Neuchâtel, Septembre 2001

Statistique de la valeur ajoutée 1997-1998 : Procédure d'extrapolation des données

Rudi Peters, Jean-Pierre Renfer et Beat Hulliger

Service de méthodes statistiques
Dienst Statistische Methoden
Servizio metodi statistici

_____

Pour tout renseignement, contacter:

Jean-Pierre Renfer

Office fédéral de la statistique
Service de méthodes statistiques

Espace de l'Europe 10, CH-2010 Neuchâtel
Tel.: (+41 32) 71 36662
Fax: (+41 32) 71 36093
Bestellnummer: 338-0005
© BFS/OFS

La section de la production et du chiffre d'affaire (PROD) de l'Office fédéral de la statistique (OFS) mène chaque année auprès des entreprises l'enquête de la valeur ajoutée.

La statistique de la valeur ajoutée observe, au moyen de résultats comptables, la structure et l'évolution des activités des entreprises dans des secteurs d'activités économiques donnés. Elle permet tout d'abord de calculer des *données structurelles* et des *ratios micro-économiques* au niveau des branches. Les données de l'enquête de la valeur ajoutée permettent également d'estimer les *principaux agrégats macro-économiques du compte de production, du compte d'exploitation et, en partie, du compte de revenu pour la section des comptes nationaux (CN).*

Le service de méthodes statistiques (METH) a étudié les méthodes d'estimation, sur la base des données de l'enquête de la valeur ajoutée de 1997 et de 1998 (exercice d'activités 1997, respectivement 1998 des entreprises). Ce travail a été réalisé par MM. Rudi Peters et Beat Hulliger pour les données de 1997 puis repris par MM. Jean-Pierre Renfer et Beat Hulliger pour celles de 1998. Les résultats de leur travail sont rassemblés dans le présent rapport.

Les auteurs de ce rapport remercient particulièrement M. Lorenzo Cascioni et Mme Franziska Steiner de la section PROD. Leurs remerciements vont également à Mme Liliane Möri et M. Gabriel Gamez (CN) ainsi qu'à leurs successeurs MM. Gregory Rais et Philippe Pegoraro pour leur précieuse collaboration. Ils sont également reconnaissants à M. Markus Eichenberger de l'OFIT (Office fédéral de l'informatique et de la télécommunication) pour son soutien à la mise en application informatique des procédures d'estimation.

Table des matières

1 Introduction **1**

2 Pondération initiale **1**
2.1 **Poids de sondage** 3
2.2 **Traitement initial de la non-réponse** 3
2.3 **Calcul des poids initiaux** 4

3 Méthodes d'estimation **4**
3.1 **Forme de l'estimateur de base** 5
3.2 **Agglomération des petites strates** 5
3.3 **Traitement des très grandes entreprises** 6
3.4 **Hypothèses de modélisation** 7

4 Variance de l'estimateur **10**

5 Estimateurs pour les sous-domaines **12**
5.1 **Robustification de l'estimateur** 13
5.2 **Procédure de robustification** 14
5.3 **Hypothèses sous-jacentes à la robustification** 15
5.3.1 **Choix de l'estimateur à robustifier** 15
5.3.2 **Choix d'un modèle** 17
5.3.3 **Choix des variables à robustifier** 17
5.3.4 **Choix du domaine de robustification** 17
5.3.5 **Choix du degré de robustification** 18

6 Variance de l'estimateur robustifié **19**

7 Précision **21**

8 Bibliographie **22**

9 Annexe : macro SAS d'extrapolation **23**

1 Introduction

Les données de l'enquête sur la statistique de la valeur ajoutée résultent d'un plan d'échantillon stratifié (décrit dans le rapport de méthodes "Etablissement du plan d'échantillonnage pour l'enquête sur la statistique de la valeur ajoutée", Rudi Peters et Jean-Pierre Renfer, à paraître ainsi que dans le document technique "Enquête sur la production et la valeur ajoutée WS 99 : échantillonnage complémentaire", Jean-Pierre Renfer (mars 2000)). Pour pouvoir représenter la population de référence, les données doivent être pondérées. Dans un premier temps, nous décrivons le modèle de pondération initial en indiquant respectivement les corrections du biais de sélection et de non-réponse conduisant au calcul des poids initiaux. Les méthodes d'estimation appropriées sont ensuite décrites, notamment la procédure d'estimation des totaux. La forme de l'estimateur est présentée et discutée dans ce cadre. Certaines améliorations de l'estimation font aussi partie de ce paragraphe. La variance de l'estimateur retenu fait également l'objet d'un paragraphe dans lequel nous montrons comment l'estimer et comment l'adapter à des estimateurs par sous-domaine. De manière à se prémunir contre les effets des valeurs extrêmes dans les données, une robustification de l'estimateur est proposée qui conduit à des choix sur le modèle, le domaine, les variables et enfin sur le degré de robustification. Les deux derniers paragraphes sont consacrés à la variance de l'estimateur robustifié et à la précision ainsi obtenue.

2 Pondération initiale

Pour l'exercice d'activités 1997, respectivement 1998, nous disposons de 5'602, resp. 5'687 observations exploitables (entreprises d'au moins 2 personnes occupées, d'une activité économique et d'une forme juridique couvertes par l'enquête). Celles-ci se répartissent au plus en trois classes de taille du plan d'échantillonnage : pour cela une borne M a été calculée qui fixe le seuil de ce qui est considéré comme une grande entreprise (qui sont recensées), et ceci en adoptant le critère du nombre minimal d'observations nécessaires pour garantir un coefficient de variation de l'estimateur de 3%, au niveau de chaque division économique NOGA2 (activité économique exprimée par les deux premiers chiffres dans la nomenclature NOGA). Les trois classes de taille obtenues sont : 2-19 personnes occupées (petites entreprises), 20-M personnes occupées (moyennes entreprises), plus de M personnes occupées (grandes entreprises) ; à noter que la borne M est également calculée pour chaque division économique au niveau NOGA2.
La figure 1 ci-dessous permet de visualiser le taux de sondage par strate (nombre d'unités tirées/nombre d'unités de la population) et ceci par ordre croissant du taux de sondage, par classe de taille et selon l'activité économique NOGA2. Le groupe des 255 (en 1997) et 201 (en 1998) observations de l'activité économique 55.1 (l'hôtellerie) qui résultent d'un plan d'échantillonnage particulier, ne sont pas représentées dans ce graphique (cette catégorie a été codée en 56 au niveau NOGA2 du fait que 56 n'est pas réservé dans cette nomenclature, permettant ainsi de distinguer l'hôtellerie).
La figure 1 montre que les grandes entreprises ont été recensées, les moyennes entreprises

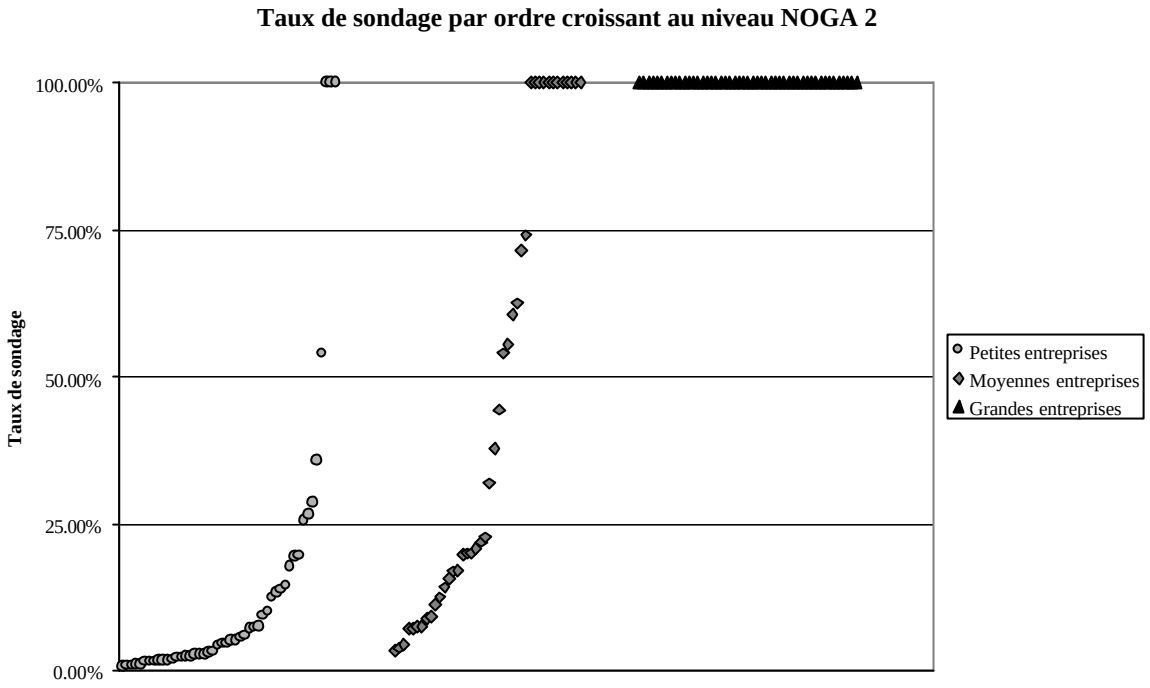


Figure 1: Taux de sondage en ordre croissant par strate.

ont été recensées dans certaines strates et tirées aléatoirement avec des taux de sondage relativement importants dans les autres strates. Enfin, les petites entreprises ont été tirées avec des taux de sondage inférieurs à 25%, à quelques exceptions près, notamment dans les strates comprenant peu d'observations (moins de 100 unités). Finalement, dans les strates comptant moins de 10 unités, les petites entreprises ont également été recensées.
Aux observations de l'échantillon, on associe des poids w_i. Ces derniers "redressent" les différences de probabilité d'inclusion et de taux de participation entre les entreprises. Les entreprises de l'échantillon sont en effet issues d'un tirage aléatoire stratifié et seule une partie des entreprises tirées ont accepté de répondre à l'enquête. Les poids permettent de redresser les données, de manière à tenir compte du plan d'échantillonnage (paragraphe 2.1) et de la non-réponse (paragraphe 2.2).

2.1 Poids de sondage

Les entreprises de l'échantillon ont été tirées aléatoirement par strate. Les strates ont été définies en fonction de l'activité économique (division NOGA2) et de la taille de l'entreprise (nombre de personnes occupées) telles qu'attribuées aux entreprises dans la base de sondage. Les erreurs de classification de l'activité économique dans la base de sondage ont pour effet de rendre les strates moins homogènes, ce qui a tendance à augmenter la variabilité des estimateurs mais n'affecte pas le biais des estimateurs développés ci-dessous.
Pour prendre en considération les taux de tirage différents par strate, nous redressons chaque observation i en lui attribuant un poids égal à l'inverse de la probabilité d'inclusion de la strate h à laquelle elle appartient :

$$w_{1i} = \frac{1}{\pi_i} = \frac{N_h}{n_h^b},$$

où n_h^b est le nombre (brut) d'entreprises tirées dans la strate h d'échantillonnage et N_h le nombre total d'entreprises présentes dans la strate h. Ainsi, dans l'exemple suivant, si 10 entreprises ont été tirées dans une strate de 50 entreprises, on attribue à chaque entreprise de la strate un poids de sondage $w_{1i} = \frac{1}{\pi_i} = \frac{50}{10} = 5$.

