Google en sait-il plus que l’Insee sur les Français ?

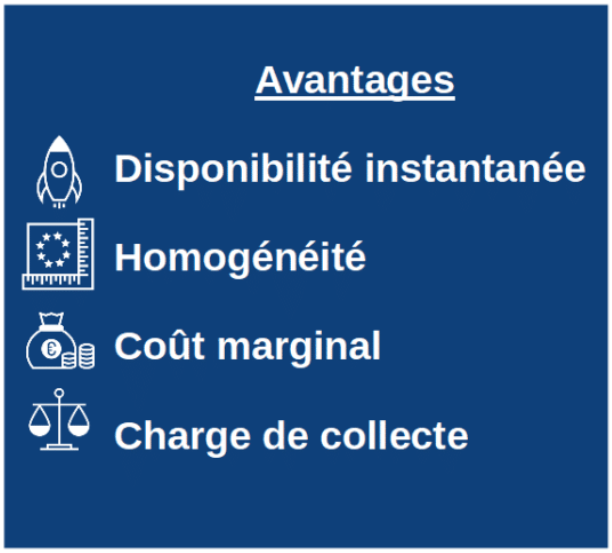
Publié le 18 décembre 2020 sur le blog de l’Insee : https://blog.insee.fr/google-en-sait-il-plus-que-linsee-sur-les-francais/

Temps de lecture : 10 minutes

Benoît Ourliac, d’après l’intervention de Jean-Luc Tavernier, directeur général de l’Insee, aux Journées de l’économie (JECO) 2020

Les grandes entreprises numériques, dont Google, recueillent des volumes considérables de données sur leurs clients. Ces nouvelles sources de données présentent des attraits indéniables, et viennent défier la statistique publique. L’Insee peut et doit naturellement tirer avantage des possibilités qu’elles ouvrent, comme il a déjà commencé à le faire depuis plusieurs années. Néanmoins, l’apport de ces nouvelles sources de données ne peut être jugé en bloc, toutes ayant leurs spécificités : la donnée ne fait pas l’information statistique, et encore moins la compréhension de phénomènes économiques ou sociaux complexes pour éclairer les débats publics. C’est autant grâce aux données qu’elle traite qu’au cadre institutionnel qui entoure ces traitements que la statistique publique remplit cette mission, fondamentale pour la vie démocratique.

Les évolutions technologiques entraînent une inflation des « traces numériques » des activités humaines, dans des champs de plus en plus étendus. Les volumes de données enregistrées par les entreprises, au premier rang desquelles Google, sont considérables, et sans commune mesure avec ce que l'Insee collecte traditionnellement par le biais d'enquête ou de données administratives.



Il y a toutefois une différence fondamentale entre Google et l'Insee dans le traitement de l'information. L'objet social de l'Insee est de diffuser des chiffres et des analyses sur des données agrégées, qui ne peuvent être reliées à un individu particulier au-delà des caractéristiques qu'il partage avec un groupe plus large (âge, niveau de vie, catégorie sociale, lieu de résidence, etc.). L'Insee ne sait pas ce que vous recherchez sur internet, de quel produit vous avez envie, quels sont vos centres d'intérêts ou avec qui vous êtes en relation… Le modèle économique de Google est à l'inverse d'exploiter de l'information au niveau individuel. Mais l'exploitation agrégée de ces données individuelles peut être un produit peu coûteux et porteur d'information statistique : il est donc légitime de se demander si toutes ces « traces numériques » peuvent se substituer aux collectes classiques de la statistique publique, voire si Google peut remplacer l'Insee dans ses missions.

Des « traces numériques » pour prévoir l'activité à très court terme ?

Le premier atout des « traces numériques » est leur disponibilité quasi instantanée. Le suivi de l'activité économique est ainsi un candidat naturel pour leur utilisation dans le champ de la statistique publique. L'Insee exerce ainsi depuis plusieurs années une veille sur la capacité des données moissonnées sur Internet pour mesurer les phénomènes économiques en temps réel (nowcasting) [Données massives, statistique publique et mesure de l'économie, *L'économie française* – Comptes et dossiers, Édition 2017]. Il l'a notamment fait à partir des fréquences des mots-clés recherchés sur Google mesurées par *Google Trends*. Les résultats sont assez mitigés, pour des raisons bien identifiées :

- Les séries disponibles sur Google Trends ne correspondent pas à un comptage exhaustif des termes retenus, mais à un échantillonnage avec des retraitements (par exemple des redressements pour corriger de la diffusion croissante d'Internet et de Google) qui ne sont

pas documentés et peuvent introduire de l'instabilité. Après tout, Google n'a pas pour objectif la cohérence temporelle, mais une amélioration constante de son moteur de recherche, ce qui conduit forcément à une instabilité de la source.

