MODELES D'ESTIMATION DES ERREURS D'ECHANTILLONNAGE[1]

P.D. Ghangurde[2]

Ce document présente les résultats d'une étude empirique sur l'adjustement de modèles linéaires logarithmiques en fonction des estimations de caractéristiques de l'enquête sur la population active et de leurs coefficients de variation (CV). Ces caractéristiques ont été regroupées en fonction des effets du plan d'échantillonnage, et des modèles ont été ajustés en fonction des estimations de leur valeur totale et de leurs CV calculés sur une période de douze mois. De tels modèles peuvent être utilisés lorsqu'on doit estimer les CV de nouvelles caractéristiques ou pour fournir des estimations plus précises de la fiabilité des estimations fondées sur des données antérieures. Le problème de l'évaluation de la validité de l'ajustement des modèles est également examiné dans ce document.

1. INTRODUCTION

Ce document présente les résultats d'une étude d'évaluation des modèles de calcul des coefficients de variation (CV) d'estimations de caractéristiques de l'enquête sur la population active du Canada (EPA). L'EPA est une enquête mensuelle menée auprès des ménages, qui est basée sur un plan d'échantillonnage aréolaire stratifié à plusieurs degrés et dont l'échantillon comprend environ 55 000 ménages.

On estime chaque mois les CV d'un ensemble de caractéristiques en appliquant la méthode d'estimation de la variance de Keyfitz selon une approximation de la série de Taylor [4], [5].

---

1 Exposé présenté à l'assemblée annuelle de l'American Statistical Association, en août 1981, à Détroit.

2 P.D. Ghangurde, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.

Cependant, pour des raisons de temps et d'argent, il est impossible de calculer la variance approximative de toutes les estimations obtenues à partir d'une enquête à grande échelle comme l'EPA. Les estimations des CV basées sur des modèles peuvent être utilisées lorsqu'on veut obtenir des estimations provisoires de la fiabilité de nouvelles caractéristiques fondées sur des données antérieures ou lorsqu'il faut étendre la période d'observation (une année par exemple). Les modèles peuvent également être utilisés pour obtenir des estimations précises de la fiabilité, notamment des indicateurs alphabétiques des intervalles des CV.

Dans la partie 2, on explique les modèles linéaires et non linéaires appliqués au calcul des estimations totales et proportionnelles. Les parties 3 et 4 portent sur les facteurs de groupement des caractéristiques, l'ajustement des modèles et l'évaluation de la validité des ajustements.

2. MODÈLES

L'EPA est une enquête mensuelle menée auprès des ménages dans laquelle le logement est l'unité d'échantillonnage du dernier degré. Chaque province du Canada est divisée en régions économiques qui sont formées de groupes de comtés ayant une structure économique semblable. Les régions économiques sont subdivisées en strates géographiques et on tire des échantillons aréolaires sans remise à plusieurs degrés, soit à deux degrés dans les strates autoreprésentatives dans les grands centres urbains et à trois ou quatre degrés dans les strates non autoreprésentatives dans les régions rurales. Aux premiers degrés, la sélection de l'échantillon se fait selon une probabilité proportionnelle à la taille de la population et, au dernier degré, où les logements sont choisis dans les grappes, la sélection est systématique.

Pour produire les estimations relatives aux strates (basées sur le plan d'échantillonnage), on applique aux données un coefficient de pondération qui équivaut à l'inverse des probabilités de sélection. Les estimations finales sont obtenues à la suite d'un ajustement du coefficient de pondération de base pour tenir compte de la non-réponse et par l'estimation

par quotient à l'intérieur des groupes âge/sexe qui sont des strates établies à posteriori. Les projections démographiques du recensement par groupe âge/sexe et par province sont les variables auxiliaires utilisées dans l'estimation par quotient. Pour plus de détails sur le plan d'échantillonnage et l'estimation, voir la publication [5].

