No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

APPROCHES ASSISTÉES PAR MODÈLE DE L'ÉCHANTILLONNAGE DANS LES ENQUÊTES COMPLEXES À PARTIR DE POPULATIONS FINIES AU MOYEN DE RÉSEAUX BAYÉSIENS : UN OUTIL D'INTÉGRATION DE DIVERSES SOURCES

Marco Ballin, Mauro Scanu et Paola Vicard[1]

RÉSUMÉ

Une classe d'estimateurs basés sur la structure de dépendance entre une variable multidimensionnelle d'intérêt et le plan d'enquête est définie, soit la structure de dépendance décrite par les réseaux bayésiens. Il s'agit d'estimateurs par quotient formant une sous-classe identifiée par une structure de dépendance particulière. Nous démontrons, à l'aide d'une simulation de Monte Carlo, comment l'emploi d'un estimateur correspondant à la structure de population est plus efficace que les autres. Nous soulignerons également la manière dont cette classe s'adapte au problème de l'intégration de l'information provenant de deux enquêtes grâce au système d'actualisation des probabilités des réseaux bayésiens.

MOTS CLÉS : modèles graphiques; actualisation des probabilités.

1. INTRODUCTION

L'ensemble de l'information recueillie lors d'une ou de plusieurs enquêtes représente sans aucun doute un système complexe. Pour cette raison, il est non seulement important d'organiser ces enquêtes en un système approprié (par exemple, pour permettre la coordination des échantillons ou la cohérence des définitions), mais aussi de fournir les outils adéquats pour soutenir la représentation, l'estimation et la mise à jour.

Dans le présent exposé, nous démontrerons comment les réseaux bayésiens (RB) peuvent favoriser la réalisation de tels objectifs en tenant compte du plan d'enquête. Les RB ont été appliqués avec succès dans plusieurs contextes tels que : l'intelligence artificielle, les statistiques juridiques, la génétique, le diagnostic de pannes d'ordinateur et d'autres domaines où l'on doit traiter couramment un grand nombre de variables liées par une structure de dépendance complexe (Neapolitan, 2004). Les RB ont aussi déjà servi à traiter des statistiques officielles, par exemple dans la description de certains résultats d'enquête (Getoor *et coll.*, 2001) et dans l'imputation de valeurs manquantes (Thibaudeau et Winkler, 2002, Di Zio *et coll*., 2005 et dans les renvois qui s'y trouvent).

Bien que, dans les analyses précédentes, l'hypothèse i.i.d. était naturelle et tout à fait justifiée, il convient de souligner que cette hypothèse ne tient plus dans un contexte de plan d'enquête complexe. Notre propos consiste à représenter explicitement le plan de sondage dans le RB et de mettre en valeur sa relation statistique avec les variables d'intérêt. Cette démarche définit une classe d'estimateurs de la distribution conjointe de ces variables qui inclut les estimateurs par quotient. D'ailleurs, les propriétés de propagation de l'information des RB (Cowell *et coll*., 1999) seront utilisées afin de mettre à jour les estimations d'un sondage lorsqu'un choc informationnel découlant d'une autre enquête aura eu lieu. En ce sens, les RB permettent de visualiser et de comprendre l'interaction de l'information provenant de différentes enquêtes.

[1] Marco Ballin, Istat, via A. Ravà 150, 00100 Rome, Italie; Mauro Scanu, Istat, via C. Balbo 16, 00184 Rome, Italie (scanu@istat.it); Paola Vicard, Università Roma Tre, via Ostiense 139, 00154 Rome, Italie (vicard@uniroma3.it).

2. LES ESTIMATEURS BASÉS SUR DES RÉSEAUX BAYÉSIENS

Soit $\mathcal{P}$, une population finie de taille n, et soit $X_1,\ldots,X_k$, k variables d'intérêt. Par souci de simplicité, considérons ces variables comme étant catégoriques, avec une distribution statistique en $\mathcal{P}$

$$F(x_1,...,x_k) = \sum_{i=1}^{N} \frac{I_{x_1\ldots x_k}(x_{i1},...,x_{ik})}{N}, \tag{1}$$

où $I(.)$ est la fonction indicatrice. L'équation (1) est le paramètre d'intérêt. Soit $\mathcal{S}$ l'échantillon tiré de $\mathcal{P}$ selon un plan d'enquête complexe défini par les facteurs de pondération d'enquête w_i, $i \in \mathcal{S}$. Un estimateur naturel et intuitif de la distribution (1) est :

$$\hat{F}(x_1,...,x_k) = \sum_{i\in S} I_{x_1\ldots x_k}(x_{i1},...,x_{ik}) \frac{w_i}{\sum_{i\in S} w_i}. \tag{2}$$

L'estimateur (2) est un estimateur par quotient.

