LES RENCONTRES DE STATISTIQUE APPLIQUÉE

Utiliser des données administratives dans le cadre de la recherche en sciences sociales

Mardi 24 novembre 2015, 9h30 - 17h00

Institut Henri Poincaré (amphithéâtre Hermite)
11 rue Pierre et Marie Curie • 75231 Paris cedex 05

Service des Méthodes Statistiques (SMS) de l'Ined et
Société Française de Statistique

Résumés des interventions

Matin

9h45 : Kamel Gadouche (CASD) • **Présentation du Centre d'Accès Sécurisé aux Données confidentielles : données, actualités, perspectives**

Le CASD est un équipement permettant aux chercheurs de travailler à distance, de manière hautement sécurisée, sur des données individuelles très détaillées. On peut qualifier ces données de confidentielles car elles sont le plus souvent couvertes par un secret : secret professionnel, secret des affaires, secret statistique, secret fiscal, secret médical etc. Les données présentes sur le CASD sont donc toutes d'une grande précision, identifiantes ou indirectement identifiantes, et contiennent une grande richesse d'information. La mise à disposition de ces données ne peut se faire que dans des conditions de sécurité très élevée garantissant leur confidentialité ainsi que leur traçabilité.

Le CASD met aujourd'hui à disposition des données des ministères de la justice, de l'éducation, de l'agriculture, des finances pour les données fiscales... Pour ces dernières, il a été nécessaire de modifier la loi (loi ESR de 2013) et qu'un décret soit publié en 2014 pour qu'elles puissent être mises à disposition des chercheurs. Le décret d'application précise explicitement que l'accès ne peut s'effectuer qu'au moyen du centre d'accès sécurisé aux données (CASD) du Genes. De nouvelles sources sont en permanence ajoutées pour les besoins de la recherche. Dans le domaine de la santé, le besoin de sécurité est au moins le même que dans les autres domaines, mais la nature et le volume des données changent ainsi que les usages associés. Ceux-ci sont beaucoup plus diversifiés. C'est ainsi que l'accès aux PMSI (données détaillées des séjours hospitaliers) sera mis à disposition très prochainement sur le CASD.

Depuis près de 3 ans, le CASD travaille sur l'amélioration des conditions de réalisation d'appariements de données (croisement de diverses sources de données). Un article de loi est en cours d'examen par l'assemblée : son objectif est de simplifier la procédure juridique d'interconnexion de fichiers pour les chercheurs grâce à un niveau de sécurité accru.

Les nouvelles possibilités de croisement de données, permises grâce à la loi de santé ou la loi sur le numérique, augmentent aussi mécaniquement le volume des données à traiter. Le CASD a d'ores et déjà commencé à intégrer dans son architecture sécurisée des technologies issues du monde du big data, ouvrant ainsi de nouvelles possibilités d'exploitation des données de gros volume pour la recherche scientifique.

10h15 : Raphaëlle Fleureux et Alexia Ricard (Adisp, Centre Maurice Halbwachs) • **Les données administratives accessibles à l'ADISP (Archives de Données Issues de la Statistique Publique)**

L'Adisp diffuse gratuitement à l'ensemble de la communauté scientifique en France et à l'étranger des enquêtes et bases de données produites par la statistique publique afin de développer l'usage des données statistiques dans les sciences sociales. Le catalogue de l'Adisp est alimenté grâce à des conventions signées avec l'INSEE, plusieurs services statistiques ministériels et des institutions publiques. Parmi les références mises à disposition, environ un quart sont des données dites "administratives" qui peuvent être divisées en deux catégories : des données issues de fichiers administratifs d'une part, et des données résultant d'un appariement entre une enquête et un ou plusieurs fichier(s) administratif(s) d'autre part. Parmi celles-ci, on peut citer les Déclaration Annuelle de Données Sociales (DADS) ou les enquêtes Revenus Fiscaux et Sociaux (ERFS). Ces sources sont accessibles sous forme de fichiers détails, de tableaux agrégés construits à la demande, parfois les deux. La collecte en contenu et le caractère exhaustif de ces données en font des sources incontournables du catalogue de l'Adisp.

