONNÉES DE CAISSES »

BUT: Améliorer les statistiques de prix à la consommation

- ✓ Fonder les publications de l'Insee (indice d'ensemble, sous-indices et prix moyens) sur un volume d'informations considérablement accru pour en améliorer d'autant la précision et l'efficacité de la collecte

- ✓ Publier de nouvelles statistiques plus détaillées, c'est-à-dire selon une résolution géographique et de nature de produits consommés renforcée

• Besoin exprimé dans le rapport Quinet sur la mesure du pouvoir d'achat (2008 - CNIS)

• Modernisation de l'IPC prônée et encouragée par la Commission européenne (Eurostat)

UN PROJET EN TROIS PHASES

2010

Étude faisabilité

- premiers contacts avec des enseignes de la Grande distribution
  ➔ *accès à 3 ans (2007-2009) de données mensuelles sur 10 familles de produits*
- premiers calculs d'indices sur ces données

2011

Expérimentation

- inscription de l'enquête au programme des enquêtes de l'Insee

2012

- conventionnement avec des enseignes de la Grande distribution volontaires pour la transmission de données quotidiennes
- marché d'achat d'un référentiel d'articles et d'un référentiel de points de vente à la société IRI (qui fait également office de tiers de transmission des données de caisses)

2013

- début des transmissions quotidiennes des données (données de caisses + référentiels)
- mise en place de la base de données

2014

- premiers calculs d'indices sur les données quotidiennes (non inclus dans l'IPC)

201..

Production

LES GARANTIES APPORTÉES AUX ENSEIGNES DANS LA PHASE EXPÉRIMENTALE

- Conventions bilatérales Insee-enseignes (2 ans): les engagements de l'Insee sont
  - de « n'utiliser ces données que dans le cadre de l'élaboration de l'IPC à l'exclusion de tout autre usage »;
  - « à ce que seuls les agents de la division des prix et ceux des services informatiques chargés de l'indice des prix aient le droit d'utiliser ces données, à l'exclusion de tout autre agent »;
  - « à détruire les données sous un délai de 24 mois ».
- Transmission et stockage des données hautement sécurisés:
  - Communication par FTP SSL des fichiers bruts reçus par IRI (sans charge pour les enseignes), chiffrement par échange de clés asymétriques (outil GNUPG)
  - Dépôt dans un coffre à accès nominatif administré par le chef de la division des prix (à ce jour, les données sont accessibles par 6 personnes)
- La réception de ces données s'inscrit dans le cadre de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques
  - Les données sont couvertes par le « secret statistique »
  - Adoption d'un arrêté spécifique aux données de caisses (JORF NOR:EFIS1134706A) qui reprend les termes des engagements de l'Insee tels qu'indiqués ci-dessus.

Annexe 4 - Présentation Insee sur les données de la téléphonie mobile :

Projet d'utilisation des données de téléphonie mobile à l'Insee





Direction de la Méthodologie et de la Coordination Statistique et Internationale

Insee

Mesurer pour comprendre

9/09/2015

Un intérêt au niveau européen

Les instituts de statistiques nationaux européens s'intéressent aux données de la téléphonie

- Une analyse de la faisabilité pour les statistiques du tourisme diffusée par Eurostat en avril 2014

- Au niveau Français :

- Des contacts techniques INSEE / opérateurs depuis mi 2014

Quelle utilisation ?

L'Insee souhaite estimer la population présente sur un territoire :

- Pour calibrer des équipements et des services d'urgence : offre de soin, risques sanitaires, moyens de secours et de sécurité,
- Indicateurs envisagés :

  Nombre moyen et maximal de personnes présentes en nuitée au niveau communal (semaine/w.e)

  Indicateur mensuel
  Diffusion annuelle
  Environ deux trimestres après année de référence

Quel existant ?

Données disponibles

- Population résidente (recensement), communal
- Mobilité domicile travail ou études (RP, DADS), communal
- Déplacements touristiques des français (SDT) région, type agglo
- Visiteurs étrangers (EVE), France
- Enquêtes fréquentation des hébergements collectifs, département

