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

EXAMEN DE L'ÉLABORATION ET DE LA MISE À L'ESSAI D'UNE MÉTHODOLOGIE DE MISE À JOUR DES INDICATEURS DU RECENSEMENT

Mohammed Yar, Neil Higgins, Philip Clarke et Patrick Heady[1]

RÉSUMÉ

Ce document porte sur l'application de modèles logistiques hiérarchisés à la mise à jour de données régionales de recensement, ainsi que sur les façons de combiner deux sources d'information, à savoir les données du recensement même et celles d'une enquête à reprises. Les questions étudiées sont celles de la caractérisation des tendances locales et de la prédominance (ou de la non-prédominance) de la tendance nationale dans les tendances locales. On y propose aussi une procédure de production d'estimations postcensitaires. Pour illustrer son application, nous citons deux exemples où les ensembles de données d'une enquête à reprises sont indépendants dans le temps. Par ces exemples, nous constatons que la tendance nationale peut être bien représentée par une série de valeurs annuelles à l'origine et que rien n'indique qu'il existe des tendances locales linéaires distinctes. Il y a cependant des indications selon lesquelles la modélisation des résidus par autorégression (AR) améliorerait les estimations.

1. INTRODUCTION

Dans une planification et une répartition efficientes des ressources, une exigence fondamentale, surtout à l'échelon local, est que l'on dispose de données régionales à jour. Au Royaume-Uni comme dans bien d'autres pays, les données censitaires servent largement à l'affectation des ressources financières, mais on y recense la population tous les dix ans, et le dernier recensement a eu lieu en 2001. À mesure que l'on s'éloigne de la date de recensement, les données censitaires perdent de leur actualité, voire de leur utilité, et l'incertitude s'accroît d'autant à mesure que l'on s'éloigne de la date du dernier recensement. Cette constatation vaut particulièrement pour les régions où l'évolution socio-économique est rapide. Cela risque de rendre peu efficiente la répartition des ressources d'après les données de recensement et de soulever des inquiétudes sur le caractère équitable de la distribution des ressources dans les années intercensitaires. Voilà pourquoi les spécialistes de la statistique officielle au Royaume-Uni cherchent à évaluer d'autres méthodes de production d'estimations des variables du recensement pour ces années. C'est ce que l'on appelle les estimations postcensitaires ou encore les mises à jour des données censitaires.

Dans le contexte propre au Royaume-Uni, notre but est d'établir des estimations postcensitaires pour un certain nombre de variables du recensement de 1991 dans la période intercensitaire (année 2001 comprise) par la combinaison de données transversales (celles du recensement de 1991) et de données d'enquête à reprises (enquête sur la population active de 1993 à 2001). Ces cinq dernières années environ, l'Office for National Statistics du Royaume-Uni a mis au point une méthodologie de production d'estimations régionales (ce que l'on appelle la méthodologie SAEP). Avec cette méthodologie, une limitation est l'exploitation d'ensembles de données pour un seul point temporel. Le problème de mise à jour de données du recensement exige en revanche que nous utilisions des données d'enquête à reprises. Il se pose donc naturellement la question de savoir si on doit traiter chaque période séparément ou essayer de tirer de la force de la combinaison des années. Dans cette étude, nous avons donc voulu étendre l'application de la méthodologie SAEP à des ensembles de données d'enquête à reprises. En représentant les tendances locales comme des variations aléatoires autour de la tendance nationale, nous élaborons un modèle logistique mixte hiérarchisé décrivant l'évolution temporelle des caractéristiques du recensement. En prenant les données régionales de recensement et les estimations des tendances locales, ce modèle produit des estimations postcensitaires.

[1] Office for National Statistics, Royaume-Uni.

Voici comment se présente le reste de notre exposé : à la section 2, nous présentons l'extension aux opérations de mise à jour du recensement de 1991 de l'application de la méthodologie du Small Area Estimation Programme (SAEP) de l'ONS; à la section 3, nous vérifions la valeur de cette méthodologie en l'appliquant à une simulation avec les données du registre à 100 % de la population finlandaise pour les années 1987-1996, dont nous connaissons les valeurs réelles; à la section 4, nous indiquons les résultats d'une application SAEP aux données du recensement de 1991 au Royaume-Uni et ceux d'une variable type du recensement, celle de la proportion de ménages d'une seule personne; à la section 5 enfin, nous décrivons les travaux en cours et mettons en évidence les secteurs où s'impose un complément de recherche.

2. ÉLABORATION DE LA MÉTHODOLOGIE DE PRODUCTION D'ESTIMATIONS À JOUR DU RECENSEMENT

Dans la démarche SAEP, la pièce maîtresse est l'élaboration d'un modèle où on établit le lien entre une variable d'enquête (données d'enquête) d'intérêt et des variables auxiliaires (sources de données administratives et recensements) et utilisons les éléments de modélisation pour produire des estimations régionales. Dans le cas de variables binaires, ce sont les probabilités de présence d'une caractéristique chez une personne ou dans un ménage que nous modélisons :

$$\begin{aligned} & y_{ij} \sim B(1,\pi_j) \\ & y_{ij} = \pi_{ij} + e_{ij} \\ & \text{logit}(\pi_j) = \alpha + \beta\overline{X}_j + u_j \end{aligned} \qquad (1)$$

