Statistique appliquée en sciences sociales

Révolution des données, évolution des pratiques

Jeudi 31 mai 2018

Ined (9h30-17h)

10 ANS
LES RENCONTRES DE STATISTIQUE APPLIQUÉE
Journée spéciale | 31 mai 2018
Ined – Paris 20e
Statistique appliquée en sciences sociales : révolution des données, évolution des pratiques

Ces dernières années ont été marquées par l'abondance des données mises à disposition du statisticien, particulièrement dans la recherche en Sciences Sociales. Dans un contexte général d'ouverture, l'accès aux données administratives et aux données de santé représente une évolution majeure que nous devons intégrer.

Comme dans de nombreux domaines, l'apport de ces « Données massives » conduit à nous interroger : comment intégrer cet afflux de données, quelle est la qualité et pertinence de ces données hétérogènes, la méthode à utiliser...

Nous essaierons au cours de cette journée de brosser un panorama sur cet éventail de données potentiellement accessibles, mais aussi sur les avancées depuis 10 ans sur quelques techniques développées notamment à l'Ined qui nous permettent aujourd'hui de les intégrer dans nos analyses.

Inscription sur : https://www.ined.fr/fr/actualites/rencontres-scientifiques/seminaires-colloques-ined/10ansrsa/

Programme et résumés des présentations

Président de séance : France Guérin-Pace (Ined)

9h30	**Arnaud Bringé (Ined) et Bénédicte Garnier (Ined)** Accueil de la journée
9h45	**Magda Tomasini (Ined)** Introduction par la Directrice de l'Ined
10h00	**Javier Nicolau (Drees) • L'accès aux données du système national des données de santé (SNDS)** La loi de modernisation de notre système de santé du 26 janvier 2016 a réformé la procédure d'accès aux données de santé et créé, par son article 193, le Système national des données de santé (SNDS), constitué de cinq grandes sources de données : les données de remboursement de l'assurance maladie, les données de séjours hospitaliers, les causes médicales de décès, les données relatives au handicap et un échantillon des données des organismes de l'assurance maladie complémentaire. Les trois premiers flux sont d'ores et déjà chaînés, ce qui constitue une base considérable (elle couvre toute la population) pour mieux connaître la santé des citoyens et permettre son amélioration. Depuis septembre 2017, une procédure d'accès a été mise en place. Elle permet à tous les acteurs publics et privés qui veulent mener des recherches, études ou évaluations avec les données du SNDS de demander une autorisation à la CNIL.
10h35	**Lidia Panico (Ined) • Families over the lifecourse: Quels apports des cohortes de naissance ?** Les études de cohortes sont utilisées depuis longtemps pour étudier les phénomènes sociaux et leurs impacts sur les enfants et leurs familles. Elles ont l'avantage d'inclure le temps comme facteur critique et ont donc été des excellents outils pour étudier les trajectoires des individus et comprendre l'enchaînement causal des relations. Dans cette présentation, je mettrai en avant des utilisations novatrices des données de cohortes, avec des exemples de comment ces études ont évolué pour aborder des questions émergentes en sciences sociales et en santé publique ; comment des nouvelles sources de données ont été ajoutées aux études de cohortes pour enrichir leur valeur ; et comment les études de cohortes peuvent être une composante essentielle de notre infrastructure de données. Je parlerai également des nouveaux risques et des défis auxquels est confronté ce type de collecte de données.
11h10	Pause café
11h25	**Laurent Toulemon (Ined) • Accéder aux données administratives : nouvelles données, nouveaux outils** La « loi pour une république numérique » organise, entre autres, la mise à disposition des données des administrations à des fins de recherche. Elle facilite également les appariements de fichiers. La loi offre donc des possibilités nouvelles et nombreuses pour la recherche. Reste à construire et documenter des fichiers utilisables à des fins de recherche, à partir des fichiers administratifs de gestion. Pour cela, les utilisateurs de ces données doivent participer à la construction et la documentation de ces fichiers. Deux exemples concrets seront présentés.

→

12h00	**Sophie Pennec (Ined) • Générer des données par microsimulations : Apports en démographie** Dans cette présentation, les fondements du développement de la micro simulation seront rappelés ainsi les types de modèles et les principales applications de la méthode. Sur la base de l'exemple du modèle DynoptaSim en cours de développement, seront présentées les différentes phases de construction de ce type de modèle.
Président de séance : Léonard Moulin (Ined)	
14h15	**Elise Coudin (Insee, SSP LAB) •Données massives et statistique publique : quelques retours d'expérience** Les Big Data interpellent la statistique publique dans sa production d'indicateurs et d'études à partir de sources traditionnelles (enquêtes, données administratives). Disponibles plus rapidement, délivrant de l'information à une échelle très fine, a priori sans coût de collecte, mais au prix d'informations au format complexe, parfois changeant, de biais de représentativité, ces nouvelles données ouvrent-elles des opportunités de remplacement des sources traditionnelles dans la production statistique ou plutôt de complément ? Au travers l'exemple des données de téléphonie mobile, cette présentation mettra en évidence l'apport et les limites de ce type de données pour l'estimation localisée de population résidente et pour l'analyse de la ségrégation sociale.
14h50	**Julien Bolaert (Sage Univ. de Strasbourg / iPOPs Ined) • Apprentissage statistique en sciences sociales** L'objet de cet exposé est de présenter un panorama de ce que l'apprentissage statistique (machine learning) peut apporter aux sciences sociales. Au-delà de la reconnaissance d'image ou la maîtrise du jeu de go, il offre en effet de nombreuses méthodes aisément transposables à la recherche. D'une part, des méthodes désormais bien établies offrent des alternatives aux méthodes plus standard, qu'il s'agisse de régression ou d'analyse factorielle. D'autre part, l'apprentissage pourrait à court terme contribuer significativement à l'analyse de séquences, à l'analyse textuelle, voire à la production de données.
15h25	Pause café
15h40	**Nicolas Robette (Ensae, Crest-LSQ, Ined) •L'analyse de séquences pour étudier les trajectoires individuelles: un état des lieux, quelques limites et des promesses** Au cours des vingt dernières années, l'analyse de séquences s'est imposée comme l'un des principaux outils pour étudier les parcours de vie (life courses) et les carrières. Si elle a fourni beaucoup de résultats intéressants, ceux-ci ont rarement marqué une rupture avec les connaissances existantes. Cette limite est peut-être pour partie dûe à des usages qui se sont progressivement standardisés. Il existe cependant des approches alternatives à la construction de typologies de séquences, et des innovations méthodologiques récentes semblent ouvrir de nouvelles voies prometteuses.
16h15	**Valérie Golaz (Ined / Lped) et Fabrice Yameogo (Aix-Marseille Univ. / Lped) • Articuler données individuelles et données exogènes grâce aux méthodes multiniveaux: De nouvelles possibilités mais toujours des écueils** L'approche multiniveau permet d'associer à des données individuelles des variables agrégées issues d'autres sources (enquêtes, données administratives, données issues de l'analyse spatiale, etc.). De plus en plus de données sont accessibles facilement, et parfois directement via internet. Néanmoins il demeure difficile de trouver une adéquation des données dans le temps et dans l'espace. La présentation sera illustrée d'exemples est-africains (Kenya, Ouganda) élaborés dans le cadre de la préparation d'un atelier du Work Package 5 *education* du projet DEMOgraphy-Statistics-for-AFrica (https://demostaf.site.ined.fr/).
16h50	Discussion et conclusion

En savoir plus sur ce séminaire du service des méthodes statistiques https://statapp.site.ined.fr/