- Il n'est pas possible de connaître le contexte des requêtes effectuées, qui peuvent avoir des motivations très diverses et sans rapport avec l'activité économique que l'on cherche à mesurer. Par exemple, on peut taper « automobile » parce qu'on veut s'acheter un véhicule, mais on peut aussi vouloir se renseigner sur le dernier scandale dans le secteur. Les requêtes peuvent aussi réagir à l'actualité médiatique, voire n'être que l'écho des propres publications des instituts statistiques !

Une alternative à la collecte traditionnelle de données par les instituts statistiques ?

Même si cela ne concerne pas spécifiquement Google, le recours aux « traces numériques » peut être envisagé comme une alternative à la collecte traditionnelle des instituts nationaux statistiques (INS).

Ainsi dans le domaine de la mesure de l'inflation, une des missions les plus emblématiques des instituts statistiques nationaux, un projet de collecte automatisées des prix sur Internet, le *Billion Prices Projec*t, a été lancé à la fin des années 2000 au MIT. Cette initiative est d'autant plus notable qu'elle a été développée en dehors du champ de la statistique publique, et justement pour pallier les carences de celle-ci en Argentine (puis ultérieurement au Venezuela). Sans surprise, les défaillances de la statistique argentine ont été confirmées. Sur les pays les plus développés, il n'y a pas d'écart systématique avec ce que mesurent les INS sur le champ restreint des produits et des points de ventes couverts : par définition, ce projet ne couvre que les prix de vente en ligne de produits vendus en ligne, alors que l'inflation mesurée par les INS est représentative de la totalité de la consommation des ménages[1].

Beaucoup d'INS — dont l'Insee bien sûr — se sont mis à collecter des prix sur Internet, pour tenir compte du commerce électronique et des tarifs de plus en plus différenciés (en France, 160 000 prix sont ainsi collectés chaque mois pour le transport aérien, 350 000 pour le transport ferroviaire). Mais l'abandon des relevés de prix dans les points de vente physiques n'est pas envisageable pour continuer à mesurer précisément l'inflation. D'autres modes de collecte, comme les données de caisse de la grande distribution, ont été développés pour tirer avantage des « traces numériques » laissés par les consommateurs dans ces points de ventes physiques.

Au mieux un complément

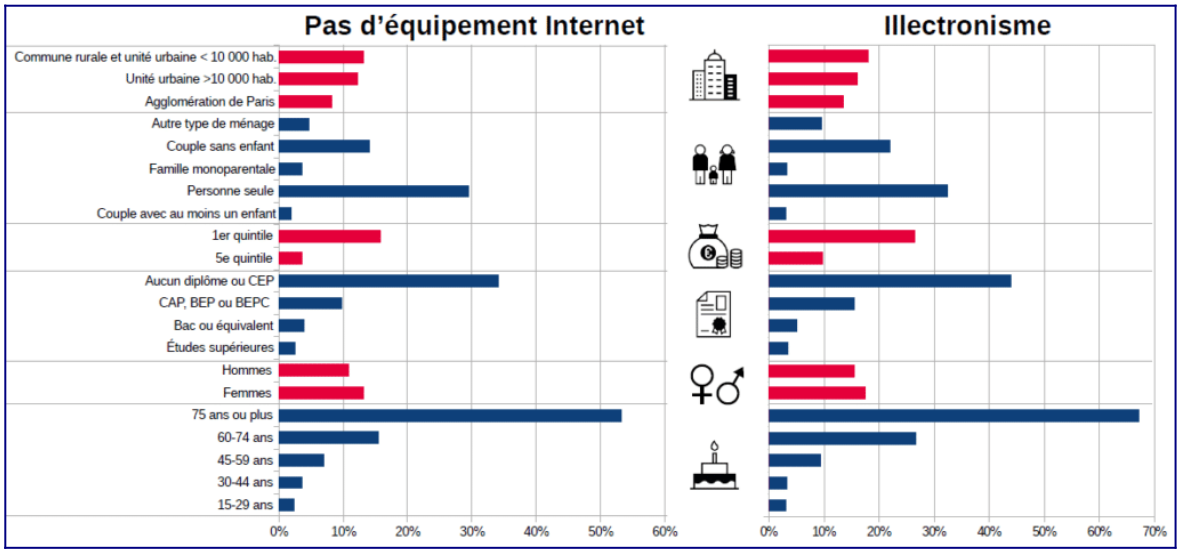
De fait, peu de statistiques publiques seraient aisément remplaçables exclusivement par une moisson de données disponibles sur le web, ou plus largement par le recueil indirect de « traces numériques » laissées par les ménages ou les entreprises. Au mieux cela peut compléter ou enrichir une collecte traditionnelle. En effet, les enquêtes menées auprès des entreprises ou des ménages par l'Insee reposent souvent sur des protocoles assez lourds, des questionnaires relativement longs car il s'agit de mesurer des phénomènes complexes qu'il n'est ni possible ni souhaitable de déléguer à une observation indirecte.