Les estimations de la variance de diverses caractéristiques au niveau provincial sont obtenues par une approximation de la série de Taylor, en supposant que les unités primaires d'échantillonnage (UPÉ) à l'intérieur des strates non autoreprésentatives sont choisies de façon indépendante. Dans les strates autoreprésentatives, les grappes échantillonnées sont réparties en deux groupes qui deviennent des pseudo-UPÉ et sont supposées avoir été choisies de façon indépendante. L'estimation de la variance du total estimé d'une caractéristique à l'échelle nationale consiste à faire la somme des estimations correspondantes de la variance à l'échelon provincial [5]. La variance d'une estimation $\hat{X}$ du total X d'une caractéristique dans une province donnée prend donc la forme:

$$V(\hat{X}) = F\ (W-1)\ X\ (1 - \frac{X}{P}),$$

où P = la population de la province

W = l'inverse du taux de sondage

F = l'effet du plan d'échantillonnage pour la caractéristique

n = la taille de l'échantillon (nombre de personnes).

L'expression (1) pour calculer $V(\hat{X})$ relie la variance de l'estimation complexe par quotient qui est fondée sur un plan d'échantillonnage stratifié à plusieurs degrés à la variance de l'estimation basée sur un échantillon aléatoire simple de même taille tiré de la population finie de taille P. La variance d'échantillonnage d'une estimation de la valeur totale d'une caractéristique fondée sur un échantillon aléatoire simple de taille $n(=\frac{P}{W})$ correspond à la variance binomiale courante corrigée en fonction de la population finie. Le terme F, l'effet du plan d'échantillonnage, est un facteur d'accroissement de la variance qui réagit à des facteurs comme la méthode d'échantillonnage à chaque degré, le niveau de stratification et de stratification à posteriori, la taille des unités aux divers degrés et la

constitution de grappes à partir des totaux obtenus pour la caractéristique dans la province. On doit souligner ici que la stratification et la stratification à posteriori ont habituellement pour effet de réduire la variance d'une estimation, alors que la constitution de grappes l'augmente.

De façon générale, les effets du plan d'échantillonnage tendent à être plus importants que les effets dus à l'échantillonnage par grappes des unités de l'EPA. Les effets du plan tendent à être moins élevés dans le cas des catégories de la population active comme celles des "personnes occupées" et des "chômeurs" par groupe âge/sexe, en raison de la stratification à posteriori par âge et sexe qui réduit leur variance. Lorsque la population active est classée selon le secteur d'activité, les effets du plan tendent à être plus élevés à cause de la concentration dans des régions précises. On sait que les effets du plan d'échantillonnage sont fonction des mesures d'homogénéité et de la taille moyenne des grappes. Des modèles fondés sur de tels rapports ont été conçus pour nombre d'enquêtes. Dans une étude sur les composantes de la variance intervenant dans l'EPA, on a analysé les effets du plan d'échantillonnage et les mesures d'homogénéité d'un certain nombre de caractéristiques [2].

Le coefficient de variation nous donne une mesure de précision des estimations sans égard à leur niveau et à l'échelle. Le CV $(\hat{X})$ est donné par

$$CV\ (\hat{X}) = F(W - 1)(\frac{1}{X} - \frac{1}{P})\ . \qquad (2)$$

En utilisant des logarithmes pour déterminer la base e des deux membres de l'expression (2), nous obtenons une équation reliant CV, X et P qui prend la forme:

$$\log CV\ (\hat{X}) = \frac{1}{2} \log F(W - 1) - \frac{1}{2} \log X + \frac{1}{2} \log (1 - \frac{X}{P})\ . \qquad (3)$$

À cause du troisième terme du deuxième membre, l'équation (3) n'est pas linéaire sous la forme log X, même si F(W-1) est supposé constant.