où u_j est une variable aléatoire à moyenne 0 et à variance σ_u^2. La barre ($\bar{}$) qui surmonte la lettre désigne les covariables au niveau des régions, c'est-à-dire les moyennes ou les proportions disponibles pour toutes les régions j qui sont habituellement des données centrées. Voici l'estimation de petite région et l'intervalle de confiance à 95 % pour la j^e proportion régionale π_j :

$$\hat{\pi}_j = (1 + \exp[-(\hat{\alpha} + \hat{\beta}\overline{X}_j)])^{-1}. \qquad (2)$$

$$(1 + \exp[-(\hat{\alpha} + \hat{\beta}\overline{X}_j \pm 2\hat{\sigma})])^{-1}$$

$$\hat{\sigma}^2 = VAR(\hat{\alpha} + \hat{\beta}\overline{X}_j + \hat{u}_j)$$

On ajuste les modèles à l'aide du progiciel MLwiN. Pour mieux connaître la méthodologie SAEP, on consultera Heady et coll. (2001).

3. EXTENSION DE L'APPLICATION DE LA MÉTHODOLOGIE SAEP À LA MISE À JOUR DES DONNÉES DE RECENSEMENT

Nous décrirons une extension de l'application du modèle à la mise à jour des données de recensement. Nous posons que, initialement au moment t=1, et la variable d'enquête y et la variable correspondante de recensement Y mesurent la même entité. Cette enquête est périodiquement reprise à t=2…T. Le problème est de produire des estimations régionales pour la variable Y à t=T par les données de recensement à t=1, ainsi que par les données d'enquête à t=1,2…T et peut-être aussi par les données d'une ou de plusieurs variables auxiliaires X à t=1,2…T, si de telles données sont disponibles.

À supposer que le modèle (1) vaille pour la variable d'enquête y_{ijt} et la covariable Y, et plus précisément la covariable $\text{log}it(\Pi_{jt})$, pour chacun des points temporels t=1...T, on obtient un système de T équations :

$$\begin{aligned} & y_{ijt} \sim B(1,\pi_{jt}) \\ & \text{logit}(\pi_{jt}) = \alpha_t + \gamma\ \text{logit}(\Pi_{j1}) + u_{jt} \end{aligned}.$$

À l'aide des fonctions indicatrices $ind\_k(t)$ qui prennent la valeur 1 si t égale k et la valeur 0 dans les autres cas et après une certaine reparamétrisation, ce système d'équations peut s'écrire comme désignant la fonction indicatrice suivante :

$$
\begin{aligned}
&y_{ijt} \sim B(1, \pi_{jt}) \\
&\text{logit}(\pi_{jt}) = \beta_{j0} + \gamma\ \text{logit}(\Pi_{j1}) + \beta_{j2}\text{ind\_2(t)} + \ldots\ldots\ldots\ldots + \beta_{jT}\text{ind\_T(t)} \\
&\beta_{j0} = \beta_0 + u_{j0} \\
&\beta_{j2} = \beta_2 + u_{j2} \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
&\beta_{jT} = \beta_T + u_{jT}
\end{aligned}
\qquad (3)
$$

Les variables π_{jt} et Π_{j1} sont respectivement les variables d'enquête et de recensement. Ce sont les probabilités qu'une personne présente une caractéristique. On peut supposer en principe que, au niveau des régions, les résidus $(u_{j0}, u_{j2}, \ldots\ldots\ldots, u_{jT})$ seront en corrélation dans le temps, mais nous posons ici qu'ils sont indépendants et forment une distribution normale multidimensionnelle où la moyenne est nulle (0,0,……..,0) et où les variances et covariances sont en matrice diagonale avec le ii^{e} élément σ_{ii} (u_{jt}). Ainsi, pour la j^{e} région, la moyenne de population marginale (sans la covariable) pour l'année t à l'échelle des logits est donnée par :

$$
\begin{aligned}
&\beta_{j0} + u_{j0} \\
&\beta_{j0} + u_{j0} + \beta_{jt} + u_{jt}, t = 2, \ldots\ldots\ldots T
\end{aligned}
$$

Voici la matrice des variances-covariances pour le π_{jt} marginal (à l'échelle des logits) avec le ik-th élément

$$
\begin{aligned}
&\sigma_{00}, i = k = 1 \\
&\sigma_{00} + \sigma_{ii}' i = k = 2, \ldots, T \\
&\sigma_{00}, i \neq k
\end{aligned}
$$

Cette matrice des variances est du type « symétrie composée » avec des éléments arbitraires en diagonale. On suppose la corrélation constante pour les diverses années.

Dans le cas des coefficients fixes individuels des années, le coefficient β_t mesure la variation de la proportion nationale π pendant la période (0, t) à l'échelle des logits. Ainsi, le coefficient β_{jt} peut s'interpréter comme la mesure de la variation de la proportion π de t=0 à t=t à l'échelle des logits pour la région j. Une importante caractéristique du modèle (3) est qu'il n'y a pas d'hypothèse explicite au sujet de la forme de la tendance principale sous-jacente.