2.2 Traitement initial de la non-réponse

L'enquête n'est pas obligatoire et seule une partie des entreprises interrogées y répond. La non-réponse a été particulièrement importante au sein des strates des petites entreprises, augmentant parfois de façon très importante les poids des entreprises qui répondent. Pour pouvoir corriger la non-réponse, les entreprises répondantes d'une strate sont supposées former un sous-échantillon aléatoire des entreprises interrogées. Pour corriger la non-réponse, nous multiplions les observations d'une strate par l'inverse du taux de réponse observé dans la strate :

$$w_{2i} = \frac{1}{\tau_i} = \frac{n_h^b}{n_h},$$

avec n_h^b le nombre d'entreprises tirées dans la strate h et n_h le nombre d'entreprises de la strate h qui ont répondu à l'enquête.
A noter que l'hypothèse faite sur les entreprises répondantes (sous-échantillon aléatoire des entreprises interrogées) n'est en général satisfaite que partiellement. Cependant le facteur w_{2i} proposé permet de tenir compte des taux de réponse différents dans les strates et ainsi de tendre à réduire le biais induit par la non-réponse.
Un certain nombre de problèmes techniques se posent dans l'estimation du nombre d'entreprises répondantes :

- Certaines entreprises ont répondu à l'enquête mais ont vu leur nombre d'emplois équivalents plein temps (EPT, selon la définition de la statistique de l'emploi STATEM) passer sous le seuil de 2 unités ou leur division économique et/ou leur forme juridique non couverte par l'enquête ; ces entreprises doivent être écartées, lors de l'extrapolation, comme des entreprises ne faisant pas partie de la population d'étude, en forçant leur poids à 0 mais elles doivent compter comme entreprises répondantes dans l'estimation des taux de réponse.
- Les entreprises qui ont cessé leur activité ne sont pas enregistrées dans le fichier disponible à l'extrapolation ; pour des raisons de commodité de calcul, ces entreprises ne sont pas comptabilisées comme entreprises répondantes ; nous supposons que leur nombre équivaut, au niveau de chaque strate, au nombre de nouvelles entreprises (entreprises qui n'étaient pas présentes dans la base de sondage) et nous considérons que la sous-évaluation du taux de réponse compense la sous-couverture du cadre de sondage.

2.3 Calcul des poids initiaux

Le poids initial est le produit de l'inverse du taux de sondage par l'inverse du taux de réponse. Il s'exprime comme le rapport entre le nombre total d'entreprises dans la strate et le nombre d'entreprises répondant à l'enquête :

$$w_i = w_{1i} w_{2i} = \frac{1}{\pi_i} \frac{1}{\tau_i} = \frac{N_h}{n_h^b} \frac{n_h^b}{n_h} = \frac{N_h}{n_h}.$$

Ainsi, en reprenant l'exemple précédant, si sur les 10 entreprises interrogées dans la strate, 5 entreprises répondent à l'enquête, le poids $\frac{1}{\pi_i}$ de chacune des 5 observations est encore à multiplier par le facteur $\frac{10}{5} = 2$ pour tenir compte de la non-réponse au niveau des strates, soit un poids $w_i = \frac{50}{10} \frac{10}{5} = 10$.

3 Méthodes d'estimation

L'estimateur utilisé est l'estimateur par le quotient, qui se base sur une estimation de la grandeur moyenne par EPT et sur la connaissance que l'on a du nombre d'EPT au niveau de la population dans chaque division économique.
L'estimation se fait au niveau de la division économique NOGA2. Celle-ci correspond à l'activité principale NOGA2 de l'entreprise telle qu'elle ressort des informations récoltées par l'enquête ; celle-ci peut différer de celle inscrite dans la base de sondage, dont l'actualité ne peut être garantie. Nous considérons séparément le groupe des hôtels, étant donné qu'il a été traité séparément lors du tirage.

3.1 Forme de l'estimateur de base

L'estimateur par le quotient au niveau de la division économique c est calculé par le quotient des moyennes pondérées :

$$\hat{\beta}_c = \frac{\left(\frac{\sum_{S_c} w_i y_i}{\sum_{S_c} w_i}\right)}{\left(\frac{\sum_{S_c} w_i x_i}{\sum_{S_c} w_i}\right)} = \frac{\sum_{S_c} w_i y_i}{\sum_{S_c} w_i x_i},$$

où :
w_i est le poids associé à l'entreprise i ;
y_i est la variable d'étude pour l'entreprise i mesurée par l'enquête ;
x_i est le nombre d'EPT de l'entreprise i, calculé à partir du questionnaire ;
S_c est le sous-échantillon d'entreprises de division économique c.
En multipliant ce quotient par le nombre X_c d'EPT, on obtient le total Y_c de la variable étudiée au niveau de la division économique (estimateur par le quotient) :

$$\hat{Y}_c = \hat{\beta}_c X_c.$$

Du point de vue de la terminologie, X_c est parfois appelée valeur de calibration puisque l'estimateur par le quotient "calibre" les données sur le nombre d'EPT dans les différentes divisions économiques. A noter que ce calage permet de diminuer un éventuel biais de non-réponse dans le cas où celle-ci serait corrélée avec X_c.
Pour pouvoir appliquer cet estimateur, il faut pouvoir disposer du nombre total d'EPT X_c dans les différentes divisions économiques avec suffisamment de précision : nous utilisons ici les résultats de la STATEM, qui sont issus d'une enquête auprès d'environ 40'000 établissements. L'éventuelle différence dans les définitions d'EPT ainsi que la variabilité dûe à l'enquête STATEM n'est pas prise en compte, ni dans l'estimateur ni dans l'estimation de la variance.
Le total au niveau Suisse s'obtient comme la somme des totaux sur les différentes divisions économiques :

$$\hat{Y} = \sum_c \hat{Y}_c.$$

3.2 Agglomération des petites strates

La non-réponse a été importante, surtout auprès des petites entreprises. Nous avons récolté pour certaines strates très peu d'observations, voire aucune. Ceci rend problématique la correction de la non-réponse (nécessité d'au moins 1 observation), ainsi que le calcul de la variance (nécessité d'au moins 2 observations).

Nous regroupons les strates sans ou avec une seule observation à une ou plusieurs strates "voisines". C'est ainsi que pour les données 1997 toutes les strates de division NOGA2='16' et toutes les strates de division NOGA2='23' sont regroupées ; les strates de classes de taille 1 (petites entreprises) et 2 (moyennes entreprises) sont regroupées pour les divisions NOGA2='19', '37', 60', '61', 62' et '80'. Pour les données 1998, toutes les strates de division NOGA2='11', '16' et '71' sont regroupée. Les strates de classes de taille 1 et 2 sont regroupées pour les divisions NOGA2='20', '23', '41', '61', et '62'. Les poids des observations et la variance des estimateurs sont calculés sur la nouvelle stratification.

3.3 Traitement des très grandes entreprises

Nous considérons comme très grandes entreprises, les 268 entreprises suisses avec plus de 424 personnes occupées dans la population de sondage pour l'année 1997 et les 275 entreprises avec plus de 414 personnes occupées (correspondant aux 5% plus grandes entreprises de l'échantillon net, c'est-à-dire après non-réponse). Ces entreprises sont issues de strates interrogées de façon exhaustive (classe de taille égale à 3).
Ces très grandes entreprises nécessitent un traitement particulier :

- pour toutes ces entreprises, le mécanisme de réponse n'est pas nécessairement comparable aux autres entreprises de leur strate (erreur de modélisation de la non-réponse) ;
- les valeurs $\frac{y_i}{x_i}$ de ces très grandes entreprises s'écartent plus ou moins fortement des autres valeurs de leur strate et de leur division économique et augmentent la variabilité des estimateurs ;
- leur contribution au total est important.

Pour remédier à ces problèmes, il convient :

- de bien vérifier les valeurs fournies pour ces entreprises ;
- de créer pour chacune des divisions économiques NOGA2 possèdant de telles entreprises, une strate supplémentaire de classe de taille 4 ; les très grandes entreprises sont retirées de leur strate initiale (classe de taille 3) et affectées à cette nouvelle strate avec un poids égal à 1 (entreprises considérées comme seulement "représentatives" d'elles-mêmes) ;
- de sortir ces entreprises de l'estimateur par le quotient :

$$\hat{Y}_c = \hat{\beta}_{c^*} X_{c^*} + \sum_{c \setminus c^*} y_i,$$

où c^* désigne l'ensemble des entreprises de division économique c qui ne sont pas des très grandes entreprises. Quelques entreprises qui ne font pas partie de la population initiale et

dont leurs données ont été recueillies via leur société-mère, sont adjointes à la strate supplémentaire de la division économique NOGA2 correspondante. Cette procédure consistant à former des strates "surprises" a été proposée par Hidiroglou et Srinath (1981).
Nous prenons comme valeurs X_{c^*} de calibrage, le nombre d'EPT par division économique (d'après la STATEM) moins le nombre d'EPT dans les très grandes entreprises (d'après les tailles fournies sur les formulaires de réponse).

3.4 Hypothèses de modélisation

L'estimateur par le quotient suppose un modèle linéaire entre la variable d'étude y_i et le nombre x_i d'EPT : $y_i = \beta_{c^*} x_i + \varepsilon_i$. L'estimateur est asymptotiquement non biaisé sous le modèle $E(\epsilon_i) = 0$, c'est-à-dire si la droite de régression passe par l'origine. En effet, sous la distribution d'échantillonnage en question, il y a un petit biais que l'on peut négliger si l'échantillon est suffisamment grand. A noter que le biais augmente à mesure que l'ajustement du modèle se péjore. A titre d'exemple, nous montrons sur la figure 2 la relation dans une division économique c^* entre la variable T017, production totale, et la variable s14, le nombre d'EPT, observée sur les entreprises de l'échantillon de 1997.
A noter que sur la figure 2 certaines observations ne suivent pas la relation linéaire entre les deux variables et apparaissent comme des données aberrantes. Ces données seront traitées selon une procédure décrite dans le paragraphe consacré à la robustification. Il faut également souligner que la validité de cette relation peut être contestée pour certaines variables d'étude, rendant l'estimateur seulement asymptotiquement non biaisé conditionnellement au modèle (taille d'échantillon suffisamment grande).
Il est donc bon de s'assurer de disposer d'un nombre suffisant d'observations par division économique. Ceci est également nécessaire pour pouvoir estimer $\hat{\beta}_{c^*}$ avec suffisamment de fiabilité. La figure 3 fournit l'information sur les tailles disponibles dans chacune des divisions économiques (divisions NOGA sans les très grandes entreprises). Par exemple, l'échantillon de 1997 contient 5 divisions économiques (les divisions de code NOGA valant 23, 30, 41, 62, 64) de moins de 15 observations et de taux de couverture (le nombre d'entreprises dans l'échantillon divisé par le nombre total d'entreprises dans la population) inférieur à 50%, pour lesquelles les estimations doivent être prises avec la plus grande réserve. Exactement le même problème se pose avec l'échantillon de 1998.