Toutefois, dans le cas des petites valeurs de X, le troisième terme a un effet négligeable. Un modèle fondé sur (3) est exprimé par

$$\log CV(\hat{X}) = A + B \log X + E, \tag{4}$$

où A et B sont les paramètres du modèle et E le terme de l'erreur. L'estimation du paramètre B sera différente de $-\frac{1}{2}$ selon que B log X est plus ou moins une approximation de $\frac{1}{2} \log [X/(1 - \frac{X}{P})]$ dans l'étendue de X. Dans une évaluation des ajustements du modèle (4) et du modèle de rechange (5) donné par

$$\log CV(\hat{X}) = A + B \log \frac{X}{(1 - \frac{X}{P})} E, \tag{5}$$

on a constaté que la validité de l'ajustement, exprimée par R^2, soit le rapport de la somme des carrés de la régression sur la somme des carrés totale, était sensiblement la même pour les deux modèles. Le modèle (4) est linéaire sous la forme log X et log CV et peut être appliqué plus facilement que le modèle (5).

Un modèle non linéaire correspondant à (4) prend la forme:

$$CV(\hat{X}) = A' X^{B'} + E, \tag{6}$$

où A' et B' sont les paramètres du modèle et ε est le terme de l'erreur. Les deux modèles (4) et (6) ont été ajustés en fonction des estimations mensuelles et du (CV) de 90 caractéristiques dans chacune des 10 provinces.

3. GROUPEMENT DES CARACTÉRISTIQUES

On a calculé la moyenne des effets du plan mensuels des estimations de l'EPA pour la période de janvier à décembre 1980 et pour chacune des 90 caractéristiques, sauf la population totale de chaque province et du Canada, puis on les a portées sur un graphique afin de déterminer l'étendue

des deux groupes. Dans chaque province, le premier groupe comprenait les caractéristiques affichant des effets plus grands que D.

Le tableau 1 indique les valeurs limites D des groupes I et II dans chaque province et pour le Canada, de même que le nombre de caractéristiques incluses dans le groupe II. Pour grouper les caractéristiques, on les a classées par ordre croissant de la moyenne des effets du plan. La valeur limite D a été choisie de façon que l'hypothèse de l'égalité des effets du plan soit satisfaite le plus possible dans le groupe I. Le deuxième groupe comprend toutes les autres caractéristiques pour lesquelles l'égalité des effets du plan est supposée plus rudimentaire. La plupart des caractéristiques relatives à la situation vis-à-vis de l'activité, classées par groupe âge/sexe se retrouvent dans le groupe I. Les personnes occupées par secteur d'activité et la durée de la période de chômage entrent habituellement dans le groupe II. Les effets moyens varient beaucoup entre les provinces et pour le Canada. On examine actuellement l'application d'un groupement des caractéristiques plus détaillé basé sur des modèles des effets du plan.

Il convient de souligner qu'environ 80 % des caractéristiques dans chaque province et pour le Canada ont été classées dans le groupe I. Pour produire une estimation prudente du CV de nouvelles caractéristiques, on peut utiliser des modèles basés sur le groupe II. Dans le cas d'une caractéristique dont les estimations mensuelles du CV sont calculées régulièrement, on peut appliquer les modèles construits pour le groupe dans lequel cette caractéristique est classée, si on veut obtenir une estimation approximative du CV qui sera plus précise que celle basée sur les données mensuelles.

Dans la partie suivante, nous examinons les hypothèses appliquées à l'ajustement des modèles (4) et (6) et nous évaluons ces ajustements.

4. ÉVALUATION DES MODÈLES

Pour ajuster le modèle linéaire logarithmique (4), il faut essentiellement considérer ce modèle comme un modèle de régression linéaire simple sous la forme $y = \log CV(\hat{X})$ et $x = \log X$, et calculer les estimations des paramètres A et B selon la base de régression linéaire. Les hypothèses usuelles d'indépendance des erreurs et de variance constante s'appliquent et, selon ces hypothèses, R^2 fournit une mesure de l'ajustement du modèle. Les valeurs des paramètres estimés et les coefficients de détermination, R^2, dans les groupes I et II, relativement aux 10 provinces et au Canada, figurent au tableau 2. L'ajustement effectif de ces modèles a été fait à l'aide du programme SAS.