On peut penser par exemple à la situation sur le marché du travail au sens du BIT : pour se passer de l'interrogation directe des ménages, il faudrait notamment disposer d'informations détaillées (et disproportionnées au regard de la finalité) sur leur emploi du temps, pour bien mesurer les démarches pour chercher un emploi et la disponibilité pour en occuper un ; quant au souhait de trouver un emploi, troisième critère pour caractériser les chômeurs au sens du BIT, impossible de le mesurer sans interroger directement les personnes concernées. De plus, cette implication des ménages peut contribuer à la confiance dans les statistiques sur un sujet aussi important dans le débat public, plutôt que des estimations issues d'observations indirectes sur lesquelles les citoyens n'auraient aucune prise [Chômage : les Français mentent-ils aux Français ?, *Note de blog*].

Par ailleurs, la demande sociale ne faiblit pas pour ce type d'enquêtes, et la révolution numérique des données de ces dernières années n'a en rien freiné les besoins de connaissance de phénomènes sociaux, qui ne laissent pas de « traces numériques » directes : pour citer quelques exemples, handicap et dépendance, séparations familiales, sans domicile fixe, conditions de travail, trajectoire des immigrés et descendants d'immigrés, etc.

À cela s'ajoutent les problèmes classiques, mais essentiels pour un institut statistique, de représentativité. Les capacités et compétences numériques de la population sont très inégalement réparties et peuvent entraîner de fait des biais de sélection sévères [Une personne sur six n'utilise pas Internet, plus d'un usager sur trois manque de compétences numériques de base, *Insee première* n°1780]. Il faut également mentionner le peu d'information dont on dispose sur les parts de marché des entreprises numériques et les caractéristiques de leurs clients : les données d'une entreprise ne sont jamais que représentatives des seuls clients de cette entreprise.

Absence d'équipement Internet et illectronisme en 2019
en % de la population des 15 ans et plus



Source : enquête TIC Ménages 2019
(cliquez sur le graphique pour l'agrandir)

Un intérêt renouvelé dans le contexte de la crise sanitaire actuelle

Dans le contexte actuel de crise sanitaire et des mesures de confinement prises par les pouvoirs publics, la statistique publique s'est mobilisée pour répondre au besoin de mesurer rapidement l'ampleur du choc économique et social [Nouvelles données pour suivre la conjoncture économique pendant la crise sanitaire : quelles avancées ? quelles suites ?, *Note de blog*]. Le recours à des indicateurs à haute fréquence s'est imposé, pour leur disponibilité quasi instantanée naturellement mais aussi parce que la collecte traditionnelle ne pouvait pas toujours être assurée de façon satisfaisante. Les problèmes de qualité évoqués précédemment ont pu temporairement être relégués au second plan, du fait de l'ampleur du choc et parfois de l'absence d'alternative.

