Données émergentes : comment les aborder, quelles possibilités et quelles limites ? quelques retours d'expérience de statistique publique

Elise Coudin
Responsable du SSP Lab
Lab de sciences des données
DMCSI, Insee

Insee Mesurer pour comprendre

Cette intervention est faite en toute indépendance vis-à-vis de l'organisateur de la manifestation. Je n'ai pas de conflit d'intérêts en lien avec le sujet traité.

Données émergentes : de quoi parle-t-on ?

- Traces numériques générées par l'activité des individus :
  a. contenu web et réseaux sociaux, textes
  b. enregistrements automatiques (téléphonie mobile, géolocalisation, capteurs routiers, données de caisse, transactions bancaires, données de gestion d'acteurs privés, compteurs intelligents, objets connectés IoT),
  c. images satellites, photos aériennes, plans cadastraux …

- Big Data : Vélocité (accès quasi immédiat), Volumineuses , et de formats Variés (données non structurées)

  → Données “émergentes” en référence à pays/marché émergent : “pays dont le niveau de richesse est inférieur à celui des pays développés mais qui connait une croissance économique rapide (wikipedia)”

Données émergentes : à quels potentiels et limites s'attendre?

- Potentiel
  a. Enregistrement continu laissant espérer une disponibilité rapide (timeliness)
  b. Granularité spatiale et temporelle fine
  c. Réduction du fardeau de réponse
  d. Mesures objectives sans biais de mémoire ou autres biais de réponse (dépenses, déplacements)
  e. Aborder des phénomènes non mesurables par les outils traditionnels

- Limites
  a. données secondaires : non créées avec l'objectif direct de répondre à la question que l'on se pose
  b. souvent peu pertinentes, bruitées, incomplètes, non représentatives (biais),
  c. Nécessité d'avoir de grandes capacités techniques d'accueil et de pré-traitement (volume, variété),
  d. formats changeants (web) demande des investissements/adaptations en continu
  e. accès restreint (privé), méthodologie non transparente, articulation transparence/secret des affaires
  f. Cadre juridique pas toujours présent pour permettre ou pérenniser l'accès
  g. et pour les informations s'appuyant sur des données personnelles des questions d'acceptabilité sociale

- Défis pour l'Insee et la statistique publique
  a. Placement dans l'écosystème avec de nouveaux acteurs/producteurs d'indicateurs issus de ces données
  b. Méthodologies, technologies et infrastructures d'accueil et de traitement, des données, analyse, (datascience)
  c. Modes de travail collaboratifs, open source

Insee Mesurer pour comprendre

Big data, traces numériques et statistique publique

Opportunités

niveau de détail dans le temps
niveau de détail dans l'espace
fraîcheur
nouvelles informations
accessibilité
charge de réponse

Défis

formats non standards
pas faites pour mesurer (secondaires)
biais
non pérennes
bruit

Statistique publique

Qualité

pertinence, pérennité
indépendance, transparence
confidentialité, impartialité
objectivité,
fiabilité
méthodo solide

Expérimentation

La réponse de l'Insee

Pour explorer ces nouvelles sources de données, gagner en compétences, concourir à une culture d'innovation, en recourant à des outils collaboratifs … l'Insee a crée deux unités (en 2018/2019)

- Le SSP Lab inséré dans la direction de la méthodologie et coordination stat et inter

- l'Unissi (unité Innovation et stratégie du système d'information) dans la direction des systèmes d'information, en particulier la division Innovation et instruction technique
  en charge d'apporter les infrastructures, outils et pratiques pour accompagner la transformation numérique de l'Insee favorisant, `collaboration, reproductibilité, open-source, datascience, dataops, SSPCloud/datalab`
  `https://datalab.sspcloud.fr/`

Insee Mesurer pour comprendre

Un SSP Lab ...

