RÉSUMÉS

Séance 1 -- Discours principal

(A) La qualité et les statistiques officielles : le présent et l'avenir

Paul P. Biemer, RTI International, États-Unis

Le mot « qualité » est aussi ambigu qu'il est omniprésent. Nous tenterons de dissiper cette ambiguïté quelque peu et de répondre à un certain nombre de questions courantes. Qu'est-ce que la qualité? Comment est-elle évaluée? Quelles sont les stratégies éprouvées pour l'améliorer? Comment peut-elle être gérée efficacement dans un organisme statistique? Il sera surtout question de « statistiques officielles », c'est-à-dire des statistiques produites par les organismes nationaux de statistique (ONS) à des fins liées aux politiques publiques. Nous examinerons les concepts de la qualité organisationnelle, de la qualité des processus et de la qualité des produits, des dimensions de la qualité et de leurs utilisations, de l'erreur d'enquête totale par rapport à l'incertitude statistique totale, des risques intrinsèques par rapport aux risques résiduels pour la qualité, ainsi que des méthodes pratiques d'évaluation de la qualité. Nous discuterons aussi de certains problèmes de qualité associés à l'intégration de données d'enquête et d'autres sources de données pour la production de produits de données hybrides, une activité de plus en plus courante. Nous traiterons aussi de l'incidence de ces concepts pour les ONS et les autres organismes de statistique, question de réfléchir à l'avenir des statistiques officielles. Pour illustrer les idées, nous utiliserons des exemples d'initiatives récentes axées sur la qualité au sein de plusieurs organismes statistiques.

Séance 2A -- Estimation sur petits domaines

(A) Mégadonnées, mégapromesse, mégadéfi : l'estimation pour les petits domaines (EPD) a-t-elle sa place dans un monde centré sur les mégadonnées?

Partha Lahiri, Lijuan Cao, Ying Han, Kartik Kaushik et Cinzia Cirillo, University of Maryland, États-Unis

La demande pour diverses statistiques socioéconomiques, sur le transport et sur la santé pour de petites régions géographiques s'accroît sans cesse, alors que les organismes d'enquête cherchent désespérément des moyens de réduire leurs coûts pour respecter leurs exigences budgétaires. Dans l'environnement d'enquête actuel, l'application de méthodes d'enquête-échantillon normalisées pour de petites régions, qui nécessitent un grand échantillon, n'est généralement pas faisable lorsqu'on tient compte des coûts. L'accessibilité de mégadonnées tirées de sources multiples offre aux statisticiens de nouvelles occasions d'élaborer des méthodes novatrices d'EPD. Nous expliquerons d'abord comment extraire des renseignements pertinents à partir des mégadonnées de médias sociaux pour résoudre un problème d'EPD. Nous décrirons ensuite comment l'EPD peut aider à résoudre un problème en apparence différent, qui consiste à prévoir la circulation en temps réel en exploitant les riches mégadonnées tirées des sondes de véhicules.

(A) La correction d'erreurs de mesure dans les grands registres de population par l'estimation pour les petits domaines

Danny Pfeffermann et Dano Ben-Hur, Central Bureau of Statistics, Israël

Comme bien des pays, Israël possède un registre de population assez précis au niveau national, qui compte environ 8,5 millions de personnes. Toutefois, le registre est beaucoup

moins précis pour de petits domaines avec une erreur moyenne du dénombrement d'environ 13 %. La principale raison des inexactitudes au niveau du domaine est que les gens qui quittent une région ou viennent s'y établir tardent souvent à déclarer leur changement d'adresse. Afin de corriger les erreurs au niveau du domaine, nous examinons la procédure en trois étapes que voici.

A- Prélever un échantillon du registre afin d'en tirer des estimations initiales du nombre de personnes résidant dans chaque domaine le « jour du recensement ».

B- Appliquer le modèle de Fay-Herriot aux estimations initiales pour en améliorer l'exactitude.

C- Calculer une estimation définitive pour chaque domaine en tant que combinaison linéaire des estimations obtenues à l'étape B et du chiffre du registre.

Nous discuterons des considérations motivant cette approche et nous présenterons des résultats empiriques fondés sur le dernier recensement mené en Israël. Nous examinerons et illustrerons une procédure permettant de tenir compte de la non-réponse non aléatoire à l'étape A.

(A) L'estimation pour les petits domaines fondée sur des modèles faisant appel à des mégadonnées

Stefano Marchetti, Université de Pise, Italie

Les organismes nationaux de statistique ont pour mission de produire des statistiques pour les citoyens et les décideurs. Il est reconnu que l'échantillonnage constitue un moyen efficace d'obtenir des estimations actuelles fiables pour une région particulière dans des domaines socioéconomiques. En général, il est important d'estimer des paramètres de population à un niveau régional plus pointu, où la taille de l'échantillon est habituellement restreinte et ne permet pas d'obtenir des estimations directes fiables. Les méthodes d'estimation pour les petits domaines (EPD) — qui reposent sur le couplage de modèles en fonction de variables auxiliaires permettant d'« emprunter de l'information » parmi les petits domaines — ont pour objectif d'obtenir des estimations fiables lorsque les estimations directes ne sont pas fiables. Les méthodes d'EPD peuvent être classées en modèles au niveau de l'unité et au niveau du domaine : les modèles au niveau de l'unité nécessitent un ensemble commun de variables auxiliaires entre l'enquête et le recensement ou les registres qui sont connues pour toutes les unités de population; les modèles au niveau du domaine sont fondés sur des estimations directes et des variables auxiliaires agrégées. Les politiques de protection de la vie privée et les coûts élevés du recensement rendent difficile d'utiliser des données au niveau de l'unité. Il est facile d'obtenir des variables auxiliaires agrégées de différentes sources et elles peuvent être utilisées dans des modèles au niveau du domaine. Les mégadonnées — qui donnent accès à des ensembles de données de plus en plus variées et abondantes, et ce, de plus en plus rapidement — peuvent être utilisées comme variables auxiliaires dans des modèles au niveau du domaine si elles sont traitées convenablement. Nous présentons deux applications de l'EPD : d'abord, nous utilisons les données sur la mobilité pour estimer l'incidence de la pauvreté au niveau local en Toscane, en Italie; ensuite, nous utilisons les données de Twitter pour estimer la part des dépenses de consommation alimentaire au niveau de la province en Italie. En outre, nous discutons de l'utilisation de sources de mégadonnées dans les statistiques officielles en mettant l'accent sur les défis à venir.

Séance 2B -- Intégration de données I

(A) Estimation de la répartition annuelle du patrimoine dans le Système de comptabilité nationale du Canada

Margaret Wu et Cilanne Boulet, Statistique Canada, Canada

Au cours des dernières années, il y a eu un intérêt grandissant envers les mesures de répartition du bien-être économique. Pour répondre à ce besoin, Statistique Canada est en train d'élaborer une série de tableaux annuels qui intègrent des données macroéconomiques sur les comptes nationaux avec des microdonnées d'enquête sur le patrimoine, le revenu et la consommation. Ce produit, Répartition des comptes économiques des ménages, ajoute une composante distributionnelle aux comptes macroéconomiques du Canada, donnant ainsi un portrait plus complet du bien-être économique des ménages canadiens dans le cadre des comptes nationaux.

Cette présentation décrira la méthodologie utilisée pour établir les tableaux de patrimoine. Par le passé, Statistique Canada ne menait des enquêtes sur le patrimoine qu'occasionnellement. L'un des principaux défis de ce projet est de trouver un moyen de combler l'absence relativement longue de données entre les années d'enquête. La modélisation, l'étalonnage, l'analyse comparative et le ratissage sont combinés afin de relever les défis pour combler ces absences de données et assurer la cohérence avec les totaux des comptes macroéconomiques et les données d'enquête. L'intégration de nouvelles sources de données administratives sur les passifs dans les tableaux de patrimoine est également à l'étude.

(A) Système de traitement des données transactionnelles

Agnes Waye, Serge Godbout et Nathalie Hamel, Statistique Canada, Canada

Les données transactionnelles sont de plus en plus utilisées comme source de données administratives ou dans les enquêtes. La richesse et le volume des données permettent à l'utilisateur d'obtenir des renseignements précieux et d'effectuer une analyse plus approfondie des tendances. Toutefois, de tels ensembles de données de grande taille et de structures complexes posent des défis uniques en matière de traitement et d'estimation de données, et les méthodes classiques de traitement de données nécessitent des solutions adaptées. À Statistique Canada, il y a une lacune dans l'infrastructure statistique pour le traitement des données transactionnelles. Comme un niveau élevé de flexibilité est nécessaire, nous avons identifié le besoin de mettre au point un système plus robuste pour traiter les données transactionnelles. Un système de traitement des données transactionnelles a été élaboré pour les enquêtes sur le transport, qui comprennent de nombreuses enquêtes comportant des données transactionnelles. Une enquête a été intégrée à ce système jusqu'à présent (Enquête de base tarifaire), et, graduellement, d'autres enquêtes des programmes de statistiques sur l'aviation, le transport ferroviaire et le camionnage seront également intégrées. Ce système met en œuvre les étapes de la phase du processus, telles que définies dans le Modèle statistique général du processus opérationnel (MSGPO), y compris des caractéristiques comme l'importation, la vérification et l'imputation, l'intégration, l'équilibrage et l'estimation de données. Ce document présentera la définition et les caractéristiques particulières des données transactionnelles, la façon dont elles sont traitées, les leçons apprises, les défis auxquels nous avons fait face ainsi que les problèmes futurs à résoudre dans le système de données transactionnelles.

(A) Mesurer la mobilité en combinant les mégadonnées et les sources de données administratives au Centre for Big data statistics de Statistique Pays-Bas

Marko Roos, Statistique Pays-Bas, Pays-Bas

Au Centre for Big data statistics de Statistique Pays-Bas, des travaux sont effectués, entre autres thèmes, sur la mesure des tendances de la mobilité. À cette fin, plusieurs sources de données administratives sont combinées à des sources de mégadonnées. Plus précisément, les mégadonnées issues des données sur les boucles de trafic et des données des

fournisseurs de téléphonie mobile sont comparées aux données de sources administratives sur les salaires. Les données administratives sur les salaires indiquent les habitudes de navettage, car elles contiennent les adresses des entreprises pour les employés. En combinant ceux qui ont un lieu de résidence et en projetant les itinéraires de navettage sur les matrices d'origine-destination qui en résultent, on obtient des tendances possibles de navettage. Les sources de mégadonnées provenant des boucles de trafic et des fournisseurs de téléphonie mobile sont utilisées pour corroborer et affiner ces tendances. Cette approche pourrait permettre d'avoir un aperçu rapide et détaillé de la mobilité au sein des Pays-Bas.

Dans le présent document, les premiers résultats de cette approche sont présentés dans le contexte d'un résumé d'autres thèmes du Centre for Big data statistics des Pays-Bas.

(A) Un coup d'œil à l'intérieur de la boîte : Combiner les distributions agrégées et marginales pour déterminer les distributions conjointes

Marie-Hélène Felt, Banque du Canada, Canada

Ce document propose une méthode pour estimer la distribution conjointe de deux variables ou plus lorsque seules leurs distributions marginales et la distribution de leur agrégat sont observées.

La détermination non paramétrique est obtenue en modélisant la dépendance à l'aide d'une structure latente de facteurs communs. De multiples exemples sont donnés de paramètres de données pour lesquels des échantillons multivariés de la distribution conjointe d'intérêt ne sont pas facilement disponibles, mais certaines mesures agrégées sont observées.