Soit S une variable additionnelle (catégorique) avec autant de catégories qu'il y a de différentes valeurs de pondération, disons $w_{(h)}$, $h=1,\ldots H$, et une distribution de probabilité marginale

$$F(S=h) = \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}}, \quad h=1,\ldots,H, \tag{3}$$

où n_h est la taille du sous-ensemble d'unités s_h de $\mathcal{S}$ avec un facteur de pondération égal $w_{(h)}$, $h=1,\ldots H$. En utilisant S dans (2), il est possible de masquer le rôle des facteurs de pondération qui conditionnent S, par exemple :

$$\hat{F}(x_j \mid S=h) = \sum_{i\in s_h} \frac{I_{x_j}(x_{ij})}{n_h}, \quad \hat{F}(x_j \mid X_l = x_l, S=h) = \sum_{i\in s_h} \frac{I_{x_j x_l}(x_{ij}, x_{il})}{\sum_{i\in s_h} I_{x_l}(x_{il})} \tag{4}$$

pour tout $j = 1,\ldots K$, et $l \neq j$, et pour mettre en valeur le rôle de S dans l'estimateur (2) en utilisant la règle d'enchaînement (Cowell *et coll*. 1999). Cela est rendu possible en réécrivant l'estimateur (2) à l'aide d'une factorisation récursive faisant appel aux facteurs en (3) et en (4), c'est-à-dire

$$\begin{aligned}\hat{F}(x_1,...,x_k) &= \sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i\in s_h} \frac{I_{x_1}(x_{i1})}{n_h} \sum_{i\in s_h} \frac{I_{x_1x_2}(x_{i1},x_{i2})}{\sum_{i\in s_h} I_{x_1}(x_{i1})} \cdots \sum_{i\in s_h} \frac{I_{x_1\ldots x_K}(x_{i1},...,x_{iK})}{\sum_{i\in s_h} I_{x_1\ldots x_{K-1}}(x_{i1},...,x_{iK-1})} \\ &= \sum_{h=1}^{H} F(S=h)\hat{F}(x_1 \mid S)\hat{F}(x_2 \mid X_1, S)\cdots\hat{F}(x_K \mid X_1,\cdots,X_{K-1},S). \end{aligned} \tag{5}$$

La factorisation (5) montre que l'estimateur par quotient habituel (2) repose sur un modèle de dépendance particulière entre les variables $X_1,\ldots,X_K, S$. Le modèle de dépendance dont il est question correspond à la situation de dépendance complète entre ces variables, c'est-à-dire le modèle *saturé*. Cette situation de dépendance peut être représentée graphiquement comme un RB appelé *clique*. Par exemple, lorsqu'il y a trois variables d'intérêt *X, Y,* et *Z* en vertu du plan *S*, la clique correspond au réseau a) de la figure 1. Il est à noter que l'ordre des variables dans la factorisation (5) n'est pas unique et qu'il débouche sur des structures graphiques autres, mais équivalentes.

