AMÉLIORATION DU PARTAGE DES DONNÉES AU MOYEN DE « PLANS SÉCURITAIRES »

Production de connaissances pour éclairer la pratique scientifique

Kristine Witkowski

Inter-university Consortium for Political & Social Research, Université du Michigan

CONTEXTE DE PARTAGE DES DONNÉES

- Aux termes de la politique aux États-Unis, il faut soumettre un plan de partage des données lorsque l'on demande du financement pour la recherche
- Effort actuel en vue de revoir le processus pour protéger les sujets humains (ANPRM 7/22/2011)
- Approche multidimensionnelle pour la formulation de données, en vue d'une utilisation sécuritaire et optimale (Lane, 2007)
- Nécessité de réfléchir au partage des données rapidement et fréquemment, au moyen de connaissances spécialisées

PRINCIPE DIRECTEUR

- Les producteurs doivent être capables de tirer parti efficacement de la recherche sur la divulgation pour déterminer avec précision les travaux requis pour répondre de façon optimale aux objectifs du partage des données

OBJECTIF

- Améliorer la valeur et l'utilisation sécuritaire des données en sciences sociales, particulièrement les microdonnées contextualisées
- Simuler la pratique scientifique pour produire des connaissances, en vue d'un usage généralisé et adapté

PROJET DE RECHERCHE

- Projet sur cinq ans appuyé par le National Institute for Child Health & Development
- Dan Brown — Université du Michigan
  Michael Elliott
  Trivellore Raghunathan
  Kristine Witkowski
  Kevin Leicht — Université de l'Iowa

CONSEIL CONSULTATIF

- John Abowd, Université Cornell
- Marc Armstrong, Université de l'Iowa
- Jerry Reiter, Université Duke
- Natalie Shlomo, Université de Southhampton
- Christopher Skinner, London School of Economics & Political Sci.
- Laura Zayatz, U.S. Census Bureau

SIMULATIONS DE DIVULGATION

- Simuler des travaux de divulgation pour des séries représentatives de fichiers de microdonnées artificiels
- Estimer les résultats de la divulgation, mesurés pour un ensemble exhaustif d'éléments de risque, d'utilité et de coût
- Conformément à d'autres spécifications de paramètres d'échantillonnage et de conception de bases de données
- Contrôle d'ensembles itératifs d'emplacements d'enquête (ou un ensemble précis ciblé pour la collecte)

SIMULATIONS DE DIVULGATION

- Les microdonnées restreintes de l'American Community Survey fournissent de l'information géographique particulière utilisée tout au long du projet
- Les fichiers artificiels offrent une souplesse méthodologique et permettent d'assurer la confidentialité des données
- Dans le cadre du projet, on procède à des expériences, afin d'évaluer l'exactitude des estimations calculées à partir de données artificielles

MODÈLES POUR LES DONNÉES ARTIFICIELLES ET LES PROBABILITÉS DE RÉIDENTIFICATION DE LA POPULATION

- Estimer la composition des participants probables, ainsi que la population générale à l'étude
- Imputation multiple
- Distributions de probabilités conjointes pour des pixels de 1-km^2
  - Détermination d'attributs personnels et de résultats en matière de santé non nominatifs
  - LandScan, recensement décennal, microdonnées de l'ACS, BRFSS
  - Méthodes de pondération aréolaire pour estimer les données de pixel à partir de données plus agrégées (c.-à-d. des groupes d'îlots)
  - Contrôle de la non-réponse (pondérée et non pondérée)

MÉTADONNÉES

$$\mu^{m}_{a} ; \sigma^{m}_{a} ; \delta^{m} = f [s, r, d]$$

Pour tout résultat de divulgation donné (m) découlant des éléments d'échantillon (s) , de diffusion (r) et de CDS (d) estimés à partir de la reproduction de fichiers artificiels (a,f)

Où :

μ^{m}_{a} = Résultat estimé (moyen)

σ^{m}_{a} = Variance du résultat estimé (fiabilité, précision)