L'estimateur du ratio appliqué est le quotient des moyennes pondérées. Nous supposons une variance des résidus proportionnelle à la taille de l'entreprise ($var(\epsilon_i) = x_i \sigma^2_{c^*}$), au niveau de chaque division économique, c'est-à-dire une dispersion des résidus augmentant de manière "parabolique" avec la taille de l'entreprise. D'autres hypothèses sur la variabilité, nous auraient donné à préférer, pour des raisons de variance, d'autres estimateurs du ratio. Par exemple, il aurait été préférable (moindre variance) d'utiliser :

- l'estimateur $\hat{\beta}_{c^*} = \frac{\sum_{S_{c^*}} w_i x_i y_i}{\sum_{S_{c^*}} w_i x_i^2}$ des moindres carrés pondérés, dans le cas de résidus donnés par $\varepsilon_i = y_i - \beta_{c^*} x_i$ de variance constante ($var(\epsilon_i) = \sigma^2_{c^*}$), c'est-à-dire de dispersion constante quelle que soit la taille de l'entreprise ;

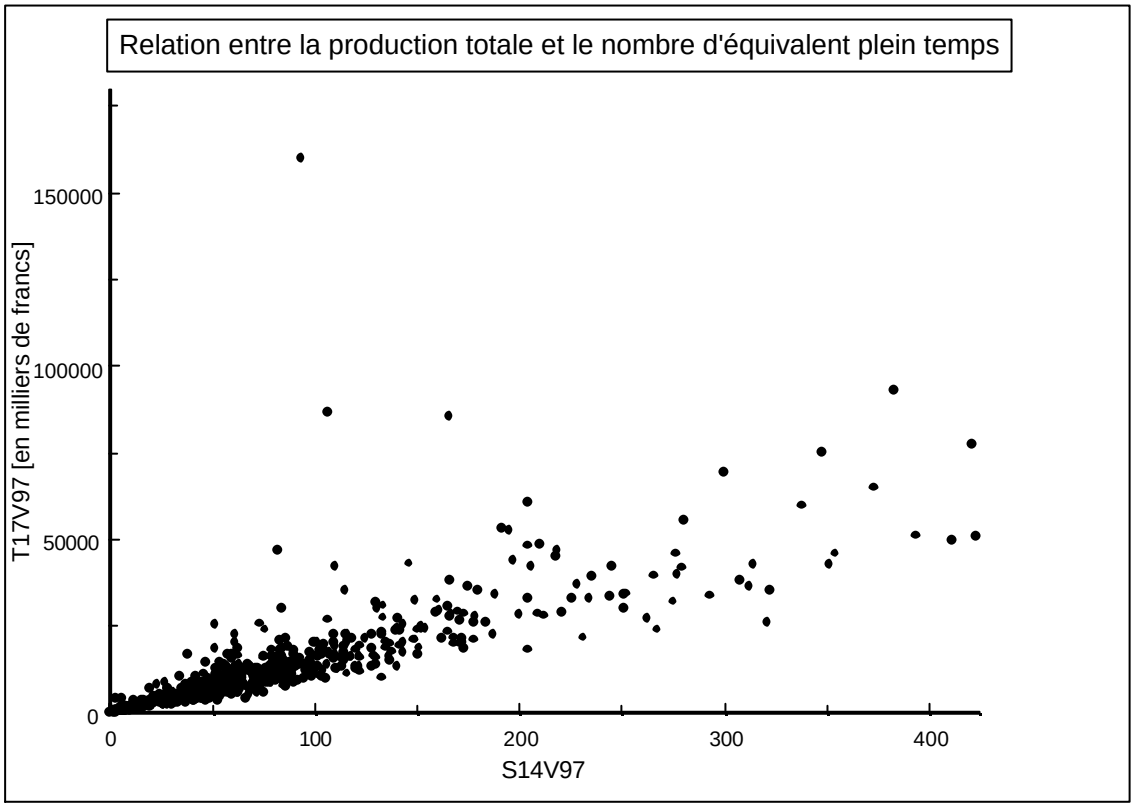


Figure 2: Production totale et nombre d'emplois équivalents plein temps dans une division économique.

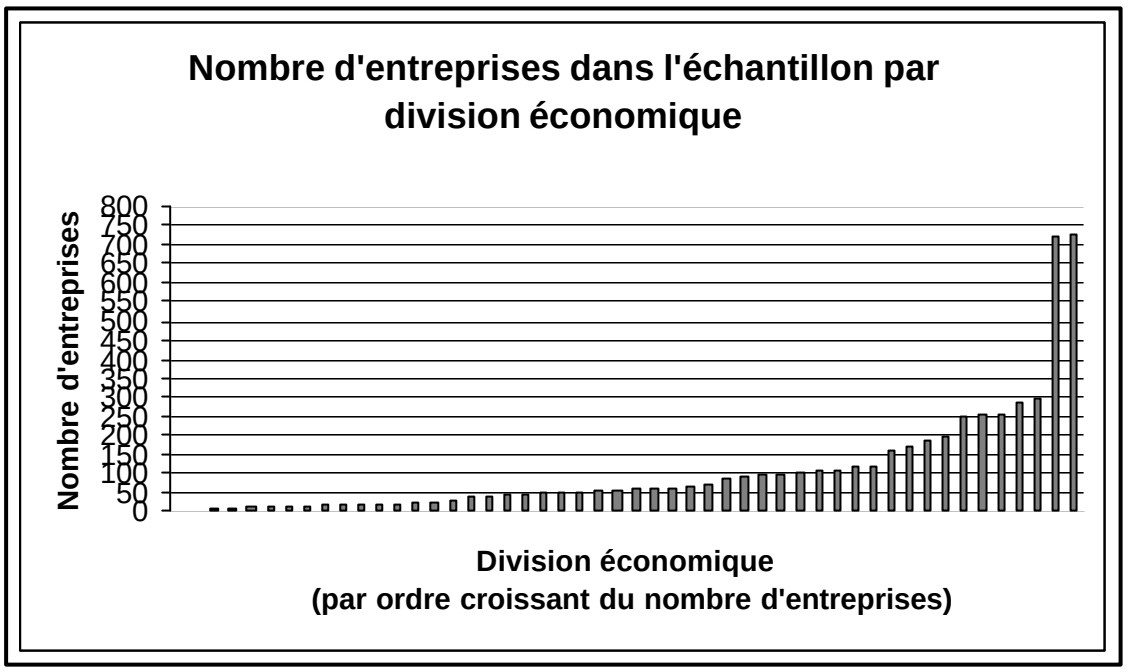


Figure 3: Nombre d'entreprises de l'échantillon de 1997 par division économique.

- l'estimateur $\hat{\beta}_{c^*} = \frac{\sum_{S_{c^*}} w_i \frac{y_i}{x_i}}{\sum_{S_{c^*}} w_i}$ de la moyenne pondérée des quotients, dans le cas de résidus $\varepsilon_i = y_i - \beta_{c^*} x_i$ de variance proportionnelle au carré de la taille de l'entreprise ($var(\epsilon_i) = x_i^2 \sigma_{c^*}^2$), c'est-à-dire de dispersion augmentant linéairement avec la taille de l'entreprise ;
- l'estimateur $\hat{\beta}_{c^*} = \exp(\frac{\sum_{S_{c^*}} w_i \ln(\frac{y_i}{x_i})}{\sum_{S_{c^*}} w_i}) = \prod_{S_{c^*}} (\frac{y_i}{x_i})^{\frac{w_i}{\sum_{S_{c^*}} w_i}}$ de la moyenne géométrique pondérée, dans le cas de résidus ε_i', définis par la relation $y_i = \beta_{c^*} x_i \varepsilon_i'$, de variance constante en logarithme ($var(\ln(\epsilon_i)) = \sigma_{c^*}^2$), c'est-à-dire de dispersion augmentant exponentiellement avec la taille de l'entreprise.

En général, on peut accepter une augmentation de la dispersion des résidus de manière parabolique. Nous avons choisi, par souci d'uniformité, l'estimateur du quotient des moyennes pondérées, quelle que soit la division économique et la variable d'étude considérées ; il se peut toutefois que, pour certaines divisions économiques ou certaines variables d'étude, un autre estimateur aurait été de moindre variance, sans compter que la relation linéaire entre x_i et y_i n'est pas toujours appropriée.

4 Variance de l'estimateur

La variance mesure l'erreur aléatoire de l'estimateur, dûe au caractère aléatoire de :

- l'échantillon : un échantillon aléatoire stratifié au niveau des classes économiques NOGA2 et des classes de taille a été tiré (voir le rapport de méthodes sur le plan d'échantillonnage, Rudi Peters et Jean-Pierre Renfer, à paraître) ;

- la non-réponse : les entreprises répondantes sont supposées former par strate un sous-échantillon aléatoire des entreprises tirées ; la variance de l'estimateur est calculée conditionnellement aux tailles de l'échantillon net (après non-réponse) dans les différentes strates (cette estimation néglige le caractère également aléatoire du nombre d'entreprises répondantes et des déviations de l'hypothèse de l'échantillon aléatoire).

- la mauvaise classification de certaines entreprises dans la base de sondage (environ 5% des entreprises ont une division économique différente de celle mentionnée dans la base de sondage : ces différences apparaissent sur la figure 4 pour les données de 1997) ; la variabilité dûe à la mauvaise classification est prise en compte dans l'estimation de la variance en utilisant une variable indicatrice I_{ic^*} valant 1 si l'entreprise i appartient à la division considérée c^* et 0 sinon.

La variance du total pour la division économique c se déduit de la variance du ratio $\hat{\beta}_{c^*}$:

$$var(\hat{Y}_c) = var(\hat{\beta}_{c^*} X_{c^*} + \sum_{c \setminus c^*} y_i) = var(\hat{\beta}_{c^*} X_{c^*}) = X_{c^*}^2 var(\hat{\beta}_{c^*}).$$

Puisque les très grandes entreprises sont supposées complètement recensées, ces entreprises n'interviennent pas dans la formule de la variance.
La variance du total $\sum_c \hat{Y}_c$ au niveau Suisse est donnée par la somme des variances sur les différentes divisions économiques, c'est-à-dire :

$$var(\hat{Y}) = \sum_c var(\hat{Y}_c).$$

La variance de l'estimateur $\hat{\beta}_{c^*}$ est calculée sur la forme linéarisée de l'estimateur :

$$var(\hat{\beta}_{c^*}) \approx \frac{1}{(\sum_S w_i x_i I_{ic^*})^2} var(\sum_S w_i e_i I_{ic^*}),$$

où $e_i = (y_i - \hat{\beta}_{c^*} x_i)$ est le résidu associé à l'entreprise i et S_{c^*} est l'échantillon des entreprises, à l'exclusion des très grandes.
En considérant les observations recueillies comme des observations issues d'un tirage aléatoire stratifié, la variance de la somme pondérée $\sum_S w_i e_i I_{ic^*}$ des erreurs peut être estimée par :