Dans l'application, les distributions intra-ménages sont récupérées en combinant les données d'enquête au niveau individuel et des ménages. La Banque du Canada surveille le comportement de paiement des Canadiens au moyen de deux enquêtes, soit la Methods-of-Payment (MOP) Survey et la Canadian Financial Monitor (CFM). Les deux enquêtes recueillent de l'information sur les choix de paiement au point de vente ainsi que sur les habitudes de gestion de la trésorerie, mais ils diffèrent par rapport à leur unité d'observation. Dans la MOP, l'unité d'observation est le répondant individuel et toutes les questions se rapportent aux caractéristiques et aux comportements individuels du répondant. Dans la CFM, l'unité d'observation principale est le ménage et les caractéristiques démographiques sont observées pour les chefs de famille de sexe féminin et masculin, mais l'argent comptant et les autres méthodes de paiement sont recueillis au niveau du ménage agrégé : on demande au répondant de déclarer le total familial mensuel.

En appliquant ma méthodologie, je peux étudier les influences intra-ménages, en ce qui concerne les pratiques de paiement et de gestion de trésorerie, en l'absence de données intra-ménages. Je trouve que, pour les personnes vivant en couple, les pratiques personnelles de gestion de trésorerie en ce qui concerne les retraits et les avoirs sont grandement influencées par l'utilisation de l'argent comptant et des cartes à valeur unique du partenaire.

(A) Ajustement des estimations de l'enquête sur les transports par des capteurs routiers installés en permanence à l'aide de techniques de capture-recapture

Jonas Klingwort, Bart Buelens et Rainer Schnell, Université de Duisburg-Essen, Allemagne et Statistique Pays-Bas, Pays-Bas

Récemment, l'intégration des données provenant des capteurs est devenue de plus en plus pertinente pour les statistiques officielles.

L'intégration des données des capteurs est particulièrement utile si elles peuvent être couplées aux données d'enquête et aux données administratives. L'application présentée ici combine de tels jeux de données pour quantifier et ajuster la sous-déclaration dans les estimations ponctuelles d'enquête.

Nous utilisons la Dutch Road Freight Transport Survey (RFTS) et les données des capteurs routiers produites par les postes de pesage automatisés (technologie « weigh-in-motion [WIM] ») installés sur les autoroutes néerlandaises. Le RFTS est un échantillon probabiliste de propriétaires de camions enregistrés qui déclarent les déplacements et le poids de la cargaison pour le camion échantillonné au cours d'une semaine donnée. Les neuf stations WIM mesurent en continu chaque camion qui passe et utilisent un système de caméra qui analyse les plaques d'immatriculation pour identifier les camions. Les jeux de données peuvent être couplés un à un en utilisant la plaque d'immatriculation comme identificateur unique. Comme les registres administratifs fournissent des renseignements sur le poids à vide de chaque camion et remorque, le poids de la charge peut être calculé. Ainsi, l'enquête et les capteurs mesurent indépendamment la même variable cible : le poids de la cargaison transportée.

Les méthodes de capture-recapture sont utilisées pour estimer la sous-déclaration dans le RFTS. L'hétérogénéité des véhicules en ce qui a trait aux probabilités de saisie et de recapture est modélisée au moyen d'une régression logistique et de modèles log-linéaires. Différents estimateurs seront comparés et discutés. Les résultats montrent que l'approche est prometteuse en ce qui concerne la validation et l'ajustement des données d'enquête à l'aide de données de capteurs externes. Cette application constitue un nouvel exemple de l'utilisation des statistiques multisources pour améliorer les avantages offerts par les données de capteurs dans le domaine des statistiques officielles.

Séance 3A -- Prioriser l'utilisation des données administratives pour la statistique officielle

(F) Élaboration d'une méthodologie de recensement combinant des données administratives et des données obtenues par collecte traditionnelle

Jean-François Simard, Roxanne Gagnon, Georgina House, Francis Demers, Christian Nadeau, Statistique Canada, Canada

Statistique Canada a lancé en juin 2016 le Projet de transformation du Programme du recensement qui vise à élaborer une méthodologie de recensement faisant un usage optimal des données administratives disponibles tout en maintenant la qualité et la pertinence des produits du recensement. La méthodologie privilégiée est celle de recensement combiné s'appuyant fortement sur les registres statistiques construits en majeure partie à partir de données administratives. Dans un recensement combiné, les statistiques sont obtenues à partir de registres ou d'autres sources de données administratives, auxquelles sont ajoutés des renseignements provenant d'un échantillon pour certaines variables ou d'un dénombrement partiel pour certaines autres variables.

Des simulations sont menées afin de mesurer la fiabilité des comptes de populations et de logements produits à partir de données administratives supplémentées de données obtenues par collecte traditionnelle. L'objectif des simulations est d'évaluer différentes options méthodologiques. Ces simulations sont les premières d'une série qui devrait culminer en la simulation complète d'un recensement combiné en 2021.

Les simulations combinent les données de la Base statistique de données démographiques canadiennes et les données du recensement de 2016. Les comptes de population obtenus sont comparés aux comptes équivalents du recensement traditionnel à différents niveaux de

géographie, et pour des caractéristiques démographiques telles que l'âge, le sexe et la composition des ménages. Diverses possibilités pour améliorer et mesurer la couverture d'un recensement combiné sont intégrées aux simulations.

Cet exposé traitera des méthodes proposées pour l'éventuelle mise en œuvre d'un recensement combiné au Canada de même que des résultats des premières simulations.

(A) Les données administratives d'abord comme paradigme statistique pour les statistiques officielles canadiennes : signification, défis et possibilités

Eric Rancourt, Statistique Canada, Canada

Depuis des décennies, les bureaux nationaux de la statistique affirment qu'ils ont l'intention d'utiliser de plus en plus de données administratives, ce qu'ils ont d'ailleurs fait à divers degrés d'un programme à l'autre. Cependant, avec l'avènement de la révolution des données, il ne s'agit plus d'un souhait, d'une question secondaire, d'une méthode marginale ou d'une tendance croissante : c'est devenu le centre d'attention pour l'avenir des programmes. Que l'objectif soit d'accroître la pertinence, de réduire le fardeau de réponse, d'accroître l'efficacité ou de produire plus rapidement avec plus de détails, l'utilisation des données administratives (au sens le plus large) prolifère à un rythme effréné dans les systèmes statistiques et sans systèmes statistiques. Statistique Canada fait face au nouveau monde des données en se modernisant et en adoptant un paradigme axé sur les données administratives d'abord. Le présent document tente d'expliquer ce que cela signifie, de mettre en évidence certains des défis sur les plans pratiques et théoriques et de signaler les éventuelles possibilités.

(A) L'utilisation de données administratives dans le recensement de 2018 de la Nouvelle-Zélande — l'avenir se joue maintenant

Nathaniel Matheson-Dunning, Anna Lin et Christine Bycroft, Statistics New Zealand, Nouvelle-Zélande

Le recensement de 2018 de la Nouvelle-Zélande est un recensement complet modernisé qui privilégie le mode de réponse numérique d'abord et une utilisation accrue de données administratives en complément aux réponses à l'enquête. Pour la première fois, des sources administratives (y compris le recensement précédent) feront partie intégrante de la collecte des données du recensement.

Le recensement de 2018 est une étape vers un avenir à long terme où le recensement serait fondé sur des sources administratives, appuyées par des enquêtes, conformément à l'objectif de notre organisme de recourir à des données administratives avant tout.

Nous avons la chance de disposer d'un riche éventail de sources de données administratives et d'enquête couplées dans l'infrastructure de données intégrées (IDI) de Stats NZ. Des évaluations de la qualité ont permis de déterminer quelles variables dérivées de sources administratives produiront des renseignements de grande qualité à l'appui du recensement. Pour le recensement de 2018, des sources administratives, ainsi que les renseignements fournis dans le recensement précédent de 2013, servent à imputer les réponses manquantes pour une personne donnée.

En plus de la non-réponse partielle, il arrive dans certains cas que des ménages entiers ne répondent pas au recensement. Dans nos évaluations de la qualité, nous avons constaté qu'il est plus difficile de placer les bonnes personnes dans les ménages construits à partir des adresses administratives. En nous appuyant sur les travaux du U.S. Census Bureau, nous avons élaboré un modèle permettant d'évaluer pour quels ménages dérivés de l'IDI la

composition du ménage est susceptible d'être la plus fiable. Les données administratives peuvent ensuite servir à dériver les réponses pour chaque membre du ménage.

Nous présenterons les méthodes utilisées pour les deux formes d'imputation. Nous résumons les constatations établies jusqu'à maintenant et nous discutons des incidences probables pour les produits finaux du recensement de 2018.

Séance 3B -- Méthodes novatrices

(A) Le jeu d'imitation : Aperçu d'une approche d'apprentissage automatique pour coder la classification industrielle

Javier Oyarzun, Statistique Canada, Canada

Le Registre des entreprises (RE) de Statistique Canada joue un rôle fondamental dans le mandat de Statistique Canada. Le RE est un répertoire qui indique toutes les entreprises opérant au Canada. Près d'une centaine d'enquêtes-entreprises utilisent le RE de différentes façons. Toutefois, il sert principalement pour établir une base de sondage, tirer des échantillons, recueillir et traiter des données ainsi que pour produire des estimations.

Le RE a une incidence directe sur l'efficacité du processus d'enquêtes-entreprises, sur la fiabilité des données produites par les programmes de statistiques des entreprises et sur la cohérence du Système de comptabilité nationale. Au début de 2018, Statistique Canada a commencé à établir une nouvelle méthodologie pour coder de façon probabiliste la classification industrielle des entreprises. Cette méthodologie, qui utilise l'exploration de données, l'exploration de textes et l'apprentissage automatique, fournira à Statistique Canada un outil pour coder les classifications industrielles manquantes et améliorer la qualité globale des classifications industrielles dans le RE.

Ce document traitera du Système de classification des industries de l'Amérique du Nord (SCIAN) et de son utilisation dans les programmes statistiques à Statistique Canada. Il présentera également les approches actuelles et nouvelles de codage du SCIAN. Enfin, le document traitera des défis liés à la partie du RE qui n'est pas codée et présentera des cas complexes de codage du SCIAN.

(A) Modélisation des erreurs de mesure afin d'assurer la cohérence entre les taux de croissance du chiffre d'affaires mensuels et trimestriels

Arnout van Delden, Sander Scholtus, et Nicole Ostlund, Statistique Pays-Bas, Pays-Bas

Pour un certain nombre de secteurs économiques, Statistique Pays-Bas (SPB) produit des taux de croissance du chiffre d'affaires des entreprises, c'est-à-dire des chiffres mensuels fondés sur une enquête-échantillon et des chiffres trimestriels fondés sur des données fiscales administratives. SPB vise à comparer les taux de croissance mensuels aux taux trimestriels afin de produire des résultats uniformes, en particulier pour les comptes nationaux qui utilisent les deux séries comme intrants.

Les résultats préliminaires de l'analyse comparative ont montré que les taux de croissance mensuels étaient ajustés différemment d'un trimestre à l'autre. En fait, le chiffre d'affaires trimestriel des données administratives s'est révélé relativement importante au cours du quatrième trimestre comparativement aux données de l'enquête, alors que l'inverse était vrai au premier trimestre. Cet effet est probablement causé par les tendances trimestrielles des erreurs de mesure, par exemple en raison des processus administratifs au sein des entreprises. Ces tendances peuvent également se produire dans d'autres pays qui utilisent

ou visent à utiliser des données administratives à court terme et à comparer les résultats avec les données d'enquête.

Nous présentons une méthodologie qui vise à détecter et à corriger automatiquement de telles erreurs de mesure. En commençant par un vaste ensemble de variables de base, que nous avons sélectionnées lors de discussions avec les bureaux administratifs, nous vérifions lesquelles de ces variables sont associées aux tendances trimestrielles des erreurs de mesure en utilisant des arbres de décision. Par la suite, nous construisons des variables composites qui sont les mieux associées aux erreurs de mesure saisonnières. Ces variables composites sont utilisées dans un modèle de régression mixte qui décrit la relation entre le chiffre d'affaires trimestriel des données administratives et des données d'enquête. Le modèle est composé de plusieurs groupes d'unités, chacun saisissant une structure d'erreur de mesure différente, dont un qui utilise la variable composite pour expliquer les effets de mesure trimestriels. Les erreurs de mesure, qui seraient autrement un obstacle à une analyse comparative raisonnable, pourraient être corrigées à l'aide des estimations des paramètres de ce modèle de mélange.