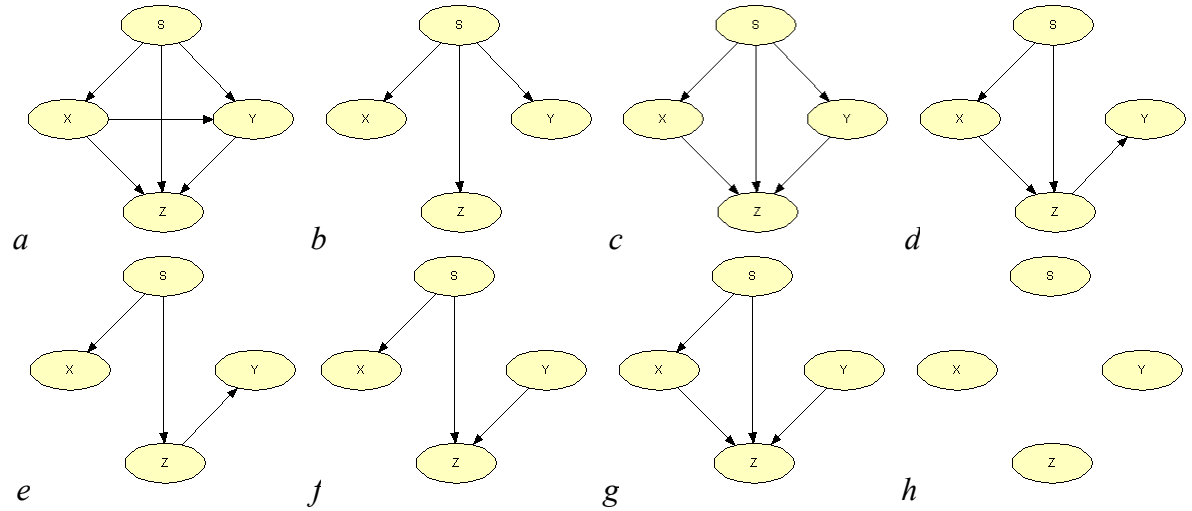

Figure 1 – Huit RB possibles représentant la structure de dépendance entre trois variables d'intérêt *X, Y, Z* et le plan d'enquête *S*.

Il peut arriver que la clique soit un modèle surparamétré lorsque certaines variables sont marginalement ou conditionnellement indépendantes. Par exemple, les réseaux b)-h) dans la figure 1 présentent quelques RB simplifiés. La structure de dépendance d'un RB laisse entrevoir un estimateur basé sur la règle d'enchaînement, dont la forme générale est :

$$\hat{F}_{BN}(x_1,...,x_K) = \sum_{h=1}^{H} F(S) \prod_{j=1}^{K} \hat{F}(x_j \mid pa(X_j)) \tag{6}$$

Par exemple, l'estimateur tiré d'un RB du RB f) de la figure 1 :

$$\hat{F}_f = \left[\sum_{h=1}^{H} F(S) \hat{F}(x \mid S) \hat{F}(z \mid Y, S) \right] \hat{F}(y) = \left[\sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i \in s_h} \frac{I_x(x_i)}{n_h} \sum_{i \in s_h} \frac{I_{yz}(y_i, z_i)}{\sum_{i \in s_h} I_y(y_i)} \right] \sum_{i \in S} \frac{I_y(y_i)}{n}$$

Il convient de noter que $pa(X_j)$ n'inclut pas nécessairement S, comme dans les réseaux d)-h) de la figure 1. Dans un tel cas, l'estimateur de la distribution conditionnelle n'emploie pas de poids d'échantillonnage. En outre, la classe d'estimateurs (6) basés sur différentes structures en RB est finie pour un plan d'enquête établi S, et inclut toujours l'estimateur (2).

Bien que chaque facteur estimé en (6) ne soit pas biaisé, au chapitre du plan, lorsque la somme des poids d'échantillonnage dans chaque catégorie S est constante par rapport à la variabilité de l'échantillon (par exemple, dans le plan fondé sur l'échantillonnage stratifié), l'absence de biais n'est assurée que pour l'estimateur relié à la clique, c'est-à-dire l'estimateur (2) (la seule condition étant que la somme des poids d'échantillonnage soit égale à N). Les autres estimateurs peuvent être biaisés par rapport à la distribution statistique conjointe à plusieurs variables de $X_1,...X_K$ en $\mathcal{P}$. Néanmoins, il existe des preuves empiriques selon lesquelles l'exploitation de la structure de dépendance des variables d'intérêt et du plan d'enquête pourrait mener à de meilleurs estimateurs (section 3).

Une classe différente d'estimateurs basés sur des RB est abordée dans Ballin *et coll*. (2005). Elle consiste à produire des estimations de poids également pour les facteurs de l'équation (6) qui n'admettent pas S dans la conditionnelle $pa(X_j)$. La comparaison entre ces deux définitions sera présentée ailleurs.

3. LA SIMULATION DE MONTE CARLO

Une simulation de Monte Carlo a été menée dans le but d'évaluer la performance des estimateurs suggérés par les RB et les conséquences des éventuelles erreurs de spécification du modèle. Plus particulièrement, huit populations formées de 10 000 unités ont été générées selon les huit RB de la figure 1. De chaque population, on a tiré 500 échantillons de taille 1 000 conformément au plan fondé sur l'échantillon stratifié décrit dans le tableau 1 (il est à noter que la fraction de sondage n'est pas proportionnelle à la taille de la strate).

