Recuejl du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

À LA RECHERCHE DE PLANS DE SONDAGE OPTIMAUX

Paul W. Ludington[1]

RÉSUMÉ

Le Programme de plan optimal représente une solution de rechange pour le remaniement décennal des enquêtes démographiques auquel procède à l'heure actuelle le U.S. Census Bureau. Le programme a pour but de minimiser l'erreur quadratique moyenne (EQM) des estimations d'enquête par optimisation de la sélection des échantillons annuels de ménages.

Les premiers travaux se sont concentrés sur l'utilisation de systèmes multi-agents (connus sous le nom d'intelligence artificielle distribuée) pour produire des échantillons annuels optimaux pour toutes les enquêtes démographiques. Le premier système multi-agents optimise les inputs du remaniement. Il représente chaque ménage comme un agent logiciel autonome, projette les données chronologiques sur les ménages pour l'année courante de remaniement sous forme de densités de probabilité et utilise ces densités de probabilité pour simuler les caractéristiques courantes des ménages qui sont calées * de façon optimale + sur les estimations d'enquête courantes. Le processus de calage doit tenir compte des priorités concernant les densités de probabilité (en ce qui a trait à la qualité des données) tout en minimisant les ajustements par simulation.

Le deuxième système multi-agents sélectionne des échantillons optimaux pour toutes les enquêtes démographiques. Il applique l'algorithme bayésien d'optimisation (ABO) à chaque étape du plan de sondage pour répartir les unités d'échantillonnage en sous-ensembles échantillonnés et non échantillonnés, sans qu'il soit nécessaire de stratifier au départ les unités d'échantillonnage. Les échantillons sont * optimaux + en ce sens que l'ABO minimise un critère qui inclut à la fois la différence quadratique moyenne des estimations d'échantillon simulées par rapport aux valeurs de population et la différence quadratique moyenne du réseau bayésien d'échantillon par rapport au réseau de population.

Le présent article, qui présente les résultats de travaux de recherche et d'analyse entrepris par les employés du Census Bureau, a été soumis à un examen de portée plus limitée que celui auquel sont soumises les publications officielles du Bureau. Ce rapport est diffusé afin de tenir les parties intéressées au courant des travaux de recherche en cours et de favoriser la discussion.

MOTS CLÉS : Programme de plan optimal, systèmes multi-agents, algorithme bayésien d'optimisation, réseaux bayésiens, erreur quadratique moyenne

1. INTRODUCTION

Le Programme de plan optimal vise à réaliser trois objectifs stratégiques du U.S. Census Bureau. Le premier est de rendre les processus d'enquête plus efficaces. Le deuxième est de produire des statistiques exactes et à jour grâce à l'élaboration de nouveaux échantillons qui reflètent les caractéristiques courantes et la répartition géographique de la population. Le troisième est d'améliorer l'actualité, l'exactitude, la pertinence, l'accessibilité et la rentabililté des enquêtes et des recensements grâce à l'adoption de nouvelles applications techniques et méthodologiques (U.S. Census Bureau, 2002).

La société comporte des **processus démographiques**, comme le vieillissement et la migration, qui déterminent les caractéristiques et les interactions des individus. Ces processus deviennent de plus en plus complexes et intégrés (comme les communications par Internet). Les **systèmes démographiques** sont créés pour suivre et contrôler les processus démographiques. Ces systèmes (comme les enquêtes par sondage) deviennent de plus en plus compliqués, intégrés, coûteux, automatisés et intelligents. Un objectif est de construire un système optimal d'échantillonnage. Les **données démographiques** sont produites par les systèmes démographiques. Ces données deviennent de plus en plus détaillées, essentielles et confidentielles (comme les données pour la sécurité intérieure). Un deuxième objectif est d'optimiser les inputs du processus de conception des enquêtes. Les **estimations démographiques** sont calculées d'après les données

1U.S. Bureau of the Census, salle 3723-3, Washington, DC 20233

démographiques. Ces estimations doivent être de plus en plus fines (pour des petites régions et des petits domaines), intégrées et exactes. Le troisième objectif est de sélectionner des échantillons qui minimisent l'erreur quadratique moyenne des estimations d'enquêtes. Les applications démographiques intègrent des estimations démographiques exactes. Ces applications sont de plus en plus essentielles à la survie de l'homme.

