Insee
Mesurer pour comprendre

Big Data et Statistique Publique

Lettre d'information
N°10 – décembre 2020

Crise du Covid et confinement : les méthodes de la statistique publique à l'épreuve du feu

Alors que la fin de l'année 2020 approche, une année tellement singulière qui nous a tous chamboulés tant sur le plan professionnel que personnel, cette lettre Big data vient à point. Elle nous permet de mesurer combien l'Insee et les SSM ont non seulement réussi à s'adapter rapidement pour continuer à assurer leurs missions mais ont également innové, notamment sur le plan méthodologique. Certaines productions habituelles ont dû être adaptées pour pallier les problèmes de remontées d'information habituelle (arrêt des enquêtes en face à face, taux de réponse dégradés) ; d'autres ont également été mobilisées selon une fréquence plus importante qu'habituellement et plus rapidement, y compris sur de nouveaux champs d'information inexploités jusque-là. De nouvelles enquêtes originales ou remontées d'informations ad hoc ont été mises sur pied et exploitées en un temps record. Le suivi de la conjoncture s'est intensifié et renouvelé, en mesurant l'impact en temps réel des conséquences économiques et sociales de la crise (*nowcasting*). De nouvelles sources haute fréquence ont ainsi été mobilisées, pour certaines de manière inédite (données de cartes bancaires, données de moteurs de recherche, données de téléphonie mobile…). Cette lettre n'a pas l'ambition de couvrir l'ensemble des innovations menées pendant la crise. Elle se fait l'écho d'un cycle de quatre séminaires organisés par le SSP Lab et le DMS entre mi-octobre et mi-novembre tout en proposant des pistes d'approfondissement. Il est encore trop tôt pour définir précisément comment ces innovations vont se poursuivre mais il est certain que notre activité va s'en trouver transformée dans la durée. La crise a servi de catalyseur au lancement de nombreuses innovations en statistique publique, en France comme dans la plupart des autres pays. Il nous reste à transformer l'essai et à installer durablement de nouveaux processus de travail, l'exploitation de nouvelles sources et de nouveaux partenariats : un beau challenge pour 2021 et au-delà !

Sylvie Lagarde, directrice de la méthodologie et de la coordination statistique et internationale

Conjoncture et crise Covid : comment l'Insee a adapté ses méthodes ?

Comme la plupart des missions de l'Insee, l'élaboration de la conjoncture économique a été bouleversée par la crise. La fermeture de nombreuses d'entreprises a grandement perturbé les conditions de collecte des données traditionnelles. Devant ce choc inédit et soudain, il a fallu en outre répondre à un besoin d'informations plus précoces qu'à l'habitude. Entre adaptation et innovation, les conjoncturistes de l'Insee n'ont pas hésité à sortir des sentiers battus durant la période. En premier lieu, les enquêtes de conjoncture auprès des entreprises ont vu leur taux de réponse chuter après la mise en place du confinement à la mi-mars et l'arrêt brutal d'une grande partie de l'activité économique. Or les soldes d'opinion, qui estiment chaque mois si la tendance est attendue à la hausse ou à la baisse, sont calculés sur un échantillon constant. Cet échantillon contient essentiellement les unités ayant répondu le mois précédent. Pour les non-répondantes au mois N, c'est la réponse du mois N-1 qui est reportée. Lors d'un choc négatif brutal sur l'activité le mois N, la méthode sous-estime grandement l'impact négatif et s'est donc révélée très inadaptée à la situation. Elle a été modifiée depuis avril 2020 et ne s'appuie désormais que sur la base des seules entreprises répondantes. Néanmoins, une réflexion autour du traitement de la non-réponse est prévue et la période a été propice à l'accélération du passage à la production automatisée des questionnaires.

L'indice de la production industrielle qui donne une mesure précoce du suivi de l'activité dans l'industrie est lui calculé à partir des enquêtes mensuelles de branches. Ces dernières ont également connu une chute vertigineuse de leur taux de réponse (- 20 points par rapport au niveau habituel). Là encore, la méthode d'imputation de la non-réponse n'était pas adaptée en cas de non-réponse importante associée à un choc majeur puisqu'elle consistait à reporter en partie l'évolution de l'année précédente. Pour estimer la réponse des entreprises non répondantes, les données sur les heures rémunérées issues de la déclaration sociale nominative (DSN), ainsi que des données de consommation d'électricité d'établissements gros consommateurs ont été utilisées. En outre, la désaisonnalisation appelait aussi un traitement particulier : intégrer sans précaution des points atypiques liés à la crise sanitaire aurait conduit à distordre de façon inadaptée ce traitement. Les chocs ont ainsi été neutralisés du point de vue de la saisonnalité, comme recommandé par les experts méthodologues de la DMCSI.