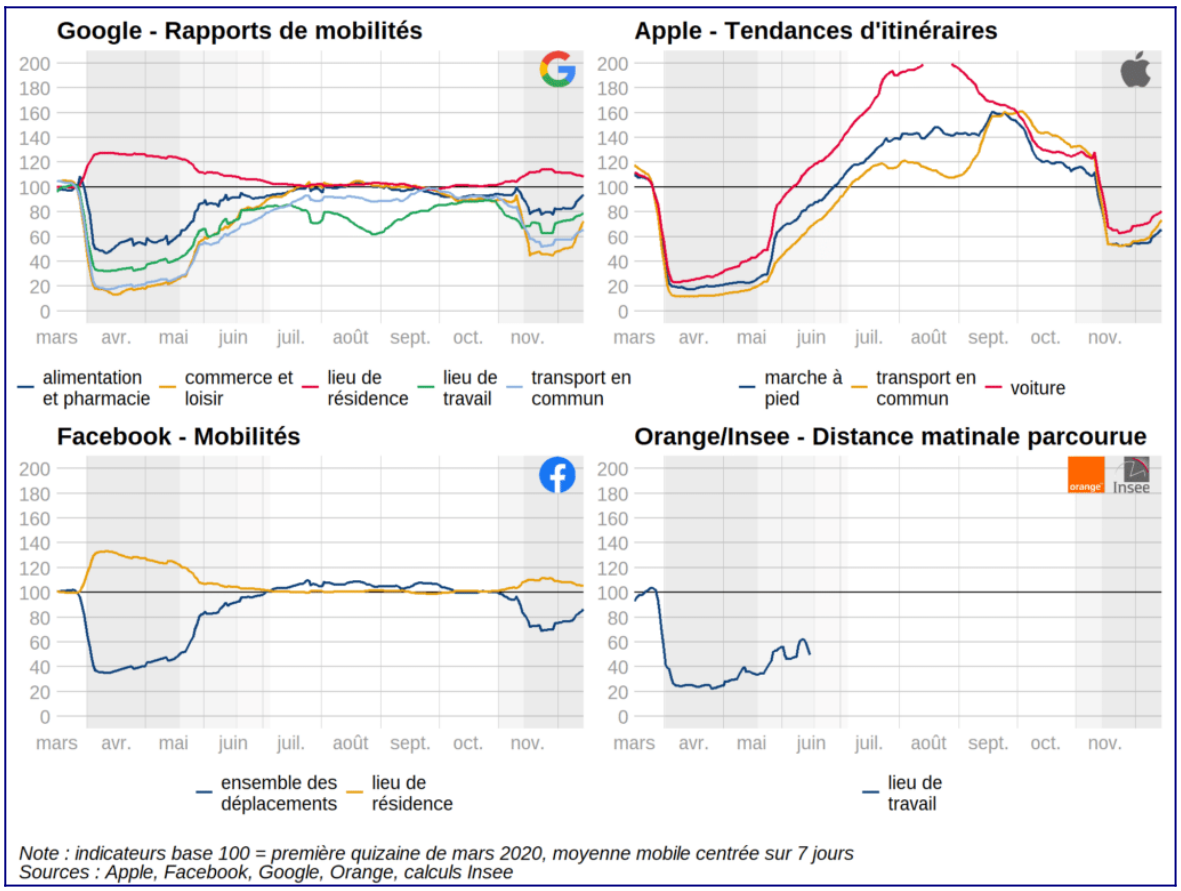
L'Insee a privilégié pour cela les indicateurs les plus directs de l'activité économique (les transactions par carte bancaire, et dans une moindre mesure les consommations d'électricité ou encore des données sur le fret), mais pas de façon exclusive. La confrontation entre les informations, assez qualitatives, qui remontaient des fédérations professionnelles ou des grandes entreprises d'une part, et ces données quantitatives à haute fréquence d'autre part, qui se sont révélées très cohérentes, a permis de donner un ordre de grandeur de la chute de PIB et de consommation dès la dernière semaine de mars.

Mais il n'y a pas eu de travail équivalent de la part de la plupart des autres instituts nationaux statistiques, et il faut bien reconnaître que pour prendre la mesure du choc et le comparer entre pays, les indicateurs de mobilité publiés par Google (et d'autres entreprises numériques) peuvent apparaître séduisants pour combler ce vide.