(A) Intégrer les techniques d'échantillonnage dans les mégadonnées

Antoine Rebecq, Ubisoft Montréal, Canada

Les outils de mégadonnées sont avancés et complexes et, par conséquent, la culture informatique est plus répandue que la culture statistique parmi les équipes de science des données. Néanmoins, l'analyse de grands jeux de données présente des défis statistiques dont les professionnels de la science des données et de l'apprentissage automatique doivent être conscients. Nous présentons quelques exemples de la façon dont l'ajout d'outils d'échantillonnage dans les mégadonnées a aidé les analystes de données d'Ubisoft à mieux comprendre ce que les joueurs aiment dans les jeux.

À Ubisoft, la principale source de données sur les joueurs provenant des jeux est les événements qui sont envoyés par le client du jeu aux serveurs, puis passent par un réseau de mégadonnées pour les mettre à la disposition des analystes. Le volume des données est si important qu'il faut parfois encore échantillonner les événements. Par conséquent, certains événements sont observés avec des probabilités d'inclusion complexes. Les scientifiques des données ne sont souvent pas conscients des biais, généralement lorsqu'ils entrainent des modèles d'apprentissage automatique. Nous démontrerons que nous pouvons améliorer le rendement des algorithmes en utilisant l'étalonnage en marge.

D'autres données sont également recueillies à l'aide d'échantillons conventionnels de possibilités de commercialisation. Nous discuterons de la façon dont nous pouvons créer des plans d'échantillonnage et de repondération efficaces pour de telles données en tirant parti des mégadonnées recueillies lors d'événements dans le jeu. Les techniques d'explorationde textes sont également devenues de plus en plus populaires ces dernières années. Les entreprises regardent ce que les utilisateurs disent sur les médias sociaux pour cibler les secteurs de leurs produits qui doivent être améliorés. Nous montrons que les techniques d'échantillonnage peuvent également aider à traiter les biais (p. ex., biais de sélection automatique) inhérents à ces jeux de données.

Enfin, les outils d'échantillonnage sont utilisés dans le cadre de projets plus exploratoires. Les données du réseau sont incroyablement précieuses, mais aussi très difficiles à gérer. Nous montrons comment l'échantillonnage peut être utilisé pour rendre les analyses des données du réseau moins chères et plus fiables.

(A) Developper une base de données d'immeubles ouverte et exploratoire

Alessandro Alasia, Jean Le Moullec et Marina Smailes, Statistique Canada, Canada

À venir

(A) Les défis de la production d'estimations nationales de la maltraitance des enfants à l'aide de données administratives provenant de différents secteurs de compétence
David Laferrière et Catherine Deshaies-Moreault, Statistique Canada, Canada

Ila été demandé à STC de réaliser une étude de faisabilité sur la façon d'élaborer un système de surveillance de la maltraitance infantile. Le Système Canadien de Surveillance des Signalements d'Enfants Victimes de Maltraitance (SCSSEVM) intégrerait les données des organismes de protection de l'enfance de chaque province et territoire pour calculer les estimations annuelles de la maltraitance infantile dans cinq catégories : la violence physique, la violence psychologique, la violence sexuelle, la négligence et l'exposition à la violence entre partenaires. Afin de réduire le fardeau des travailleurs du bien-être de l'enfance, la principale source de données serait un recensement des données administratives des provinces et des territoires (PT).

Il y a plusieurs défis à relever pour mettre en œuvre le SCSSEVM. Chaque province et territoire a sa propre loi sur ce qui constitue la maltraitance infantile et sur la façon de la catégoriser. Les PT ont également différents systèmes pour repérer et suivre les cas de maltraitance. Le contenu et l'exhaustivité des données administratives, pouvant comprendre des microdonnées et des textes narratifs, varient considérablement.

Traditionnellement, il faudrait des codeurs pour repérer les cas de maltraitance à partir des récits. Toutefois, pour le SCSSEVM, les techniques de traitement du langage naturel au moyen de l'apprentissage automatique seront explorées afin de déterminer si les cas de maltraitance pourraient être automatiquement repérés et classés à partir des rapports narratifs.

Un autre défi est que les données administratives pourraient ne pas être disponibles de tous les PT. Lorsque les données administratives ne sont pas disponibles, une solution de rechange consiste à demander aux travailleurs des services de protection de l'enfance de remplir des enquêtes.

Nous discutons de l'étude de faisabilité du SCSSEVM, un projet qui explore les défis pratiques et techniques de l'utilisation des approches traditionnelles ainsi que des techniques plus modernes pour créer des estimations nationales cohérentes à partir des données administratives et des données d'enquête de 13 PT.

Séance 4 -- Discours du gagnant du Prix Waksberg

(A) Le statisticien sage et le calage conditionnel
Donald Rubin, Université Harvard, USA

À venir

Séance 5A -- Données géospatiales

(A) Surveillance du développement spatial durable : Analyse (semie-) automatisée des images satellites et aériennes pour la transition énergétique et les indicateurs de durabilité

R.L. Curier, T.J.A. de Jong, D. Iren et S. Bromuri, Statistics Netherlands and Business Intelligence and Smart Services Institute, Pays-Bas

L'Europe vise à remplacer d'ici 2050 au moins 30 % de la demande de combustibles fossiles par des ressources renouvelables, nécessitant ainsi des systèmes énergétiques urbains qui émettent moins de carbone et utilisent moins d'énergie. De nos jours, l'énergie solaire joue un rôle important dans la transition énergétique, et les décideurs politiques et les exploitants de réseaux sont très intéressés par la cartographie des panneaux solaires. Les statistiques actuelles sur l'énergie solaire sont basées sur des enquêtes sur l'importation de panneaux solaires et ne fournissent que des chiffres nationaux sur une base annuelle, tandis que la transition énergétique crée une demande d'information à l'échelle régionale et locale avec des échelles de temps plus courtes.

L'étude actuelle vise à produire une carte des panneaux solaires ainsi que des statistiques sur le nombre de panneaux solaires en automatisant la détection de ceux-ci. À cette fin, le contenu de l'information provenant des images aériennes et satellitaires à haute résolution est analysé au moyen de l'intelligence artificielle afin de permettre la détection et la classification automatiques des panneaux solaires. Deux approches d'apprentissage machine, soit les machines à vecteurs de support et les réseaux neuronaux à convolution profonds, seront utilisées pour identifier les panneaux solaires dans des images de diverses résolutions. En outre, le projet utilisera également des registres existants tels que des informations sur les retours de TVA des propriétaires de panneaux solaires et des informations acquises auprès fournisseurs d'énergie pour l'algorithme d'apprentissage automatique.

Dans cette présentation, les résultats préliminaires pour la province de Limburg (Pays-Bas), la Flandre (Belgique) et la Rhénanie-du-Nord-Westphalie (Allemagne) feront l'objet d'une discussion et la valeur ajoutée de l'utilisation des données de télédétection pour inférer l'information sera abordée.

(A) De bonnes clôtures font-elles vraiment de bons voisins? Une comparaison par juxtaposition de la composition téléphonique aléatoire et des méthodes de confinement géographique en utilisant des enquêtes de facteur de risque

James Dayton et Matt Jans, ICF, États-Unis

À venir

À venir

Séance 5B -- Enquêtes non probabilistes

(A) Étalonner les échantillons non probabilistes à l'aide d'échantillons probabiliste utilisant la méthode LASSO

Jack Chen, Michael R. Elliott et Rick Valliant, Université du Michigan, États-Unis

L'une des applications les plus importantes de la recherche d'enquêtes est le sondage électoral. En raison de la diminution de la couverture téléphonique terrestre et de l'amélioration de la technologie de filtrage téléphonique, il est devenu très difficile pour les sondeurs électoraux de saisir les intentions de vote en temps opportun. Cela a donné lieu à

une utilisation accrue et à un accès facile à des échantillons de personnes provenant d'enquêtes en ligne non probabilistes. Cependant, les échantillons non probabilistes sont à risque de biais de sélection en raison des différences d'accès, des degrés d'intérêt et d'autres facteurs. L'étalonnage est une méthode normalisée utilisée dans les statistiques officielles et d'autres contextes qui utilise des poids pour ajuster les estimations de totaux d'un échantillon aux totaux connus dans une population. Comme les échantillons non probabilistes n'ont pas de propriétés inférentielles solides, nous envisageons d'utiliser des méthodes d'étalonnage assistées par modèle qui permettent une estimation robuste des totaux de population. En particulier, nous considérons l'étalonnage à des totaux estimés de la population à l'aide d'une régression adaptative LASSO – estimation-contrôle LASSO (ECLASSO). La méthode LASSO adaptative peut produire un estimateur cohérent d'un total de population dans la mesure où un sous-ensemble des vrais prédicteurs est inclus dans le modèle de prévision, ce qui permet d'inclure un grand nombre de covariables possibles sans risque de surcharge. Cela offre la possibilité d'étalonnage à des estimations provenant d'échantillons probabilistes de meilleure qualité avec des tailles d'échantillon modestes. Nous appliquons le modèle ECLASSO pour prédire le résultat du vote de onze élections de gouverneurs et de huit élections sénatoriales à l'élection de mi-mandat aux États-Unis en 2014. Puisque les résultats réels de l'élection sont publiés, nous pouvons comparer le biais et l'erreur quadratique moyenne de la racine du modèle ECLASSO avec les méthodes traditionnelles de correction de la pondération.

(A) Inférence statistique avec des échantillons d'enquête non probabilistes

Changbao Wu, Yilin Chen et Pengfei Li, Université de Waterloo, Canada

Nous établissons un cadre général pour les inférences statistiques avec des échantillons d'enquête non probabilistes lorsque des renseignements auxiliaires pertinents sont disponibles à partir d'un échantillon d'enquête probabiliste. Nous élaborons une procédure rigoureuse pour estimer les scores de propension pour les unités de l'échantillon non probabiliste et construisons un estimateur doublement robuste pour la moyenne de population finie. L'estimation de la variance est abordée dans le cadre proposé. Les résultats des études de simulation montrent la robustesse et l'efficacité de nos estimateurs proposés par rapport aux méthodes existantes. La méthode proposée est utilisée pour analyser un échantillon d'enquête non probabiliste recueilli par le Pew Research Center avec des renseignements auxiliaires provenant du Système de surveillance des facteurs de risque comportementaux et de l'Enquête sur la population courante. Nos résultats illustrent une approche générale de l'inférence avec des échantillons non probabilistes et soulignent l'importance et l'utilité des renseignements auxiliaires provenant des échantillons d'enquête probabiliste.

(A) Comprendre les effets du couplage d'enregistrements sur l'estimation du total lors de la combinaison d'une source de mégadonnées avec un échantillon probabiliste

Benjamin Williams et Lynne Stokes, Université Southern Methodist, États- Unis

Le couplage d'enregistrements est un outil utile qui relie les enregistrements de deux listes qui renvoient à la même unité, mais qui n'ont pas d'identificateur unique. L'effet de l'erreur d'appariement du couplage d'enregistrements n'a pas été pris en considération pour le cas de l'estimation du total d'une population à partir d'un modèle de capture-recapture.

Le National Marine Fisheries Service (NMFS) estime le nombre total de poissons pêchés par les pêcheurs sportifs. Il y arrive en estimant l'effort total (le nombre de voyages de pêche) et les captures par unité d'effort (CPUE) (le nombre de poissons capturés par espèce par voyage), puis en les multipliant ensemble. Les données sur l'effort sont recueillies au moyen d'une enquête postale auprès des pêcheurs potentiels. Les données du CPUE sont recueillies au moyen d'interceptions en personne des voyages de pêche. L'enquête sur l'effort a un taux

élevé de non-réponse et est rétrospective, ce qui cause un long processus d'estimation et empêche la gestion en saison.