Ce modèle peut dégager les écarts locaux par rapport aux tendances nationales régionales. On y pose toutefois que ces écarts consistent seulement en a) des différences d'une même incidence sur toutes les années (il s'agit en réalité de différences entre les données de recensement et d'enquête) et b) en termes aléatoires influant sur une année en particulier. On ne tient aucun compte des éléments possibles d'autocorrélation temporelle entre années successives. Ainsi, l'estimateur tiré de ce modèle (nous en présentons la formule à l'échelle des logits) ne tire pas de force de la combinaison des années.

$$\text{logit}(\hat{\pi}_{jt}) = \hat{\beta}_{j0} + \hat{\gamma}\ \text{logit}(\Pi_{j1}) + \hat{\beta}_{j2}\text{ind\_2(t)} + \ldots\ldots\ldots\ldots + \hat{\beta}_{jT}\text{ind\_T(t)}$$

$$\hat{\beta}_{j0} = \hat{\beta}_0 + \hat{u}_{j0}$$
$$\hat{\beta}_{j2} = \hat{\beta}_2 + \hat{u}_{j2}$$
$$\text{.....} \qquad (3A)$$
$$\text{.....}$$
$$\hat{\beta}_{jT} = \hat{\beta}_T + \hat{u}_{jT}$$

En fait, cet estimateur diffère de l'estimateur SAEP seulement par l'inclusion d'une estimation de la variable aléatoire de région pour l'année en question. Il ne tire pas de force de la combinaison des périodes.

Dans une étude de faisabilité antérieure, Heady, Ruddock et Goldstein (1997) ont proposé la généralisation suivante de séries chronologiques pour le modèle (1) comme moyen de prise en compte de l'autocorrélation temporelle dans le contexte des problèmes d'actualisation des données de recensement :

$$y_{ijt} \sim B(1, \pi_{jt})$$
$$\log it(\pi_{jt}) = \beta_{0j} + \beta_{1j} t + \gamma \log it(\Pi_{j1})$$
$$\beta_{0j} = \beta_0 + u_{j0} \qquad (4)$$
$$\beta_{1j} = \beta_1 + u_{j1}$$
$$u_{j0} \sim N(0, \sigma_0^2), u_{j1} \sim N(0, \sigma_1^2), Cov(u_{j0}, u_{j1}) = 0$$

Les formules pour les estimations de proportions analogues à celles de la formule (3A) en découlent naturellement par substitution des estimations d'effets fixes et aléatoires dans (4).

Le modèle (4) se justifie du fait qu'on puisse calculer, à l'aide de données de séries chronologiques d'enquête, des estimations sûres des tendances de grandes régions et que, en rendant les coefficients aléatoires au niveau des petites régions ou dans le temps, on puisse établir des estimations de tendances locales. Une grande caractéristique de ce modèle est qu'on y suppose que la tendance (nationale) sous-jacente est connue et linéaire dans le temps comme dans le cas présent. Les auteurs ont fait observer que, dans l'estimation de tendances temporelles à spécificité locale, le modèle (4) est d'un rendement peu satisfaisant.

3. VALIDATION DE LA MÉTHODOLOGIE DE MISE À JOUR DES DONNÉES DE RECENSEMENT

Il importe au plus haut point de valider la méthodologie de mise à jour des données de recensement que nous avons sommairement décrite à la section précédente. Au Royaume-Uni, il n'y a guère, pour les variables étudiées, de source d'information censitaire qui soit encore de bonne qualité au niveau des administrations locales, aussi une évaluation externe indépendante de la qualité de cette mise à jour a-t-elle tout d'un problème épineux. Il reste que, à l'occasion de ce projet, l'ONS a eu accès aux données agrégées du registre à 100 % de la population finlandaise pour un certain nombre de variables sociodémographiques au niveau régional NUTS 5 (taille moyenne en 1987 : 11 000 habitants) de 1987 à 1996. Nous avons ainsi pu effectuer une simulation pour juger de la qualité des estimations. Il s'est agi de tirer des échantillons des ensembles des données en question, d'appliquer la méthode, de produire des estimations et de comparer celles-ci aux valeurs réelles.

Pour chaque année de la période 1987-1996, nous avons tiré des échantillons aléatoires distincts de 10 000 personnes pour la variable « situation d'emploi ». Nous avons constitué des ensembles de données de séries chronologiques en simulation par combinaison des échantillons individuels des années (en concaténation). Ainsi, les ensembles de données en simulation de la i^{e} série chronologique comprenaient les i^{es} échantillons individuels d'années en couplage temporel. Nous disposions au total de 100 000 enregistrements, 10 000 par an de 1987 à 1996. Dans le contexte des données finlandaises, les variables étaient les suivantes :

(i) y_{ijt} = i^{e} personne de la j^{e} région NUTS 5 ayant un emploi ou non au moment t (variable binaire d'échantillon);

(ii) Y_{j0} = proportion de la population locale de 15 à 74 ans ayant un emploi dans la région NUTS 5 en 1987 (registre finlandais de 1987).

On peut voir si les modèles (3) et (4) conviennent en consultant au tableau 1 les corrélations bidimensionnelles entre les estimations d'échantillon pour la proportion de gens ayant un emploi (échelle linéaire).