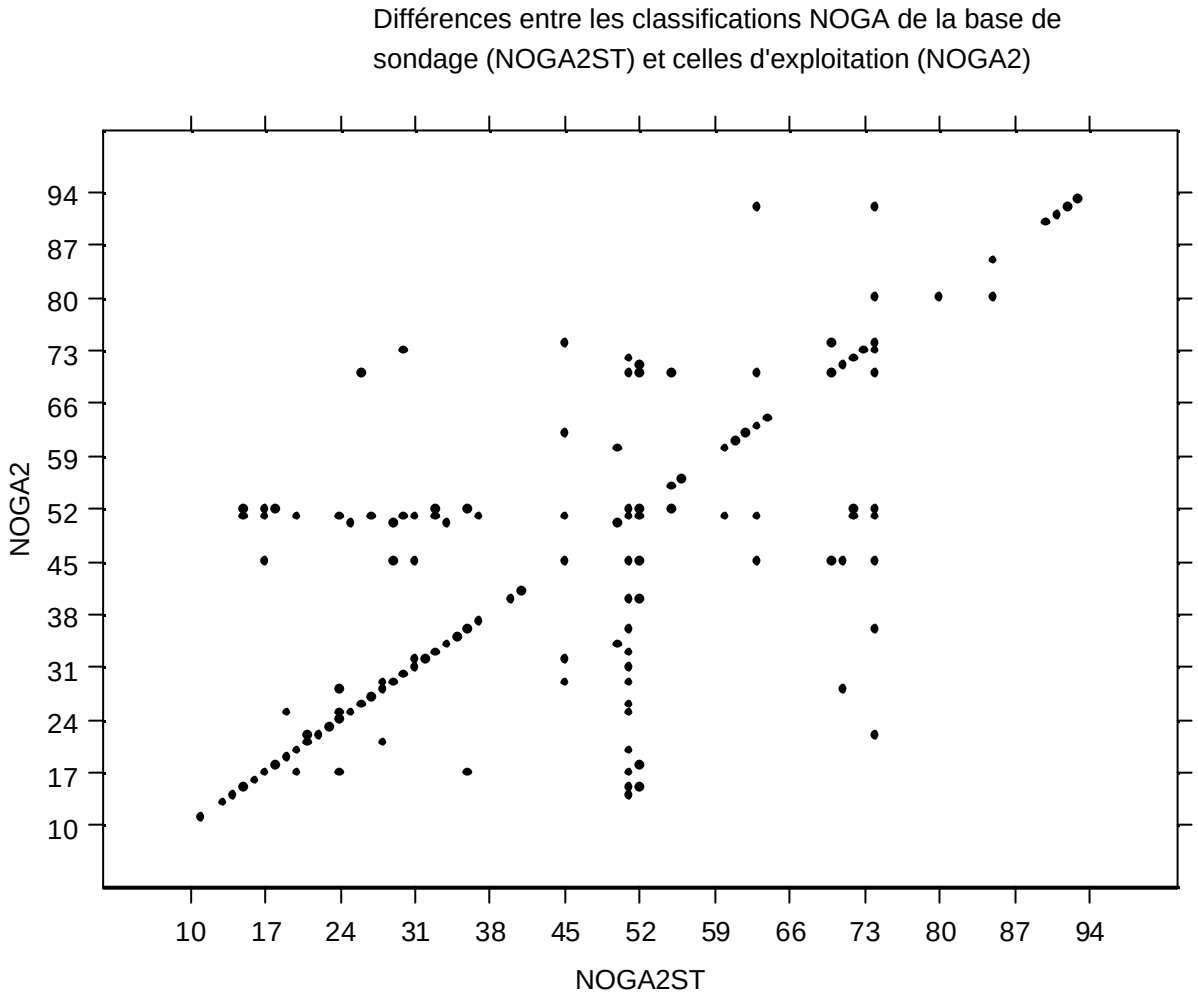


Figure 4: Classification des activités économiques pour les données de 1997.

$$var(\sum_{S} w_i e_i I_{ic^*}) = \sum_{h=1}^{H} (1 - f_h) n_h D^2_{h,we},$$

où $D^2_{h,we}$est la variance estimée de la variable $w_i e_i$ sur la strate h :

$$D^2_{h,we} = \frac{1}{n_h - 1} \sum_{S_h} (w_i e_i I_{ic^*} - \frac{1}{n_h} \sum_{S_h} w_i e_i I_{ic^*})^2,$$

et où le facteur $1 - f_h$ est le facteur qui prend en considération la dimension finie de la population de tirage et $S^*_h = S_h \cap S_{c^*}$.

Si le tirage est sans remise, que le taux de tirage est constant par strate et que w_i est constant dans S_h^*, la valeur f_h correspond au taux de tirage dans la strate, soit $f_h = \frac{n_h}{N_h}$; dans ce cas, cette expression peut aussi s'écrire de façon plus compacte comme :

$$\begin{aligned} var(\sum_S w_i e_i I_{ic^*}) &= \sum_{h=1}^{H} \sum_i w_i(w_i-1)\frac{n_h}{n_h-1}(e_i I_{ic^*} - \frac{1}{n_h}\sum_{S_h} e_i I_{ic^*})^2 \\ &\approx \sum_{h=1}^{H} \sum_i w_i(w_i-1)(e_i I_{ic^*} - \frac{1}{n_h}\sum_{S_h} e_i I_{ic^*})^2. \end{aligned}$$

Si le tirage est supposé avec remise, l'expression se simplifie en :

$$\begin{aligned} var(\sum_S w_i e_i I_{ic^*}) &\approx \sum_{h=1}^{H} \frac{n_h}{n_h-1} \sum_{S_h}(w_i e_i I_{ic^*} - \frac{1}{n_h}\sum_{S_h} w_i e_i I_{ic^*})^2 \\ &\approx \sum_{h=1}^{H} \sum_{S_h}(w_i e_i I_{ic^*} - \frac{1}{n_h}\sum_{S_h} w_i e_i I_{ic^*})^2. \end{aligned}$$

On peut montrer que cette dernière expression s'applique encore, dans le cas d'un tirage avec remise, si les probabilités d'inclusion $\frac{1}{w_i}$ des différentes observations de la strate ne sont pas égales.

5 Estimateurs pour les sous-domaines

Les totaux par division économique peuvent se ventiler en sous-domaines. Ainsi, certaines variables d'étude doivent être analysées :

- au niveau de la forme juridique (les entreprises sont classées suivant 4 groupes de formes juridiques différentes) ;
- au niveau de la classe de taille (les entreprises sont classées suivant les 4 classes de taille : 2-49, 50-99, 100-249 et 250+ personnes occupées).

Le total par sous-domaine k s'estime avec les même estimateurs que précédemment, en prenant simplement comme variable d'étude la variable y'_i, qui vaut y_i si l'entreprise appartient au sous-domaine k et 0 sinon.

Cet estimateur correspond à l'estimateur qui ventile le total d'une division économique entre les différents sous-domaines, en fonction de la part estimée dans chacun des sous-domaines :

$$\begin{aligned}
\hat{Y}_{k,c^*} &= \frac{\sum_{S_{c^*}} w_i y'_i}{\sum_{S_{c^*}} w_i x_i} X_{c^*} \\
&= \frac{\sum_{S_{k,c^*}} w_i y_i}{\sum_{S_{c^*}} w_i x_i} X_{c^*} \\
&= \frac{\sum_{S_{k,c^*}} w_i y_i}{\sum_{S_{c^*}} w_i y_i} \frac{\sum_{S_{c^*}} w_i y_i}{\sum_{S_{c^*}} w_i x_i} X_{c^*} \\
&= \widehat{p}_{k,c^*} \hat{Y}_{c^*} = \widehat{p}_{k,c^*} \hat{\beta}_{c^*} X_{c^*}.
\end{aligned}$$

Cet estimateur a l'avantage de garantir l'additivité des estimations. En effet, la somme des estimations sur les différents sous-domaines d'une division économique est égale à la valeur estimée sur toute la division économique : $\sum_k \hat{Y}_{k,c^*} = \hat{Y}_{c^*}$.
Pour autant que la fiabilité des nombres X_{k,c^*} d'EPT au niveau croisé du sous-domaine et de la division économique peut être assurée, une alternative consisterait à utiliser l'estimateur quotient au niveau du sous-domaine k et de la division économique c^* :

$$\hat{Y}_{k,c^*} = \hat{\beta}_{k,c^*} X_{k,c^*} = \frac{\sum_{S_{k,c^*}} w_i y_i}{\sum_{S_{k,c^*}} w_i x_i} X_{k,c^*}.$$

Le nombre d'observations par division économique et par sous-domaine peut être trop fin pour assurer une fiabilité suffisante à l'estimateur du ratio $\hat{\beta}_{k,c^*}$ (avec le risque d'une augmentation à la fois du biais et de la variance). Une possibilité pour remédier à ce problème consiste à utiliser un estimateur "synthétique" qui supposerait les ratios $\hat{\beta}_{k,c^*} = \hat{\beta}_{c^*}$ identiques dans les différents sous-domaines d'une même division économique. L'additivité $\sum_k \hat{Y}_{k,c^*} = \hat{Y}_{c^*}$ y est également vérifiée. Nous pouvons en fait considérer l'approche envisagée ici comme cet estimateur synthétique, dans lequel le nombre X_{k,c^*} d'EPT (au niveau croisé du sous-domaine et de la division économique) est non pas supposé connu mais également estimé par $\hat{X}_{k,c^*} = \frac{\sum_{S_{k,c^*}} w_i y_i}{\sum_{S_{c^*}} w_i y_i} X_{c^*}$.

5.1 Robustification de l'estimateur

Quelques entreprises peuvent avoir une influence "dominante" dans l'estimation parce que leurs valeurs s'écartent fortement de celles des autres entreprises. Ces valeurs "aberrantes" peuvent être considérées soit comme des valeurs "correctes" (dont d'autres pourraient appartenir à la population) soit comme des valeurs "incorrectes" dûes à des erreurs de traitement. Chambers (1986) parle dans le premier cas de valeurs aberrantes "représentatives" et dans le second cas de valeurs aberrantes "non-représentatives".

Les estimateurs de totaux et les variances correspondantes sont très sensibles à ces valeurs extrêmes. Les valeurs qui s'écartent "fortement" de celles observées sur des entreprises "similaires" vont être modifiées. Nous expliquons ci-après une procédure de robustification développée par Hulliger (1995) et qui sera appliquée dans notre cas.

5.2 Procédure de robustification

Nous supposons que les valeurs observées y_i suivent un modèle $y_i = \beta_{c^*} x_i + \varepsilon_i$, avec $E(\epsilon_i) = 0$ et $var(\epsilon_i) = x_i \sigma^2_{c^*}$, au niveau de chaque division économique c^*. Nous appliquons ensuite les estimateurs à un pas adaptés à l'échantillonnage décrits par Hulliger (1995, 1999).
Une valeur est considérée comme "extrême" si le résidu $\varepsilon_i = y_i - \beta_{c^*} x_i$ dépasse une certaine limite donnée en valeur absolue par $k_{c^*}\sqrt{x_i}\sigma_{c^*}$, où k_{c^*} est la constante de robustification, à fixer pour chaque division c^*.
Nous réduisons l'influence des valeurs extrêmes dans l'estimateur du ratio en bornant leur résidu à $k_{c^*}\sqrt{x_i}\sigma_{c^*}$ en valeur absolue. Pour cela, nous multiplions les valeurs x_i et y_i des valeurs extrêmes par un facteur u_i égal à $\frac{k_{c^*}\sqrt{x_i}\sigma_{c^*}}{|\varepsilon_i|}$; les valeurs non extrêmes sont laissées inchangées (en gardant le facteur u_i de correction égal à 1). L'estimateur du ratio robustifié s'écrit ainsi de la manière suivante :

$$\hat{\beta}^r_{c^*} = \frac{\sum_{S_{c^*}} w_i u_i y_i}{\sum_{S_{c^*}} w_i u_i x_i},$$

Pour estimer β_{c^*}, ε_i, σ_{c^*} et u_i, nous utilisons des estimateurs robustes aux valeurs extrêmes :

- β_{c^*} est estimé par le rapport des médianes pondérées de y_i et x_i :

$$\hat{\beta}^{(1)}_{c^*} = \frac{med(y_i, w_i)}{med(x_i, w_i)},$$

où la médiane pondérée est définie par $med(y_i, w_i) = y_{(j_d)}$, avec : $j_d = \min\{j : k_j \geq 0.5\}$ et $k_j = \sum_{i=1}^{j} w_{[i]} / \sum_{i=1}^{n} w_i$. A noter que les $w_{[i]}$ sont ici ordonnés selon l'ordre croissant des y_i.