En raison de ces limites, le NMFS tente de remplacer l'enquête sur l'effort par l'autodéclaration électronique. Par ce système, les pêcheurs signalent les détails de leur voyage au moyen d'un appareil électronique et demeurent éligibles à être échantillonnés lors de l'interception à quai.

Dans ce scénario, les estimateurs proposés du total utilisent les autodéclarations (un important échantillon non probabiliste) parallèlement à l'interception à quai (un échantillon probabiliste), au moyen de la méthode de capture-recapture (Liu et coll., 2017). Pour que les estimateurs soient valides, les données des voyages autodéclarés et échantillonnés dans l'interception doivent être couplées. Les estimateurs actuels supposent un appariement parfait. En pratique, cela est toutefois difficile en raison des erreurs d'instrument et de mesure.

Dans ce document, nous proposons plusieurs estimateurs supplémentaires et élaborons un algorithme de couplage d'enregistrements pour apparier les voyages. Nous examinons l'effet des erreurs d'appariement sur les estimateurs et illustrons ces effets à l'aide des données de l'une des expériences de déclaration électronique.

(A) Le moissonnage Web comme source de données de rechange pour prédire les indicateurs du commerce électronique

José Márcio Martins Júnior, Marcelo Trindade Pitta, João Victor Phacheco dias, Predro Luis do Nascimento Silva, Brazilian Network Information Center and National School of Statistical Science from IBGE, Brésil

Estimer les indicateurs de commerce électronique, comme la proportion de pages Web qui vendent des produits ou des services, au moyen d'enquêtes traditionnelles et en tenant compte de la nécessité de fournir des données désagrégées et actuelles, est coûteux et demande beaucoup de temps. Pour estimer de tels indicateurs, le présent document propose une approche qui combine les données d'enquête avec l'information extraite du code source (HTML) des pages Web des entreprises comme source de données supplémentaire. L'échantillon des entreprises sondées comprend toutes les entreprises sélectionnées et ayant répondu à l'enquête brésilienne de 2017 sur les technologies de l'information et des communications dans les entreprises (TIC-Empresas). Étant donné que les renseignements de l'enquête comprennent les adresses des pages Web, ainsi que les réponses aux questions concernant les pratiques de commerce électronique des entreprises, l'utilisation de la base de données de l'enquête et l'accès aux pages Web des entreprises correspondantes ont permis l'élaboration d'un modèle d'évaluation HTML afin d'établir des réponses automatisées à certaines des variables requises pour estimer les indicateurs de commerce électronique requis. Pour obtenir les codes sources des pages Web, on a utilisé un robot Java personnalisé. Le modèle a tenté d'utiliser l'information de la première page (page d'accueil) comme intrant pour un modèle de régression logistique concernant les réponses aux indicateurs sélectionnés. Les variables explicatives du modèle correspondent aux mots dans les pages Web, soit des variables dérivées d'un dictionnaire des mots les plus pertinents créés pour prédire le résultat de ces indicateurs. À l'aide de cette méthodologie, l'échantillon peut être élargi en passant sur de plus grands échantillons de pages Web d'entreprises, et les données résultantes utilisées pour estimer les indicateurs requis de façon plus précise, plus rapide et plus désagrégée, ce qui exige moins de ressources que la méthode habituelle. Ensuite, il est possible de calculer des estimations plus à jour des indicateurs requis, au besoin.

(F) Le sondage indirect appliqué aux modèles de capture-recapture avec dépendance entre les sources.

Herménégilde Nkurunziza et Ayi Ajavon, Statistique Canada, Canada

Capture-Recapture est une méthode largement utilisée pour estimer la taille inconnue d'une population. La méthode consiste à tirer, de la population d'intérêt, deux échantillons indépendants. L'estimateur de Petersen de la taille de la population, souvent utilisé, est fonction de la taille et du chevauchement entre les échantillons. Lavallée et Rivest (2012) se sont intéressés au cas où les échantillons sont issus d'un sondage indirect et ont introduit une généralisation de l'estimateur de Petersen basée sur la méthode généralisée de partage des poids. En pratique, l'hypothèse d'indépendance sur laquelle repose l'estimateur n'est pas souvent vérifiée (Brenner(1995)). Dans cet article, nous nous intéressons aux modèles de Capture-Recapture avec dépendance entre les échantillons et proposons une extension de l'estimateur de Lavallée et Rivest (2012). Nous analysons les propriétés de l'estimateur obtenu et présentons une illustration de la méthode à l'aide de données simulées.

Séance 6A -- Utilisation de sources de données alternatives pour les programmes de statistiques sociales

(A) Rassembler les données sur les transactions électroniques et en examiner des combinaisons en remplacement de l'enquête sur le budget des ménages

Anders Holmberg, Statistics Norway, Norvège

Lorsque Statistics Norway a annulé son enquête sur le budget des ménages de 2018, c'était en raison de préoccupations liées au compromis entre les coûts et la qualité et parce que l'organisme doutait si une approche essentiellement traditionnelle fondée sur une enquête et l'utilisation d'un journal serait satisfaisante. La décision d'annuler l'enquête a déclenché une intensification des recherches en vue d'acquérir d'autres sources de données sur la consommation des ménages. Par conséquent, le paysage national des données sur les transactions peut maintenant être bien décrit, et trois sources différentes de données sur les transactions électroniques et la façon de les combiner sont examinées à titre expérimental. L'une des sources comprend des données sur les transactions provenant d'un important fournisseur de services de paiement de la Norvège (les transactions par carte et autres transactions électroniques, ainsi que les transactions interentreprises). Le taux de couverture est assez élevé puisque les données englobent la plupart des transactions par carte effectuées en Norvège. Nous examinons également les données des caisses enregistreuses des chaînes de magasins de détail et des échantillons de données sur les membres des chaînes de magasins de détail. Toutes ces sources de données sont intéressantes en soi, mais ce qui ajoute vraiment de la valeur, du moins du point de vue de l'enquête sur le budget des ménages, c'est de trouver des façons de combiner les différentes sources. Il peut s'agir par exemple de méthodes de couplage des enregistrements de transaction par carte aux données des caisses enregistreuses pour combiner la dimension démographique des dépenses et la dimension de la consommation selon le niveau détaillé de classification de la consommation individuelle par objet. Le document traite des possibilités et des expériences méthodologiques et techniques découlant de ces travaux.

(F) Modernisation du Programme des dépenses des ménages

Christiane Laperrière, Denis Malo et Johanne Tremblay, Statistique Canada, Canada

L'Enquête sur les dépenses des ménages recueille des informations qui sont des rouages importants de l'Indice des prix à la consommation et du Système de comptabilité nationale. Ces données servent également à une large communauté d'utilisateurs généralement intéressés à l'analyse des dépenses en fonction de caractéristiques socio-économiques des ménages. Le contenu détaillé et les pratiques traditionnelles de collecte de données axées sur une entrevue personnelle et un journal de dépenses imposent un lourd fardeau aux répondants et engendrent des coûts de collecte élevés. Dans le cadre de l'initiative de

modernisation de Statistique Canada, le programme des dépenses des ménages explore le potentiel de nouvelles sources de données. L'exploitation de données aussi volumineuses engendre de nouveaux défis notamment le recours à divers algorithmes d'apprentissage automatisé pour classifier les données en catégories de dépenses pertinentes pour le programme. Le cadre d'intégration de ces multiples sources de données doit également être établi et nécessitera le développement de nouvelles méthodes. Des études ont été amorcées pour évaluer comment adapter les méthodes d'intégration aux besoins des différents types d'utilisateurs. Dans cette présentation, nous discuterons des idées novatrices envisagées pour la classification et l'intégration, des défis rencontrés dans l'exploration de ces nouvelles sources de données et des résultats des évaluations en cours.

(A) Combinaison de sources de données multiples pour améliorer les statistiques fédérales américaines

Brian Harris-Kojetin et Robert M. Groves, National Academies of Sciences, Engineering, and Medicine et Université de Georgetown, États-Unis

Les enquêtes par échantillonnage probabiliste à grande échelle sont depuis longtemps à la base de la production de nombreuses statistiques nationales aux États-Unis, mais les coûts de la réalisation de telles enquêtes ont augmenté tandis que les taux de réponse ont diminué, et de nombreuses enquêtes ne répondent pas à la demande croissante d'information locale plus opportune et détaillée. Le Committee on National Statistics aux National Academies of Sciences, Engineering, and Medicine a réuni un comité d'experts en recherche en sciences sociales, en sociologie, en méthodologie d'enquête, en économie, en statistique, en protection des renseignements personnels, en politique publique et en informatique afin d'examiner la possibilité d'un virage vers une approche combinant des sources de données pour fournir aux utilisateurs des jeux de données plus riches et plus fiables. Le groupe a mené une étude de deux ans et produit deux rapports contenant des conclusions et des recommandations à l'intention des organismes statistiques fédéraux. Le premier rapport, Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, étudie l'approche actuelle de production de statistiques fédérales et examine d'autres sources de données, comme les données administratives gouvernementales et les sources de données du secteur privé, y compris Internet et d'autres sources de mégadonnées, qui pourraient également être utilisées pour les statistiques fédérales. Le deuxième rapport, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps, examine plus en profondeur l'infrastructure, les méthodes et les compétences nécessaires pour mettre en œuvre un paradigme de sources de données multiples pour les statistiques officielles. Ensemble, les rapports donnent un aperçu des méthodes statistiques et des cadres de qualité qui ont été utilisés pour combiner l'information et décrivent la recherche nécessaire pour élaborer des méthodes permettant de combiner les sources de données et de protéger la vie privée. La présentation donnera un aperçu des principales conclusions et recommandations de ces rapports.

Séance 6B -- Couplage d'enregistrements

(A) Évaluation de l'exactitude des couplages ou des paires de candidats dans les analyses de couplage d'enregistrements

Dean Resnick et Lisa Mirel, Université de Chicago et National Center for Health Statistics, États-Unis

Toute stratégie complète de couplage d'enregistrements exige une approche permettant d'évaluer l'exactitude des paires de candidats ou des couplages au cas par cas ou sous forme agrégée. Le document de référence sur le couplage d'enregistrements de Fellegi et Sunter suggère d'examiner manuellement les paires entre le seuil de rejet et le seuil d'acceptation

pour établir le statut de couplage. Quoi qu'il en soit, la question de savoir si les examinateurs administratifs peuvent déduire avec exactitude le statut de concordance, sans parler des dépenses engagées pour effectuer cet examen pour un grand nombre de paires, reste entière. Nous présumons que lorsque l'examinateur humain peut ajouter de la valeur en voyant une transposition ou une substitution de nom ou en étant capable de tenir compte de la rareté des noms dans l'analyse, il est souvent (mais pas toujours) vrai que ces types de techniques auraient déjà dû être codés dans la routine d'évaluation des couplages programmés.

Par ailleurs, si un jeu d'essai de grande qualité est disponible, il peut être utilisé pour mesurer la qualité des liens, mais cela n'existera généralement pas. En son absence, un champ d'identification très précis (comme le numéro de sécurité sociale) peut être utilisé à la place. Le taux de couplage valide correspondrait approximativement au niveau d'accord pour ce champ dans les paires couplées. Toutefois, même si ceci n'était pas disponible, nous pouvons utiliser une approche qui applique la théorie du couplage d'enregistrements pour estimer la validité de l'appariement de la paire et ensuite utiliser cela pour estimer l'exactitude du couplage. Cette présentation abordera différentes méthodes d'évaluation de la qualité du couplage d'enregistrements qui vont au-delà de l'examen administratif.