Toutes les valeurs R^2 sont significatives et assez élevées, ce qui démontre que les ajustements sont très précis. Le graphique des erreurs ne révèle pas de tendance qui permette de conclure que l'hypothèse de la variance constante n'est pas satisfaite. En vertu de ces hypothèses et compte tenu de la normalité des erreurs, $CV(\hat{X})$ présente une distribution log-normale avec un CV constant pour n'importe quelle valeur de X.

Le modèle non linéaire (6) a été ajusté selon la méthode Gauss-Newton à l'aide du programme SAS. On a supposé que les valeurs initiales des paramètres A' et B' étaient 1.00 et -0.50 respectivement. Le nombre d'itérations requises pour atteindre la convergence a été de 8 au maximum pour chaque province et pour le Canada, la norme de convergence étant que l'écart relatif entre plusieurs sommes des carrés résiduelles doit être inférieur à 10^{-8}. Le tableau 3 contient les valeurs des paramètres estimés et la somme des carrés résiduelle pour le Canada, dans le groupe II. Les erreurs affichent une distribution approximativement normale puisque les points sur le graphique révèlent une distribution normale de probabilités.

Comme la comparaison des ajustements du modèle non linéaire appliqué aux provinces, au Canada, pour les deux groupes, présente un certain intérêt, il est nécessaire de définir une norme de validité de l'ajustement. Dans le modèle non linéaire, la somme des carrés totale n'est pas égale au total de

la régression et de la somme des carrés résiduelle. La norme R'^2 peut être exprimée

$$R'^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2},$$

où les $\hat{Y}_i$ sont les CV estimés basés sur le modèle, les $\hat{Y}_i$ sont les CV observés et $\bar{Y}$ leur moyenne. La sommation s'étend à N, qui est le nombre de caractéristiques dans le groupe, multiplié par 12, le nombre de mois. Dans le modèle linéaire, $R^2 = R'^2$. Cependant, dans le modèle non linéaire, R^2 est différent de R'^2, car la somme des carrés totale n'est pas égale à la somme des carrés de la régression plus la somme des carrés résiduelle parce que le produit n'est pas zéro.

Les erreurs $(Y_i - \hat{Y}_i)$ sont petites lorsque l'ajustement est précis et produit une valeur de R'^2 proche de 1; les erreurs $(Y_i - \hat{Y}_i)$ sont importantes lorsque l'ajustement est faible et produit une petite valeur de R'^2. Lorsque tous les points se trouvent sur la courbe ajustée, c'est-à-dire que $Y_i = \hat{Y}_i$ pour tous les i, alors $R'^2 = 1$. Toutefois, règle générale, il ne semble pas y avoir de borne inférieure à R'^2. Les valeurs de R'^2 présentées dans le tableau 4 semblent être plus élevées dans le groupe I que dans le groupe II qui comprend de 13 à 21 caractéristiques sur 90.

Bien que le modèle linéaire logarithmique (4) ait été ajusté en fonction des logarithmes des estimations et de leur CV et que l'ajustement semble précis, les modèles ajustés pour les provinces et le Canada sont appliqués à l'estimation des CV des estimations. Afin de pouvoir comparer l'ajustement du modèle transformé aux estimations initiales et à leur CV, les données et le nouveau modèle (4) ont été portés sur un graphique, et ce pour les deux groupes relatifs aux 10 provinces et au Canada. À partir de ces diagrammes, on peut conclure que ce nouveau modèle (4) peut ajuster les estimations et leur CV de façon plus précise que le modèle non linéaire (6), surtout dans

le cas des petites valeurs. Les graphiques 1 et 2 illustrent ces modèles appliqués aux valeurs du groupe II à l'échelle du Canada.