Tableau 1 : Estimations d'échantillon des corrélations bidimensionnelles entre les proportions de gens (de 15 à 74 ans) ayant un emploi dans chacune des années de la période 1987-1996

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
1987	1	0,979	0,965	0,954	0,920	0,895	0,837	0,867	0,890	0,893
1988	0,979	1	0,982	0,971	0,939	0,906	0,844	0,870	0,891	0,890
1989	0,965	0,982	1	0,979	0,937	0,900	0,841	0,865	0,884	0,877
1990	0,954	0,971	0,979	1	0,961	0,934	0,883	0,900	0,914	0,904
1991	0,920	0,939	0,937	0,961	1	0,971	0,932	0,934	0,937	0,925
1992	0,895	0,906	0,900	0,934	0,971	1	0,967	0,964	0,959	0,945
1993	0,837	0,844	0,841	0,883	0,932	0,967	1	0,974	0,960	0,942
1994	0,867	0,870	0,865	0,900	0,934	0,964	0,974	1	0,984	0,971
1995	0,890	0,891	0,884	0,914	0,937	0,959	0,960	0,984	1	0,986
1996	0,893	0,890	0,877	0,904	0,925	0,945	0,942	0,971	0,986	1

Il ressort de ce tableau que les corrélations entre les estimations éloignées de plus de quatre ans les unes des autres s'établissent en gros à une valeur constante. Quant aux corrélations entre estimations que séparent moins de quatre ans, elles sont en décroissance progressive à mesure que s'accroît l'écart temporel. Pour des corrélations tenues pour constantes entre les estimations d'années différentes, on constate que les corrélations sont faibles entre les estimations que séparent moins de quatre ans.

Si on met en courbe les tendances réelles des proportions de gens ayant un emploi (données non présentées), on remarque que les tendances locales (régions NUTS 5) sont à peu près semblables à la tendance nationale et aussi que cette dernière ne décrit habituellement pas une courbe monotone de croissance ou de décroissance au fil des ans. On en a la confirmation lorsqu'on regarde la tendance nationale à la figure 3. On a l'impression que le modèle (3) pourrait mieux convenir à cet ensemble de données que le modèle (4).

Comme on connaît les valeurs réelles de population de la variable y pour chacune des années et des régions NUTS 5, il est possible d'évaluer le rendement prévisionnel des modèles (3) et (4) en comparant les valeurs estimées par modélisation aux valeurs réelles. La figure 1 représente les valeurs effectives et les estimations de modélisation en 1996. La figure 2 fait de même pour les variations de 1995-1996.

Figure 1 : Comparaison des valeurs réelles et des estimations de modélisation des proportions d'habitants des régions NUTS 5 ayant un emploi en Finlande en 1996

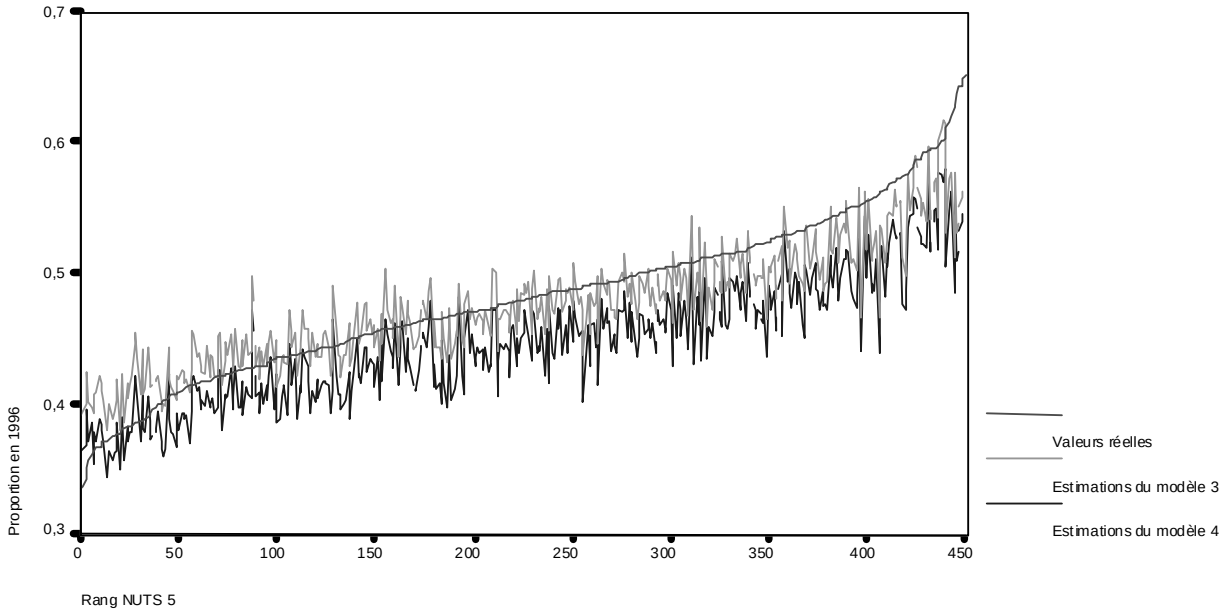


Figure 2 : Comparaison des valeurs réelles et des estimations de modélisation des variations de la proportion d'habitants des régions NUTS 5 ayant un emploi en Finlande de 1987 à 1996