Pourquoi ont-ils de la pertinence ? Parce qu'il s'agit avec le confinement de chocs très particuliers, dont la première manifestation est la diminution brutale des déplacements (domicile-travail, ou dans des centres commerciaux) ; et les déplacements, c'est que ce Google mesure le plus directement, à travers les services utilisant la géolocalisation de ses clients. De plus, les services de Google sont les mêmes dans tous les pays, ce qui pourrait éventuellement conduire à des données homogènes et comparables entre pays.

Les indicateurs de mobilité : il se passe quelque chose…mais quoi exactement ? Et de quelle ampleur ?

Les indicateurs de mobilité de Google montrent ainsi remarquablement la chute des déplacements à la mi-mars, la reprise plus graduelle au mois de mai, autour du déconfinement, et le nouveau choc consécutif au couvre-feu puis au 2^{e} confinement. Il en va de même pour les autres indicateurs de mobilité rendus disponibles par Apple ou Facebook.



(cliquez sur le graphique pour l'agrandir)

Il faut signaler que la méthodologie de construction des indicateurs de mobilité de Google est inconnue, mais qu'à ce « petit » détail près, la documentation à l'usage des utilisateurs est très bien faite, et d'ailleurs remarquablement prudente :

« Ces données représentent un échantillon de nos utilisateurs. Elles ne reflètent donc pas nécessairement le comportement exact d'une population plus importante. » (et de fait on ignore si les données sont redressées, et si oui comment).
« Évitez de comparer des lieux entre plusieurs régions. Les régions peuvent présenter des différences locales au niveau des données, ce qui peut induire en erreur. Nous ne recommandons pas d'utiliser ces données pour comparer les variables entre pays ».

Pour les tendances de mobilité d'Apple, les données sont générées à partir des recherches d'itinéraire faites dans l'application Apple Plans. Le fort accroissement de l'usage de la voiture à l'été est singulier. La méthodologie n'est pas là non plus détaillée, mais la documentation se veut tout aussi prudente :

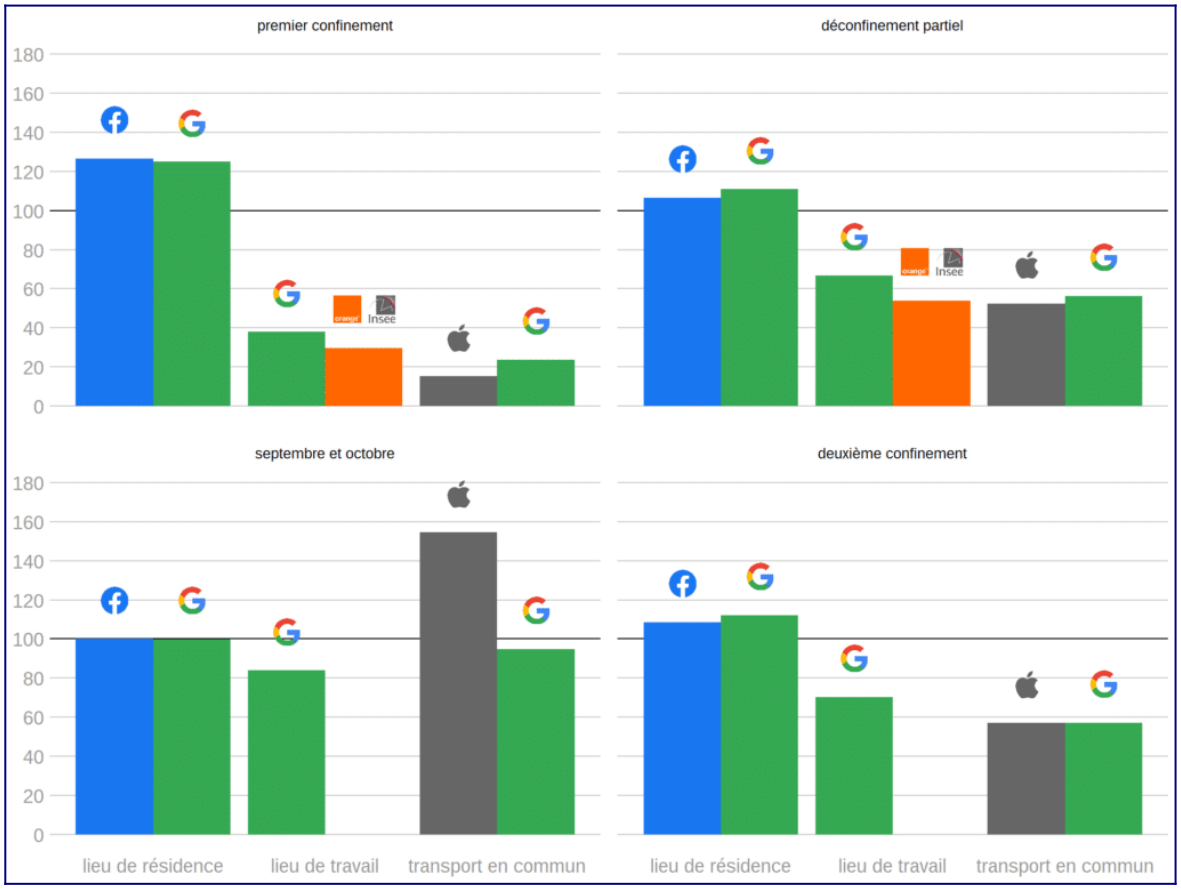
« Comme Apple Plans ne recueille pas de données démographiques sur ses utilisateurs, il est impossible de dire si les chiffres issus de l'application sont représentatifs de l'ensemble de la population. »