Le Programme de plan optimal vise à améliorer l'information utilisée pour le remaniement des enquêtes démographiques, la façon dont cette information est traitée et le calendrier du remaniement. À l'heure actuelle, les inputs pour l'échantillonnage comprennent principalement des totalisations de données du recensement. Les inputs d'échantillonnage proposés incluent des données provenant de recensements, d'enquêtes et de dossiers administratifs et englobent des opérations de prévision, de simulation et de totalisation. À l'heure actuelle, les bases de sondage incluent les listes d'unités, de régions, de logements de groupe et de constructions neuves. La base de sondage proposée est le fichier maître d'adresses. Le calendrier de remaniement sera annuel au lieu de décennal et les échantillons proposés (comme les unités primaires d'échantillonnage mises à jour) ne seront mis en application lorsqu'il sera rentable de le faire. Les domaines actuels d'estimation varient selon l'enquête et incluent l'ensemble du pays, ou des régions ou des États individuels. Le domaine d'estimation proposé sera systématiquement l'État individuel. L'algorithme actuel d'échantillonnage est la stratification et la sélection. L'algorithme proposé d'échantillonnage, appliqué tant aux unités primaires d'échantillonnage qu'aux ménages, est l'algorithme bayésien d'optimisation (ABO). Cet algorithme minimise les erreurs quadratiques moyennes des estimations d'enquête simulées, contrairement à la pratique courante qui consiste à minimiser les coefficients de variation des variables de stratification. L'ABO est un algorithme de recherche génétique qui sélectionne des unités d'échantillonnage directement, sans qu'il soit nécessaire de procéder à une stratification a priori. Les poids d'échantillonnage courants refléteront les probabilité de sélectionner des échantillons optimaux plutôt que la méthode actuelle de pondération par des mesures de taille et d'intervalles d'échantillonnage. On recourt à l'estimation de la variance durant l'échantillonnage pour calculer les erreurs quadratiques moyennes des estimations simulées d'enquête, au lieu de calculer les coefficients de variation des variables stratifiées.

Le 'Programme de plan optimal s'améliore constamment grâce à l'intégration rapide de méthodes et de techniques de pointe, comme les systèmes multi-agents. Ces systèmes s'appuient sur l'intelligence artificielle pour améliorer le processus d'échantillonnage. À la section II, nous donnons un aperçu des systèmes multi-agents. À la section III, nous discutons d'un système multi-agents pour optimiser les inputs du processus de remaniement. À la section IV, nous décrivons un deuxième système multi-agents pour la sélection d'échantillons optimaux pour toutes les enquêtes démographiques. À la section V, nous mentionnons l'orientation de futurs travaux de recherche pour le Programme de plan optimal. Le tableau 1 donne les raisons qui motivent l'élaboration de plans de sondage optimaux. Le tableau 2 compare les méthodes de conception optimale aux méthodes courantes. Le tableau 3 décrit dans les grandes lignes les techniques de conception optimale des systèmes multi-agents et l'algorithme bayésien d'optimisation.

2. SYSTÈMES MULTI-AGENTS

Un système multi-agents (SMA) est un ensemble d'agents logiciels intelligents et autonomes. Un * agent + est une entité logicielle qui possède des connaissances et des objectifs et qui agit sur son environnement pour réaliser ces objectifs. La présente section donne un aperçu des systèmes multi-agents (Sen, 2000; Wooldridge, 2002). Typiquement, l'architecture des agents inclut des récepteurs pour capter et entrer l'information sur l'environnement, des effecteurs pour agir sur l'environnement, un modèle d'autres agents dans le système et des modules de coordination, d'apprentissage, de planification et d'inférence. Les systèmes multi-agents sont classés en fonction de la relation entre leurs agents. Les agents coopératifs poursuivent conjointement des objectifs communs. Chaque objectif est décomposé en sous-objectifs qui sont affectés à des agents individuels d'après leurs capacités et leur accès aux ressources. Les agents coopératifs sont organisés selon une hiérarchie de l'autorité et utilisent des protocoles pour partager l'information. Les agents non coopératifs sont soit adversaires soit simplement indifférents les uns aux autres. Les systèmes multi-agents décrits dans le présent article comprennent des agents coopératifs.