Centre de ressource et d'animation pour promouvoir l'innovation statistique

Avec pour missions

- explorer les nouvelles sources de données, les nouvelles méthodes de statistiques et de sciences des données, nouveaux sujets de statistiques au travers d'expérimentations concrètes

- assurer une veille technologique et diffuser les méthodes statistiques et de sciences des données innovantes pertinentes pour la production de statistiques officielles

Notre façon de travailler : expérimenter, explorer, collaborer, mutualiser, animer, former, diffuser

Nos partenaires : les unités métiers de l'Insee, des services statistiques ministériels, Eurostat, Unece et les instituts statistiques nationaux, producteurs de données (privées), chercheurs

Site Internet : http://ssplab.lab.sspcloud.fr/index/

Insee Mesurer pour comprendre

Et en pratique ? 3 retours d'expérience

1. Les données de caisse des enseignes de la grande distribution

2. L'attirail de données émergentes mobilisées pour décrire l'évolution de la situation économique et sociale pendant la crise

3. Le cas complexe des données à caractère personnel avec la téléphonie mobile

1. Les données de caisse des enseignes de la grande distribution

Identifiant du point de vente	EAN	Description de l'article	Date des ventes	Quantités vendues	Prix de vente (en €)	Chiffre d'affaires (en €)
723	3275770004817	[illegible] 150G	20140108	10	1.89	18.90
723	3155230040286	[illegible] BACON 150G	20140108	7	2.38	16.66
986	3185670001080	[illegible] STRAINED SOFT 6%MG 1KG	20140128	25	2.59	64.75

Issu du Courrier des statistiques n3, décembre 2019

Données de caisse

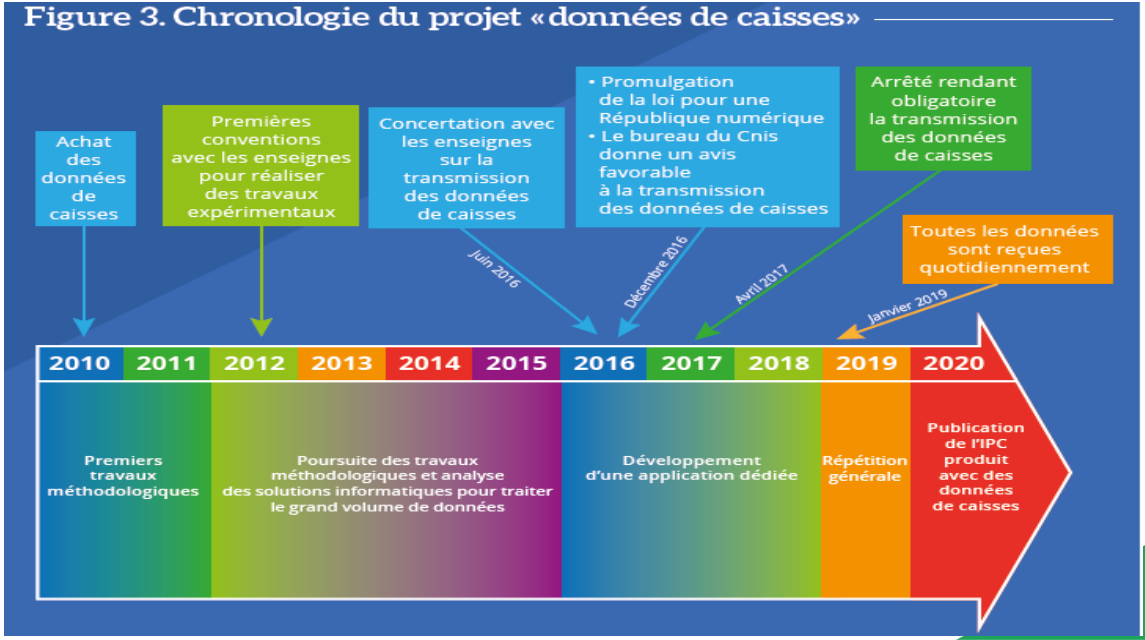
Depuis, janvier 2020 - indice des prix à la consommation (IPC) intègre les données de caisse des enseignes de la grande distribution

- Meilleure couverture des points de vente, des produits, connaissance en détail des volumes (panier), meilleur suivi des prix (jour), moins de charge de collecte des prix en magasin,
- Le code-barre permet d'identifier fiablement dans le temps les produits,et de les classer grâce à un référentiel de codes-barre.