δ^{m} = Différence par rapport au résultat observé (validité, exactitude)

$$o^{m}_{ra,f} = m(o^{m}_{--,-}) + m(o^{m}_{ra,-}) + e(o^{m}_{ra,f})$$

Où :

f = Fichier compilé à partir d'une itération d'échantillon particulière

ra = Expérience à partir de données réelles (r) ou artificielles (a)

m = Mesures différente des résultats de la divulgation

$o^{m}_{ra,f}$ = Résultat de la divulgation pour le fichier

$m(o^{m}_{--,-})$ = « Total » du résultat moyen des fichiers

$m(o^{m}_{ra,-})$ = Résultat moyen des fichiers réels ou artificiels

$e(o^{m}_{ra,f)}$ = Variation parmi les fichiers réels ou artificiels

Exactitude du résultat estimé

F_o^m = CMT(Entre les groupes) / CME (À l'intérieur des groupes)

$\delta_\mu^m = [m(o^m_{a,-}) - m(o^m_{r,-})] / m(o^m_{r,-})$

$\delta_\sigma^m = [s(o^m_{a,-}) - s(o^m_{r,-})] / s(o^m_{r,-})$

$\phi^m = s(o^m_{r,-}) / s(o^m_{a,-})$

$\theta^m = m(o^m_{r,-}) - [\phi^m * m(o^m_{a,-})]$

Résultat estimé (corrigé)

$\mu^m_a = E(\theta^m) + [E(\phi^m) * m(o^m_{a,-})]$

Variance du résultat estimé (corrigée)

$\sigma^m_a = E(\phi^m) * s(o^m_{a,-})$

MÉTADONNÉES

$$\mu^{m}_{a} ; \sigma^{m}_{a} ; \delta^{m} = f [s, r, d]$$

Pour tout résultat de divulgation donné (m) découlant des éléments d'échantillon (s) , de diffusion (r) et de CDS (d) estimés à partir de la reproduction de fichiers artificiels (a,f)

Où :

μ^{m}_{a} = Résultat estimé (moyen)

σ^{m}_{a} = Variance du résultat estimé (fiabilité, précision)

δ^{m} = Différence par rapport au résultat observé (validité, exactitude)

ÉLÉMENTS D'ÉCHANTILLON (s)

- Étude d'une population d'adultes (18 ans et plus)
- Région d'étude limitée : Indiana, Illinois, Michigan, Ohio, Wisconsin
- Enquête auprès des ménages fondée sur un échantillon à deux degrés de secteurs de recensement et d'unités de logement regroupées à l'intérieur de ces secteurs
- Taille totale de l'échantillon
- Plan d'échantillonnage détaillé – emplacements, populations cible et taux d'échantillonnages

ÉLÉMENTS DE DIFFUSION (r)

- Au niveau de la personne
  - Caractéristiques d'identification des répondants (p. ex., âge, sexe, race/origine ethnique, obésité, composition du ménage, attributs du conjoint)
  - Résultats en matière de santé non nominatifs : santé autodéclarée, problèmes de santé chroniques (p. ex., diabète)
  - Ensembles de 6 ou 10 attributs, maintenus constants

ÉLÉMENTS DE DIFFUSION (r)

- Au niveau géographique
  - Identificateurs directs de la région, de l'État et de la densité de population (p. ex., statut de RSM)
  - Identificateurs indirects ou variables contextuelles
    - Unités spatiales administratives et géoréférencées : comtés, secteurs de recensement, groupe d'îlots, et pixels de 1-km^2
    - Données destinées au public : recensement, EPA, NASA, autres
    - Ensembles de variables d'intérêt général (listes de souhaits)
    - Échantillons représentatifs de tous les ensembles possibles

ÉLÉMENTS DE DIFFUSION (r)

- Au niveau géographique
  - Identificateurs indirects ou variables contextuelles
    - Domaine ou mesure : caractéristiques de la population et des logements, qualité de l'air, superficie boisée, proximité d'incinérateurs, miles de routes
    - Type ou taille aréolaire des régions géographiques sous-jacentes :
      Pixels, groupes d'îlots, secteurs de recensement et comtés
    - Nombre de variables à diffuser
    - Entropie