(A) Règles de décision et estimation du taux d'erreur pour le couplage d'enregistrements au moyen d'un modèle de probabilité

Clayton Block, Élections Canada, Canada

Depuis 1997, Élections Canada tient à jour le Registre national des électeurs, une base de données sur les Canadiens de 18 ans et plus utilisée pour administrer les élections fédérales. Cette base de données est mise à jour à partir de plusieurs sources administratives fédérales et provinciales, liées aux électeurs dans la base de données à l'aide de renseignements personnels comme le nom, la date de naissance, le sexe et l'adresse. Au départ, un logiciel de couplage commercial fondé sur la théorie de Fellegi-Sunter a été utilisé pour ces activités de couplage. Graduellement, la méthodologie et le logiciel utilisés se sont orientés vers des solutions personnalisées, offrant plus de souplesse sur la façon dont les paires potentielles sont traitées, et réduisant les taux d'erreur de classification associés au processus de couplage. L'une des améliorations clés à la méthodologie est la reformulation de la règle de décision bien connue de Fellegi-Sunter, maintenant exprimée en termes de probabilité d'intérêt et comparée aux tolérances d'erreur réelles. Pour l'appariement des renseignements personnels, les probabilités requises sont calculées à partir des paires observées à l'aide d'un simple modèle de probabilité de correspondance fortuite pour la date de naissance. Les hypothèses du modèle devraient être très réalistes. Les probabilités calculées pour chaque paire peuvent aussi être simplement additionnées pour produire des estimations des deux types d'erreurs d'appariement, ce qui n'exige aucun logiciel spécialisé et aucune procédure mathématique complexe. Les méthodes décrites ont été utilisées pour divers processus de couplage à Élections Canada, chacun ayant des taux d'appariement prévus différents. Dans tous les cas, les taux d'erreur produits semblent crédibles. À l'avenir, ces résultats pourraient être comparés à ceux obtenus à partir de méthodes d'estimation du taux d'erreur concurrentes et plus compliquées.

(A) Combiner les données des entreprises commerciales avec les données administratives des employeurs et des employés – défis méthodologiques du couplage, de la préparation et de la représentativité

Manfred Antoni et Marie-Christine Laible, Institute for Employment Research (IAB) Research Data Center, Allemagne

Nous décrivons le couplage des données sur les entreprises commerciales du Bureau van Dijk (BvD) avec les données administratives sur l'emploi du Centre de données de recherche (FDZ) de l'Agence d'emploi fédérale de l'Allemagne à l'Institute for

Employment Research(IAB). BvD est un fournisseur commercial de données d'entreprise et ses bases de données ont été utilisées principalement pour analyser les renseignements d'affaires. Parallèlement, le FDZ fournit aux chercheurs un accès gratuit aux données administratives et d'enquête depuis plus de 10 ans.

Pour combiner le potentiel de recherche des deux sources de données, le FDZ a effectué un couplage d'enregistrements des entreprises (unité indépendante) fournies dans la base de données de BvD, Orbis, avec les établissements (sous-unités dépendantes) dans la base de données Establishment History Panel du FDZ. Premièrement, nous décrivons le processus de couplage et les méthodes appliquées par le FDZ. Jusqu'à présent, aucun couplage à grande échelle entre les données BvD et les données administratives n'a été réussi. Le principal obstacle est que les deux sources de données ne contiennent pas d'identificateur commun qui permettrait un couplage direct. Le FDZ a donc effectué le couplage en comparant, entre autres, les noms des entreprises et des établissements donnés dans les bases de données originales. Deuxièmement, nous présentons les étapes de la création d'un jeu de données de recherche à partir du tableau de correspondance entre les entreprises et les établissements généré par le couplage d'enregistrements. Nous décrivons les défis rencontrés, comme les affectations multiples, et les méthodes appliquées pour les surmonter. Le jeu de données qui en résulte contient des renseignements longitudinaux sur les entreprises, leurs établissements dépendants et tous leurs employés. Troisièmement, nous présentons des analyses de représentativité qui examinent la sélectivité du jeu de données de recherche.

(A) Évaluation de la qualité des couplages entre les données administratives cliniques des hôpitaux et des données sur les décès de la Statistique de l'état civil du Canada.

Nancy Rodrigues, Institut canadien d'information sur la santé, Canada

Trois bases de données cliniques-administratives des hôpitaux du Canada ont été couplées à la base de données sur les décès de la Statistique de l'état civil (BCDECD) du Canada afin de fournir des renseignements sur les patients qui sont décédés après leur congé de l'hôpital, ainsi que des renseignements supplémentaires sur les patients qui sont décédés à l'hôpital.

Les jeux de données couplés ont été créés pour élaborer et valider des indicateurs de soins de santé et des mesures de rendement et effectuer des analyses des résultats. Il était donc impératif d'évaluer l'adéquation des données à leur utilisation. La qualité a été évaluée en calculant la couverture des décès pour tous les contributeurs couplés, en créant un profil du jeu de données couplées et en analysant les problèmes soulevés par les utilisateurs. Ces analyses ont été guidées par un outil d'évaluation des sources de données existant, qui fournit un ensemble de critères permettant l'évaluation selon cinq dimensions de la qualité.

Les variables présentaient une bonne disponibilité des données, avec des taux de 95 % ou plus. Il y avait des écarts dans la date de décès saisie dans les deux sources liées pour 1,4 % des décès aigus à l'hôpital; la grande majorité d'entre eux ne présentaient qu'une seule journée de différence. Un groupe d'utilisateurs et un processus de suivi des problèmes ont été créés pour partager l'information sur les données couplées et veiller à ce que les problèmes soient triés par la partie appropriée et permettre un suivi en temps opportun avec le fournisseur de données.

Une évaluation guidée des données a permis de déterminer les forces et les limites et de les partager pour appuyer une utilisation appropriée. La rétroaction au fournisseur de données appuie l'amélioration continue de la méthodologie de couplage.

(A) Pondération inverse de la probabilité pour corriger la classification erronée des résultats à l'aide des jeux de données administratives couplées

Christopher A. Gravel, Kristian B. Filion, Pauline M. Reynier et Robert W. Platt, McGill University et Lady Davis Institute, Canada

De grandes données observationnelles réelles sur les soins de santé, comme les données sur les réclamations, peuvent être utilisées pour des études sur l'innocuité et l'efficacité des médicaments après leur mise en marché. Il est possible d'utiliser des poids de probabilité inverse de traitement (score de propension) pour aborder la confusion mesurée dans les études de cette nature, sous l'hypothèse d'une mesure exacte de la variable de résultat.

Bon nombre de ces jeux de données souffrent d'une classification erronée systématique des résultats en raison d'erreurs de codage et/ou d'enregistrement. À titre d'exemple, des études ont démontré la possibilité d'une évaluation incomplète des résultats cardiovasculaires dans Clinical Practice Research Datalink (CPRD) - un dépôt d'information sur la médecine générale du Royaume-Uni. Nous introduisons un nouvel ensemble de poids qui peut être utilisé conjointement avec les poids des scores de propension pour produire une estimation uniforme du rapport de probabilité causale marginal en présence d'une classification erronée des résultats binaires (erreur de diagnostic) à l'aide de l'information de validation interne.

Pour obtenir une source de validation interne pour les résultats hébergés dans le CPRD, nous utilisons des dossiers hospitaliers couplés tirés de l'Hospital Episode Statistics (HES), une base de données contenant les dossiers d'hospitalisation des patients au Royaume-Uni. Nous illustrons ensuite l'approche pondérée proposée à l'aide d'un exemple d'étude de l'utilisation de statines après un infarctus du myocarde et du risque d'accident vasculaire cérébral pendant un an. Nous comparons nos résultats à ceux d'une méta-analyse d'essais cliniques randomisés. Enfin, nous présentons des études de simulation portant sur les propriétés de l'estimateur pondéré proposé, y compris l'incidence de la sélection du modèle sur la réduction du biais et de la variabilité.

Séance 7A -- Fusion de données et couplage d'enregistrements

(A) Méthodes de régression linéaire simple des données de couplage à des fins d'analyse secondaire

Li-Chun Zhang et Tiziana Tuoto, Statistics Norway et Istat, Norvège et Italie

À moins qu'un identificateur unique ne soit disponible, le couplage de deux ensembles de données distincts produira des erreurs qui peuvent introduire un biais dans l'analyse subséquente, si les données couplées sont traitées comme si elles avaient été réellement observées. Nous examinons la régression linéaire du point de vue des analystes secondaires, qui reçoivent seulement l'ensemble de données couplées, mais pas les enregistrements non couplés des deux ensembles de données. De plus, nous supposons que l'analyste n'a pas accès à toutes les variables clés de couplage, ni aux détails ou aux outils de la procédure de couplage même, mais qu'il reçoit plutôt certains renseignements non confidentiels à propos de la précision du couplage d'enregistrements.

Nous élaborons certaines méthodes de régression linéaire simple des données de couplage et nous les comparons aux méthodes fréquentistes courantes. En particulier, ces méthodes existantes partent de l'hypothèse selon laquelle les deux ensembles de données distincts forment un espace de correspondance complète, de sorte qu'il y a appariement vrai entre chaque enregistrement d'un des ensembles de données et un enregistrement unique de l'autre ensemble de données. Notre approche consiste à assouplir cette hypothèse de manière à permettre la situation plus réaliste où l'un ou l'autre des ensembles de données peut contenir certains enregistrements non appariés.

Étant donné le manque d'information sur les mécanismes d'erreur de mesure sous-jacents qui ont causé les erreurs de couplage, toutes les méthodes d'analyse secondaire doivent comporter une hypothèse d'erreurs de couplage non informatives. Nous proposons un test de diagnostic des erreurs de couplage non informatives, que peuvent effectuer les analystes secondaires. Ce test peut être utile en pratique pour décider si l'analyse de régression même est acceptable.

Toutes les méthodes proposées seront illustrées au moyen d'ensembles de données de couplage accessibles au public.

(A) L'inférence statistique à partir de multiples fichiers de données faisant l'objet d'un couplage d'enregistrements probabiliste

Partha Lahiri et Ying Han, University of Maryland, États-Unis

Les organismes statistiques gouvernementaux utilisent fréquemment des méthodes de couplage d'enregistrements probabiliste (CEP) pour relier rapidement et avec précision deux fichiers volumineux ou plus qui comprennent des renseignements sur les mêmes personnes ou entités, à partir des données disponibles, qui ne comprennent habituellement pas de codes d'identification uniques sans erreur. Comme le CEP fait appel à des bases de données déjà existantes, il permet de réaliser de nouvelles analyses statistiques sans devoir investir le temps et les ressources considérables nécessaires à la collecte de nouvelles données. Les erreurs de couplage sont inévitables lorsqu'on combine plusieurs fichiers par CEP en raison de l'absence d'identificateurs uniques sans erreur. Même un faible nombre d'erreurs de couplage peuvent introduire un biais important et une variabilité accrue dans l'estimation des paramètres d'un modèle statistique. On ne saurait trop insister sur l'importance de tenir compte de l'incertitude due au couplage d'enregistrements dans l'analyse statistique. Nous élaborons un cadre théorique d'inférence statistique à l'aide d'un modèle intégré général qui comprend un modèle de couplage pour prendre en compte l'incertitude due au processus de couplage d'enregistrements probabiliste et un modèle mixte pour estimer les paramètres du modèle de couplage. Une version simplifiée de la méthodologie proposée et de l'algorithme numérique est fournie, question de réduire le nombre de calculs à effectuer. Enfin, nous terminerons l'exposé en soulevant plusieurs problèmes épineux qui se posent dans ce domaine de recherche de plus en plus important.

(F) Équations estimantes au niveau de la paire pour l'analyse primaire de données couplées

Abel Dasylva, Statistique Canada, Canada

Une nouvelle méthodologie d'équation estimante est proposée pour l'analyse primaire de données couplées, c.-à-d. une analyse par quelqu'un ayant un accès total aux microdonnées et informations de projet associées. Elle est décrite lorsque les données proviennent du couplage de deux registres ayant une couverture exhaustive de la même population, ou du couplage de deux échantillons probabilistes, qui se chevauchent, comme c'est le cas lorsque lesdits registres ont de la sous-couverture. Cette méthodologie prend en compte l'incertitude concernant le statut d'appariement des paires d'enregistrements, à partir d'un modèle de mélange de la distribution marginale du vecteur de concordances dans une paire. Elle s'appuie sur l'hypothèse d'indépendance conditionnelle entre les vecteurs de concordances et les réponses étant donné les variables explicatives.