Code de la strate h	Taille de la strate N_h	$F(S = h)$	Taille d'échantillon n_h
$S = 1$	5 995	0,5995	100
$S = 2$	2 959	0,2959	200
$S = 3$	1046	0,1046	700

Tableau 1 : Tailles de strate et d'échantillon

Pour chaque population et pour chaque répétition de Monte Carlo, les estimations de la distribution conjointe découlant de l'estimateur suggéré par le RB correspondant ont été comparées avec la distribution conjointe réelle. La performance de l'estimateur a été mesurée par la moyenne des estimations de Monte Carlo des erreurs quadratiques moyennes (MSE) relatives de chaque élément de la distribution conjointe :

$$MSE\left(\hat{F}_{BN}\right) = \frac{1}{V}\sum_{v=1}^{V}\sum_{x,y,z}\frac{\left[F(x,y,z) - \hat{F}_{BN.v}(x,y,z)\right]^2}{F(x,y,z)},$$

où V représente le nombre de répétitions de Monte Carlo et $\hat{F}_{BN.v}(x,y,z)$ est l'estimation de $F(x,y,z)$ obtenue avec le v^e échantillon. Afin d'avoir une idée de la robustesse des estimations par rapport à une erreur de spécification de modèle, la distribution conjointe a été estimée à l'aide des estimateurs suggérés par des RB associés aux sept autres structures en RB. Les résultats sont présentés dans le tableau 2, où chaque rangée renvoie à une population et chaque colonne renvoie à un estimateur.

Pop	MSE($\hat{F}_a$)	MSE($\hat{F}_b$)	MSE($\hat{F}_c$)	MSE($\hat{F}_d$)	MSE($\hat{F}_e$)	MSE($\hat{F}_f$)	MSE($\hat{F}_g$)	MSE($\hat{F}_h$)
a	0,32	0,54	0,36	3,07	3,51	2,71	2,57	8,59
b	0,27	0,15	0,22	3,16	3,14	10,41	10,43	11,93
c	0,29	0,45	0,25	10,17	11,29	9,17	9,12	11,07
d	0,28	6,25	0,45	0,10	0,47	0,63	0,52	15,34
e	0,30	0,37	0,25	0,12	0,11	0,18	0,22	7,12
f	0,28	0,25	0,23	1,44	1,43	0,11	0,15	7,25
g	0,30	0,31	0,25	1,30	1,46	0,28	0,18	6,98
h	0,37	0,14	0,31	0,14	0,11	0,14	0,25	0,03

Tableau 2 – Les estimations de Monte Carlo des erreurs quadratiques moyennes relatives des estimateurs basés sur un RB suggérés par les RB de la figure 1.

Dans le tableau 2, l'estimateur est identifié par le même indice que la population correspondante (la première colonne renvoie à l'estimateur basé sur un RB suggéré par le RB a) de la figure 1; la deuxième, à l'estimateur basé sur un RB suggéré par le RB b) de la figure 1, et ainsi de suite). Cette simulation montre que les estimateurs suggérés par le RB correspondant fonctionnent toujours mieux que l'estimateur (2) (il suffit de comparer les valeurs contenues dans la diagonale avec celles de la première colonne). Même s'il est difficile de mesurer la « distance » entre le modèle de génération de données et le modèle découlant de l'estimateur par quotient habituel $\hat{F}_a$, la comparaison entre les valeurs de la diagonale et les valeurs de la première colonne suggère que le gain d'efficacité dépend d'une telle « distance ». Par exemple, dans le cas d'une population b) (caractérisée par seulement trois liens), le gain d'efficacité est plus élevé (0,15 comparé à 0,27) que dans le cas de la population c) (0,25 vs 0,29), qui diffère du modèle habituel seulement en ce qui concerne le lien entre X et Y.

Le résultat précédent est plutôt général. Pour la population h), les différents estimateurs semblent ordonnés en fonction du nombre de flèches supplémentaires. Le pire estimateur est en réalité $\hat{F}_a$, dont la structure est la plus distante. À partir des autres rangées, il est possible de constater que l'ajout de flèches semble avoir un effet moins prononcé que le retrait de flèches. En effet, une structure simplifiée est incapable de prendre en considération les dépendances réelles, tandis que, lorsque les données indiquent une indépendance, cette dernière est pratiquement préservée lors de l'ajout de (quelques) flèches.