- ε_i est estimé par :

$$\hat{e}^{(1)}_i = y_i - \hat{\beta}^{(1)}_{c^*} x_i,$$

- σ_{c^*} est estimé par la médiane pondérée des résidus standardisés en valeur absolue (en supposant la distribution proche de la loi normale) :

$$\hat{\sigma}^{(1)}_{c^*} = \frac{1}{0.67} med\left(\frac{|\,\hat{e}^{(1)}_i\,|}{\sqrt{x_i}}, w_i\right),$$

- u_i est estimé par

$$\begin{aligned} \hat{u}_i^{(1)} &= \frac{k_{c^*}\sqrt{x_i}\hat{\sigma}_{c^*}^{(1)}}{\mid \hat{e}_i^{(1)} \mid}, \text{ si } \left|\hat{e}_i^{(1)}\right| > k_{c^*}\sqrt{x_i}\hat{\sigma}_{c^*}^{(1)} \\ &= 1, \text{ sinon.} \end{aligned}$$

Les estimations des paramètres β_{c^*}, ε_i, σ_{c^*}et u_i peuvent être raffinées de façon itérative, en remplaçant $\hat{\beta}_{c^*}^{(1)}$ par $\hat{\beta}_{c^*}^{r}$, $\hat{e}_i^{(1)}$ par $\hat{e}_i^{r}$, $\hat{\sigma}_{c^*}^{(1)}$ par $\hat{\sigma}_{c^*}^{r}$ et $\hat{u}_i^{(1)}$ par $\hat{u}_i^{r}$. On montre qu'en procédant ainsi, l'estimateur du ratio robustifié converge vers un estimateur robuste de type M. Pour des raisons de commodité, nous nous limitons aux estimateurs initiaux (one-step ratio algorithm). Au sein d'une même division économique, les poids diffèrent suivant la classe de taille, voire la classe économique (entreprise mal codée dans la base de sondage), d'où est tirée l'entreprise. Certaines entreprises (principalement les petites) ont des poids largement dominants par rapport aux entreprises de leur division économique comme le montre la figure 5 et vont donc influencer énormément les estimateurs $\hat{\beta}_{c^*}^{(1)}$et $\hat{\sigma}_{c^*}^{(1)}$. En fait, le poids de certaines entreprises peut même être plus grand que la moitié du poids total de sa division économique et ainsi réduire à néant la robustesse de la médiane pondérée!
Afin de "robustifier" les estimateurs $\hat{\beta}_{c^*}^{(1)}$et $\hat{\sigma}_{c^*}^{(1)}$ face aux poids dominants, nous avons finalement décidé de les estimer avec les médianes non pondérées (même si l'inconvénient de ces estimateurs est leur non-consistance) :

$$\widetilde{\beta}_{c^*}^{(1)} = \frac{med(y_i)}{med(x_i)}$$

et

$$\widetilde{\sigma}_{c^*}^{(1)} = \frac{1}{0.67} med(\frac{\mid \hat{e}_i^{(1)} \mid}{\sqrt{x_i}}).$$

Pour des sous-domaines k, on peut utiliser $\hat{\beta}_{c^*}^{r}$ dans l'estimation de $\hat{Y}_{k,c^*} = \widehat{p}_{k,c^*}\hat{\beta}_{c^*}X_{c^*}$ au lieu de $\hat{\beta}_{c^*}$. L'inconvénient majeur de cet estimateur est son manque de robustesse face aux valeurs de y_i qui dominent dans $\widehat{p}_{k,c^*} = \frac{\sum_{S_{k,c^*}} w_i y_i}{\sum_{S_{c^*}} w_i y_i}$ (même si $\widehat{p}_{k,c^*}$ est évidemment borné par 1).
Il est possible d'utiliser une robustification pour chaque sous-domaine mais malheureusement au détriment de l'additivité des estimateurs.

5.3 Hypothèses sous-jacentes à la robustification

5.3.1 Choix de l'estimateur à robustifier

La robustification proposée est conçue pour améliorer l'estimateur $\hat{\beta}_{c^*}$ du quotient au niveau de la division économique. Une autre robustification aurait été nécessaire si un autre estimateur avait été utilisé.

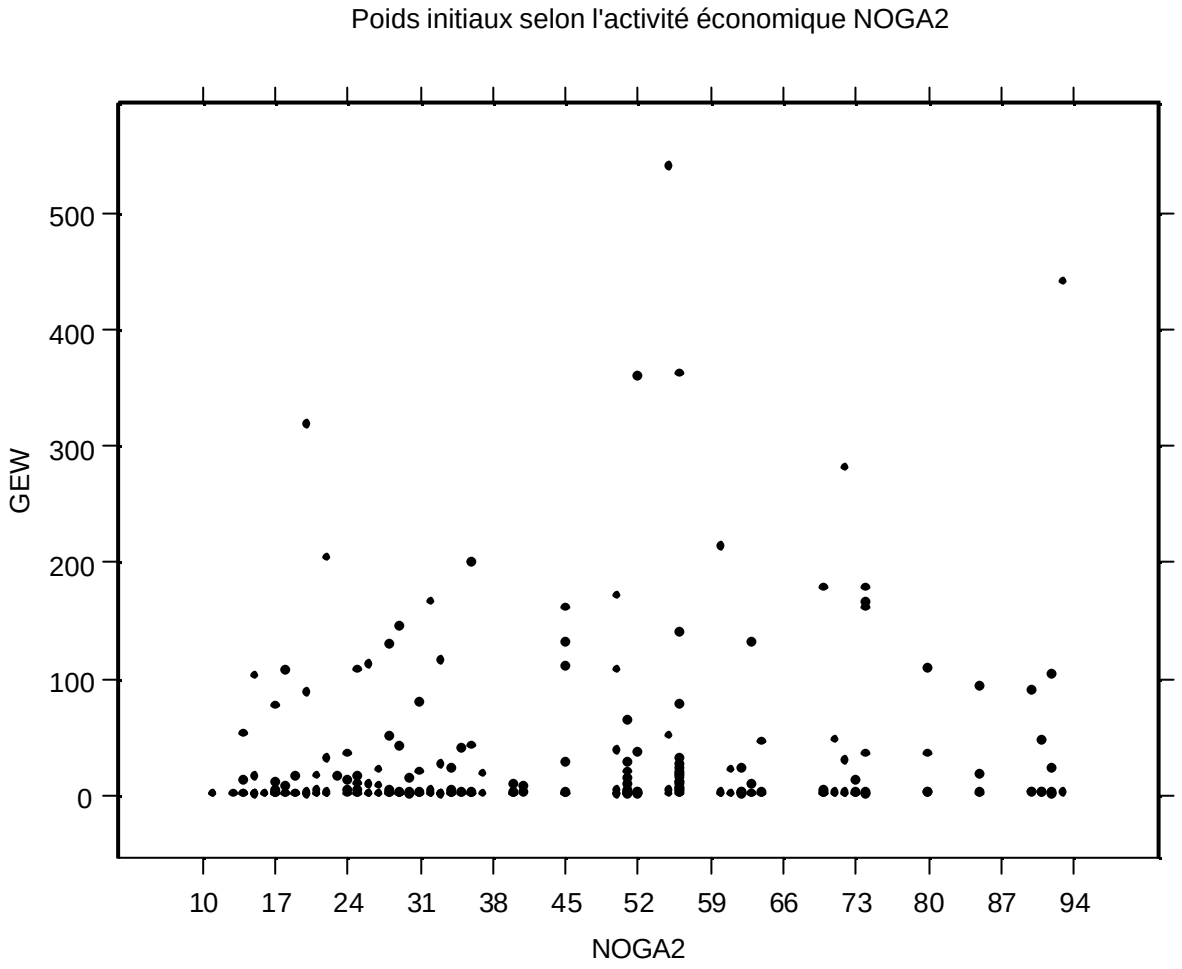


Figure 5: Poids initiaux des entreprises selon l'activité économique pour les données de 1997.

5.3.2 Choix d'un modèle

La robustification repose sur une modélisation de la variable d'étude : $y_i = \beta_{c^*} x_i + \varepsilon_i$, avec $E(\epsilon_i) = 0$ et $var(\epsilon_i) = x_i \sigma^2_{c^*}$. Si la réalité s'écarte fortement de cette modélisation (par exemple, si une division économique englobe des activités assez différentes), la robustification devient inefficiente et peut même introduire un biais relativement important. Notons qu'en principe, pour pouvoir estimer avec fiabilité les différents paramètres, il faut disposer d'un nombre suffisant d'observations (disons au moins 15 observations par domaine de robustification). Or, certaines divisions économiques ont peu d'observations et les résultats de la robustification sont alors à considérer avec précaution.
Nous ne prenons pas en considération la partie excédentaire des résidus dans l'extrapolation des données ; nous la considérons soit comme suspect, résultant d'une plausibilisation insuffisante, soit comme résultant d'une réalisation d'une distribution normale, possible mais trop "rare" pour devoir intervenir dans l'estimation, soit comme négligeable parce que leur nombre est faible et que les résidus extrêmes négatifs pourraient partiellement compenser les résidus extrêmes positifs. Cependant d'autres approches sont possibles.

5.3.3 Choix des variables à robustifier

La correction u_i de robustification dépend de la variable y_i sur laquelle la robustification est appliquée. On voudrait cependant disposer d'un seul jeu de facteurs de correction u_i quelque soit la variable d'étude car cela facilite l'implémentation informatique et permet une meilleure cohérence économique entre les différentes estimations réalisées.
Nous calculons les facteurs de correction u_i sur les 3 variables, qui sont jugées d'intérêt primordial pour l'enquête, à savoir la valeur de la production brute $u_i^{(1)}$, la rémunération des salariés $u_i^{(2)}$ et la consommation intermédiaire $u_i^{(3)}$. Nous prenons comme facteur de correction le minimum des facteurs de correction $u_i^{(1)}$, $u_i^{(2)}$ et $u_i^{(3)}$ de manière à corriger les influences trop importantes sur ces trois variables : $u_i = \min(u_i^{(1)}, u_i^{(2)}, u_i^{(3)})$. Nous espérons que ces facteurs permettent une robustification raisonnable sur les autres variables (pour une discussion sur ce point, voir Hulliger (1999)).

5.3.4 Choix du domaine de robustification

La correction u_i de robustification dépend du domaine sur lequel la robustification est appliquée. Nous choisissons comme domaine de robustification la division économique NOGA2. Nous excluons toutefois de la robustification les très grandes entreprises, que nous supposons parfaitement plausibilisées et qui ne devraient pas poser de problèmes de robustification dans l'estimateur (ces dernières sont comptées de façon exhaustive avec un poids unitaire et ne sont ainsi pas prises en compte pour estimer les quotients).