Facebook détaille un peu plus sa méthodologie mais fournit un avertissement similaire (en anglais uniquement) :

« It is important […] to consider their limitations. Different data sources will produce mobility estimates that reflect the demographic and geographic coverage of the owners of devices they come from. We expect these to be biased, both across different geographies and among different demographic groups, and must therefore be interpreted in their particular social, economic and political context. »

L'Insee a eu pour sa part temporairement accès à des données agrégées sur l'activation des antennes relais de l'opérateur Orange, et a pu construire des indicateurs de mobilités à partir des distances parcourues sur une plage matinale où les déplacements domicile-travail sont les plus fréquents. Les données ne vont malheureusement pas au-delà de la première phase du déconfinement, fin mai, Orange n'ayant pas souhaité prolonger l'accès gratuit à ses données pour l'Insee.

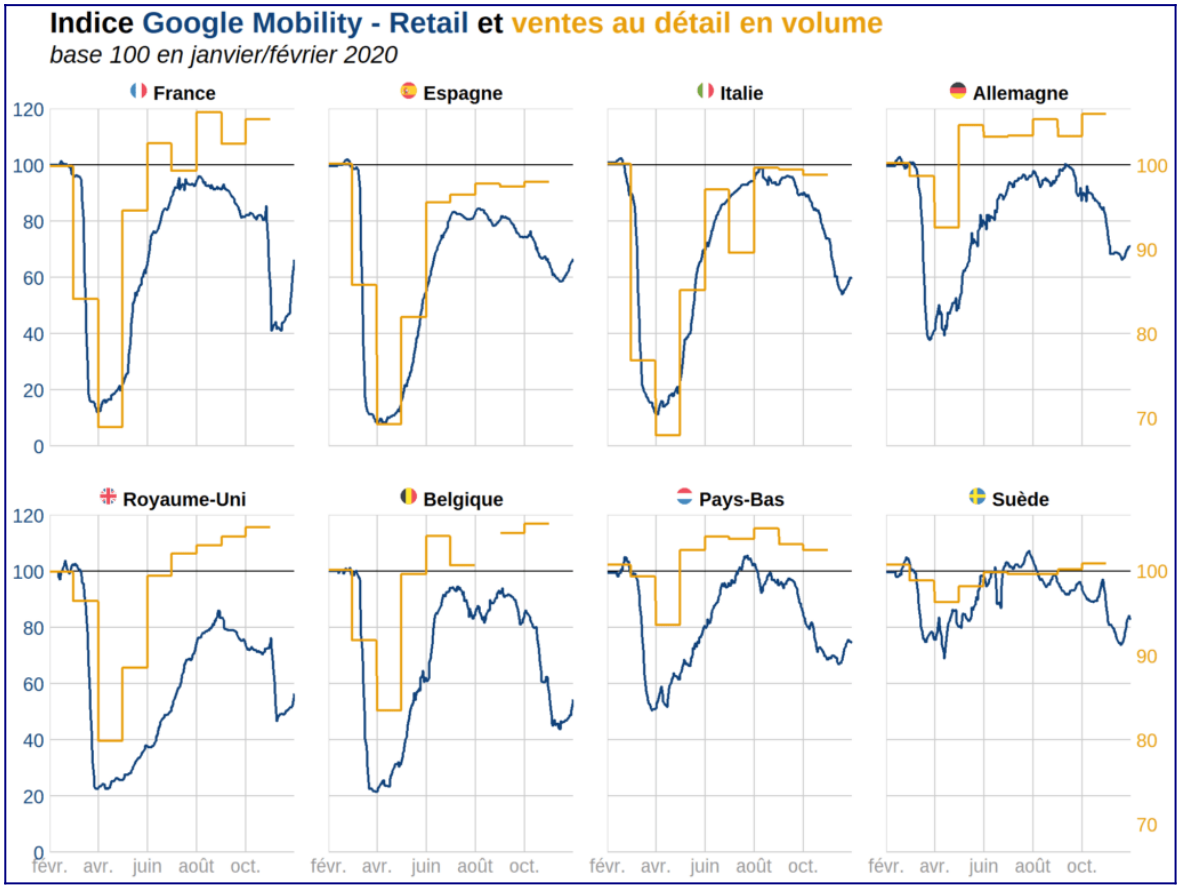
Les courbes sont similaires, mais lorsque l'on essaie de comparer la quantification de l'évolution des déplacements selon les différents indicateurs, les ordres de grandeur sont loin d'être convergents. Il est donc malaisé d'en retirer autre chose comme information qu'il s'est produit des modifications de grande ampleur dans les déplacements de la population.