Le tableau 3 présente les contributions des biais estimés et de la variance aux erreurs quadratiques moyennes relatives des estimateurs. Il est à noter que, dans le cas de l'estimateur habituel $\hat{F}_a$ (qui est non biaisé), la contribution des biais estimés à la MSE de Monte Carlo est négligeable. La contribution des biais est également négligeable le long de la diagonale. Ce résultat donne à entendre que les estimateurs qui intègrent une structure de dépendance connue sont approximativement non biaisés. À l'extérieur de la diagonale, la contribution des biais est habituellement plus élevée.

Pop	$Biais_a$	Var_a	$Biais_b$	Var_b	$Biais_c$	Var_c	$Biais_d$	Var_d	$Biais_e$	Var_e	$Biais_f$	Var_f	$Biais_g$	Var_g	$Biais_h$	Var_h
a	2,0	98,0	74,0	26,0	25,9	74,1	57,8	42,2	70,4	29,6	69,3	30,7	48,1	51,9	85,1	14,9
b	1,9	98,1	3,0	97,0	1,9	98,1	57,8	42,2	59,1	40,9	88,7	11,3	85,3	14,7	94,9	5,1
c	1,6	98,4	71,5	28,5	2,6	97,4	82,9	17,1	85,5	14,5	79,2	20,8	69,4	30,6	91,2	8,8
d	0,4	99,6	97,4	2,6	41,2	58,8	1,6	98,4	68,5	31,5	61,4	38,6	31,1	68,9	97,7	2,3
e	3,8	96,2	62,5	37,5	1,9	98,1	5,2	94,8	7,1	92,9	17,8	82,2	12,8	87,2	90,7	9,3
f	3,9	96,1	49,9	50,1	1,7	98,3	28,3	71,7	30,5	69,5	4,6	95,4	1,8	98,2	83,3	16,7
g	3,1	96,9	56,9	43,1	2,1	97,9	12,1	87,9	44,9	55,1	44,2	55,8	1,8	98,2	51,0	49,0
h	4,1	95,9	6,2	93,8	1,7	98,3	2,5	97,5	3,9	96,1	4,0	96,0	1,7	98,3	6,3	93,7

Tableau 3. Les estimations de Monte Carlo de la contribution des biais et de la variance à la MSE relative

Ces résultats valent aussi dans le cas des distributions marginales estimées.

4. LES ENJEUX LIÉS À L'INTÉGRATION

Dans un système complexe et intégré formé de deux enquêtes ou plus, il est important de disposer des outils permettant d'actualiser les estimations calculées dans une enquête lorsque les résultats d'autres enquêtes ou de nouvelles archives deviennent disponibles. Dans un tel cas, nous nous trouvons en présence d'un *choc informationnel* et nous devons diffuser l'information supplémentaire parmi les résultats d'enquêtes précédentes afin d'assurer la cohérence entre les enquêtes (cohérence externe). Ce problème peut être abordé dans les statistiques officielles au moyen d'estimateurs de calibration qui permettent d'estimer les paramètres d'intérêt sous des contraintes linéaires. Nous suggérons ici l'emploi de RB, qui représentent un outil naturel de propagation de l'information (Cowell *et coll.*, 1999).

Les RB peuvent être actualisés lorsqu'un choc informationnel se produit. Par nouvelle information, nous entendons ici une nouvelle distribution statistique pour une ou plusieurs variables d'intérêt obtenues à partir d'archives ou d'une nouvelle enquête. La relation au sein des variables d'un RB (c'est-à-dire les flèches qui les relient entre elles) est le trajet par lequel se propage ce type d'information. Par souci de simplicité, considérons un RB composé de deux nœuds, X_1 et X_2, joints par la flèche $X_1 \rightarrow X_2$. Par conséquent, les distributions de probabilité $F(X_1 = x_1)$, $F(X_2 = x_2 \mid X_1 = x_1)$ peuvent être associées au RB. Changeons la distribution de probabilité marginale de X_2 en $F^*(X_2 = x_2)$. Afin que le réseau puisse absorber la nouvelle distribution $F^*(X_2=x_2)$ sans affecter la relation entre les variables, c'est-à-dire la distribution conditionnelle de X_2 sachant X_1, il est nécessaire d'actualiser la distribution marginale de X_1 :

$$F^*(X_1 = x_1) = \sum_{x_2} F(X_1 = x_1 \mid X_2 = x_2) F^*(X_2 = x_2) = \sum_{x_2} F(X_1 = x_1, X_2 = x_2) \frac{F^*(X_2 = x_2)}{F(X_2 = x_2)} .$$

Autrement dit, l'ancienne distribution conjointe $F(X_1 = x_1, X_2 = x_2)$ est actualisée à l'aide du ratio $F^*(X_2 = x_2)/F(X_2 = x_2)$ (appelé *ratio d'actualisation*) faisant la conversion entre les anciennes et les nouvelles distributions marginales de X_2.