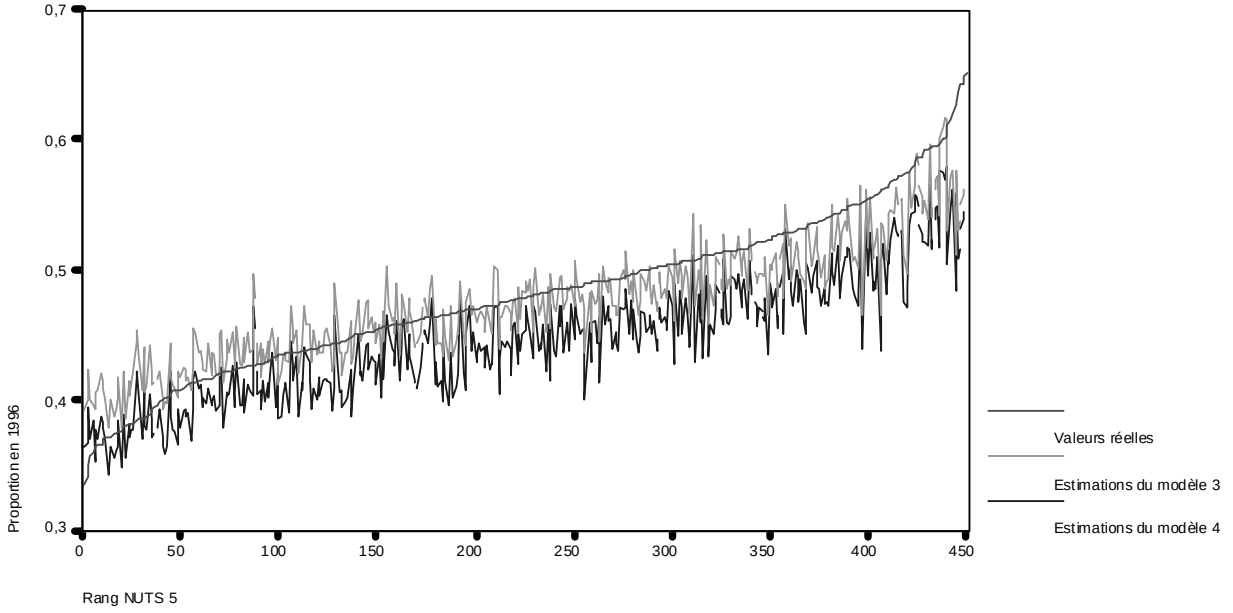
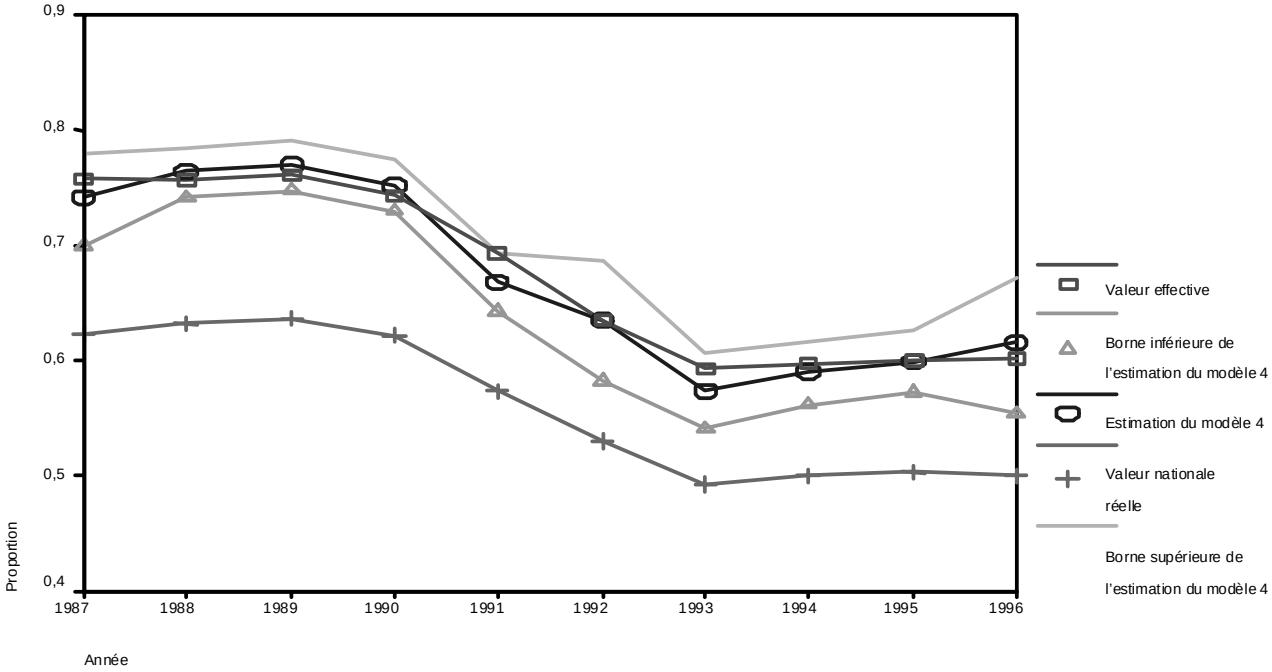


Figure 3 : Valeurs réelles et intervalles de confiance à 95 % des estimations de modélisation de la proportion d'habitants de la région NUTS 257 ayant un emploi en Finlande de 1987 à 1996



On peut voir aux figures 1 et 2 que les estimations de proportions ou de variations issues du modèle (3) sont plus proches des valeurs réelles que celles du modèle (4). L'examen des modèles ajustés (3) et (4) (données non présentées) indique que :

- les estimations par le modèle (3) des variances des termes aléatoires étaient très petites sauf celles de la valeur à l'origine et de l'année 1996 (en fait, les variances de ces termes ont été d'une estimation nulle en 1989, 1990, 1991, 1994 et 1995);

- les estimations par le modèle (4) de la variance du terme aléatoire de la pente étaient proches de zéro.