Issu du Courrier des stat n3, dec 2019

Voir aussi Les big data dans l'indice Des prix à la consommation, Economie et Statistique, 2019

2016 modif de loi statistique de 1951



Les enseignements ?

Données de caisse - cas de données “simples” structurées, marché sans valorisation de ces données par les producteurs, agrégateur présent…

Investissement long rapidement mobilisé/étendu

Crise confinement - extension du champ d’utilisation des données de caisse pour continuer à construire l’IPC sans les enquêtes prix, extensions aux Dom

Utilisation des données de caisse pour suivre l’évolution de la **consommation** des ménages (quantités) dans les exercices de conjoncture, des comptes trimestriels, mais champ réduit

Nouvelles perspectives d’utilisation

- Nouvelle source pour les indices de chiffre d’affaire du commerce de détail
- Disparités spatiales de consommation, de prix
- Nouvelles méthodes : classification des libellés dans des nomenclatures autres que celles liées au référentiel des codes-barres

2. L’attirail de données émergentes mobilisées pour décrire l’évolution de la situation économique et sociale pendant la crise

Attirail des données émergentes mobilisées pendant la crise

-Crise sanitaire et confinement général de population - besoin de données quotidiennes reflétant au mieux l'activité économique, d'autant plus que les collectes traditionnelles sont perturbées

→ **Stratégie de l'Insee** [Tavernier, 2020] de fournir rapidement toutes les informations possibles pour suivre la crise et ses effets (mortalité, conjoncture économique) en acceptant de **s'appuyer sur des indicateurs expérimentaux, moins fiables, plus volatils mais disponibles plus rapidemen**t.

- Mobilisation de multiples sources - montants agrégés des **transactions par carte bancaire, données de caisse des grandes surfaces, fret ferroviaire, consommation d'électricité des entreprises, activations des réseaux de téléphonie mobile, nombres de recherches d'itinéraires sur internet, google trends, exploitation mensuelle des déclarations sociales nominatives, textes des articles de presse**…

→ données agrégées, dont on ne maîtrise pas toujours la construction

- utilisation combinée entre elles, aux dires d'experts, remontées fédérations professionnelles, dans le cadre économique, pour alimenter l'analyse économique…

Quels enseignements ?

- Des données hétérogènes dont les apports doivent être évalués au cas par cas

- Données à forte volatilité informatives pour un choc, mais pas pour des temps plus normaux (bruit)

- Les données “éloignées” du phénomène à mesurer peu utiles pour quantifier ce phénomène (ex les images satellites des usines vs paiements par carte bleue), jolie dataviz,

- Les investissements collaboratifs avec les producteurs de données privées ont permis de réagir vite.

- Méthodologie non transparente, non pérenne, peu exploitable (google trends)

- Les données mesurant des quantités (données de caisse et paiements par carte bancaire) utilité plus pérenne

- …aux côtés des outils plus standard (enquêtes)

https://blog.insee.fr/nouvelles-donnees-pour-suivre-la-conjoncture-economique-pendant-la-crise-sanitaire-quelles-avancees-quelles-suites/

3 - Le cas complexe des données à caractère personnel avec la téléphonie mobile

Insee et téléphonie mobile : contexte

- Insee a initié des **collaborations méthodologiques** avec certains opérateurs de téléphonie mobile (Orange Lab le plus ancien depuis 2016). Objectif évaluer le **potentiel** et la **qualité** des données issues de la téléphonie mobile pour la statistique publique - [Sakarovitch et al 2018, Vanhoof et al, 2017, 2018, MobiTic]
- S'inscrit dans une démarche poussée par **Eurostat** [Ricciato et al, 2020]
- S'inscrit dans la demande d'information complémentaire aux statistiques usuelles relayée par le **Cnis** (Moyen terme) → **population présente en complément de** la population légale ou résidente
- Les données de téléphonie ne sont pas collectées dans le but de produire une information statistique. Elles sont le **produit accessoire d'un service**, ici de téléphonie mobile. Les utiliser pour construire des statistiques fiables, robustes, transparentes tout en garantissant le secret reste un **champ de recherche actif**…
- Ce sont des **données à caractère personnel**, la géolocalisation est par nature identifiante. S'appliquent le RGPD et réglementation sur les télécoms (e-privacy).
- Et puis la **crise Covid-2019** est arrivée...