(A) Problèmes pratiques de couplage de données: Expériences avec l'Environnement de fichiers couplés (LFC) et sa future infrastructure.
Peter Timusk et Julio Rosa, Statistique Canada, Canada

À venir

Passer du recensement aux sources fiscales : un changement de base de sondage pour une meilleure coordination des échantillons de l'Insee
Thomas Merly-Alpa, Institut national de la statistique et des études économiques, France

Grâce à une plus grande disponibilité et une meilleure qualité des sources administratives issues des services fiscaux, l'Institut National de la Statistique et des Études Économiques (Insee) a développé une alternative pour ses bases de sondage : le Fichier Démographique sur les Logements et les Individus (Fidéli), qui regroupe les informations fiscales, corrigées des doublons et des règles de gestion administrative.

L'utilisation de Fidéli à la place du recensement de la population, traditionnellement utilisé comme base de sondage à l'Insee, permet de diminuer l'étendue des zones d'enquêtes et d'améliorer leur plan de sondage. Cependant, ce changement entraîne la disparition d'informations, la modification de concepts et l'apparition de nouvelles variables, notamment pour le repérage des enquêtes qui pourra se baser sur une géolocalisation plus précise.

Le renouvellement de l'Échantillon-Maître, prévu pour 2020 dans le cadre du projet de Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes (Nautile), ainsi que celui de l'échantillon de l'enquête Emploi en Continu (EEC), sont basés sur ce fichier, qui présente les caractéristiques d'une bonne base de sondage, en particulier l'exhaustivité.

La coordination de ces deux tirages a été étudiée afin de limiter les déplacements des enquêteurs. Les unités primaires de l'Échantillon-Maître sont tirées par échantillonnage spatialement équilibré ; l'échantillon de l'EEC est un ensemble de grappes compactes d'une vingtaine de logements également sélectionnées par sondage spatialement équilibré sur des variables proxy de la situation vis-à-vis du marché du travail. Enfin, leur coordination se fait par l'introduction d'unités de coordination (UC) sélectionnées par sondage indirect via les unités primaires, les grappes de l'EEC étant sélectionnées dans les UC dans une dernière phase.

(A) Une histoire de voyageur transfrontalier : Dois-je rester ou partir?
Tanja Armenski et Zdenek Patak, Statistique Canada, Canada

Les intempéries peuvent influer sur la circulation transfrontalière des véhicules de diverses façons. Les phénomènes météorologiques comme la pluie, la pluie verglaçante, la neige et les températures extrêmes peuvent entraîner d'importantes fluctuations du volume de véhicules qui entrent au pays et qui en sortent. Cet article vise à mieux comprendre l'incidence de la météo sur les flux transfrontaliers de trafic en intégrant les données météorologiques et les données sur le trafic.

La recherche est menée à l'aide des données sur le trafic recueillies par l'Agence des services frontaliers du Canada (ASFC) et utilisées par Statistique Canada comme source administrative des comptes de dénombrement à la frontière. Les données météorologiques ont été obtenues d'Environnement et Changement climatique Canada. Des paramètres météorologiques comme les températures moyennes, les précipitations et les chutes de neige sont obtenus à partir de stations météorologiques individuelles partout au Canada.

Pour expliquer les variations de la série sur le trafic transfrontalier, des modèles ARIMA avec des régresseurs liés aux conditions météorologiques et aux vacances ont été utilisés. Le trafic sortant, les véhicules canadiens revenant des États-Unis et le trafic entrant, les véhicules américains entrant au Canada, sont analysés séparément. Les implications pratiques sont discutées et des recommandations pour d'autres recherches sont fournies.

(A) Utilisation de multiples sources de données pour créer et affiner des agrégations géographiques pour la surveillance des sous-comtés

Angela K Werner et Heather Strosnider, Division of Environmental Health Science and Practice, National Center for Environmental Health, Centers for Disease Control and Prevention, États-Unis

La mission du National Environmental Public Health Tracking Program (Programme de suivi) du Center for Disease Control and Prevention est de fournir de l'information à partir d'un réseau national de données intégrées sur la santé et l'environnement, ce qui oriente les mesures visant à améliorer la santé des collectivités. Le programme de suivi prévoit la diffusion régulière de données à une résolution géographique plus élevée afin d'améliorer la surveillance de la santé environnementale et de favoriser les changements à l'échelle locale. Lors de l'affichage des données à plus haute résolution, il faut tenir compte de plusieurs facteurs, comme la stabilité et la suppression de ces données. Pour ce faire, il faut un regroupement temporel et/ou de nouvelles agrégations de régions géographiques afin de réduire au minimum la suppression et l'instabilité des affichages tout en veillant à ce qu'ils restent classés comme sous-comtés. La méthode de création de ces régions géographiques doit être normalisée afin que les unités géographiques soient comparables d'un État à l'autre et utilisées dans un système national de surveillance.

À l'aide de sources de données multiples, y compris les limites des secteurs de recensement, les données sur la santé et les données sur la population, des agrégations optimales ont été créées pour deux schémas d'agrégation (c.-à-d. un schéma d'agrégation des résultats rares et un schéma d'agrégation des résultats plus communs) pour un ensemble d'États pilotes. Un examen initial des nouvelles agrégations et des consultations avec les États a révélé plusieurs problèmes, comme la fusion entre comtés, des variations dans les fusions et des unités géographiques avec des populations plus importantes que nécessaire. Une autre méthode de fusion utilisant des centroïdes pondérés en fonction de la population a été explorée après l'établissement de seuils de population appropriés pour les deux schémas d'agrégation. Les travaux futurs comprennent l'amélioration des régions géographiques agrégées en abordant certains des défis rencontrés et en explorant l'utilisation de facteurs supplémentaires dans le processus d'agrégation.

(A) Technologies de gouvernance des données de la prochaine génération utilisées dans la statistique officielle

Ryan White, Statistique Canada, Canada

Cet article présentera les résultats de la validation de principe des principales technologies de traitement informatique, d'analyse et de modélisation des données qui sont utilisées dans un cadre de prétraitement des données administratives et de validation des données. Les technologies et le cadre de prétraitement favorisent une bonne gouvernance des données et une science des données reproductibles fondée sur les principes des technologies de logiciel libre et de l'informatique en nuage. Les pipelines de données de référence pour le couplage d'enregistrements à l'aide de sources de données synthétiques sont présentés en se fondant sur la technologie des conteneurs et des grappes de nuage informatique élastique. Des outils modernes et novateurs de la science des données sont utilisés pour la gestion des données et des processus de traitement qui assurent la reproductibilité et la provenance des données

et des processus. Le pipeline de données de référence est un processus de données automatisé et reproductible qui sert d'outil d'établissement de profils et de recherche pour l'analyse de données en mémoire et l'analyse de données distribuées et évolutives. L'article présente des comparaisons du traitement des données à l'aide d'ensembles de données traditionnels orientés en rangées (CSV) avec le format de données en colonnes établi par la norme de l'industrie, organisé pour des opérations analytiques efficaces dans du matériel informatique moderne. Le cadre prototype de prétraitement et de validation des données sert de moteur de modélisation, d'établissement de profils et de conversion des données administratives, qui démontre une stratégie de données pour la prochaine génération de la science des données reproductibles et de l'analyse des données à Statistique Canada.

Séance 8A -- Méthodes en démographie

(F) Projection du niveau de compétences en littératie à l'aide d'un modèle de microsimulation

Samuel Vézina et Alain Bélanger, Institut national de la recherche scientifique, Canada

Cet article a pour objectif de présenter un module qui permet de projeter le niveau de compétences en littératie des adultes (âgés entre 25 et 64 ans) à l'aide du modèle de microsimulation LSD-C. Ce modèle projette la population canadienne selon des variables démographiques (âge, sexe, lieu de résidence, lieu de naissance, statut de génération et statut d'immigration), ethnoculturelles (langue maternelle, langue le plus souvent parlée à la maison, connaissance des langues officielles, groupes de minorités visibles, religion) et socio-économiques (éducation, statut d'activité sur le marché du travail).

Une analyse approfondie des données de trois enquêtes transversales sur les compétences des adultes au Canada a été effectuée en amont de cet exercice de modélisation et de projection. Cette analyse, stratifiée selon le statut d'immigration, a permis d'identifier les facteurs qui déterminent le niveau de littératie de la population et de juger de la comparabilité des données dans le temps (mesure pseudo-longitudinale de l'effet de cohorte).

La méthode retenue est simple car les données disponibles ne permettent pas de modéliser de façon dynamique la trajectoire du niveau de compétences en littératie de la population. Le score de littératie est imputé aux cas simulés sur la base de plusieurs caractéristiques : âge, sexe, région de résidence, éducation, compétences linguistiques, statut d'activité sur le marché du travail, âge à l'immigration, durée de résidence au Canada, pays de naissance et pays d'obtention du diplôme. Le score des individus est recalculé à chaque fois que survient un changement d'état d'une des caractéristiques susmentionnées. Il est possible de calculer le score moyen de la population adulte totale et de mesurer l'impact des changements sociodémographiques projeté sur ce score moyen.

Cette méthode a été appliquée à un autre modèle de microsimulation, PÖB, qui projette la population de l'Autriche.

(A) Mettre les données administratives au cœur du système de statistiques démographiques —l'expérience en Angleterre et au Pays de Galles

Rebecca Tinsley, Office for National Statistics, Royaume-Uni

L'Office for National Statistics (ONS) utilise des données intégrées pour mener des recherches entourant l'ambition qu'a le gouvernement de faire en sorte que les recensements postérieurs à 2021 reposent sur des sources de données alternatives. Les progrès accomplis dans le cadre du projet du recensement basé sur les données administratives ont notamment consisté à réaliser un éventail de produits de recherche qui sont habituellement issus du

recensement décennal; à comparer ces produits avec les statistiques officielles, et à obtenir les commentaires des utilisateurs.

Jusqu'ici, les divers produits de recherche portent sur la taille de la population, les statistiques sur les ménages et un éventail de caractéristiques de la population. Ces produits ont fait appel à diverses méthodes et sources de données intégrées, dont des données administratives, des données commerciales (données agrégées sur les téléphones mobiles) et des données d'enquête.

Récemment, nous avons réorienté nos recherches afin de comprendre comment les données administratives peuvent servir à produire non seulement le dénombrement de la population, mais aussi des composantes de la variation de la population, y compris des estimations de la migration internationale.

Il s'agit d'une étape clé d'un ambitieux programme de travail visant à passer, d'ici le printemps 2020, à un nouveau système essentiellement fondé sur des données administratives pour les statistiques sur la population et la migration en Angleterre et au Pays de Galles.

Ce travail prend appui sur nos recherches sur un recensement basé sur les données administratives et les pousse plus loin, l'objectif étant de parvenir à une compréhension courante de la population et de la migration à l'aide de toutes les sources disponibles.