5. CONCLUSION

Les caractéristiques étudiées sont le nombre total de personnes classées selon leur situation vis-à-vis de l'activité par groupe âge/sexe, secteur d'activité et état matrimonial, de même que le nombre total de personnes selon la durée de la période de chômage. Les modèles peuvent être utilisés aussi bien pour la production de proportions que pour celle des totaux, mais ils ne peuvent pas être appliqués aux estimations relatives aux régions infraprovinciales, notamment les centres urbains ou les groupes de régions économiques, parce que les effets du plan d'échantillonnage dans ces régions sont moins stables et peuvent être beaucoup plus élevés à cause de l'effet de la correction du quotient effectuée à l'aide des projections démographiques provinciales [1].

On a posé comme hypothèse d'application de modèles pour une nouvelle caractéristique que l'effet du plan d'échantillonnage est proche de la moyenne établie pour le groupe. Il faut alors détailler encore plus le groupement des caractéristiques en fonction de modèles qui permettent de lier les effets du plan aux mesures d'homogénéité de ces caractéristiques. Pour l'ajustement des modèles, on a supposé que les erreurs ne sont pas corrélées et que la variable indépendante est fixe. Comme on a utilisé 12 estimations mensuelles de chaque caractéristique, il est possible que les erreurs relatives aux estimations d'une caractéristique particulière soient corrélées. On examine actuellement la possibilité d'étendre cette étude de façon à englober les modèles qui tiennent compte des erreurs produites avec la variable indépendante et des erreurs corrélées.

Le problème qui se pose quant à l'évaluation de l'ajustement de modèles non linéaires, qu'ils soient effectivement ajustés aux données ou qu'il s'agisse de modèles linéaires transformés, est l'absence d'une norme de comparaison des ajustements de divers modèles. Il est probable que la norme proposée à la partie 4 convienne pour la comparaison des ajustements de modèles à

diverses séries de données, mais il est probable également que cela ne convienne pas à des modèles différents.

TABLEAU 1: VALEURS LIMITES DE L'EFFET DU PLAN D'ÉCHANTILLONNAGE ET NOMBRE DE CARACTÉRISTIQUES DANS LES GROUPES I ET II*

Province	Valeur limite (D)	Nombre de caractéristiques	
		Groupe I	Groupe II
Terre-Neuve	2.3	75	15
Île-du-Prince-Édouard	1.9	73	17
Nouvelle-Écosse	1.9	74	16
Nouveau-Brunswick	2.2	77	13
Québec	1.9	73	17
Ontario	1.7	69	21
Manitoba	2.0	76	14
Saskatchewan	2.8	76	14
Alberta	2.1	71	19
Colombie-Britannique	2.3	73	17
Canada	1.9	77	13

* Une caractéristique est classée dans le groupe I si l'effet du plan d'échantillonnage pour cette caractéristique (dont la moyenne est calculée sur une période de 12 mois, soit de janvier à décembre 1980) est inférieur ou égal à la valeur limite D. Si l'effet moyen est supérieur à D, la caractéristique est alors incluse dans le groupe II.

TABLEAU 2: COEFFICIENTS DE RÉGRESSION ET R^2 SELON LE MODÈLE LINÉAIRE LOGARITHMIQUE

Province	Groupe	Coefficient de régression A	B	R^2
Terre-Neuve	I	3.3119	-0.5723	0.9534
	II	3.7757	-0.6101	0.9377
Île-du-Prince-Édouard	I	2.7962	-0.5617	0.9485
	II	3.1796	-0.5885	0.8887
Nouvelle-Écosse	I	3.4612	-0.5837	0.9702
	II	3.6412	0.5257	0.8717
Nouveau-Brunswick	I	3.2782	-0.5545	0.9606
	II	3.7544	-0.6017	0.9357
Québec	I	4.3298	-0.5942	0.9686
	II	4.3093	-0.5216	0.9127
Ontario	I	4.3825	-0.6053	0.9736
	II	4.1796	-0.5009	0.9633
Manitoba	I	3.5155	-0.5926	0.9619
	II	3.8769	-0.5640	0.9166
Saskatchewan	I	3.3796	-0.5700	0.9544
	II	3.5478	-0.4423	0.8994
Alberta	I	3.6960	-0.5968	0.9678
	II	3.7526	-0.5090	0.9513
C.-B.	I	3.9847	-0.5750	0.9621
	II	3.9814	-0.4708	0.8410
Canada	I	4.3458	-0.5936	0.9703
	II	4.2357	-0.5191	0.9699