Le domaine des systèmes multi-agents est lié à d'autres domaines de l'intelligence artificielle, comme la représentation des connaissances, la recherche et l'apprentissage automatique. En outre, il s'appuie sur les techniques et les résultats de la sociologie, de la science cognitive, de l'économie et de la science de la gestion. Les systèmes multi-agents offrent quatre grands avantages. Premièrement, ils sont utiles pour la modélisation de domaines problématiques qui sont intrinsèquement distribués, sans grande centralisation, comme la démographie. Deuxièmement, les systèmes

multi-agents sont modulaires et comportent la conception d'agents fonctionnels. Cette propriété est utile lors de l'étude de phénomènes démographiques complexes. Troisièmement, les systèmes multi-agents sont adaptatifs, ce qui les rend plus robustes que les systèmes centralisés pour la prévision des valeurs démographiques pour les populations dynamiques. Quatrièmement, les systèmes multi-agents modélisent les aspects sociaux des comportements intelligents, comme les interactions démographiques.

Le Programme de plan optimal comprend deux systèmes multi-agents. Le premier optimise les inputs de remaniement tel que décrit à la section 3. Le deuxième optimise les échantillons de ménages tel que décrit à la section 4.

3. INPUTS DE REMANIEMENT OPTIMAUX

Le premier système multi-agents du Programme de plan optimal saisit des données provenant d'enquêtes, de recensements et de dossiers administratifs et simule un * état + démographique pour chaque ménage figurant dans le fichier maître d'adresses courant. Dans le présent contexte, le terme * état + s'entend d'un ensemble de valeurs pour des variables démographiques (éventuellement nombreuses) associées à chaque ménage. Les états simulés sont des valeurs réelles de données, si ces valeurs sont connues. Sinon, ils reflètent une * meilleure estimation + de ces valeurs.

Un agent logiciel autonome représente chaque ménage et chaque agent ménage est associé à un noeud particulier dans un réseau bayésien. Un réseau bayésien est un modèle graphique formé de noeuds liés entre eux par des arcs à chacun desquels est associée une probabilité conditionnelle. Donc, chaque noeud du réseau est associé à un agent ménage dans un état démographique particulier. Chaque arc du réseau est associé à une interaction entre ménages, codée dans un message agent, qui est transmis avec une probabilité particulière.

Une densité de probabilité est estimée pour chaque agent ménage (dans chaque répétition du plan de sondage) qui reflète les probabilités que le ménage occupe d'autres états démographiques. La densité de probabilité est estimée d'après les données chronologiques connues au sujet du ménage en question et d'après les probabilités de changements démographiques et d'interactions pour des ménages similaires. La génération d'un nombre aléatoire pour chaque agent ménage (dans chaque répétition) et la comparaison de ce nombre aléatoire aux probabilités des fonctions de densité cumulatives permettent de simuler un état démographique courant pour chaque ménage.

Les valeurs des variables démographiques dans chaque état (comme le nombre de personnes dans le ménage) peuvent être regroupées en un niveau d'agrégation géographique plus élevé. En particulier, les valeurs agrégées peuvent être calées durant la simulation sur des estimations produites d'après des enquêtes démographiques récentes. Ce calage est * optimal + en ce sens que les priorités des agents ménages reflètent la qualité des données sur lesquelles se fondent leur densité de probabilité et que les rajustements de calage visent les agents à faible priorité pour commencer. Les valeurs simulées finales des ménages sont totalisées comme il l'est requis (par exemple, par UPE) pour appuyer la sélection d'échantillons démographiques optimaux à la section 4.

4. ÉCHANTILLONS OPTIMAUX DE MÉNAGES

Le deuxième système multi-agents du Programme de plan optimal sélectionne des échantillons annuels de ménage pour toutes les enquêtes démographiques. Un agent logiciel autonome représente chaque domaine d'estimation pour chaque enquête. D'autres agents représentent les ménages dans chaque domaine d'estimation. Le système entre les données simulées pour tous les ménages, ainsi que les nombres d'UPE et de ménages dans l'échantillon pour chaque domaine d'estimation. Puis, un algorithme de sélection d'échantillon sélectionne un échantillon optimal de ménages pour chaque enquête et chaque domaine d'estimation. L'algorithme bayésien d'optimisation est intégré dans l'algorithme de sélection d'échantillon (Pelikan, 1999).