ÉLÉMENTS DE CDS (d)

- Expériences de couplage : au niveau géographique
  - Étrangers et intrus connus
  - Lien avec des sources publiques de variables contextuelles
    - Données complètes et précises
  - Appariements : régions géographiques (dans la population) comportant les mêmes attributs que les emplacements visés par l'enquête
  - Îlots : Région, État, densité de population
  - Attributs personnels couplés aux attributs géographiques utilisés pour préciser les estimations des régions particulières qui ont été intégrées à l'étude

ÉLÉMENTS DE CDS (d)

- Techniques de CDS : au niveau géographique
  - Faire l'hypothèse que les variables d'identification personnellse ne sont pas masquées
  - Appliquées après la collecte : Recodage global et valeurs synthétiques de variables contextuelles
    - Couplage déterministe, couplage probabiliste, k plus proches voisins, distance de Mahalanobis, autres
  - Appliquées avant la collecte : Le « plan sécuritaire »

PLAN SÉCURITAIRE

- Formuler une technique de CDS innovatrice s'appliquant à la réidentification des attributs personnels, en maintenant constants les attributs géographiques
- Étude qui supplémente leur échantillon et recueille des données en conséquence pour réduire le risque d'un échantillon unique (c.-à-d., k-anonymat)
- Passer outre aux contraintes de la pratique établie qui consiste à s'occuper de la divulgation une fois les données recueillies

PLAN SÉCURITAIRE

- Échantillon de base : Plan d'échantillonnage formulé pour répondre aux objectifs analytiques (U_b, C_b)
- Examen de la divulgation préemptif : Risque de divulgation de l'échantillon de base (R_b)
- Échantillon supplémentaire : Plan d'échantillonnage formulé pour respecter les objectifs de confidentialité
  ($R_s \sim 0$, $U_s > U_b$, $C_s > C_b$)

Où : R = Risque, U = Utilité, C = Coût

RÉSULTATS DE LA DIVULGATION (m)

- Risque
  - Divulgation d'identité : Probabilité de réidentification de la population et k-anonymat
    - Personnes de la population à l'étude partageant des attributs géographiques et personnels similaires
    - Répondants partageant des attributs géographiques et personnels similaires dans le contexte de la diffusion des données
  - Tailles de cellules continues; situation à risque avec seuil défini par la sensibilité du contenu
  - Par enregistrement – par sous-population cible – par plan

RÉSULTATS DE LA DIVULGATION (m)

- Utilité
  - Perte d'information : Caractérisation de la diffusion comme un tout, y compris des mesures continues et catégoriques; invariant au changement à l'échelle
    - 12 mesures fournies par Domingo-Ferrer, Torra et Mateo-Sanz
  - Biais de suppression : Régions géographiques et sous-populations les plus à risque
  - Inférence statistique : Rapports entre les résultats en matière de santé et les contextes spatiaux