11h15 : Maryam Karimi (Inserm), Grégoire Rey (Inserm), Aurélien Latouche (Cnam) • **L'utilisation des données DADS chainées aux causes de décès pour étudier l'association entre trajectoires socioprofessionnelles et mortalité par cause**

Le panel de la Déclaration Annuelle de Données Sociales (DADS), géré par le Département de l'Emploi

et des Revenus d'Activité (DERA) de l'INSEE, a été constitué à partir des déclarations annuelles des données sociales. Depuis 1976, le DERA accumule et relie au niveau individuel tous les épisodes professionnels déclarés par les employeurs, concernant des individus nés en octobre d'une année paire. Le panel DADS est donc un échantillon longitudinal représentatif de la population salariée en France du secteur semi-public et privé non agricole. La base de données Cosmop-DADS a été construite par l'Institut de Veille Sanitaire (InVS) en appariant le panel DADS et la base des causes médicales de décès enregistrée par le Centre d'épidémiologie sur les causes médicales de décès (Inserm-CépiDc). Cette base contient 957 299 hommes et 798 291 femmes.
Pour étudier l'association entre les trajectoires professionnelles et la mortalité cause-spécifique, une première approche consiste à employer une covariable dépendant du temps dans un modèle à risques proportionnels de Cox. Cependant, cette approche ne tiendra pas compte de la bi-directionnalité entre les trajectoires professionnelles et la survie cause-spécifique. Nous proposons enfin d'incorporer les professions comme le sous-model longitudinal dans le cadre d'un modèle conjoint. Lorsque les professions ont été codées selon la classification française créée par l'INSEE, sans ordre hiérarchique claire entre les catégories professionnelles, une implémentation théorique d'un modèle conjoint pour des données longitudinales nominales et de la survie cause-spécifique a été nécessaire. L'application de cette méthode proposée sur un sous-échantillon de Cosmop-DADS sera illustrée.

11h45 : Amélie Carrère et Mathieu Brunel (Drees) • **CARE ménages : Les enrichissements**
L'enquête CARE (Capacités, Aides et Ressources des seniors) ménages, volet senior est une enquête auprès des personnes âgées de 60 ans ou plus. Elle a pour but d'éclairer le débat national sur la dépendance et de répondre à trois objectifs :
- suivre l'évolution de la perte d'autonomie,
- estimer le reste à charge lié à la dépendance,
- mesurer l'implication de l'entourage auprès de la personne âgée.

Répondre à ces objectifs uniquement par interrogation peut s'avérer inefficace et lourd pour les personnes interrogées. Compléter les données d'enquêtes avec des données administratives permet de disposer de données fiables tout en limitant la charge d'enquête supportée par les personnes interrogées. C'est pour ces raisons que la DREES (Direction de la recherche, des études, de l'évaluation et des statistiques), service statistique du ministère des affaires sociales, de la santé et du droit des femmes a prévu d'enrichir l'enquête CARE en ménages, volet seniors avec de nombreuses sources de données administratives.
Cette présentation aura pour objectif de décrire l'enquête CARE ménages volet seniors ainsi que les appariements avec des données administratives associés. Il met en évidence les objectifs de ce type d'enrichissements et les contraintes qu'ils impliquent.

Après-midi

14h15 : Marie Zins, Julie Gourmelen, Marie Genreau, Gaëlle Santin, Alice Guéguen, Marcel Goldberg, Matthieu Carton (UMS 11 Inserm UVSQ Villejuif) • **Utilisation des bases de données médico-administratives : exemples dans Constances**
Les bases de données médico-administratives (BDMA) suscitent un intérêt croissant dans la communauté scientifique. Ces données, massives, quasi-exhaustives, complexes présentent des limitations importantes lorsqu'elles sont utilisées seules, sans lien avec d'autres données. En particulier, elles contiennent très peu d'information sur les caractéristiques sociales et professionnelles des personnes, et aucune information sur le mode de vie, les facteurs de risques… Constances est une cohorte généraliste, labellisée infrastructure nationale de recherche, visant à inclure 200 000 volontaires sur l'ensemble du territoire métropolitain. De très nombreuses données sont recueillies lors de l'inclusion et du suivi par des questionnaires (mode de vie, habitudes, expositions professionnelles, santé des femmes…), par des examens médicaux, paramédicaux (biométrie, vision, biologie, cardiorespiratoire…) et neuropsychologiques pour les volontaires âgés de plus de 45 ans. Par ailleurs, un appariement annuel avec les BDMA gérées par la CNAMTS, la Cnav et le CépiDc (SNIIRAM, carrières professionnelles, causes médicales de décès) permet d'enrichir les données recueillies directement auprès des volontaires. En parallèle de cette cohorte de volontaires participants, une cohorte de non-participants a également été constituée et suivie uniquement au travers de BDMA.
Les objectifs de la présentation sont d'exposer les principaux points du protocole de Constances, les caractéristiques du SNIIRAM, la méthodologie mise en œuvre par l'UMS 011 pour utiliser les données du SNIIRAM et les deux types d'utilisation de ces données dans le cadre de Constances : pondérations pour prise en compte de la non-réponse et enrichissement des données des volontaires. La prise en compte de la non-réponse à l'aide des données des BDMA sera présentée ainsi que deux exemples d'enrichissement des données dans le cadre de projets de recherche.