L'ABO est un algorithme de recherche génétique qui sélectionne des ensembles optimaux de ménages dans les ensembles optimaux d'UPE. À chaque itération dans une séquence, il construit un réseau bayésien à partir des données simulées pour l'ensemble courant d'unités d'échantillonnage. Grâce à une inférence probabiliste au sujet du réseau, il identifie les unités non échantillonnées qu'il faut échanger avec les unités échantillonnées pour améliorer l'échantillon

global. Par pondération et estimation de la variance des données d'échantillon à chaque itération, il calcule la différence quadratique moyenne entre les estimations d'échantillon simulées et les valeurs de population. Il calcule aussi la différence quadratique moyenne entre le réseau bayésien d'échantillon et le réseau de population. Puis, il minimise un critère de ces différences quadratiques moyennes.

Pour chaque enquête, l'algorithme de sélection d'échantillon complet comprend les étapes qui suivent.

1) Sélectionner un domaine d'estimation.

2) Entrer les estimations démographiques courantes pour ce domaine (EST1) et simuler des données pour tous les ménages qu'il comprend.

3) Construire un réseau bayésien pour le domaine (RB1) d'après les données simulées sur les ménages.

4) Sélectionner au hasard un ensemble initial d'UPE échantillonnés.

5) Construire le réseau bayésien d'échantillon (RB2).

6) Mettre à jour les UPE échantillonnées.

- Remplacer des UPE échantillonnées par des UPE non échantillonnées après inférence sur RB2.

7) Calculer les estimations d'échantillon (EST2) d'après les données sur les ménages pour les UPE échantillonnées.

8) Mettre à jour le réseau bayésien d'échantillon (RB2).

9) Comparer EST2 à EST1 et RB2 à RB1 :

- si les valeurs ne sont pas suffisamment proches, retourner à l'étape 6);
- si les valeurs sont suffisamment proches, continuer.

10) Sélectionner au hasard un échantillon initial de ménages pour toutes les UPE échantillonnées.

11) Construire un réseau bayésien d'échantillon (RB3) pour chaque UPE échantillonnée.

12) Mettre à jour les ménages échantillonnés pour toutes les UPE échantillonnées.

- Remplacer certains ménages échantillonnés par des ménages non échantillonnés après inférence sur chaque RB3.

13) Calculer des estimations d'échantillon (EST3) d'après les données sur les ménages échantillonnés pour chaque UPE échantillonnée.

14) Mettre à jour le réseau bayésien d'échantillon (RB3) pour chaque UPE échantillonné.

15) Comparer EST3 à EST2 et RB3 à RB2 :

- si les valeurs ne sont pas suffisamment proches, retourner à l'étape 12);
- si les valeurs sont suffisamment proches pour le domaine d'estimation final, arrêter;
- si les valeurs sont suffisamment proches pour le domaine d'estimation non final, retourner à l'étape 1).

Le tirage d'échantillons annuels pour les enquêtes démographiques ne devrait être mis en oeuvre que lorsqu'il est

rentable de le faire. Il pourrait être nécessaire de garder les UPE échantillonnées de l'année précédente et de ne sélectionner que des échantillons de ménages durant l'année courante.

5. ORIENTATION DES FUTURS TRAVAUX DE RECHERCHE

Une communication intitulée * Mathematical Analysis and Simulation of Multi-Agent Systems for Optimal Survey Design + sera présentée au colloque de recherche de l'automne 2003 du Federal Committee on Statistical Methodology à Washington, DC.

Un article intitulé * Implementing Optimal Sampling Programs in Intelligent Organizations + décrira la modélisation du U.S. Census Bureau comme un agent complexe, composite, qui met en oeuvre le Programme de plan optimal et génère des échantillons optimaux pour toutes les enquêtes démographiques. Le domaine de l'intelligence organisationnelle intègre des agents artificiels et humains et considère des questions telles que la mémoire organisationnelle et les capacités d'apprentissage.

Pour appuyer les travaux de recherche susmentionnés, on spécifiera deux petits systèmes multi-agents et on produira des prototypes de ces systèmes pour en analyser les algorithmes de contrôle, les bases de données Internet, les architectures d'agent et les protocoles de communication. Ces systèmes prototype faciliteront aussi l'évaluation des avantages et des coûts du Programme de plan optimal.

Tableau 1
Motivations de l'Optimal Survey Design

Processus démographiques

- Déterminent les caractéristiques personnelles et les interactions.
- Degré croissant de complexité et d'intégration.