Estimations de population résidente déjà réalisées

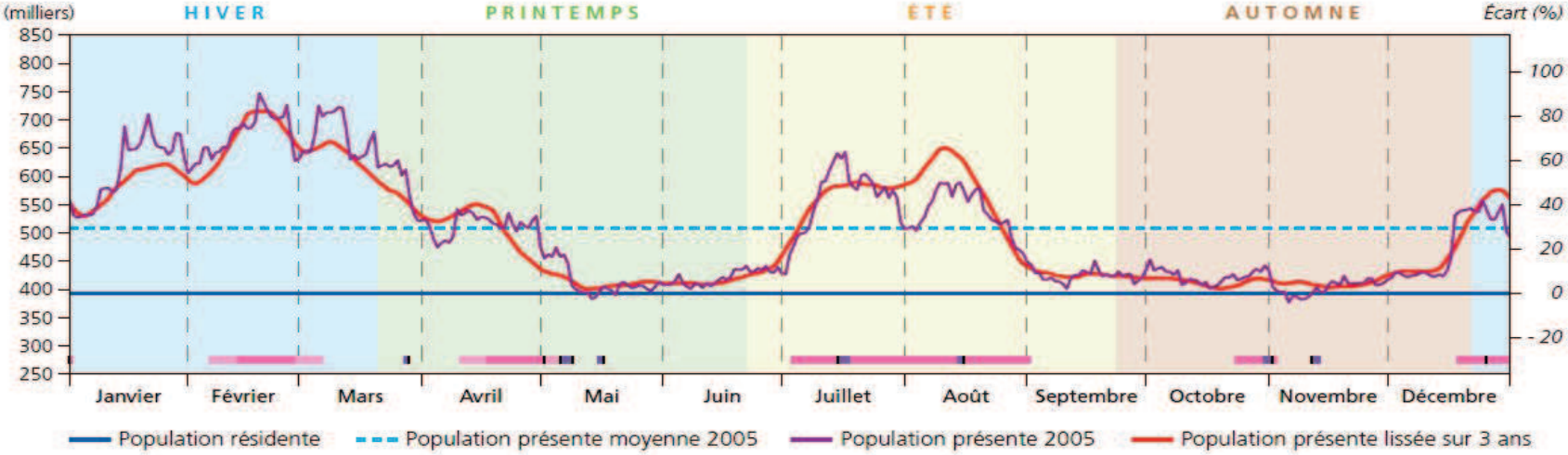
- Population résidente par jour et par département 2005, Direction du tourisme
- Population présente mensuelle par bassin de vie 2005, Insee

Savoie (73)

En moyenne annuelle, la population présente est supérieure de 29,4 % à la population résidente.
En 2005, le maximum de population présente a été atteint le 18 février avec 747 000 personnes présentes.

Population résidente au 1er janvier 2004 (INSEE)		392 300
Population présente moyenne en 2005		507 500
Population présente maximale	18 février 2005	747 000
Population présente minimale	5 novembre 2005	376 100

	Écart	(en %)
Population présente moyenne - population résidente	+ 115 200	+ 29,4
Population présente maximale - population résidente	+ 354 700	+ 90,4
Population présente minimale - population résidente	- 16 200	- 4,1



Population présente maximale, moyenne et minimale par mois en 2005

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Population présente maximale	709 800	747 000	724 900	541 500	473 400	440 800	643 200	589 300	450 100	451 900	423 000	557 300
Population présente moyenne	611 900	673 400	641 500	512 900	413 300	420 900	560 500	535 800	423 600	422 700	402 000	478 100
Population présente minimale	526 500	613 600	521 100	473 900	382 000	399 900	426 200	447 300	400 000	401 000	376 100	419 600

La fréquentation touristique totale représente l'équivalent de 133 200 habitants permanents
(dont 66,2 % de Français et 33,8 % d'étrangers).

Source : Mobilité touristique et population présente, Ouvrage réalisé sous la direction de Christophe TERRIER,Département Stratégie, Prospective, Évaluation et Statistiques/Direction du Tourisme

Quelle organisation ?

L'exemple des données de caisse :

- Une phase expérimentale sur la base de conventions depuis 2010
- Utilisation de données centralisées par la profession
- Mise en place d'un dispositif pérenne à partir de 2015

Dans le respect des principes de la statistique européenne :

- Aucune communication des données en dehors de la sphère statistique
- Respectueuse de la vie privée
- Respectueuse du secret des affaires
- Sans porter atteinte à la valeur économique des données pour les opérateurs

Les garanties apportées aux enseignes de la distribution dans la phase expérimentale

Les engagements de l'Insee sont :

- de « n'utiliser ces données que dans le cadre de l'élaboration de l'IPC à l'exclusion de tout autre usage »;
- « à ce que seuls les agents de la division des prix et ceux des services informatiques chargés de l'indice des prix aient le droit d'utiliser ces données, à l'exclusion de tout autre agent »;
- « à détruire les données sous un délai de 24 mois »

Transmission et stockage des données hautement sécurisés :

- Communication par FTP SSL des fichiers bruts reçus par IRI (sans charge pour les enseignes), chiffrement par échange de clés asymétriques (outil GNUPG)
- Dépôt dans un coffre à accès nominatif administré par le chef de la division des prix (à ce jour, les données sont accessibles par 6 personnes)

La réception de ces données s'inscrit dans le cadre de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques

- Les données sont couvertes par le « secret statistique »
- Adoption d'un arrêté spécifique aux données de caisses (JORF NOR:EFIS1134706A) qui reprend les termes des engagements de l'Insee

Quelle organisation ?

Phase de mise au point avec l'Insee

- Définition des concepts, données et éléments de métadonnées transmises par les opérateurs
- Mise au point de la méthode de consolidation au niveau Insee des redressements avec les données externes disponibles
- Définition des indicateurs diffusés après discussion avec la profession

En régime permanent

- Transmission sécurisée régulière de données pour une diffusion annuelle de l'Insee