(cliquez sur le graphique pour l'agrandir)

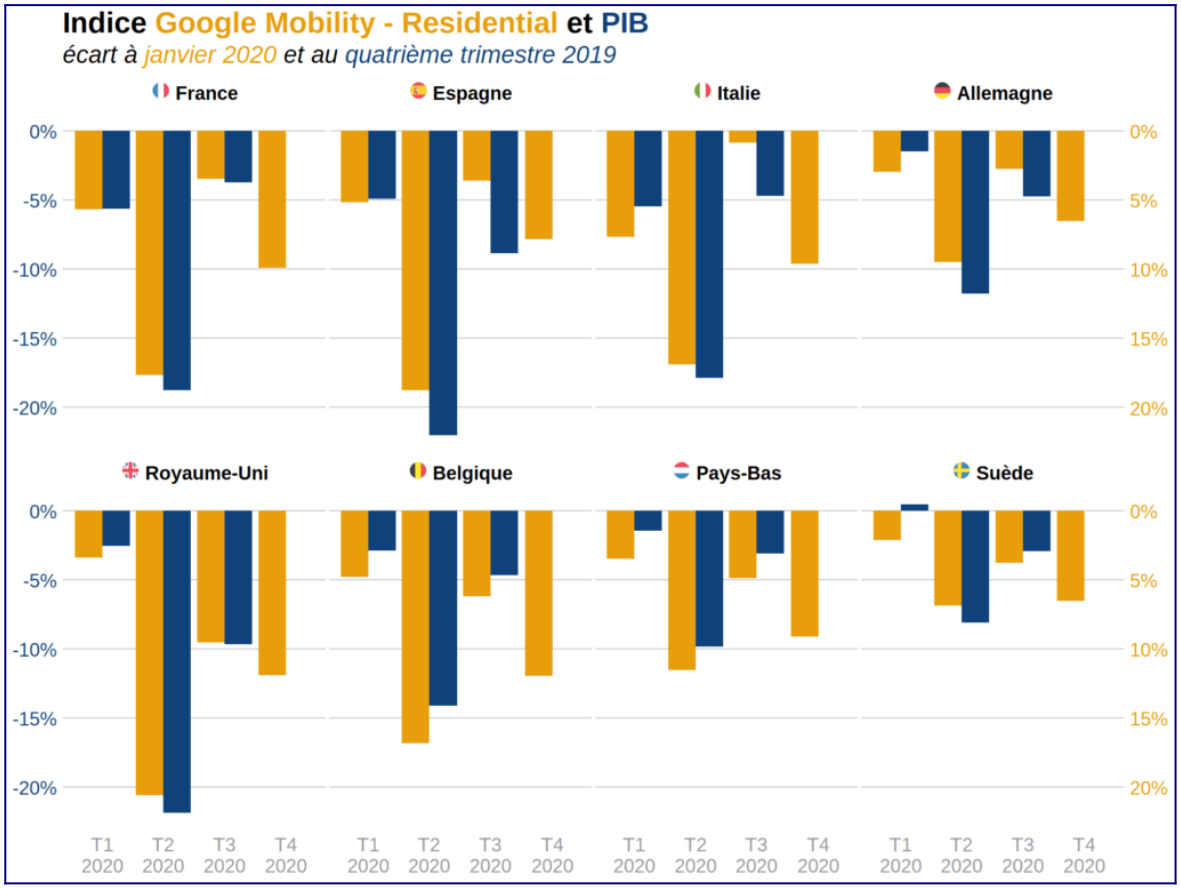
Certains types de mobilité sont liés à des comportements économiques précis et mesurés par ailleurs par la statistique publique, sans en être une observation directe. C'est le cas notamment des déplacements dans les points de vente du commerce avec les ventes au détail, bien qu'ils n'en soient ni une condition nécessaire (du fait de la vente en ligne) ni suffisante (le déplacement en