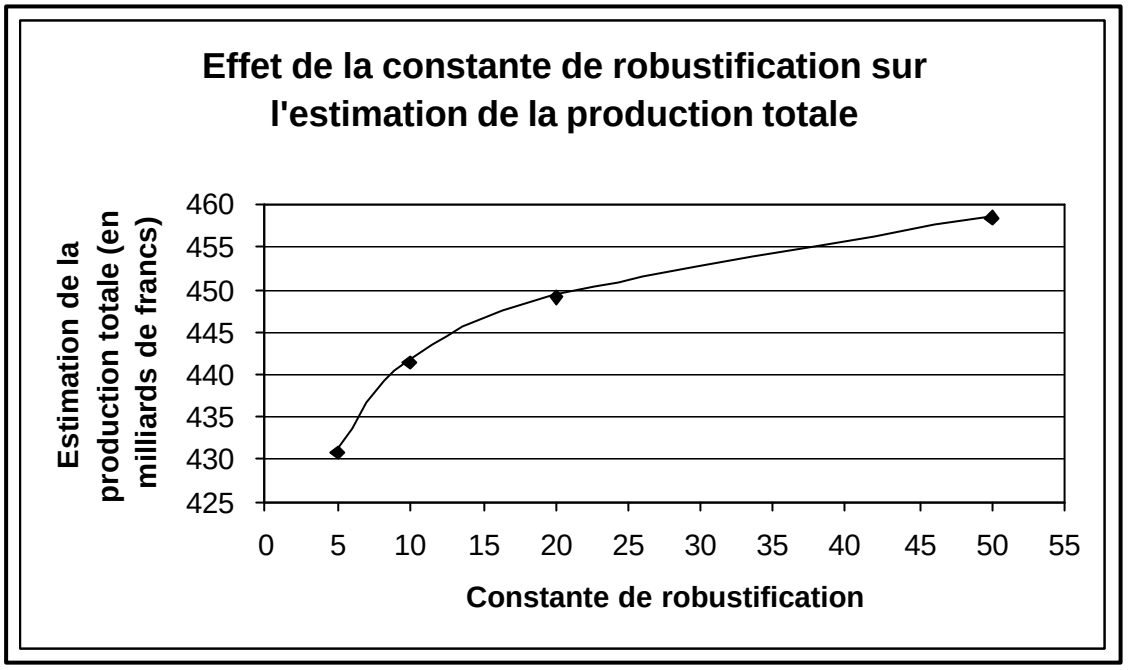


Figure 6: Impact de la constante de robustification sur le total estimé pour les données de 1997.

5.3.5 Choix du degré de robustification

Le choix de la constante k_{c^*} est un problème délicat. Si la constante k_{c^*} est trop petite, la robustification mène à de nombreuses et importantes corrections avec le risque de sous-estimer les grandeurs (les résidus extrêmes sont en général positifs). Si par contre elle est trop élevée, la robustification ne joue plus son rôle et laisse passer trop de valeurs aberrantes. Dans ce cas on se retrouve dans la situation où l'on n'opère pas de robustification et les résultats restent très sensibles aux valeurs extrêmes. Il s'agit donc de trouver un bon compromis entre le biais, qui est inconnu, et la variance !
Ci-joint deux graphiques, montrant l'évolution du montant total au niveau Suisse de l'estimation de la production (figure 6) et de son écart-type (figure 7), en fonction de la constante de robustification :
La solution "correcte" nous semble devoir se situer quelque part entre les deux extrêmes $k_{c^*} = 5$ (forte robustification) et $k_{c^*} = 50$ (pas de robustification). Un choix de $k_{c^*} = 10$ nous apparaît comme raisonnable mais pourraît encore être optimisé suivant la division économique considérée. Pour des raisons de commodité, nous fixons toutefois la constante de robustification à 10, quelle que soit la division économique.
Pour la constante de robustification égale à 10, nous représentons l'histogramme $\bar{u}_h$ des facteurs moyens de correction (figure 8), facteurs moyens $\bar{u}_h = \frac{\sum_{S_h^*} u_i}{n_{h^*}}$ calculés au niveau des différentes divisions économiques. Au total, 164 observations sur 5'602 observations sont

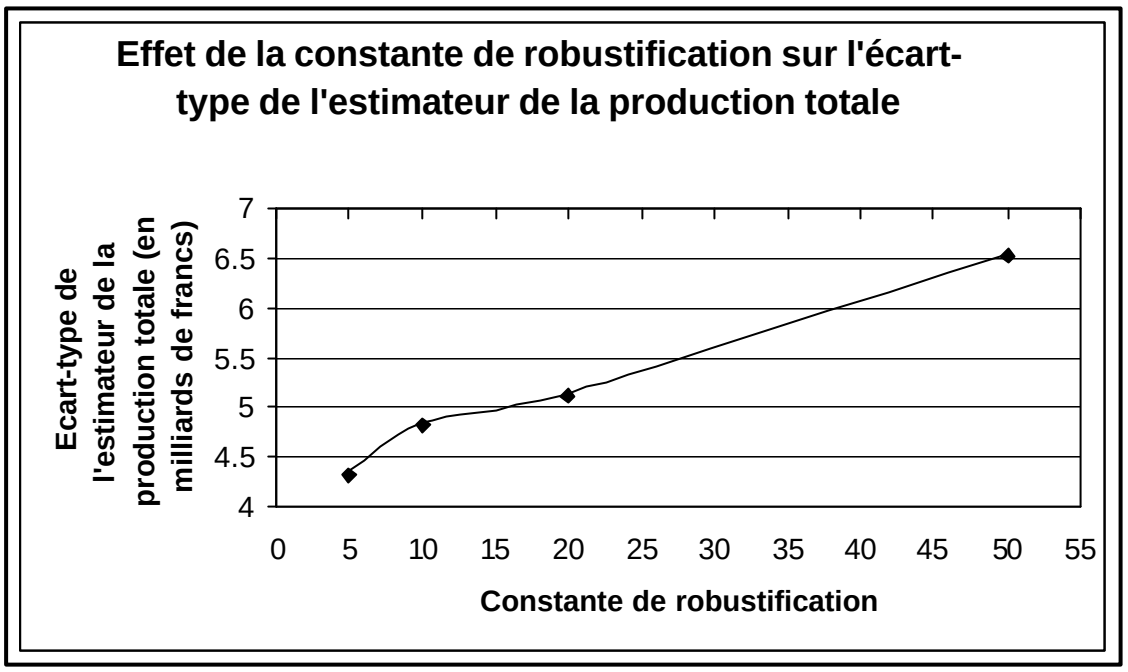


Figure 7: Impact de la constante de robustification sur l'écart-type de l'estimateur pour les données de 1997.

corrigées ($u_i < 1$), certaines divisions économiques étant proportionnellement plus touchées que d'autres.

6 Variance de l'estimateur robustifié

La modification de l'estimateur complique l'estimation de la variance. En faisant certaines hypothèses, nous pouvons toutefois nous ramener à une formule simple du même type que celle établie précédemment. Selon les hypothèses qui doivent être faites, les résultats doivent être considérés comme approximatifs et interprétés avec suffisament de précaution (Hulliger (2000)).
Si nous pouvons faire comme si les poids des observations ne dépendaient pas de l'échantillon, nous pouvons, comme précédemment, linéariser l'estimateur $\hat{\beta}^r_{c^*} = \frac{\sum_{S_{c^*}} w_i u_i y_i}{\sum_{S_{c^*}} w_i u_i x_i}$ et déduire l'estimateur de la variance du ratio :

$$var(\hat{\beta}^r_{c^*}) \approx \frac{1}{(\sum_{S_{c^*}} w_i u_i x_i)^2} var(\sum_{S_{c^*}} w_i u_i e_i),$$

où $e_i = y_i - \hat{\beta}^r_{c^*} x_i$ est l'estimation du résidu de l'observation i.

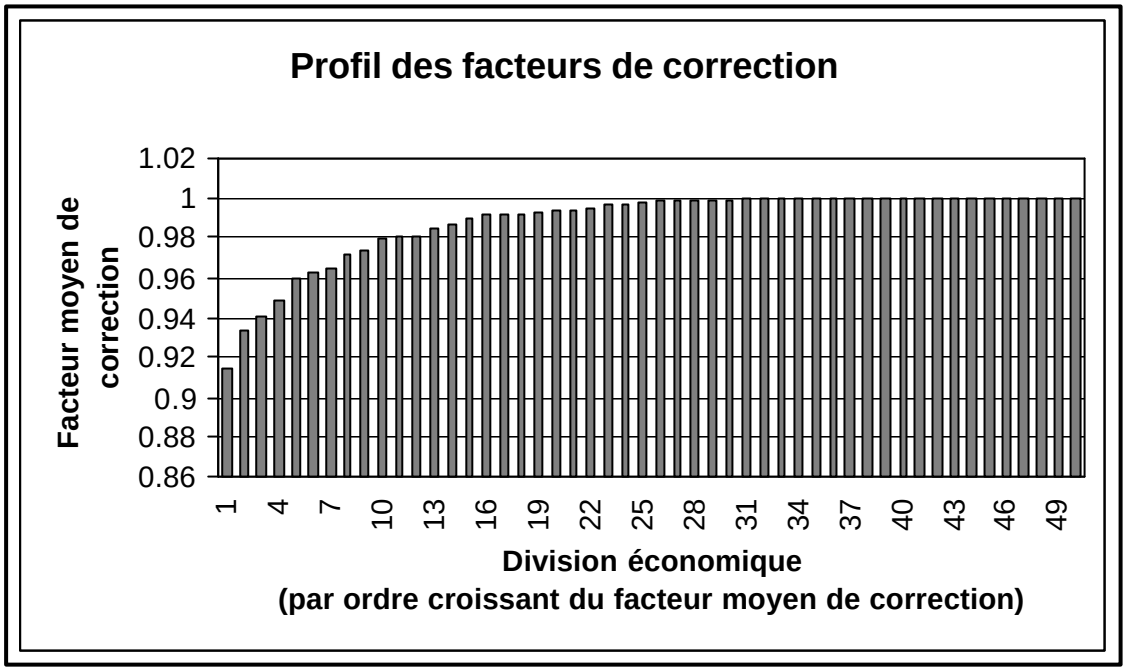


Figure 8: Représentation des facteurs moyens de correction pour $k_{c^*} = 10$.

En considérant l'échantillon comme issu d'un tirage avec remise de probabilités inégales de tirage $\frac{1}{w_i u_i}$, la variance de la somme pondérée $\sum_{S_{c^*}} w_i e_i$ des erreurs peut être estimée par :

$$var(\sum_{S_{c^*}} w_i u_i e_i) \approx \sum_{h=1}^{H} n_h D^2_{h,wue},$$

où

$$D^2_{h,wue} = \frac{1}{n_h - 1} \sum_{S_h} (w_i u_i e_i - \frac{1}{n_h} \sum_{S_h} w_i u_i e_i)^2.$$

Dans la mesure où, dans certaines strates, la taille de l'échantillon est non négligeable par rapport à celle de la population de référence, l'hypothèse d'un tirage avec remise est douteuse et aura tendance à surestimer la variance. A l'instar de la formule de variance en cas de probabilité de tirage constante par strate, nous introduisons un facteur de correction $(1 - \frac{1}{w_h})$ au niveau de chaque strate pour prendre en compte le caractère sans remise du tirage :

$$var(\sum_{S_c} w_i u_i e_i) \approx \sum_{h=1}^{H} (1 - \frac{1}{w_h}) n_h D^2_{h,wue},$$