Cette présentation traitera des points suivants :

•Les progrès réalisés jusqu'à présent

•Les difficultés à estimer des stocks et des flux de population cohérents à partir de données administratives, y compris les résultats de la collaboration à un atelier international sur le même sujet

•Un aperçu des commentaires des utilisateurs

(A) Demographic Characteristics File (DCF) (fichier des caractéristiques démographiques) : Attribuer la race et l'origine hispanique aux migrants internes en utilisant les données du recensement et d'autres sources de données administratives

Amel Toukabri, U.S. Census Bureau, États-Unis

Les migrations internes entre les États et les comtés constituent un facteur important à l'origine des variations annuelles des estimations démographiques infranationales officielles publiées par le U.S. Census Bureau. Le Census Bureau utilise une vaste gamme de données fédérales de recensement décennal ainsi que d'autres données fédérales administratives afin de produire des estimations des migrations en fonction des caractéristiques démographiques. Les données provenant des déclarations de revenus fédérales annuelles et de l'administration de la sécurité sociale sont couplées, à l'échelon individuel, aux données du recensement pour fournir un portrait démographique des personnes ayant migré d'un État à l'autre ou d'un comté à l'autre, par âge, par sexe, par race et par origine hispanique. Le résultat est un fichier maître de personnes correspondant à chaque période de migration, appelé Demographic Characteristics File (DCF) (fichier des caractéristiques démographiques). Dans cet article, nous présenterons la méthode sous-jacente à la production du DCF, en mettant l'accent sur un processus d'imputation en plusieurs étapes pour les cas où les données relatives à la race et à l'origine hispanique sont manquantes. Par rapport à la méthode précédente, le DCF améliore la distribution des âges par race pour les enfants nés après le recensement de 2010. Nous présenterons les différences de distribution selon l'âge et la race en matière de population des comtés entre le DCF et la méthode précédente et nous mettrons en évidence

quelques études de cas. Enfin, nous conclurons en évoquant les limitations, les étapes suivantes et les orientations futures de la recherche.

Séance 8B -- Intégration de données II

(A) L'utilisation des enquêtes Web pour mesurer et estimer les principaux résultats pour la santé, une étude pilote

Yulei HE, Hee-Choon Shin, Bill Cai, Van Parsons, Peter Meyers et Jennifer Parker, Centers for Disease Control and Prevention des États-Unis, États-Unis

Étant donné les contraintes croissantes en matière de coûts et de ressources des enquêtes traditionnelles, les enquêtes en ligne ont souvent été utilisées comme une solution de rechange souple sur le terrain. Cependant, la documentation antérieure a bien documenté les limites générales de l'utilisation des enquêtes en ligne. Il existe également des recherches dynamiques sur l'amélioration de leur utilisation, tant du point de vue appliqué que méthodologique. Une question importante est de savoir si les enquêtes en ligne peuvent produire (directement ou en cours de calibration) des estimations nationales officielles des principaux résultats pour la santé (p. ex., prévalence du diabète, couverture d'assurance). Le National Center for Health Statistics des Centers for Disease Control and Prevention des États-Unis a mené une étude pilote visant à évaluer l'utilisation des enquêtes en ligne pour mesurer et estimer avec précision les résultats importants en matière de santé, de concert avec la National Health Interview Survey (NHIS). La NHIS peut être considérée comme une source de données de référence pour produire des estimations nationales des principaux résultats pour la santé aux États-Unis. Cette présentation abordera le contexte et quelques résultats initiaux de l'étude. Plus précisément, nous comparons les estimations des deux sources en général et parmi des sous-groupes clés. Nous explorons également des méthodes statistiques avancées pour calibrer les estimations des enquêtes en ligne en utilisant la NHIS comme étalon.

(A) Méthode d'intégration des données : Une consolidation de l'hétérogénéité sémantique et des sources de données avec le projet d'évaluation de la politique en matière de détention en Angleterre et au Pays de Galles.

Marie-Eve Bedard, Statistique Canada, Canada

Cette recherche mettra en évidence les problèmes méthodologiques découlant de l'utilisation de multiples sources de données pour le projet de recherche réalisé sur l'évaluation du rendement des politiques de sécurité en détention en Angleterre et au Pays de Galles en 2014. Le projet a utilisé plusieurs sources de données, comme des données administratives, des données d'enquête, des indicateurs de rendement clés et des données du modèle de qualité des prisons recueillies par les chercheurs. Étant donné que les mesures conceptuelles entourant les données ont été prises à partir de sources différentes et que certaines données n'ont pas été recueillies aux deux moments, il y a un manque de certitude quant à la cause et à l'effet de la politique. L'utilisation des données administratives a également fait ressortir d'autres problèmes. L'un des principaux parmi ceux-ci était les limites de l'échantillon. Étant donné que l'échantillon est prédéfini par l'administration, le chercheur travaille avec un bassin de personnes prédéterminées et un autre à deux moments différents. Le même problème a été observé avec les données de l'enquête. Une méthode a été élaborée pour consolider ces sources de données et leur hétérogénéité sémantique aux fins de l'évaluation, au moyen d'une analyse du score de la variation résiduelle, d'une analyse factorielle des composantes principales, ainsi que d'une régression et d'échelles robustes de l'erreur type. Ces méthodes ont été combinées à d'autres méthodes, comme la distance de Cook, le calcul de y et le test du facteur d'inflation de la variance afin d'unifier ces données et de les rendre comparables pour ce genre d'évaluation. Avec ce type de méthode d'intégration des données, bien que

réussie, l'analyse des données demeure limitée à plusieurs égards, ce qui laisse place à des recherches plus poussées pour trouver un moyen de combler l'écart entre ces sources de données.

(A) Combinaison des données au niveau de l'unité et de la région pour l'estimation des petites régions au moyen de modèles multi-niveaux pénalisés

Jan Pablo Burgard, Joscha Krause et Ralf Münnich, Université de Trèves, Allemagne

L'estimation des petits domaines (EPD) est largement utilisée pour obtenir des estimations de domaines en présence de petits échantillons. Les applications modernes d'EPD sont fondées sur des modèles au niveau de l'unité ou du secteur, selon la disponibilité des données sur le sujet respectif. Un modèle au niveau de l'unité tient compte des données inférieures au niveau de la région pour l'estimation des paramètres du modèle, tandis qu'un modèle au niveau de la région utilise des données à ce niveau. Si des données sur les deux niveaux sont disponibles, elles devraient être combinées afin de maximiser la puissance explicative du modèle sous-jacent et d'améliorer les estimations quantitatives par secteur par rapport à la prise en compte d'un seul niveau de données.

Toutefois, la combinaison des données au niveau de l'unité et de la région soulève des questions méthodologiques. Le couplage de l'information exige l'estimation des paramètres du modèle pour les deux niveaux. Ainsi, le nombre d'estimations augmente, ce qui peut déstabiliser les estimations quantitatives de la région en raison d'un manque de degrés de liberté. De plus, les données au niveau des unités et des régions ont des caractéristiques de répartition différentes, comme en ce qui concerne les modèles de dispersion et les structures de covariance au sein des covariables. Par conséquent, les différentes sources de données ne devraient pas être traitées de la même façon dans le processus d'estimation.

Nous étudions les modèles à niveaux multiples pénalisés pour résoudre ces problèmes et nous combinons les données au niveau de l'unité et de la région pour améliorer les estimations quantitatives de la région par rapport aux méthodes standards d'EPD. Les pénalités nettes multivariées norme L1, norme L2 et élastiques sont utilisées pour la régularisation spécifique au niveau afin d'équilibrer les différentes sources de données et de produire des prévisions optimales du modèle. Une application empirique en médecine sociale est fournie en combinant des données d'enquête allemandes et des microdonnées de recensement.

(A) Mesures de la qualité pour le rapport mensuel sur le pétrole brut et le gaz naturel (MCONG)

Evona Jamroz, Lihua An, et Sanping Chen, Statistique Canada, Canada

Le programme mensuelle Pétrole brut et gaz naturel (PBGN) est une composante essentielle du produit intérieur brut mensuel canadien. Ce programme combine trois catégories de sources de données: les données provenant de plusieurs sondages reliés, des données administratives d'autres agences gouvernementales, et des données d'allocation historiques basées sur « l'opinion d'experts ». Un nouveau système, intégrant les données des trois sources ci-mentionnées, est en création pour le programme PBGN. Dans ce papier, nous résumons le travail en cours et les défis demeurants pour le développement de mesures de qualité pour les estimés du nouveau programme PBGN.

En ce qui a trait aux trois source de données, les données administratives sont fournies dans une structure agrégée et pour lesquelles nous assumons qu'il n'y a aucune erreur. Pour les données des sondages, la variance d'échantillonnage et/ou d'imputation peut être estimée en utilisant les méthodes conventionnelles. Un défi à souligner est d'estimer l'erreur associée à

un paramètre fondé sur le jugement d'un expert. Nous proposons une approche Bayésienne pour un tel paramètre.

Enfin, nous proposons un processus largement inspiré de la linéarisation de Taylor pour intégrer les différents composantes de la variance dans un seul coefficient de variation (CV) pour les estimés finaux du programme PBGN. Les situations pour lesquelles le CV n'est pas une mesure de qualité adéquate seront aussi discutées.

(A) Résultats des ateliers sur les données intégrées organisés par le Federal Committee on Statistical Methodology et la Washington Statistical Society

Alexandra M. Brown, University of Maryland, États-Unis

Dans l'ensemble des organismes statistiques fédéraux, on s'intéresse de plus en plus à l'intégration des données traditionnelles d'enquête et de recensement aux données auxiliaires (structurées et non structurées) afin d'accroître la qualité et l'actualité des données et des statistiques produites. Pour combler l'écart qui existe dans la compréhension de la qualité de ces jeux de données (comparativement aux enquêtes-échantillons) et améliorer la communication entre les producteurs de données et les utilisateurs de données, le Federal Committee of Statistical Methodology (FCSM) et la Washington Statistical Society (WSS) ont organisé conjointement trois ateliers qui ont exploré les pratiques actuelles pour rapporter sur la qualité des données intégrées de façon transparente. Le présent rapport résume les trois ateliers et réunit les principaux thèmes.

Séance 9A -- Défis d'accès aux données - Vie privée et confidentialité à l'ère des sources de données multiples

(A) Le cadre qui sous-tend l'outil de classification de la confidentialité et son rôle dans la modernisation

Joseph Duggan, Michelle Marquis, Claude Girard et Jack Gambino, Statistique Canada, Canada

Avec l'outil de classification de la confidentialité, Statistique Canada met en œuvre composante, certes petite mais clé, de soutiende sa récente initiative de modernisation. Les renseignements statistiques de naturesensible sont maintenant classifiés selon un continuum de risque, qui remplace la classification binaire traditionnelle qui sous-tendait deux environnements de travail distincts : le réseau A pour l'utilisation et le traitement internes de l'information protégée, et le réseau B pour la communication externe et la diffusion de nos produits statistiques. La combinaison de ce changementavec d'autres initiatives - en lien à l'utilisation de sources de données alternatives et combinées, à la modernisation de l'accès aux microdonnées et au regroupement d'un plus grand nombre de partenaires dans le cadre d'efforts de collaboration - permettra d'harmoniser nos pratiques de contrôle de la divulgation avec les infrastructures de TI actuelles et futures. L'outil vise à faciliter tout cela en sensibilisant davantage les gens aux questions et aux pratiques en matière de confidentialité, tout en aidant les gardiens des données à déterminer le niveau de confidentialité de toutesdonnées sélectionnées détenues à Statistique Canada. Le présent document décrit la méthodologie qui sous-tend l'outil de classification de la confidentialité, intentionnellement simple, et les leçons apprises au cours de son élaboration.