Ici, le mécanisme de propagation de l'information a été illustré au moyen d'un exemple à deux variables. Les résultats ci-dessus peuvent aussi servir aux cas à plusieurs variables, lorsque la nouvelle information vient actualiser une distribution à plusieurs variables. À cette fin, on a défini différents algorithmes efficaces, basés sur le concept d'*arbres de jonction* (voir Jensen, 1996).

Le processus de propagation que nous venons d'illustrer peut s'appliquer à la procédure de stratification a posteriori traditionnelle. Reprenons l'estimateur par quotient (2). Par souci de simplicité, prenons un choc informationnel qui toucherait seulement la variable X_1, dont la distribution statistique actualisée est N_{1q}^*, $q=1,\ldots,Q$. La propagation de ce choc à travers le RB revient à stratifier X_1 a posteriori. De façon plus précise, les anciens poids d'échantillonnage w_i sont changés en :

$$w_i^* = w_i \frac{N_{1q}^*}{\sum_i w_i I_{x_{1i}}(q)} = w_i \frac{N_{1q}^*}{\hat{N}_{1q}}, \quad i : I_{x_{1i}}(q) = 1, \quad q = 1,\ldots, Q$$

où $\hat{N}_{1q}$ sont les estimations de fréquence calculées selon les anciens poids de l'enquête. Le choc produit un changement dans le nœud S, qui se modifie en un nouveau nœud S^* de telle sorte que : (i) S^* catégories sont obtenues à partir du produit cartésien des catégories S et X_1, c'est-à-dire (h,q), $h=1,\ldots,H$, $q=1,\ldots,Q$; (ii) les unités provenant de la même catégorie (h,q) ont le même poids, $w_{(h,q)}^*$. Encore une fois, le théorème de Bayes permet de calculer la distribution de probabilité de S sachant X_1:

$$F(S = h \mid X_1 = q) = \frac{F(S = h)F(X_1 = q \mid S = q)}{\sum_{h=1}^{H} F(S = h)F(X_1 = q \mid S = q)}, \quad q = 1, \ldots, Q;\ h = 1,\ldots, H.$$

La stratification a posteriori ne modifie pas la distribution précédente, c'est-à-dire la relation statistique entre S et X_1 selon le plan d'enquête initial. En outre, la stratification a posteriori de la nouvelle distribution de X_1, $N_1^*(q)$, $q=1,\ldots,Q$, ou, mieux encore, de la distribution statistique relative $F_1^*(q) = N_{1q}^*/N, q = 1,\ldots,Q$, revient à considérer la nouvelle distribution conjointe suivante :

$$F(S^* = (h,q)) = F(S = h, X_1 = q) = F(S = h \mid X_1 = q)F_1^*(q) = \frac{F(S = h)F(X_1 = q \mid S = h)}{\sum_{h=1}^{H} F(S = h)F(X_1 = q \mid S = h)} F_1^*(q) =$$

$$= \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \frac{n_{hq}}{n_h} \frac{F_1^*(q)}{\hat{F}_1(q)}, \quad q = 1,\ldots,Q; h = 1,\ldots,H.$$

S^* est caractérisé par des poids constants, $w_{(h,q)}^*$, pour toutes les unités d'une même catégorie (h,q). Soit n_{hq} la taille de cette catégorie. Par conséquent,

$$w_{(h,q)}^* = \frac{\sum_{h=1}^{H} n_h w_{(h)}}{n_{hq}} F(S^* = (h,q)) = w_{(h)} \frac{F_1^*(q)}{\hat{F}_1(q)} = w_{(h)} \frac{N_1^*(q)}{\hat{N}_1(q)} \frac{\hat{N}}{N}$$

avec $\hat{N} = \sum_{i=1}^{N} w_i$.