Insee Mesurer pour comprendre

Pendant la 1ère vague, une demande d'accès à des indicateurs anonymes

Juste avant l'entrée en confinement mi-mars, des déplacements massifs de population ont eu lieu, dont l'ampleur n'était pas mesurée

Pour renseigner les acteurs publics dans la gestion de la crise, rapidement, l'Insee propose à chacun des opérateurs de lui fournir des indicateurs anonymes et agrégés de fréquentation et de mobilité, pour des durées limitées

3 opérateurs répondent favorablement, et gracieusement.

Quels **indicateurs** sont fournis à l'Insee ?

- Comptages territoriaux anonymes issus des activations du réseau (via les antennes)
  Et remontés par les systèmes de monitoring du réseau

- Comptages de mobiles pendant la nuit, potentiellement redressés à la population totale par l'opérateur, département de présence croisé par département de résidence, sur des périodes variant selon l'opérateur courant de mi-janvier 2020 à fin mai 2020

- Comptages de déplacements via matrices Origine-Destination entre EPCIs

- Indicateurs issus des **offres commerciales** des opérateurs (pour deux opérateurs)

Des données secondaires, issues du réseau qui présentent des **limites**

- Dépendent du **maillage** en antennes du réseau
- Les **remontées d'information** par les systèmes de surveillance et de contrôle du réseau ne sont pas toujours complètes
- Les **téléphones éteints** ou en mode avion ne se connectent pas au réseau la nuit
- Les **comportements** des utilisateurs peuvent changer avec le confinement
- La méthodologie de construction des indicateurs relève du **secret des affaires** pour l'opérateur et n'est pas entièrement partagée.

L'Insee

- Garde la main sur l'étape finale de redressement à la population totale : recalage chaque jour des indicateurs sur les estimations de population résidente départementales

- Combine l'information provenant de plusieurs réseaux / opérateurs / jours (via modèles économétriques)

- Compléte/ interpréte les phénomènes avec des informations issues des sources de statistiques officielles (recensement, fichiers fiscaux)...

Mais il reste que …

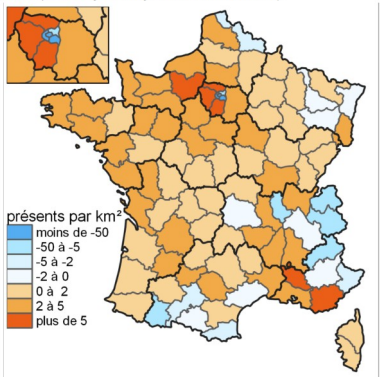
- Les comportements des porteurs de mobile sont extrapolés à l'ensemble de la population
- L'étape de redressement repose sur l'hypothèse que la population résidente en France métropolitaine égale la population présente chaque jour
- Données "pauvres" en caractéristiques
- L'anonymisation à la volée des infos (du fait de leur caractère de données personnelles obtenues sans consentement, eprivacy) réduit fortement la possibilité d'extrapolation/redressement dans un cas plus général