On garde l'impression que les valeurs réelles de 1987 et la forme supposée de la tendance sous-jacente ont très largement déterminé les estimations des modèles (3) et (4) au niveau des petites régions. Une tendance non linéaire est posée dans le modèle (3) et une tendance linéaire, dans le modèle (4). On juge qu'une meilleure modélisation de la tendance sous-jacente dans le modèle (3) explique en grande partie que le rendement d'estimation du modèle (3) soit supérieur à celui du modèle (4). C'est par le fait que la plupart des variations soient d'une application à peu près égale à toutes les régions du pays qu'on peut expliquer le fait étonnant qu'un estimateur qui ne tire pas de force de la combinaison des années n'en constitue pas moins un moyen acceptable de mise à jour des estimations locales de recensement. C'est pourquoi nous avons adopté pour l'instant le modèle (3) pour l'établissement d'estimations provisoires de mise à jour des données locales de recensement, comme nous le décrirons à la section suivante.

Avant de passer aux données britanniques, mentionnons un certain nombre de points. Pour une région type NUTS 5, la figure 3 indique à quel point le modèle (3) dégage la tendance réelle. Notons en particulier que les estimations issues de ce modèle décrivent bien la tendance effective et que les intervalles de confiance à 95 % sont satisfaisants.

Il est sûr cependant qu'il y a toujours lieu d'améliorer un modèle qui reste foncièrement incapable d'estimer des tendances à spécificité locale à cause de son hypothèse d'indépendance des termes aléatoires $\{u_{jt}, t = 1,....T\}$. Comme nous l'avons dit, des indices portent à conclure à l'existence de corrélations faibles et décroissantes entre les estimations d'échantillon que sépare une période de moins de quatre ans. On devrait donc supposer une

structure d'autocorrélation pour les termes $\{u_{jt}, t = 1,.....T\}$ (comme le montre l'AR(1)). On peut penser qu'on réussira à améliorer les estimations encore plus en modélisant les résidus de cette façon.

4. APPLICATION AUX DONNÉES DU RECENSEMENT DE 1991 AU ROYAUME-UNI

Nous avons appliqué la méthodologie exposée dans les sections qui précèdent à l'établissement d'estimations postcensitaires de trois variables (calculées) du recensement de 1991 au Royaume-Uni pour l'année 2001 et les LAD/UA (« local authority districts » et « unitary authorities ») de 1998 (taille moyenne en 1991 : 133 000 habitants), à savoir :

PHHSNG : proportion de ménages d'une seule personne au recensement de 1991;
PPLA : proportion de pensionnés vivant seuls au recensement de 1991;
PETHNIC : proportion de membres de minorités ethniques au recensement de 1991.

Par souci de concision, nous ne livrerons ici que les résultats relatifs à la variable PHHSNG. On trouvera les résultats détaillés des autres variables dans Yar et coll. (2002). Voici les ensembles de données qui ont été exploités :

données du recensement de 1991;
données de l'enquête sur la population active (EPA) pour la période 1993-2001.

On aurait pu souhaiter utiliser les données de cette enquête à partir de 1991, mais on ne peut dire au juste en quoi les résultats auraient été différents. À cause d'une importante révision de cette enquête en 1992 et d'un géocodage insuffisant des données pour cette même année, seules les données postérieures à 1992 étaient exploitables à nos fins. Ajoutons que, dans l'EPA britannique, il y a un échantillon longitudinal en renouvellement où chaque ménage sélectionné est interviewé cinq trimestres de suite, d'où la non-indépendance des données de cette enquête que séparent moins de cinq trimestres. Pour garantir l'indépendance des données et éviter toutes les complications d'effets saisonniers, nous avons donc décidé de prendre seulement les périodes 1-4 du trimestre du printemps (mars-mai) chaque année de 1993 à 2001. Nous avons raccordé les données individuelles de cette enquête aux covariables du recensement de 1991 au niveau LAD/UA.

Tableau 2. Modèle logistique à deux niveaux pour la variable de réponse HHSNG de l'enquête sur la population active

Covariable	Estimation paramétrique	Erreur-type	Valeur T
Partie fixe			
VALEUR À L'ORIGINE	0,118	0,054	2,181
LGT(PHHSNG)	1,031	0,049	20,911
IND94	0,012	0,079	0,157
IND95	0,034	0,081	0,423
IND96	-0,135	0,079	-1,709
IND97	-0,268	0,081	-3,329
IND98	-0,220	0,079	-2,779
IND99	-0,208	0,077	-2,699
IND00	-0,174	0,079	-2,211
IND01	-0,232	0,078	-2,988
LGT(PHHSNG)*IND94	-0,019	0,072	-0,262
LGT(PHHSNG)*IND95	0,008	0,074	0,105
LGT(PHHSNG)*IND96	-0,110	0,072	-1,534
LGT(PHHSNG)*IND97	-0,248	0,073	-3,403
LGT(PHHSNG)*IND98	-0,199	0,072	-2,776
LGT(PHHSNG)*IND99	-0,216	0,070	-3,091
LGT(PHHSNG)*IND00	-0,195	0,071	-2,724
LGT(PHHSNG)*IND01	-0,258	0,070	-3,672
Partie aléatoire			
Variances de niveau 2			
VALEUR À L'ORIGINE	0,001	0,000	2,545
IND94	0,006	0,004	1,758
IND95	0,002	0,004	0,547
IND96	0,005	0,004	1,421
IND97	0,007	0,004	1,812
IND98	0,003	0,003	0,997
IND99	0,000	0,000	
IND00	0,003	0,003	0,804
IND01	0,000	0,000	
Variance de niveau 1	1,000	0,000	