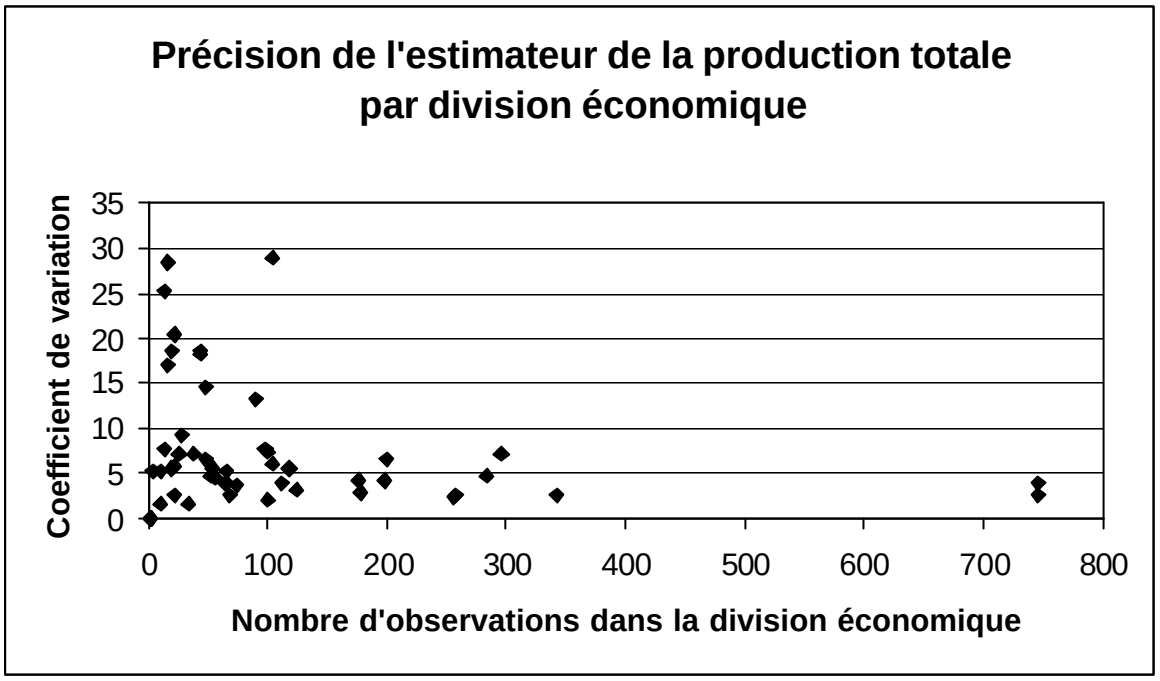


Figure 9: Coefficients de variation obtenus par NOGA2.

où w_h est le poids associé à une observation de la strate h (toutes les observations d'une même strate ont, dans cette enquête, le même poids). D'après nos expériences, le facteur de correction corrige l'estimation généralement un peu trop vers le bas et aura tendance à légèrement sous-estimer la variance.

7 Précision

Au niveau total Suisse, la précision est bonne. Ainsi le coefficient de variation relatif à l'estimation de la production totale en Suisse est de 1.50 %. Par contre, la précision devient rapidement insatisfaisante sur les estimations à des niveaux plus fins. Nous montrons sur la figure 9 la précision associée à l'estimation de la production totale par division économique NOGA2. Sont acceptables les coefficients de variation inférieurs à 5% ou, à la limite, entre 5% et 15 % mais à condition dans ce cas de présenter les résultats avec une mise en garde (par exemple en les mettant entre parenthèses) ; enfin, il faut mentionner qu'il n'est pas possible d'exclure un éventuel biais dans l'estimation des paramètres.

8 Bibliographie

Chambers, R.L. (1986), "Outlier Robust Finite Population Estimation", *Journal of the American Statistical Association,* ***81****, 1063-1069.*

Hidiroglou, M.A. and Srinath, K.P (1981), "Some Estimators of a Population Total From Simple Random Samples Containing Large Units", *Journal of the American Statistical Association,* ***76****, 690-695.*

Hulliger, B. (1995), "Outlier Robust Horvitz-Thompson Estimators, *Survey Methodology*, June 1995, **21/1**, 79-87.

Hulliger, B. (1999a), "Simple and Robust Estimators for Sampling", 1999 *Proceedings of the section on Survey Research Methods*, 54-63.

Hulliger, B. (1999b), "A Proposal for Treatment of Extreme Weights", *Technical Note,* SFSO.

Hulliger, B. (2000), "Invited Session on Outliers : Discussants Paper", Proceedings of the International Conference on Establishment Surveys II, Buffalo 2000, à paraître.

Peters, R. et Renfer, J.-P. (2000), "Etablissement du plan d'échantillonnage pour l'enquête sur la production et la valeur ajoutée", Rapport de méthodes de l'OFS, à paraître.

Renfer, J.-P. (2000), "Enquête sur la production et la valeur ajoutée WS 99 : échantillonnage complémentaire", document technique OFS, mars 2000.