→

14h45 : Stefan Lollivier (Insee) • **Le Répertoire Statistique des Logements : une nouvelle base de données**

Le projet de constitution d'un répertoire statistique des logements a été lancé par l'Insee en 2011, avec pour ambition de valoriser les données issues de l'administration fiscale sur l'impôt et les propriétés bâties et pour finalité une meilleure connaissance du parc de logement et de la démographique résidente. Ce projet part du constat selon lequel les données administratives sont de plus en plus utilisées dans l'élaboration de statistiques publiques, le Code de bonnes pratiques de la statistique européenne encourageant cette orientation.

Un premier prototype du Répertoire a été élaboré en 2013, après déclaration du traitement à la CNIL. Il a ensuite été progressivement amélioré. Les nombreux travaux méthodologiques conduits sur ce prototype ont confirmé l'intérêt et la qualité de la source afin de réaliser des études statistiques sur les logements et leurs occupants. Les effectifs de population sont cohérents avec ceux fournis par le recensement de la population, exception faite des dénombrements pour les communes de petite taille.

Le Répertoire présente un intérêt tout particulier là où l'exhaustivité de la source est nécessaire. Un domaine d'application privilégié est celui des migrations résidentielles, notamment entre des territoires de taille intermédiaire (départements, zones d'emploi,…). La source renferme également des informations sur les changements de statuts matrimoniaux, qui peuvent ainsi être examinés sur des espaces infranationaux. Plus généralement, la source est propice à l'étude des événements démographiques rares ou portant sur des sous-populations peu nombreuses.

15h45 : Jean-Marc Lazard (CEO OpenDataSoft) • **Open Data, levier de modernisation et d'innovation de l'action publique - Enjeux et exemples liés aux données administratives**

L'ouverture des données publiques est en train de devenir un puissant levier de transformation et modernisation de l'action publique, amenant toutes ses parties prenantes à interagir différemment avec la société et entre elles : exigence accrue de "rendre compte" et partage de l'information, nouvelles manières de résoudre les problèmes et d'innover. Les données administratives s'avèrent un précieux matériau au cœur de ces évolutions, et la problématique de leur ouverture révèle les enjeux liés à l'open data : conditions d'ouverture et protection de la vie privée, ouverture de données alternatives issues d'acteurs privés ... Pour illustrer cette approche générale, nous partagerons également quelques cas d'usages concrets et retours d'expérience.

16h15 : Stéphanie Combes et Pauline Givord (Département des méthodes statistiques - Insee) • **Quels usages des données massives pour les statistiques publiques? Enjeux, méthodes et perspectives**

La prolifération exceptionnelle des données numériques ces dernières années laisse penser que de nombreuses disciplines, dont l'économie et la statistique, bénéficieront de ces nouveaux gisements d'information. Les technologies permettant de traiter de tels volumes de données se sont développées à un rythme impressionnant sur la période récente. L'utilisation de ce type de données pour la production d'indicateurs statistiques ou pour mener des analyses économiques se diffuse. Les applications les plus médiatiques ont porté sur le potentiel des requêtes internet pour fournir des indicateurs avancés des épidémies de grippe (« Google Flu ») ou des indicateurs économiques. Les instituts de statistique publique s'intéressent également au potentiel de ces données, et plusieurs expérimentations sont en cours (utilisation des données de caisse pour l'indice des prix à la consommation par exemple). Les principaux apports identifiés seraient de pouvoir diffuser des indicateurs à un niveau de détail plus élevé (maille locale plus fine, sous-populations par exemple), et pour certaines d'entre elles de réduire les délais de publication grâce à un accès rapide aux flux de données. Le grand nombre d'observations peut également permettre d'améliorer la modélisation des processus économiques, en autorisant des non linéarités et/ou une meilleure prise en compte de l'hétérogénéité.

Nous commençons donc par définir et présenter rapidement les caractéristiques de ces sources, les pistes d'utilisation et les problèmes qui se posent pour les instituts de statistiques publiques (accès, confidentialité, qualité). Leur exploitation soulève néanmoins de nombreuses questions. L'utilisation des données massives est également un enjeu technique et statistique dont le praticien doit avoir une bonne compréhension pour faire des choix méthodologiques raisonnés. Nous proposons un premier aperçu de ces questions, sans viser à l'exhaustivité, ce qui serait illusoire compte tenu du rythme des innovations techniques et statistiques. Nous détaillons ensuite des aspects plus méthodologiques liés à l'exploitation de données de grande taille (analyse de données stockées en parallèle, méthodes de réduction de la dimension).

Voir https://statapp.site.ined.fr/fr/