Nous avons défini une variable binaire de cette enquête HHSNG qui prend la valeur 1 s'il s'agit d'un ménage (h/h) d'une seule personne h/h et la valeur 0 dans le cas contraire. Nous avons ajusté le modèle (3) à cette variable à l'aide du progiciel MLwiN. À cause d'une limitation de MLwiN là encore, la matrice des variances-covariances des résidus de région u_{jt} a été restreinte à la forme diagonale :

$$\text{diag}(\sigma_{00}, \sigma_{33}, \sigma_{44}, \sigma_{55}, \sigma_{66}, \sigma_{77}, \sigma_{88}, \sigma_{99}, \sigma_{1010}) .$$

Les seules covariables disponibles étaient celles du recensement de 1991 et du temps. Le tableau 2 présente le modèle ajusté à la variable de réponse HHSNG. Voici les désignations des variables du tableau 2 :

IND94 désigne pour 1994 une fonction indicatrice qui prend la valeur 1 pour 1994 et la valeur 0 dans les autres cas.

PHHSNG désigne la proportion de ménages d'une seule personne au recensement de 1991.
IND94* PHHSNG désigne l'interaction de ces deux variables.

On définit de la même façon les autres fonctions indicatrices avec leurs interactions au tableau 2.

Pour un district d'administration locale type (LAD), la figure 5 indique la tendance du modèle et les estimations directes d'enquête (EPA) avec leurs intervalles de confiance pour 1993 et 2001. À des fins de comparaison, nous présentons aussi l'estimation nationale EPA pour chacune des années (on peut voir que, dans ce cas, la tendance nationale est statique pour ainsi dire). Les estimations du modèle sont d'une plus grande efficience que les estimations directes d'enquête, à en juger par l'étendue des intervalles de confiance. Voilà un exemple où la variable étudiée « proportion de ménages d'une seule personne » a évolué lentement dans le temps. Pour cet exemple en particulier, il est aussi impossible d'écarter l'hypothèse d'une linéarité de tendance.

Diagnostics du modèle

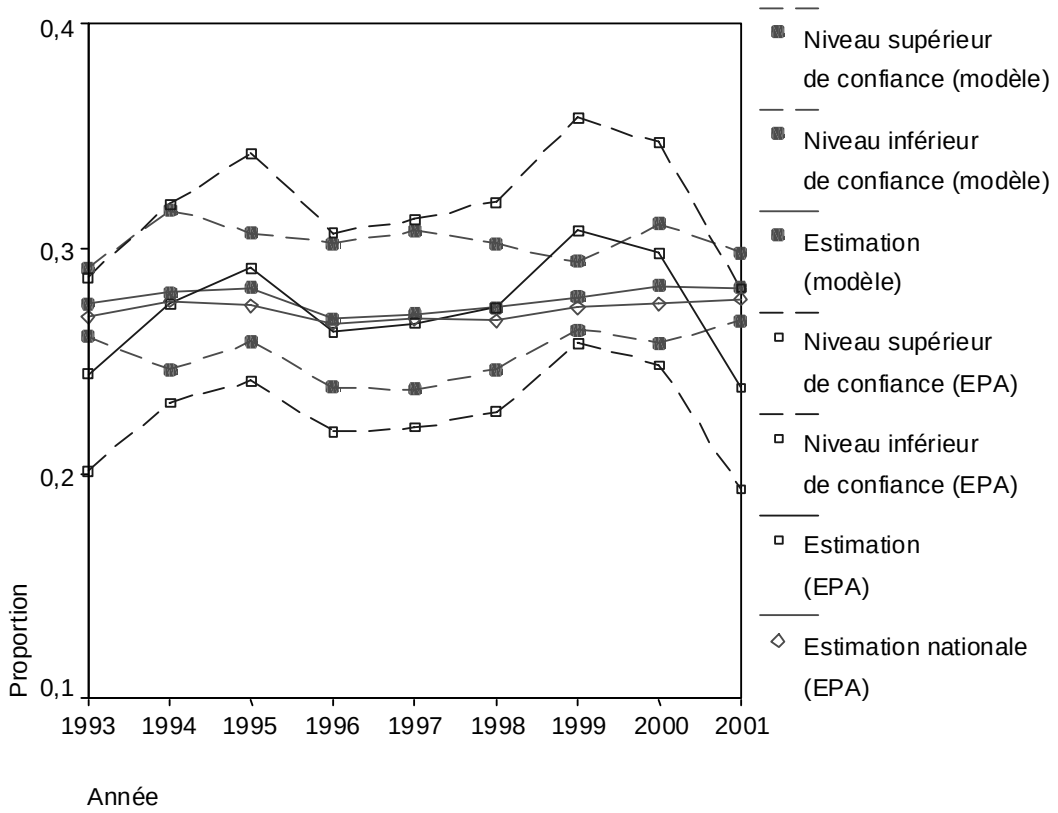
On ne peut employer dans ce cas le mode de vérification et de validation appliqué à la section 3, puisque les valeurs vraies de population sont inconnues, mais dans le contexte des estimations régionales, Brown et coll. (2001) ont conçu cinq diagnostics de validation du modèle et de vérification de la qualité de ses estimations.