Systèmes démographiques

- Surveillent et contrôlent les processus démographiques.
- Degré croissant de complexité, d'intégration, de coût, d'automation et d'intelligence.
- **Objectif :** Construire un système optimal d'échantillonnage.

Données démographiques

- Produites par les systèmes démographiques.
- Degré croissant de finesse, d'importance et de confidentialité.
- **Objectif :** Optimiser les inputs du processus de conception d'enquête.

Estimations démographiques

- Calculées d'après les données démographiques.
- Degré croissant de finesse (petites régions et petits domaines), d'intégration et d'exactitude.
- **Objectif :** Sélectionner des échantillons qui minimisent l'EQM des estimations fondées sur les données d'enquête.

Applications démographiques

- Intégrer des estimations démographiques exactes.
- Degré croissant d'importance pour la survie de l'homme.

Tableau 2
Méthodes d'élaboration de plans de sondage optimaux

Élément	Proposé	Courant
1. Calendrier de remaniement	Annuel	Décennal
2. Domaine d'estimation	État	Nation, région, État
3. Estimations, variables de conception, tailles d'échantillons	Choisies par le client	Choisies par le client
4. Base de sondage	Fichier maître d'adresses	Unités, régions, logements collectifs, constructions neuves
5. Degrés d'échantillonnage	UPE, ménage	UPE, ménage
6. Inputs d'échantillonnage	Données de recensement, données d'enquête et estimations, dossiers administratifs	Données de recensement
7. Traitement des inputs	Prévision, simulation, totalisation	Totalisation
8. Algorithme d'échantillonnage	Algorithme bayésien d'optimisation (ABO)	Stratification et sélection
9. Critère de recherche	Minimiser les EQM sur les estimations simulées	Minimiser les C.V. sur les variables de stratification
10. Espace de recherche	Tous les réseaux bayésiens	Toutes les stratifications
11. Méthode de recherche	Algorithme génétique	Escalade
12. Intervalles d'échantillonnage	Sans objet	Utilisé pour échantillonner les ménages
13. Poids d'échantillonnage	Reflète les probabilités de sélectionner des échantillons optimaux	Reflète les mesures de taille et d'intervalles d'échantillonnage
14. Estimation de la variance	Utilisée pour calculer les EQM sur des estimations simulées	Utilisée pour calculer les C.V. sur les variables de stratification

Tableau 3
Techniques pour l'élaboration de plans de sondage optimaux

Systèmes multi-agents

- * Intelligence artificielle distribuée +.
- Les agents logiciels autonomes ont des objectifs communs, possèdent des connaissances, prennent des décisions et communiquent entre eux.
- Les agents observent et modifient leur environnement extérieur.
- Utilisés pour modéliser des domaines problématiques distribués qui manquent de contrôle centralisé (comme la démographie), des phénomènes complexes et dynamiques, et des comportements sociaux intelligents.

Algorithme bayésien d'optimisation (ABO)

- Algorithme génétique élaboré à l'Université de l'Illinois à la fin des années 1990.
- Construit une série de réseaux bayésiens et utilise l'inférence sur ces réseaux pour améliorer une * population + d'entités.
- Utilisé pour modéliser les relations entre variables démographiques pour toutes les enquêtes.
- Utilisé pour répartir les unités primaires d'échantillonnage et les ménages en sous-ensembles échantillonnés et non échantillonnés optimaux.

RÉFÉRENCES

Pelikan, M., D.E. Goldberg, et E. Cantu-Paz (1999), "BOA: The Bayesian Optimization Algorithm," Illinois Genetic Algorithms Laboratory, Urbana, IL: University of Illinois at Urbana-Champaign

Pelikan, M., D.E. Goldberg, et F. Lobo (1999), "A Survey of Optimization by Building and Using Probabilistic Models," Illinois Genetic Algorithms Laboratory Report No. 99018, Urbana, IL: University of Illinois at Urbana-Champaign

Sen, S. (2000), "Multiagent Systems: Milestones and New Horizons", Tulsa, OK: University of Tulsa.

U.S. Census Bureau (2002), "U.S. Census Bureau Strategic Plan FY 2004-2008", unpublished report, Washington DC:

Wooldridge, M (2002), *An Introduction to MultiAgent Systems*, West Sussex, England: Wiley.