Enfin, malgré les aléas touchant les sources traditionnelles d'information de l'Institut, il a fallu mobiliser toutes les sources complémentaires pour éclairer au mieux la situation économique du pays et produire sept Points de conjoncture entre mi-mars et mi-juillet, dont le premier dix jours seulement après la mise en place du confinement. Transactions par carte bancaire, données de caisse, activations des réseaux de téléphonie mobile, consommation d'électricité, requêtes sur les moteurs de recherche… : ces sources ont permis de répondre à un besoin d'informations plus immédiates qu'à l'habitude. Mais elles ont aussi leurs limites et ne répondent pas toutes aux mêmes besoins. Toutes n'ont pas la précision et la couverture requises. Ces approches croisées ont permis de conforter les résultats et d'accroître la vraisemblance de l'estimation retenue sans annihiler toutefois l'imprécision de l'exercice, commensurable à l'ampleur des chocs subis. Un des enjeux majeurs des Points de conjoncture pendant cette période particulière a donc été de présenter les résultats de manière adaptée à un contexte inconnu jusqu'alors.

Nouvelles statistiques pour répondre aux enjeux de la crise

La première vague de l'épidémie de Covid-19 a suscité chez les acteurs publics de nouveaux besoins de connaissance. Face à l'ampleur et à la soudaineté des conséquences du confinement sur la société française, l'Insee a dû transformer son traitement des données traditionnelles et recourir à des données expérimentales.

L'Insee reçoit chaque mois les déclarations sociales nominatives (DSN) transmises par les entreprises contenant des informations sur leurs effectifs salariés et les rémunérations . Pendant la première vague de la Covid-19, l'institut a donc pu rapidement mobiliser celles-ci afin d'éclairer l'ampleur du choc sur l'activé. Ainsi, un suivi précoce de l'emploi et de la masse salariale a pu être réalisé, mais aussi des salariés en activité partielle, en arrêt maladie ou garde d'enfant. En interne à l'Insee, elle s'est également révélée pertinente pour accompagner la gestion des enquêtes entreprises bouleversées par la crise. Par ailleurs, l'exploitation des données de téléphonie mobile mobilise les équipes de l'Insee depuis plusieurs années en vue de la production de statistiques expérimentales d'intérêt général et d'études socio-économiques. L'expérience acquise dans ce contexte a été précieuse pour surmonter les principaux obstacles à la production et la diffusion de statistiques pertinentes (complexité des

données, partenariat avec des opérateurs de téléphonie mobile eux-mêmes très sollicités et victimes de conditions de travail dégradées, synthèse d'information provenant de plusieurs opérateurs). Deux thématiques ont été explorées : la répartition de la population et les mobilités quotidiennes. En interne, cette expérience fut un catalyseur pour mieux comprendre les attentes des acteurs publics. Enfin, au niveau régional comme au niveau national, le choc économique engendré par la pandémie et le confinement a été d'une ampleur telle qu'il a fallu apporter des repères statistiques dans un délai inhabituellement court. Au-delà d'une estimation de l'ordre de grandeur de la baisse d'activité économique, l'enjeu était également de savoir pour chaque région et chaque département, s'il avait été significativement plus impacté que la France. Les quelques statistiques régionales habituellement mobilisées pour les publications trimestrielles de conjoncture se sont largement avérées inadaptées au contexte. Ainsi, un groupe de travail sur les «impacts économiques de la crise en région», associant le Département de l'Action Régionale, les Directions Régionales et plusieurs directions statistiques, a été lancé pour apporter des éclairages régionaux ad hoc mobilisant d'autres indicateurs locaux.

Innovations méthodologiques permettant un meilleur appui aux politiques publiques en santé

La Drees était aux premières lignes pour appuyer les politiques publiques de Santé pendant la première vague. Mobilisée au centre de crise sanitaire pour ses compétences en statistiques et de cartographie, l'équipe de datascience de la Drees est partie du constat que les remontées d'informations entre les laboratoires et les établissements de Santé d'une part et les directions du ministère et les ARS étaient lentes et encore artisanales (enquêtes Flash/fichiers excel). Elle a mis en place des systèmes d'information *ad hoc* qui recueillent les données de manière sécurisée sur internet et permettent de consolider et suivre en temps réel les réponses. Les cas d'applications ? Les besoins et la gestion des stocks de respirateurs dans les établissements de santé, les capacités de test et leurs résultats dans les laboratoires. Autre système d'information mis en place en temps record en impliquant de nombreux partenaires (laboratoires, éditeurs de logiciel, APHP cyberlab, Drees…), le système d'information national de dépistage (SIDEP) renseigne sur les résultats des tests. La Drees en reçoit chaque jour une extraction de mise à jour. L'enjeu est pour elle de compiler, dédoublonner, analyser de façon la plus automatisée possible pour renseigner chaque jour sur le nombre de tests, de tests positifs ou encore le délai d'obtention des résultats. Un joli aperçu ici. Aux côtés des systèmes créés de toutes pièces, d'autres ont été adaptés à la crise sanitaire. C'est le cas de SIVIC (suivi et dénombrement des victimes en situation sanitaire exceptionnelle) mis en place par la DGS à la suite des attentats de Paris en 2015 et mobilisé ici pour suivre les hospitalisations Covid. Comme SIDEP, cette base de données est en constante évolution et la Drees réalise des traitements automatisés et quotidiens pour suivre le nombre de décès, prévoir les transferts de patients entre hôpitaux, et lorsque le temps le permet des études sur les parcours de soins. SIVIC aux côtés d'autres sources est aussi mobilisé pour piloter efficacement l'allocation du stock étatique des médicaments de réanimation dans les établissements de santé. La Drees est ainsi venue en renfort à l'agence de sécurité du médicament. Le challenge ? Évaluer rapidement la qualité de nombreuses nouvelles bases de données parcellaires et complémentaires, mettre en place et systématiser des tests de validation pour maîtriser les erreurs classiques en statistique (erreurs de couverture, de réponse, d'observation). Dernier enjeu : faciliter la diffusion et la restitution quotidienne des indicateurs de suivi épidémiologique provenant de multiples sources et utiles à une multitude d'acteurs (directions du ministère, ARS, chercheurs,…). Ceci s'est fait en mettant en place une plateforme centrale de partage des indicateurs et des métadonnées associées et un tableau de bord des indicateurs de pilotage permettant dataviz et création automatique de rapports. Ces ressources et l'application de visualisation « surveillance épidémiologique » sont accessibles sur le site covid19.sante.gouv.fr. Pour mener tous ces travaux, la Drees a fait un choix unifié d'outils : R, GitLab, GitLab CI (intégration continue), Rshiny pour la dataviz. Au final, l'appui de la Drees à la gestion de crise a permis de faire (re)découvrir aux directions "métier" du Ministère combien les statisticiens publics pouvaient apporter en amont dans la construction des systèmes d'information et en aval dans leur expertise et leur exploitation.