magasin n’entraîne pas forcément des achats). La confrontation de cet indicateur de mobilité de Google avec les ventes au détail en volume mesurées par les instituts nationaux statistiques donne des résultats très disparates : il y a beaucoup de pays où les ventes au détail ont moins chuté que ce que ces indicateurs laisseraient penser, d’autres où l’ampleur du choc initiale en mars/avril est correctement reflétée. Les divergences entre cet indicateur de mobilités et les données « en dur » s’accroissent avec le déconfinement. Il est peu probable que le plus ou moins grand dynamisme du commerce en ligne puisse expliquer ces divergences.



(cliquez sur le graphique pour l’agrandir)

Un autre indicateur, mesurant le temps passé au domicile, peut donner en creux une mesure de la présence des actifs sur les lieux de travail, dont l'activité économique dépend naturellement. Et il y de fait une très bonne corrélation entre cet indicateur et l'évolution du PIB au fil de l'année 2020. Est-ce qu'on tient là un formidable indicateur anticipé de l'activité, ou bien s'agit-il d'une simple coïncidence ? C'est impossible à savoir, mais en ces temps où le télétravail se développe massivement, le temps passé au domicile peut difficilement à lui seul permettre de prévoir l'évolution de l'activité.



(cliquez sur le graphique pour l'agrandir)

L'information statistique est indissociable de l'environnement institutionnel qui la produit

Il est incontestable que les « traces numériques » recueillies par les entreprises comme Google sur leurs clients représentent des volumes de données considérables. Mais le volume ne fait pas tout : passer de la donnée à de l'information statistique ne va pas de soi, et de l'information statistique à des savoirs utiles pour guider l'action publique ou tout simplement éclairer nos concitoyens sur l'état de l'économie et de la société non plus. Sans méconnaître l'intérêt de ces données, la statistique publique ne peut en tirer parti que par une maîtrise de toute la chaîne de production qui conduit de l'enregistrement de ces « traces » à leur restitution sous forme d'information statistique dans le débat public. Ce n'est pas aujourd'hui le cas. Leurs limites en termes d'information statistique peuvent être rapidement atteintes, même dans le contexte actuel de crise sanitaire où leurs avantages intrinsèques évoqués en introduction sont pourtant renforcés.

La statistique publique se distingue enfin avant tout par les conditions de son exercice : tout ce que l'Insee sait des Français est public, de même que la façon dont ces savoirs sont construits ; son programme de travail est orienté par les demandes sociales, exprimées à travers le Conseil national de l'information statistique (Cnis), et son indépendance vis-à-vis de toute influence extérieure est contrôlée par l'Autorité de la statistique publique (ASP). Nul ne sait aujourd'hui tout ce que Google sait des Français, encore moins pour quelles finalités il produit ces savoirs et avec qui il les partage.

Voir également :

- Google en sait-il plus que l'Insee ? (session vidéo des Journées de l'économie 2020)

Source : Insee, blog de l'Insee : https://blog.insee.fr/google-en-sait-il-plus-que-linsee-sur-les-francais/