Les outputs

Deux communiqués de presse, avec 1 puis 2 opérateurs, avril et mai

CP 18 mai - effet moyen du confinement

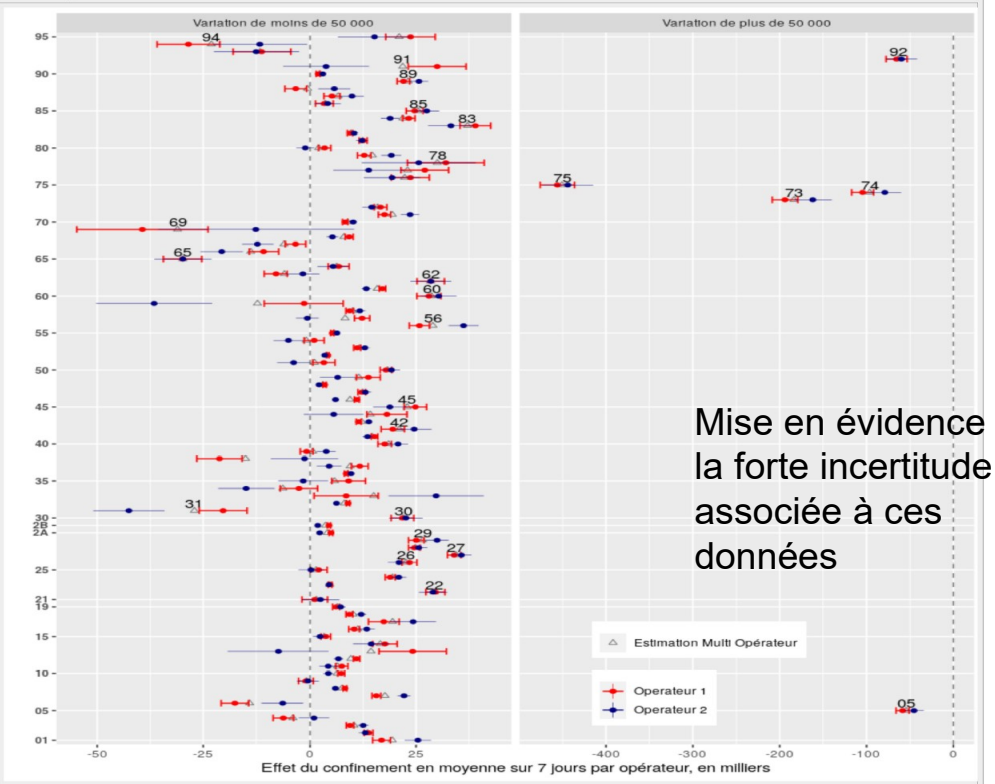
Carte 1 : évolution moyenne du nombre de métropolitains présents dans le département suite confinement (en densité, pour un jour moyen dans la semaine)



Source : Bouygues Telecom, Orange, Insee, calculs Insee.
Lecture : un jour moyen de semaine pendant le confinement, les Pyrénées-Orientales comptent pendant la nuit entre 2 et 5 personnes résidant en France métropolitaine de moins par kilomètre carré qu'avant le confinement.

Galiana, et al (2020), "Retour partiel des mouvements de population avec le déconfinement", Insee Analyses N°54, INSEE

Figure : Estimations de la variation de la population résidente française en nuitée par département suite à la mise en place du confinement, mobilisant séparément les données des deux opérateurs.



Source: Bouygues Telecom, Orange, Insee, Calculs Insee.

Mise en évidence de la forte incertitude associée à ces données

Quels enseignements ?

[Coudin, Poulhes, Suarez Castillo, 2021 Data and Policy]