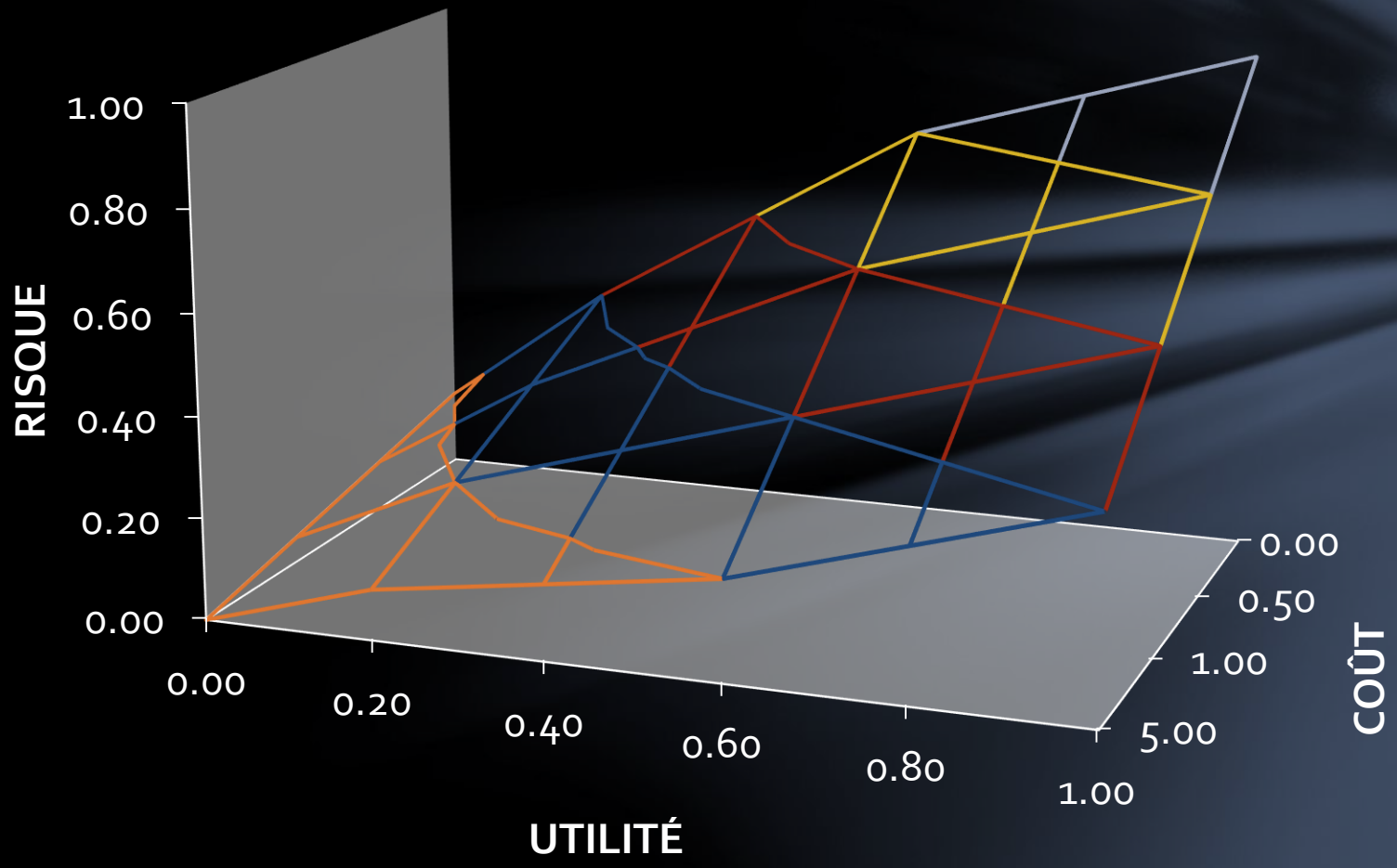
RÉSULTATS DE LA DIVULGATION (m)

- Coût
  - Valeurs monétaires moyennes des dépenses d'enquête
  - Fonction du nombre de tirages requis pour respecter les tailles d'échantillons cibles pour des sous-populations définies et détaillées de façon large
  - Reposant directement sur la pratique scientifique

AUTRES CONSIDÉRATIONS

- Valeur ajoutée et coût des échantillons dispersés au niveau spatial qui maximisent la variance des attributs géographiques (s)
- Compromis entre les données sur les attributs personnels et le niveau de détail géographique (r)
- Protection offerte par l'erreur de mesure et la concentration des populations difficiles à dénombrer (d)
- Rôle des sources de données administratives (d)

CARTE DE RISQUE-UTILITÉ - COÛT



RÉPERCUSSIONS

- Cadre souple pour la production de données empiriques, qui peuvent éclairer de façon large la prise de décisions
- Soutien du partage et de la consommation de connaissances complexes et très spécialisées
- Soutien des politiques concernant le partage des données et la protection des sujets humains
- Auditoires : Établis et nouvelles études des agences statistiques fédérales et des établissements universitaires; DRB, IRB, archives; bailleurs de fonds

MERCI. QUESTIONS?