9 Annexe : macro SAS d'extrapolation

```
%macro estvar(data=,var1=,var2=,gewvar=,idkorr=,strata=,class=,robust=,c=,
group=,uvar1=,uvar2=,uvar3=,robgew=,ident=,datpop=,clpop=,npop=) ;
/* ---------------------------------- */
/* Makro fuer robuste Varianzschaetzung eines Quotienten */
/* ---------------------------------- */
/* Eingabe Parameter : */
/* data= Name der Auswertungsdatei als SAS File */
/* var1= 1. Variable des Quotienten (Zaehler) */
/* var2= 2. Variable des Quotienten (Nenner) */
/* gewvar= Gewichtungsvariable */
/* idkorr= Variable zur Identifikation stat. Einheiten, fuer die */
/* das u Gewicht auf 1 bleibt (idkorr : 1) */
/* strata= Schichtungs(-Stratifikations)-Variable(n) */
/* class= Klassierungsvariable fuer Resultatsausgabe */
/* robust= Robustifikation (1/0 - JA/NEIN) */
/* c= Tuning Konstante fuer den univariaten one-step Schaetzer */
/* group= Klassierungsvariable fuer Schaetzer (im allg = class-Var) */
/* uvar1= Variable fuer den one-step-Algorithmus zur Berechnung der */
/* uvar2= u-Gewichte (bis zu max. 3 Variablen) */
/* uvar3= u ist dann definiert durch das Minimum */
/* robgew= 1 oder 0 ob im one-step-Algorithmus der Median */
/* mit/ohne Gewichtung berechnet wird */
/* ident= Identifikationsvariable fuer das Outfile */
/* datpop= SAS File fuer eine eventuelle Kalibrierung */
/* clpop= Klassierungsvariable fuer Kalibrierung. Der Wertebereich */
/* muss mit demjenigen der in class= angegebenen gleich sein */
/* npop= Kalibrierungsvariable */
/* Die Parameter : gewvar,idkorr,class,robust,c,group,uvar1-uvar3, */
/* ident,datpop,clpop,npop sind OPTIONAL */
/*----------------------------------------- */
/* Initialisierungen */
%if &robust eq %then %let robust=1;
%if &c eq %then %let robust=0;
%if &robust = 0 %then %do;
%let c=; %let uvar1=; %let uvar2=; %let uvar3=; %let group=;
%end;
%if &var1 eq or &var2 eq %then %do;
data _null_;
file print;
put #3 @2 " --> Fehler : Keine Auswertungsvariablen";
```

```
run;
%goto ende;
%end;
%if &data eq %then %do;
data _null_;
file print;
put #3 @2 " --> Fehler : Kein Auswertungsfile";
run;
%goto ende;
%end;
/* Kalibrierungsfile (falls vorhanden) */
%if &datpop ne %then %do;
%let clren=;
%if &clpop ne %then %do;
%let clren=rename &clpop=&class;
%end;
data _pop_;
set &datpop;
_tot_=1;
rename &npop=_npop_;
&clren;
run;
%if &clpop ne %then %do;
proc sort data=_pop_;
by &class;
run;
%end;
%end;

 /* Routine : one-step-Algorithmus */
/*-------------------------------------------*/
%macro coeff(datar=,var1=,var2=,w1=,w2=,by=,idkorr=,strata=,
ind=,robgew=,ident=);
%if &idkorr ne %then %do;
data tmpsub;
set &datar;
if &idkorr ne 1;
run;
%end;
%if &idkorr eq %then %do;
data tmpsub;
set &datar;
```

```
run;
%end;
/* Median (mit/ohne Gewichtung) */
%if &robgew = 1 %then %let fr=freq &w2;
%else %let fr=;
proc univariate data=tmpsub noprint;
var &var1 &var2;
&fr;
&by;
output out=os1 pctlpts=50 pctlpre=ma mb;
run;
/* Quotientenschätzer */
data estr;
merge &datar os1;
&by;
t=ma50/mb50;
res=abs(&var1-t*&var2);
resb=res/(&var2**0.5);
run;
%if &idkorr ne %then %do;
data tmpsub;
set estr;
if &idkorr ne 1;
run;
%end;
%if &idkorr eq %then %do;
data tmpsub;
set estr;
run;
%end;
proc univariate data=tmpsub noprint;
var resb;
&fr;
&by;
output out=os2 pctlpts=50 pctlpre=sig;
run;
/* Berechnung der u-Gewichte mit tuning Konstante c */
data estr;
merge estr os2;
&by;
u&ind=1;
sigro=sig50/0.67;
```

```
%if &idkorr ne %then %do;
if &idkorr ne 1 and res > co*sigro*(&var2**0.5) and res > 0 then do;
u&ind=co*sigro*(&var2**0.5)/res;
*u&ind=u&ind+(1-u&ind)/w;
end;
%end;
%if &idkorr eq %then %do;
if res > co*sigro*(&var2**0.5) and res > 0 then do;
u&ind=co*sigro*(&var2**0.5)/res;
*u&ind=u&ind+(1-u&ind)/w;
end;
%end;
run;
data gewu&ind;
set estr;
keep recn _tot_ u&ind &group &ident;
run;
%mend;
/* Einlesen und Aufbereitung der Auswertungdaten */
data tmp;
set &data;
%if &c eq %then %do;
keep &var1 &var2 &gewvar &strata &class &group &idkorr
_tot_ w wi &uvar1 &uvar2 &uvar3 &ident;
%end;
%if &c ne %then %do;
co=&c;
keep &var1 &var2 &gewvar &strata &class &group &idkorr
_tot_ w wi &uvar1 &uvar2 &uvar3 &ident co;
%end;
_tot_=1;
%if &gewvar ne %then %do;
%let flgw=1;
%let gwa=Ja;
w=&gewvar;
wi=int(w*1000);
%end;
%if &gewvar eq %then %do;
%let flgw=0;
%let gwa=Nein (w=1);
%let gewvar=w;
w=1;
```

```
wi=1;
%end;
if &var1 ne . and &var2 ne . and &gewvar ne . then do;
output;
end;
run;
data gewu;
set tmp;
keep _tot_ u;
u=1;
run;
%let rba=NEIN;
%if &robust = 1 and &uvar1 ne %then %do;
%let rba=JA;
%if &group eq %then %let by=by _tot_;
%if &group ne %then %do;
%let by=by &group;
proc sort data=tmp;
&by;
run;
data tmp;
set tmp;
recn=_n_;
run;
%end;
/* u1 - Gewichte */
%coeff(datar=tmp,var1=&uvar1,var2=&var2,w1=&gewvar,w2=wi,by=&by,
idkorr=&idkorr,strata=&strata,ind=1,robgew=&robgew,ident=&ident);
data gewu;
set gewu1;
u=u1;
run;
%end;
/* u2 - Gewichte */
%if &robust = 1 and &uvar1 ne and &uvar2 ne %then %do;
%coeff(datar=tmp,var1=&uvar2,var2=&var2,w1=&gewvar,w2=wi,by=&by,
idkorr=&idkorr,strata=&strata,ind=2,robgew=&robgew,ident=&ident);
data gewu;
merge gewu gewu2;
by recn;
u=min(u1,u2);
run;
```

```
%end;
/* u3 - Gewichte */
%if &robust = 1 and &uvar1 ne and &uvar2 ne and &uvar3 ne %then %do;
%coeff(datar=tmp,var1=&uvar3,var2=&var2,w1=&gewvar,w2=wi,by=&by,
idkorr=&idkorr,strata=&strata,ind=3,robgew=&robgew,ident=&ident);
/* u=min(u1,u2,u3) Gewicht */
data gewu;
merge gewu gewu3;
by recn;
u=min(u1,u2,u3);
run;
%end;
%if &robust = 1 %then %do;
data tmp;
merge tmp gewu;
by recn;
mod=0;
if u < 1 then mod=1;
uw=u*w;
run;
%end;
%if &robust = 0 %then %do;
data tmp;
set tmp;
u=1;
uw=u*w;
run;
%end;
/* Klassierung , Bestimmung der Werte der Klassierungsvariablen */
%if &class eq %then %do;
%let anzcla=0;
%let cl1=_tot_;
%let anzstr=1;
%let clst=cl=compress(left(trim(&cl1)),' ');
%let wert1=1;
proc means data=tmp noprint nway;
var &var2;
class _tot_;
weight uw;
output out=agg sum=s_samp;
run;
%if &datpop ne %then %do;
```

```
data agg;
drop _type_ _freq_;
merge agg _pop_;
by _tot_;
run;
%end;
data stmp;
length cl $ 20;
set tmp;
&clst;
run;
%end;
%if &class ne %then %do;
proc means data=tmp noprint nway;
var &var2;
class &class;
weight uw;
output out=agg sum=s_samp;
run;
%if &datpop ne %then %do;
data agg;
drop _type_ _freq_;
merge agg(in=in1) _pop_(in=in2);
by &class;
if in1 and in2;
run;
%end;
%let anzcla=1;
%let cl1=%scan(&class,1);
%let cl2=%scan(&class,2);
%let cl3=%scan(&class,3);
%let clst=cl=compress(left(trim(&cl1)),' ');
%if &cl2 ne %then %do;
%let anzcla=2;
%let clst=&clst||' '||compress(left(trim(&cl2)),' ');
%end;
%if &cl3 ne %then %do;
%let anzcla=3;
%let clst=&clst||' '||compress(left(trim(&cl3)),' ');
%end;
proc sort data=tmp out=stmp;
by &class;
```

```
run;
proc means data=stmp noprint nway;
class &class;
output out=str n=n;
run;
data _null_;
call symput('anzstr',trim(left(put(nobs,8.))));
if 0 then set str nobs=nobs;
stop;
run;
data stmp;
length cl $ 20;
set stmp;
&clst;
run;
data str2;
length cl $ 20;
set str;
&clst;
run;
data _null_;
set str2;
call symput('wert'!!left(_n_),left(trim(cl)));
run;
%end;
/* Schleife ueber Klassierung, Berechnung der Schaetzer und Varianzen */
%do i=1 %to &anzstr;
data tmp&i;
length cid 8;
set stmp;
indic=1;
if cl ne "&&wert&i" then do;
indic=0; u=0; mod=0;
&var1=0; /*&var2=0;*/
end;
clid="&&wert&i";
cid=input(clid,best20.);
bv1=u*w*&var1;
bv2=u*w*&var2;
run;
proc univariate data=tmp&i noprint;
var bv1 bv2 u w indic mod;
```

```
output out=utmp&i sum=sbv1 sbv2 su sw sindic smod n=anz;
run;
/* Schaetzer */
data est&i;
set utmp&i;
b=sbv1/sbv2;
perte=su/sindic;
_tot_=1;
run;
data res&i;
merge tmp&i est&i;
by _tot_;
e=&var1-b*&var2;
eb=u*w*e;
t1=u*w*&var1;
run;
proc sort data=res&i;
by &strata;
run;
/* Varianzen */
proc univariate data=res&i noprint;
var eb t1 w;
by &strata _tot_ cid;
output out=o&i n=ng var=vareb vart1 mean=m1 m2 w;
run;
data o&i;
set o&i;
ngf=ng*(1-(1/w));
run;
%if &flgw=1 %then %do;
proc univariate data=o&i noprint;
var vareb vart1;
by _tot_ cid;
weight ngf;
output out=oo&i sum=svareb svart1 mean=m1 m2;
run;
%end;
%if &flgw=0 %then %do;
proc univariate data=o&i noprint;
var vareb vart1;
by _tot_ cid;
output out=oo&i sum=svareb svart1 mean=m1 m2;
```

```
run;
%end;
data fin&i;
merge est&i oo&i;
by _tot_;
varb=svareb/(sbv2**2);
stdb=varb**0.5;
stdt1=svart1**0.5;
run;
%if &i = 1 %then %do;
proc append base=fin&i; run;
%end;
%else %do;
proc append base=fin1 data=fin&i; run;
%end;
%end;
proc sort data=fin1 nodupkey;
by cid;
run;
/* Resultatausgabe */
title1 "Einfache und robuste Varianzschätzung";
title2 "Robuste Schätzung : &rba / Gewichtung : &gwa / Gew_Var : &gewvar";
title3 "Dataset : &data / Schichtvariablen : &strata / Klassen_Var : &class";
title4 "Variable für Beta_Koeff : &group / c_Konstante= &c / Robust mit Gewicht : &robgew";
title5 " ";
%if &uvar1 ne or &uvar2 ne or &uvar3 ne %then %do;
title5 "Variablen für Berechnung u_min : &uvar1 &uvar2 &uvar3, &var2";
%end;
/* mit Kalibrierung */
%if &datpop ne %then %do;
data fin;
merge fin1 agg;
t_samp=b*s_samp;
t_pop=b*_npop_;
t_stdpop=stdb*_npop_;
varpop=t_stdpop**2;
if t_pop ne 0 then t_cv=(t_stdpop/t_pop)*100;
else t_cv=.;
cidc=left(put(cid,5.));
run;
/* Total - alle Klassen */
proc means data=fin nway noprint;
```

```
var sbv1 sbv2 sindic smod varb s_samp t_samp _npop_ t_pop svart1 varpop;
output out=total sum=sbv1 sbv2 sindic smod varb s_samp t_samp _npop_
t_pop svart1 varpop;
run;
data total;
set total;
cidc='Total';
stdt1=svart1**0.5; t_stdpop=varpop**0.5; stdb=varb**0.5;
b=sbv1/sbv2;
if t_pop ne 0 then t_cv=(t_stdpop/t_pop)*100;
else t_cv=.;
run;
/* Klassen und Total */
data fin;
set fin total;
keep cidc sindic smod perte s_samp b stdb t_samp _npop_
stdt1 t_pop t_stdpop t_cv sbv1 sbv2;
label cidc="Klassen_Eint"
sindic="Anz_Obs"
smod="Anz_Korr."
perte="Indikator"
s_samp="Gew.Sum &var2"
b="&var1 / &var2"
stdb="Stdabw"
t_samp="Total &var1 samp"
_npop_="Kalibrierung"
stdt1="Stdabw &var1 samp"
t_pop="Total &var1 pop"
t_stdpop="Stdabw &var1 pop"
t_cv="CV_pop"
;
run;
options linesize=132 pagesize=30 nodate pageno=1 missing='-';
proc print data=fin noobs uniform label split='*';
var cidc sindic smod perte b stdb s_samp t_samp stdt1 _npop_
t_pop t_stdpop t_cv;
label cidc="Klassen*Eint :*&class"
sindic="Anz*Obs"
smod="Anz*Korr."
perte="Indi-*kator*Korr."
s_samp="Gew.*Sum_&var2"
b="Ratio :*&var1/&var2"
```

```
stdb="Std_abw"
t_samp="Total*&var1*samp"
_npop_="Kali-*brierung"
stdt1="Std_abw*&var1*samp"
t_pop="Total*&var1*pop"
t_stdpop="Std_abw*&var1*pop"
t_cv="CV*pop"
;
format cidc $5. smod 5. perte 6.3 b 8.3 stdb 8.3 sindic 5.
s_samp 9. t_samp 9. stdt1 9. _npop_ 9. t_pop 9.
t_stdpop 9. t_cv 6.3;
run;
%end;
/* ohne Kalibrierung */
%if &datpop eq %then %do;
data fin;
merge fin1 agg;
t_samp=b*s_samp;
cidc=left(put(cid,5.));
run;
/* Total */
proc means data=fin nway noprint;
var sbv1 sbv2 sindic smod varb s_samp t_samp svart1;
output out=total sum=sbv1 sbv2 sindic smod varb s_samp t_samp svart1;
run;
data total;
set total;
cidc='Total';
stdt1=svart1**0.5; stdb=varb**0.5;
b=sbv1/sbv2;
run;
/* KLassen und Total */
data fin;
set fin total;
keep cidc sindic smod perte s_samp b stdb t_samp
stdt1 sbv1 sbv2;
label cidc="Klassen_Eint"
sindic="Anz_Obs"
smod="Anz_Korr."
perte="Indikator"
s_samp="Gew.Sum &var2"
b="&var1 / &var2"
```

```
stdb="Stdabw"
t_samp="Total &var1 samp"
stdt1="Stdabw &var1 samp"
;
run;
options linesize=120 pagesize=30 nodate pageno=1 missing='-';
proc print data=fin noobs uniform label split='*';
var cidc sindic smod perte b stdb s_samp t_samp stdt1;
label cidc="Klassen*Eint :*&class"
sindic="Anz*Obs"
smod="Anz*Korr."
perte="Indi-*kator*Korr."
s_samp="Gew.*Sum_&var2"
b="Ratio :*&var1/&var2"
stdb="Std_abw"
t_samp="Total*&var1*samp"
stdt1="Std_abw*&var1*samp"
;
format cidc $5. smod 5. perte 6.3 b 8.3 stdb 8.3 sindic 5.
s_samp 9. t_samp 9. stdt1 9.;
run;
%end;
proc datasets lib=work;
delete tmp1-tmp&anzstr
utmp1-utmp&anzstr
o1-o&anzstr
oo1-oo&anzstr
est1-est&anzstr
res1-res&anzstr
fin1-fin&anzstr
os1 os2 str str2 estr;
run; quit;
title1; title2; title3; title4; title5;
%let var1=; %let var2=; %let uvar1=; %let uvar2=; %let uvar3=;
%let gewvar=; %let strata=; %let class=; %let robust=; %let c=; %let group=;
%let rba=; %let gwa=; %let ident=; %let robgew=;
%ende:
%mend;
```

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodological reports of the Statistical Methods Unit

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Peters, R., Renfer, J.-P. et Hulliger, B., (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Rapports de méthodes du Service de méthodes statistiques

Une liste complète des titres de cette série de publications figure à la fin de ce rapport. Les rapports de méthodes peuvent être commandés auprès de:

Office fédéral de la statistique
Service des publications
Espace de l'Europe 10
CH-2010 Neuchâtel
Tél.: (+41 32) 71 36060
Fax: (+41 32) 71 36061
Internet: Order@bfs.admin.ch

Methodenberichte des Dienstes Statistische Methoden

Eine vollständige Liste der Titel dieser Serie befindet sich am Ende dieses Berichts. Die Methodenberichte können bestellt werden bei:

Bundesamt für Statistik
Publikationsdienst
Espace de l'Europe 10
CH-2010 Neuchâtel
Tel.: (+41 32) 71 36060
Fax: (+41 32) 71 36061
Internet: Order@bfs.admin.ch