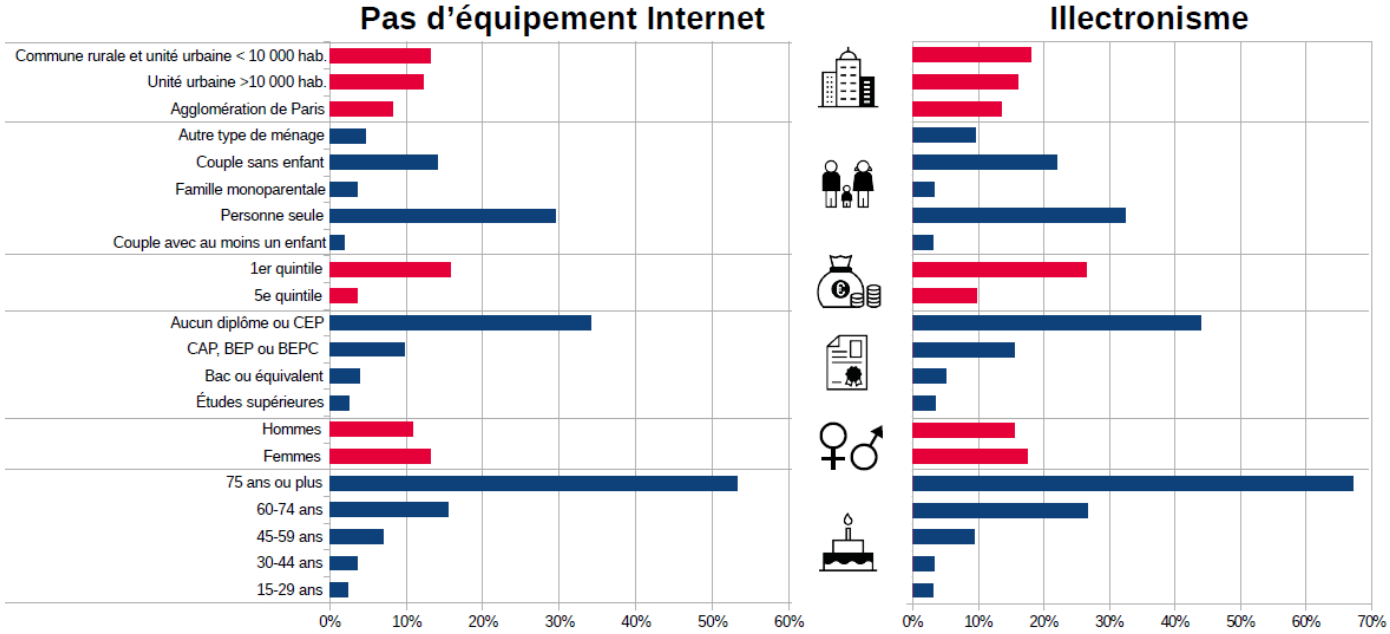
- Données informatives, mais le cadre juridique/écosystème associé réduit actuellement leur potentiel
- Renforcer les collaborations méthodologiques entre opérateurs, recherche et instituts de statistique pour tirer des statistiques d'intérêt général de la téléphonie mobile : investissement à poursuivre en mobilisant des données brutes…
- Faire évoluer l'encadrement juridique de ces données tout en respectant la vie privée (alignement e-privacy, RGPD), investir dans des techniques de calculs préservant la confidentialité (coût pour l'opérateur)
- Promouvoir l'acceptabilité sociale d'un usage raisonné des données issues de la téléphonie mobile. Post de Blog : Sémécurbe, F, Suarez Castillo, M. Galiana, L., Coudin, E. Poulhes, M. (2020)
  Que peut faire l'Insee à partir de données de téléphonie mobile? Mesure de population présente en temps de confinement et statistiques expérimentales
  , INSEE Blog post.
- Articuler les intérêts commerciaux et l'intérêt général, Articuler transparence et secret des affaires
- Construire un éco système et une gouvernance où chaque partie assure son rôle (opérateur, institut de statistique)

Réflexions et perspectives

https://blog.insee.fr/google-en-sait-il-plus-que-linsee-sur-les-francais/

- La donnée ne fait pas l'information statistique, et encore moins la compréhension de phénomènes économiques ou sociaux complexes pour éclairer les débats publics

- Les phénomènes d'intérêt ne sont pas toujours bien représentés dans les traces numériques (issus enquêtes Tic)



Réflexions et perspectives

- Non pérennité des relations mises en évidence - ex mobilité et baisse d'activité (et télétravail qui se développe)
- Non transparence/ non représentativité : ce que dit google de ses indicateurs de mobilité : *« Ces données représentent un échantillon de nos utilisateurs. Elles ne reflètent donc pas nécessairement le comportement exact d'une population plus importante. »*

⇨ Quelques sources très prometteuses, (d'autres beaucoup moins) : revenir à l'objet à mesurer et à la façon dont la donnée est créée (à quoi elle répond en priorité) pour évaluer son potentiel

⇨ Favoriser la combinaison des informations, des sources de données. Considérer les données traditionnelles / resp les nouvelles données comme compléments

⇨ Pérenniser l'usage des données émergentes passe par

- Garantir l'accès, convention/collaboration avec acteurs privés
- Développer les infrastructures et les méthodologies de traitements associées (webscraping, traitement du langage, intelligence artificielle…)
- Acceptabilité sociale d'un usage raisonné

Merci
Des questions?

insee.fr



Insee

Mesurer pour comprendre