TABLEAU 3: MOINDRES CARRÉS NON LINÉAIRES: MÉTHODE DE GAUSS-NEWTON

CANADA (GROUPE II)

Itération	A'	B'	Somme des carrés résiduelle
0	1.00000000	-0.50000000	3401.93232121
1	15.22076853	-0.23647629	461.76322678
2	26.47981387	-0.36743343	322.67707190
3	51.94184546	-0.51147529	248.68405130
4	57.29455529	-0.47434886	99.32440727
5	58.32558100	-0.48419609	96.57832290
6	58.28627964	-0.48409502	96.57810754
7	58.28746710	-0.48409960	96.57810746

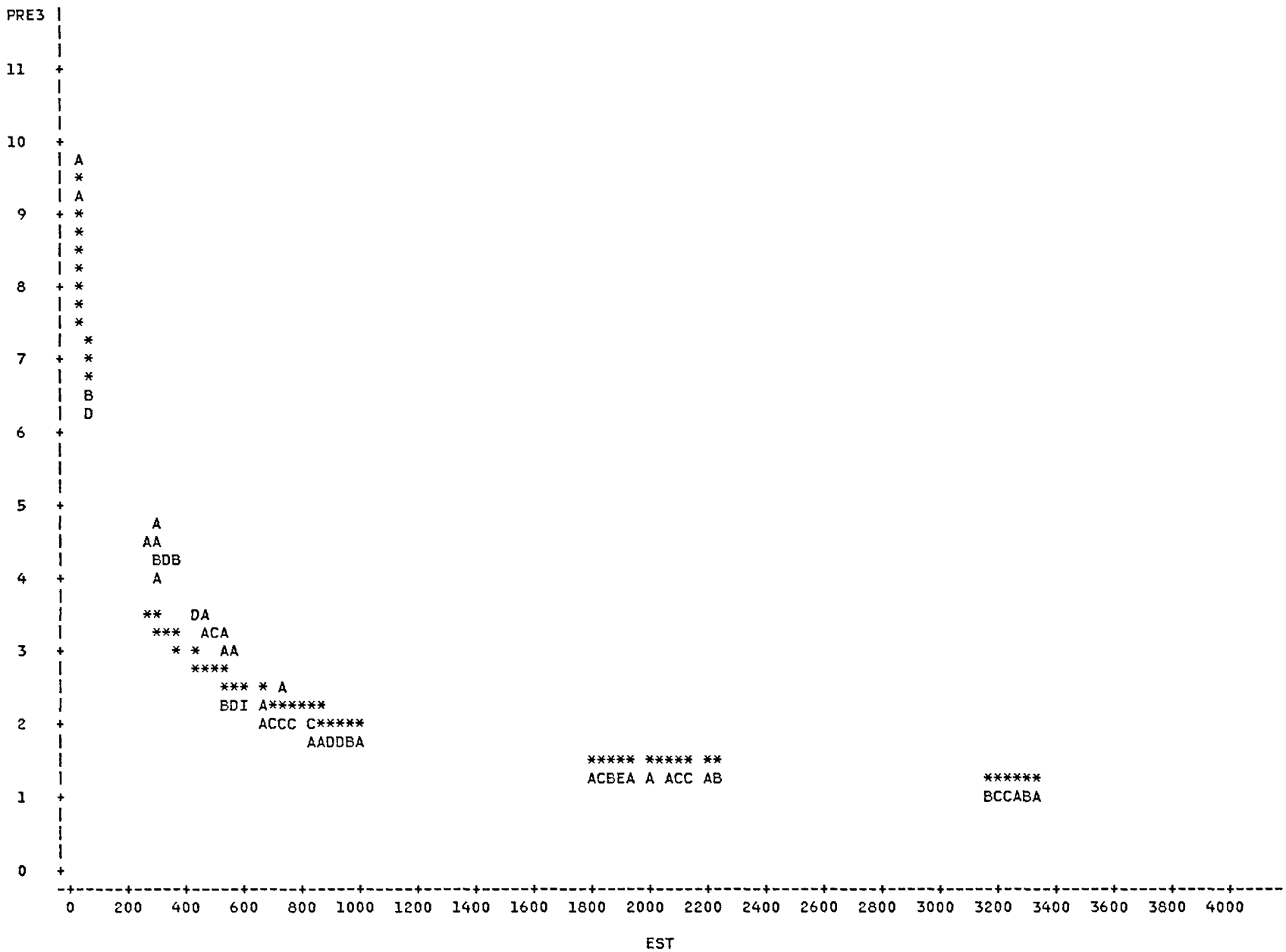
TABLEAU 4: R'^2 DANS LES GROUPES I ET II