La procédure de stratification a posteriori décrite ci-dessus peut être directement appliquée lorsque le choc informationnel concerne la distribution conjointe de deux variables reliées ou plus. Cette procédure est simple et directe dans le cas de l'estimateur (2). Pour toute autre structure en RB, la préservation du choc sur la distribution conjointe d'un sous-ensemble de variables n'est assurée que lorsque ce sous-ensemble est une clique (il faudrait considérer que cette contrainte vaut pour toute stratification a posteriori).

Dans cette section, nous avons montré comment mettre à jour les estimations produites à partir d'une enquête à la lumière de résultats provenant d'une nouvelle enquête ou d'une archive. Cette propagation d'information peut être vue comme faisant partie d'un système plus général et élargi composé de plus de deux enquêtes et d'archives organisées en fonction de divers critères, dont leur fiabilité et l'ordre chronologique de leur réalisation. L'utilisation de RB permet une actualisation efficace des estimations grâce à la propagation basée sur un arbre de jonction. En ce sens, il est nécessaire d'organiser le système d'enquêtes de façon à ce que la nouvelle information disponible à propos d'une variable (X_1, par exemple) puisse rejoindre, en suivant les flèches, toutes les variables qu'on sait statistiquement dépendantes de X_1. En termes techniques, les enquêtes doivent être liées de manière à ce que la « propriété de jonction » (Lauritzen, 1996) soit conservée.

Afin de d'élaborer davantage le sujet, nous croyons qu'un réseau bayésien orienté sur l'objet (« RBOO », soit le RBOO de Koller et Pfeffer, 1997) est un outil valide pour représenter, gérer et utiliser un système d'enquêtes. Le RBOO représente une innovation récente de la technologie des RB. Elle permet la définition hiérarchique et la construction d'un BN, en utilisant de simples blocs de construction modulaire. On peut aisément y incorporer une complexité supplémentaire en ajoutant de nouveaux modules ou en raffinant ceux qui existent déjà. Dans notre cas, le RBOO serait composé d'autant de modules (instances) qu'il y aurait d'enquêtes. Ces modules seraient reliés par un système de nœuds extrants / intrants – qui représentent, dans ce cas, les relations identitaires entre les variables observées dans différentes enquêtes – à l'origine de l'actualisation. Autrement dit, le nœud extrant serait la nouvelle estimation de la variable X, par exemple, qui devrait être propagée aux enquêtes précédentes où X avait également été observée. L'information peut alors être étendue aux autres variables au moyen de la procédure illustrée ci-dessus. Il est intéressant de constater qu'avec les RBOO, la macrocohérence serait explicitement représentée et l'utilisateur final n'aurait pas à interpréter une structure très complexe. Cependant, le tableau complet de chaque enquête individuelle n'est pas perdu puisqu'il est possible, au besoin, de se pencher sur n'importe quelle d'entre elles. Nous croyons que cette voie est prometteuse (les RBOO sont élaborés à l'aide du logiciel Hugin version 6, http://www.hugin.com), bien qu'elle doive encore être approfondie. Il faudra donc procéder à d'autres recherches.

RÉFÉRENCES

Ballin, M., Scanu, M., et Vicard, P. (2005), "Bayesian networks and complex survey sampling from finite populations". Proceedings of the 2005 FCSM Symposium, Arlington (Virginie), 14-16 novembre, 2005.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., et Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, Heidelberg: Springer.

Di Zi, M., Sacco, G., Scanu, M., et Vicard, P. (2005), "Multivariate techniques for imputation based on Bayesian networks". *Neural Network World*, 2005/4, pp. 303-309.

Getoor L., Taskar B., et Koller D. (2001), "Selectivity estimation using probabilistic models". *ACM-SIGMOD*, Santa Barbara, California, États-Unis.

Jensen, F. V. (1996), *Introduction to Bayesian Networks*. Springer.

Koller, D. et Pfeffer, A. (1997), "Object-oriented Bayesian networks". *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, (ed. D. Geiger and P. Shenoy), Morgan Kaufmann Publishers, San Francisco. pp. 302-313.

Lauritzen, S. L. (1996), *Graphical Models*. Oxford University Press.

Neapolitan, R. E. (2004), *Learning Bayesian Networks*, Upper Saddle River (NJ): Prentice Hall.

Thibaudeau Y. et Winkler W. E. (2002), “Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints”. Technical report RRS2002/9, Washington D.C., É.-U.: U.S. Bureau of the Census.