Le diagnostic de couverture est une des mesures proposées du rendement des estimations de modélisation par rapport aux valeurs réelles. Comme nous considérons que les estimations directes d'enquête sont exemptes de biais, leurs intervalles de confiance à 95 % contiendront les valeurs vraies dans 95 % des cas. Il devrait en aller de même des intervalles de confiance des estimations de modélisation. Le tableau 3 indique les taux de couverture, par les intervalles de confiance des estimations directes d'enquête, des estimations de modélisation de la période 1993-2001. La figure 4 compare pour un des districts d'administration locale les intervalles de confiance respectifs des estimations d'échantillon d'enquête et des estimations de modélisation.

Tableau 3. Diagnostic de couverture des estimations de la proportion de ménages d'une seule personne en Angleterre et au pays de Galles

Année	Pourcentage d'estimations de modélisation se situant hors de l'intervalle de confiance à 95 % des estimations directes d'enquête
1993	4,9
1994	4,6
1995	4,6
1996	3,0
1997	3,5
1998	3,2
1999	4,9
2000	4,6
2001	6,2

Figure 4. Estimations de l'enquête sur la population active et du modèle pour la proportion de ménages d'une seule personne à Bradford de 1993 à 2001



5. CONCLUSIONS ET INDICATION DES SECTEURS OÙ IL FAUDRA PEUT-ÊTRE POUSSER LA RECHERCHE

Des travaux empiriques présentés ici et ailleurs, on peut conclure à la possibilité de mettre à jour les données de recensement à l'aide d'une version très légèrement modifiée de la méthodologie SAEP, mais il existe plusieurs façons d'accroître la qualité des estimations postcensitaires. Dans le cas des ménages d'une seule personne, les seules covariables employées ont été le temps et les covariables du recensement de 1991. Nous pensons que la disponibilité d'autres covariables à évolution temporelle au niveau LAD/UA pour la variable étudiée améliorera la qualité des estimations. Dans notre exemple, ce pourrait être les données « Council Tax » et le Registre électoral. À l'époque de notre étude, ces ensembles de données étaient toutefois indisponibles.

Dans le modèle appliqué à la production d'estimations postcensitaires, nous avons supposé que, au niveau des régions, les résidus étaient indépendants dans le temps. Dans ce cas, l'analyse se ramène pour l'essentiel à une analyse séparée des données individuelles des points temporels à l'aide du modèle SAEP. Ces estimations ne tirent donc pas de force de la combinaison des années et ne se trouvent pas à exploiter tout le potentiel de l'information. Nous convenons qu'une stratégie optimale d'estimation pour la mise à jour des données de recensement devra tirer de la force tant de l'espace que du temps. Dans cette optique, on peut songer pour les résidus à des modèles plus généraux comme les modèles autorégressifs, les modèles à marche aléatoire ou ceux que présentent Rao et Yu (1994) et Pfefferman (2002). Reste à savoir dans quelle mesure les estimations s'améliorent après la prise en compte de l'autocorrélation dans le temps. Dans un contexte voisin, nous sommes en train d'étudier, dans le cadre de notre projet européen EURAREA, l'utilisation de modèles de séries chronologiques dans l'établissement d'estimations régionales. Une fois que nous disposerons de ces résultats, nous pourrons les appliquer à la production d'estimations postcensitaires.

RÉFÉRENCES

Brown, G; Chambers, R; Heady, P and Heasman, D (2001). Evaluation of small area estimation methods – an application to unemployment estimates from the UK LFS. Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: a Methodological Perspective.

Ghosh, M; Nangia, N and Kim, D, H (1996). Estimation of median income of 4-person families: A Bayesian time series approach. Journal of American Statistical Association, Vol 91, no 436, 1423-1431.

Heady, P; Clarke, P et al (2002). Small Area Estimation Project Report. Model Based Small area Estimation Series No 2, Office for National statistics, UK.

Heady, P; Ruddock, V and Goldstein, H (1997). An investigation into the possible use of multilevel models based on survey data to update census estimates for small areas. Proc. of Statistics Canada Symposium 1997, 199-203.

Pfeffermann, D (2002): Small area estimation-new developments and directions. International Statistical Review, 70, no 1, 125-144.

Purcell, N and Kish, L (1980). Postcensal estimates for local areas (or domains). International Statistical Review, vol48, 3-18.

Rao, J.N.K and Yu, M (1994). Small area estimation by combining time series and cross sectional data. Canadian Journal of Statistics, 22, 511-528.

Saei, A and Chambers, R (2002). Small area estimation: a review of methods based on the application of mixed models. FP5 project EURAREA report.

Tiler, R.B (1992). Time series modelling of sample survey data from the US Current Population Survey. Journal of Official Statistics, 8, 149-166.

Yar, M and Higgins, N (2002). An investigation into the development and testing of methodology for updating the 1991 UK Census indicators. Internal ONS report.