Province	Groupe	N*	$R'^2 = 1 - \frac{\text{Somme des carrés résiduelle}}{\text{Somme des carrés totale}}$
Terre-Neuve	I	866	0.9362
	II	190	0.8835
Île-du-Prince-Édouard	I	827	0.8925
	II	294	0.7285
Nouvelle-Écosse	I	872	0.9790
	II	192	0.7813
Nouveau-Brunswick	I	908	0.9990
	II	156	0.8639
Québec	I	859	0.9800
	II	204	0.7804
Ontario	I	823	0.9632
	II	252	0.9208
Manitoba	I	895	0.9691
	II	168	0.8137
Saskatchewan	I	896	0.9436
	II	168	0.8196
Alberta	I	845	0.9701
	II	228	0.8852
C.-B.	I	868	0.9319
	II	204	0.7786
Canada	I	923	0.9665
	II	156	0.9286

* N dans le groupe I peut être inférieur à 12 (nombre de caractéristiques) à cause de l'exclusion des caractéristiques sans estimation.

#3 SIMPLE POWER MODEL
PROV=CANADA GROUP=2

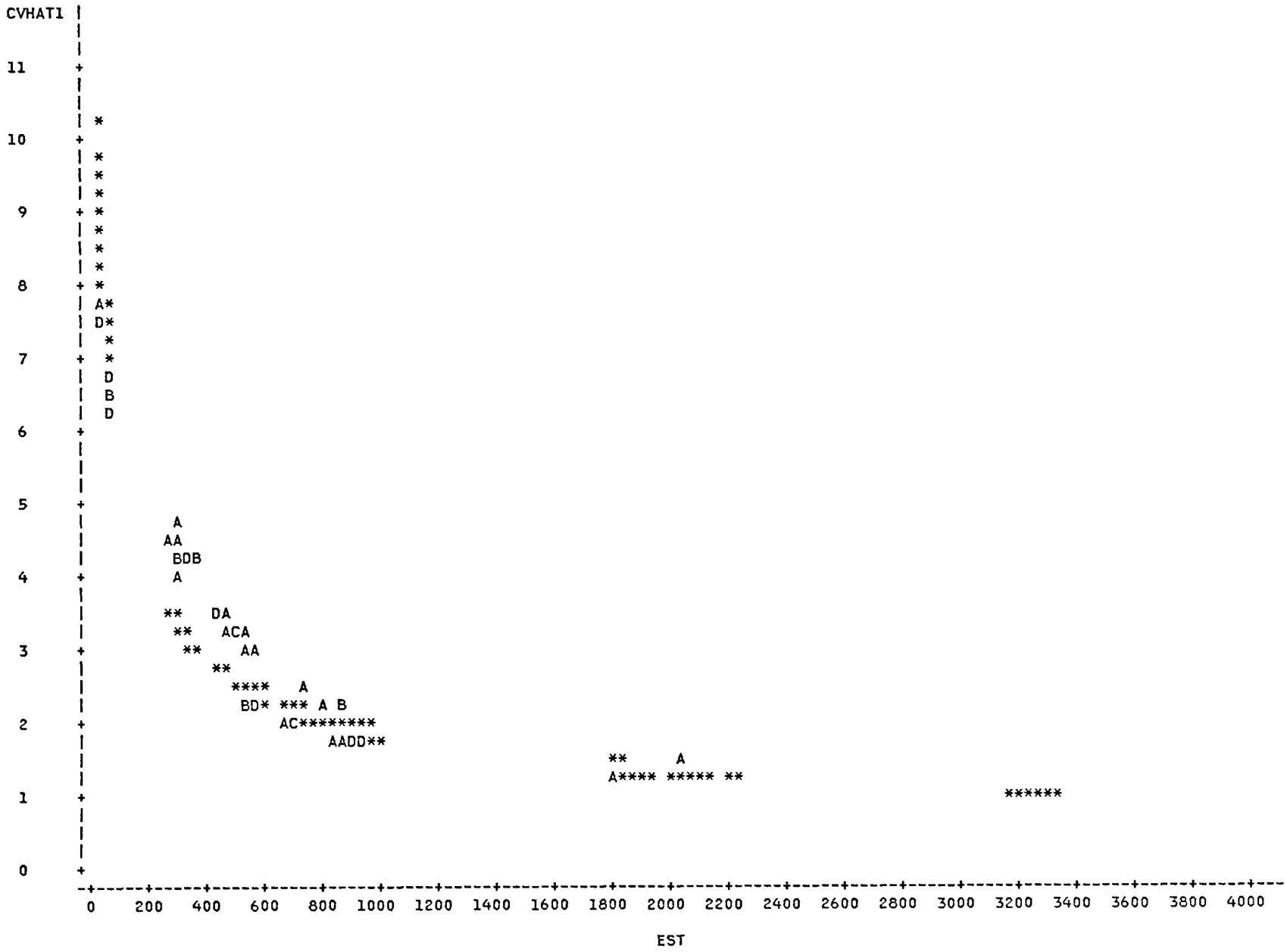
PLOT OF PRE3*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 101 OBS HIDDEN

#1 LOG-LINEAR MODEL
PROV=CANADA GROUP=2

PLOT OF CVHAT1*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 98 OBS HIDDEN

REMERCIEMENTS

L'auteur aimerait remercier l'arbitre pour ses commentaires et ses observations.

BIBLIOGRAPHIE

[1] Ghangurde, P.D. et Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics Theory and Methods, A 10(22), 2327-38.

[2] Gray, G.B. et Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Techniques d'enquête, vol. 2, n° 1, 1-30.