(A) Pourquoi le U.S. Census Bureau a adopté la confidentialité différentielle pour son recensement de la population et du logement de 2020

John M. Abowd, U.S. Census Bureau, États-Unis

Le U.S. Census Bureau reconnaît que les risques de reconstruction d'une base de données, telles qu'elles sont définies par Dinur et Nissim (2003), constituent une réelle vulnérabilité des systèmes de contrôle de la divulgation utilisés pour protéger toutes les publications statistiques issues des recensements décennaux précédents. Nos propres recherches confirment cette vulnérabilité, qui est maintenant désignée comme un problème d'entreprise. La confidentialité différentielle a été inventée en 2006 (Dwork et coll., 2006) en tant que moyen de remédier à cette vulnérabilité et de veiller au respect des principes de protection des statistiques produites à partir de sources de données confidentielles. Ces systèmes officiels de protection de la vie privée ont deux propriétés que n'ont pas les systèmes traditionnels de contrôle de la divulgation : (1) ils résistent à la composition et (2) la protection ne se dégrade pas lors du post-traitement. La propriété de résistance à la composition signifie que la perte de confidentialité découlant d'une séquence d'algorithmes différentiellement privés appliqués aux mêmes données n'est pas supérieure à la somme de la perte de confidentialité associée à chaque composante. Ces propriétés sous-entendent que les statistiques protégées par la confidentialité différentielle satisfont à un paramètre de perte de confidentialité calculable connu, qui constitue la bonne mesure globale du risque de divulgation. Elles sous-entendent également que la protection contre la divulgation est « à l'épreuve du temps » — sa force ne dépend pas d'hypothèses au sujet des ensembles de renseignements actuels ou futurs ou des capacités de calcul de l'utilisateur de données. Le prochain recensement décennal des États-Unis, qui se déroulera en 2020, donnera lieu à la publication de produits de données protégés par la confidentialité différentielle. La mise au point de cette protection est assujettie à un paramètre global de perte de confidentialité qui sera déterminé par le comité de direction de la politique de gérance des données du Census Bureau. L'adéquation de chaque statistique à son utilisation sera mesurée directement, compte tenu de l'incertitude due au contrôle de la divulgation statistique.

(A) Mise en œuvre de registres nationaux de santé qui préservent la vie privée

Rainer Schnell et Christian Borgs, University of Duisburg-Essen Lotharstr, Allemagne

La plupart des pays développés exploitent des registres de santé tels que des registres de naissances. Ces types de registres sont importants pour des applications en recherche médicale, telles que pour des études de suivi des traitements contre le cancer. Le couplage de ces registres à des données administratives ou à des données d'enquêtes offre des possibilités de recherche, mais peut soulever des préoccupations liées à la protection de la vie privée. En raison de la récente harmonisation des règles en matière de protection des données en Europe avec le Règlement général sur la protection des données (RGPD), il est possible d'établir des critères sur la façon d'exploiter de tels registres tout en préservant la vie privée.

Un registre de données sur la santé utilisé à des fins de couplage doit être protégé contre les risques de réidentification, sans que la qualité du couplage soit compromise. Nous ferons la démonstration de solutions qui offrent une forte résilience contre les attaques de réidentification tout en préservant la qualité du couplage à des fins de recherche. Plusieurs techniques ultramodernes de couplage d'enregistrements qui préservent la vie privée ont été comparées au cours de l'élaboration. Pour les essais en situation réelle, nous avons apparié les données sur la mortalité provenant d'un registre administratif local (n = 14 003) avec les dossiers de santé d'un hôpital universitaire (n = 2 466). Nous avons procédé à un essai à plus grande échelle des solutions proposées en appariant 1 million d'enregistrements simulés d'une base de données nationale de noms avec un sous-ensemble corrompu (n = 205 000).

Nous discuterons du risque de réidentification associé à différentes mises en œuvre, compte tenu des nouvelles méthodes d'attaque récemment mises au point. Enfin, nous présenterons des recommandations détaillées sur l'exploitation de registres de santé préservant la vie privée qui sont susceptibles d'être utilisés à des fins de couplage, y compris des lignes directrices opérationnelles et des suggestions de pratiques exemplaires.

Séance 9B -- Utilisation de sources de données administratives ou alternatives II

(A) Combiner les données du recensement et d'Hydro-Manitoba pour comprendre la consommation d'électricité résidentielle

Chris Duddek, Statistique Canada, Canada

Statistique Canada reçoit les dossiers mensuels d'Hydro-Manitoba depuis 2015. Puisque Hydro-Manitoba est le fournisseur d'électricité de la plupart des Manitobains, il semblerait facile d'obtenir la consommation annuelle totale d'électricité résidentielle. Toutefois, lorsque l'on compare ces données aux estimations publiées, on remarque des écarts. Pour mieux en comprendre les causes, nous modifions les variables géographiques dans le fichier afin de comparer les données d'Hydro-Manitoba aux chiffres du Recensement de 2016. La comparaison nous permet de détecter des problèmes attribuables à la manière dont les données du fichier d'Hydro-Manitoba sont structurées. Une brève analyse transversale des données est suivie d'un examen des éléments longitudinaux des données. L'article se conclut en soulignant les étapes nécessaires pour corriger les problèmes afin d'estimer la consommation annuelle totale d'électricité.

(A) Évaluer les effets de bien-être avec l'Enquête sociale générale

Xavier Lemyre, Ministère du Patrimoine canadien, Canada

La méthode d'évaluation du bien-être en trois étapes est utilisée pour évaluer les impacts de la participation à des activités telles que le théâtre sur la satisfaction à l'égard de la vie. La méthode mesure le montant d'argent qui serait nécessaire pour rendre un participant à une activité aussi bien en l'absence de sa participation que lorsqu'elle se produit. En raison de problèmes d'endogénéité, l'approche de variable instrumentale est utilisée pour mesurer les effets du revenu sur le bien-être. De plus, les effets de l'activité et du revenu sur la satisfaction à l'égard de la vie sont estimés par étapes, ce qui permet d'utiliser différents ensembles de données par étapes.

Dans cette présentation, nous examinerons comment des équations simultanées sont utilisées pour estimer les valeurs monétaires qui traduisent les effets de bien-être de la participation à des activités artistiques, culturelles et sportives. De nouvelles applications rendues possibles grâce aux données administratives liées et à un accès aux microdonnées détaillées seront présentées, ainsi que les possibilités de recherches ultérieures.

(E) Vers un système statistique centré sur des registres : Éléments nouveaux à Statistique Canada

Jean Pignal, Philippe Gagné, Christian Wolfe et Tanvir Quadir, Statistique Canada, Canada

Le Projet de transformation et d'intégration des registres statistiques vise à bâtir et tenir à jour une infrastructure de registres statistiques (IRS) formée de registres de base interconnectés (population, immeubles, entreprises et activités) selon des données administratives. L'IRS a pour buts de maximiser l'utilisation des renseignements utiles tirés des données déjà recueillies, d'améliorer l'actualité des statistiques et de réduire le fardeau de réponse tout en protégeant la vie privée des Canadiens. Cette présentation portera sur ce qui suit : la méthode et le cadre pour l'élaboration continue de registres anonymes de la population et des immeubles, le remaniement du Registre des entreprises existant et le cadre conceptuel d'un registre des activités. Cette présentation comprendra également un aperçu du Système canadien des registres statistiques intégrés (SCRSI), qui réunit les registres de base sous-jacents et jette les bases pour l'infrastructure du registre.

(E) Remplacer les questions de recensement en utilisant des données administratives couplées

Scott McLeish et Athanase Barayandema, Statistique Canada, Canada

Depuis 2011, les données administratives d'Immigration, Réfugiés et Citoyenneté Canada (IRCC) sont couplées aux données de recensement afin de documenter le traitement de la vérification et de l'imputation, ainsi que la certification des résultats des recensements. Ces données administratives sur l'immigration comprennent les caractéristiques à l'arrivée de tous les immigrants arrivés au Canada depuis 1980, ainsi que certaines données qui remontent à 1952. À la suite du couplage avec ces données administratives, deux variables, la catégorie d'admission des immigrants et le type de demandeur, ont été ajoutées au Recensement de 2016 pour la première fois.

Ces couplages aux sources de données administratives ont mis en évidence la qualité des questions actuelles du recensement liées à l'immigration et ont permis de mieux comprendre les erreurs attribuables à la non-réponse et à la mesure. Par exemple, le lien entre le nombre d'années pendant lesquelles les immigrants ont vécu au Canada et l'exactitude de leurs réponses a été établi.

Cette présentation décrira une étude qui examine la faisabilité de remplacer les questions sur l'immigration dans le Recensement de 2021 en utilisant des dossiers administratifs couplés. Elle examinera les forces et les limites du statu quo et de l'utilisation des données administratives couplées, y compris les problèmes liés à l'intégration des données, comme les incohérences entre les sources de données.

(A) Méthodes automatiques de collecte de données : Points de vente au détail et centres commerciaux de Vancouver

Paul Holness, Statistique Canada, Canada

Les estimations des ventes au détail sont l'un des principaux indicateurs économiques utilisés par la Banque du Canada et la communauté des gens d'affaires pour générer une politique stratégique, guider les décisions en matière d'investissement (stratégies) et évaluer la performance économique. Les plus récents chiffres indiquent que la contribution de l'industrie de la vente au détail (SCIAN 44-45) au PIB était d'environ 90,5 milliards de dollars en 2014 et représentait 4,9 % du PIB total du Canada. Le secteur de la vente au détail représente la plus vaste industrie au Canada, employant près de deux millions de personnes. Les magasins à grande surface et les chaînes de magasins de détail comptent parmi les plus grands importateurs du Canada et au cours des dernières années, un nombre croissant de ces détaillants à points de vente fixes ont commencé à augmenter leurs offres de commerce électronique.

L'Enquête mensuelle sur le commerce de détail (EMCD) recueille des renseignements sur les ventes, les ventes du commerce électronique et le nombre de points de vente au détail selon la province, le territoire et certaines régions métropolitaines de recensement (RMR) auprès d'un échantillon de détaillants. Cet article utilise des méthodes de collecte en ligne afin de déterminer si la base de sondage de l'Enquête mensuelle sur le commerce de détail (EMCD) peut être générée automatiquement au moyen des données de l'interface de programmation d'applications (IPA) des Services Web de Google et du registre du Canadian Directory of Shopping Centres (CDSC).

La nouvelle application utilise des données de référence recueillies manuellement auprès de la RMR de Vancouver depuis juin 2017 et des données Web récemment extraites afin de créer un prototype qui présente un concept d'exécution en tableau de bord comme couche de présentation du système. Le système proposé a recours à des services aux entreprises

automatisés issus des services en ligne de Google, à des recherches à proximité, à des recherches textuelles ainsi qu'à des outils de moissonnage Web, comme Scrapy et Selenium, afin d'extraire des renseignements à partir du registre du Canadian Directory of Shopping Centres (CDSC). Ensemble, ces technologies fournissent un accès direct à des données sources et réduisent les interventions manuelles d'acquisition de données. Les données JSON extraites sont facilement analysées en ensembles de données en valeurs séparées par des virgules (CSV) qui sont faciles à lire dans SAS ou Excel. Finalement, les résultats ont démontré que les méthodes fondées sur le Web produisaient une population comparable de points de vente au détail tout en nécessitant beaucoup moins de ressources comparativement au processus manuel.

Notre étude fournit des renseignements quantitatifs détaillés qui permettent aux gestionnaires d'accomplir les tâches suivantes : (1) évaluer l'utilisation de méthodes automatiques dans la préparation, la mise à jour et l'évaluation de la base de sondage de l'EMCD; (2) évaluer les répercussions de ces méthodes sur la qualité des attributs des points de vente au détail sur le Registre des entreprises (RE) de Statistique Canada; et (3) estimer l'efficacité et les coûts globaux de ces méthodes sur le programme de la DCDIS.

Séance 11 -- Séance Plénière

(A) Mesurer l'incertitude en présence de multiples sources de données

Sharon Lohr, Arizona State University, États-Unis

Dans un échantillon probabiliste avec réponse complète, la marge d'erreur fournit une mesure fiable et théoriquement justifiée de l'incertitude. Toutefois, lorsque des estimations provenant de multiples échantillons ou sources de données administratives sont combinées, les marges d'erreur traditionnelles sous-estiment l'incertitude — les différences entre les statistiques de diverses sources dépassent souvent la variabilité d'échantillonnage estimée. Nous examinerons un certain nombre de méthodes qui ont été proposées pour mesurer l'incertitude découlant de sources de données combinées en appliquant le tout à l'estimation de la prévalence du tabagisme et des taux d'agressions sexuelles, et nous décrirons quelques orientations possibles de la recherche.