Les enquêtes: adaptation et nouveauté

La Crise du Covid a fortement mis à contribution les statisticiens qui ont dû, d'une part éclairer de leurs chiffres un phénomène nouveau et d'autre part, tenir compte des perturbations générées par les circonstances sur leurs outils d'observation statistique afin que ceux-ci continuent de produire des chiffres comparables à ceux produits hors crise. Les enquêtes auprès des ménages tombent dans les deux catégories précédentes. Ainsi, plusieurs instituts nationaux de statistique (INS), dont la France, ont contribué à des enquêtes épidémiologiques nationales pour mesurer la prévalence de l'épidémie auprès des individus. C'est le cas des INS d'Italie, et d'Espagne dont les enquêtes sont assez comparables à l'enquête française EpiCov, menée par l'Inserm et la Drees, à laquelle l'Insee a contribué, notamment en tirant l'échantillon d'enquête et en procédant aux redressements. Ceux-ci ont été rendus particulièrement complexes en raison de l'existence d'un phénomène de sélection endogène très marqué, c'est-à-dire une dépendance entre certaines variables collectées dans l'enquête, comme les symptômes déclarés, et la participation, dépendance qui n'est pas liée à des variables auxiliaires observées. Il faut noter que le protocole de collecte, essentiellement auto-administré par Internet, rend assez probable l'auto-sélection des individus en fonction de l'intérêt pour la thématique d'enquête. Dans ces circonstances, l'existence d'un lien entre variables collectées dans l'enquête directement liées à la thématique et participation est, au fond, assez naturelle. Dans EpiCov, ce lien est particulièrement marqué pour les symptômes déclarés : les personnes répondantes déclarent plus de symptômes que les non-répondants, à caractéristiques données. Un modèle d'Heckman a été ajusté pour calculer des probabilités de réponse non biaisées dans ce contexte, le plan de sondage d'EpiCov ménageant des instruments nécessaires au calcul d'un tel modèle. EpiCov présente ainsi les deux caractéristiques d'être conçue pour éclairer la crise et d'avoir suscité une adaptation assez profonde des méthodes habituellement utilisées pour redresser les enquêtes auprès des ménages.

Les enquêtes auprès des entreprises ont dû s'adapter aussi. C'est le cas, par exemple, des enquêtes de fréquentation touristique dont les modèles de correction de non-réponse ont été revus rapidement. En effet, ces modèles s'appuyaient jusqu'alors uniquement sur l'information du passé. Mais le caractère explicatif du passé peut être remis en cause dans le contexte de crise sanitaire et de ruptures de série. Le modèle modifié s'appuie désormais sur des variables auxiliaires, issues des déclarations TVA et de la DSN, en particulier permettant de déterminer si l'entreprise est ou non en activité au moment de la collecte. Au final, l'impression qui domine est que ces circonstances exceptionnelles ont poussé les statisticiens d'enquêtes à faire preuve d'une très grande réactivité et d'une réelle créativité.

Quelques références/Actualités / Brèves

La page dédiée du séminaire avec vidéos et supports : Intranet ou extranet

Pour ceux qui ne la connaissent pas encore, la nouvelle plateforme https://datalab.sspcloud.fr/

Merci à tous les intervenants du cycle de séminaires pour leur contribution.

Cette lettre est une occasion d'informer largement et d'échanger.
N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à innovation@insee.fr.

Les archives de cette lettre sont disponibles sur agora

Demande d'inscription individuelle à la lettre : dg75-l001@insee.fr