[3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", exposé présenté à une assemblée de l'International Association of Survey Statisticians.

[4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52, 503-510.

[5] Platek, R. et Singh, M.P. (1976), Méthodologie de l'enquête sur la population active, n° 71-526 au catalogue, hors-série.

[6] Sprent, P. (1969) "Models in Regresion and Related Topics", Methuen and Co.

Le comité de rédaction désire remercier M. R.E. Drover, Président, Comité de publication et le personnel de la Division des services administratifs pour leur appui constant dans le publication de cette revue.

De même que N. Brien, M. Fluet, P. Foy et G. Kriger pour leur aide lors de la révision des textes et la préparation de la version finale.

Nous remercions également Mme. D. Edirisinghe, qui a dactylographié soigneusement la revue, et qui s'est acquittée de nombreuses autres tâches associées à sa publication.

Le comité de rédaction désire enfin remercier les personnes suivantes, qui ont bien voulu faire la critique des articles présentés au cours de l'année dernière.

D.A. Binder
R.G. Carter
G.H. Choudhry
D.P. Dixon
J.D. Drew
S. Earwaker
M. Fluet
P. Foy
G.B. Gray
M.A. Hidiroglou
G. Kriger
S. Kumar
M.L. Lawes
I. Macredie
M.J. March
A. Satin
K.P. Srinath
L. Swain
P.F. Timmons