Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

MÉTHODE D'ESTIMATION À COURT TERME DE L'APPORT DE MAIN-D'OEUVRE À L'AIDE DE DONNÉES PRÉLIMINAIRES COURANTES DE SOURCES ADMINISTRATIVES COMPORTANT DES ERREURS DE COUVERTURE

Alessandro Pallara, Ciro Baldi, Piero Demetrio Falorsi, Raffaella Succi[1] et Aldo Russo[2]

RÉSUMÉ

Dans le présent document, nous proposons une méthode d'estimation des indicateurs de facteur travail à l'aide des données administratives de la base de données sur la sécurité sociale (BSS). Si nous avons conçu cette méthode, c'est que les organismes statistiques nationaux ont l'obligation de répondre aux critères de qualité du *règlement n° 1165/98* des Communautés européennes concernant les statistiques conjoncturelles. Les données prévues par ce règlement sont d'une telle finesse de désagrégation qu'il serait impossible de satisfaire à toutes les exigences par une collecte directe de données. En raison de leur actualité et de leur détail de « couverture », les données administratives constituent une précieuse source pour l'obtention d'estimations des agrégats de la population des entreprises en tout respect de ces critères de qualité.

MOTS CLÉS : Données administratives; Indicateurs à court terme; Erreur de couverture; Modèle pour surperpopulation; Biais de sélection

1. INTRODUCTION

Dans les instituts statistiques nationaux (ISN), on s'intéresse de plus en plus à l'utilisation de sources administratives à des fins statistiques, ce recours permettant d'alléger les coûts et le fardeau de réponse des enquêtes. Dans les règlements officiels des organismes statistiques internationaux, on recommande aussi l'emploi des dossiers administratifs comme solution d'appoint ou de rechange à une collecte directe de données sur les entreprises, les exploitations agricoles et les établissements en vue d'un relèvement de la qualité des données d'enquête.

Pour qu'il soit fait un fructueux usage de données administratives à des fins statistiques, il faut évidemment que de bonnes sources de données soient disponibles. En Italie, la base de données sur la sécurité sociale (BSS de l'*Istituto Nazionale della Previdenza Sociale* ou *INPS*) est la grande source administrative de données sur l'emploi et les salaires dans les secteurs privé et public. On l'a utilisée comme source première de renseignements dans le cadre de la conception d'une enquête trimestrielle sur l'emploi, la rémunération et les coûts de main-d'œuvre dans l'ensemble des entreprises des secteurs de l'industrie et des services en prévoyant une combinaison de données de collecte directe par enquête-entreprises et de données des dossiers administratifs.

Cette enquête vise principalement à satisfaire aux exigences du *RÈGLEMENT DU CONSEIL EUROPÉEN N° 1165/98 concernant les statistiques conjoncturelles* (SC), et notamment la production dans chaque État membre d'estimations trimestrielles *(i)* du nombre de personnes, *(ii)* des salaires et traitements et *(iii)* des heures travaillées.

Le projet de conception de cette enquête comporte deux volets. Le premier a pour objet l'établissement d'une méthode d'estimation courante de deux variables, à savoir *(i)* le nombre de personnes ayant un

[1] Italian National Institute of Statistics, Via Depretis 74/b, 00184, Rome (Italie)
[2] Department of Political Institutions and Social Sciences, University Roma Tre, Rome (Italie)

emploi et *(ii)* les salaires et traitements en valeur brute, dans le cas des entreprises comptant moins de 500 salariés, et ce, à l'aide des seules données administratives. Le second volet portera surtout sur les problèmes de combinaison des estimations des entreprises de moins de 500 salariés, obtenues par source administrative, avec les renseignements des grandes sociétés auprès desquelles l'institut statistique national (ISN) italien fait directement enquête chaque mois.

On peut systématiquement juger de l'intérêt de ce projet en considérant les questions suivantes : (*i*) les données que prévoit le règlement sur les statistiques conjoncturelles (SC) sont d'une désagrégation sectorielle si fine (niveau à 2 chiffres de la classification NACE, 1re révision) qu'il serait impossible de répondre à toutes les exigences par une collecte directe de données; (*ii*) les statistiques à court terme de l'emploi et des salaires connaissent de graves problèmes de « couverture » en Italie, si bien que, pour l'instant, on ne mène qu'une enquête directe par sondage auprès des très grandes sociétés (comptant plus de 500 salariés) et sur des groupes déterminés d'activités économiques; (*iii*) le recours aux données administratives offre dans ce cas l'avantage que les définitions sont dans ce cas cohérentes avec les dispositions de ce règlement, d'où la possibilité d'obtenir la plupart des renseignements recherchés sans ajouter au fardeau imposé aux entreprises.

Dans cet exposé, nous nous concentrons sur la première partie de la conception de l'enquête précitée et présenterons la méthode conçue pour l'établissement d'estimations trimestrielles des salariés et des salaires avec les données administratives comme seule source d'information. La méthode d'estimation proposée développe certains résultats d'études antérieures (Falorsi *et coll.,* 2000). On obtient des estimations par un modèle prévisionnel reposant sur les données courantes d'un sous-ensemble d'unités de la BSS et par une extension aux valeurs *y* non observées à l'échelle des unités du registre de la sécurité sociale. L'utilisation des données BSS ne va pas sans quelques problèmes :

i. le sous-ensemble d'unités fournissant des données courantes est un échantillon *non aléatoire* de la population; dans la mesure où le mécanisme de production de données est informatif pour la procédure d'estimation, les estimations obtenues peuvent être entachées d'un biais de sélection (Royall, 1988). Un outil de compensation est l'élaboration du modèle à l'intérieur de sous-groupes homogènes (Hidiroglou et coll., 1995);

ii. la BSS peut comporter des *erreurs de couverture* à cause de problèmes de mise à jour du registre (unités introduites, retirées ou transformées) ou de retards d'exécution des tâches administratives pour certaines unités; ainsi, l'univers BSS peut différer systématiquement de la population visée; l'inclusion au registre administratif se modélise comme le résultat d'un processus de Bernoulli dont la probabilité de « succès » est constante à l'intérieur de chaque sous-population.

Le document est structuré de la manière suivante : à la section 2, nous présentons les paramètres d'intérêt, puis définissons le modèle statistique d'estimation; à la section 3, nous décrivons une forme explicite des estimations et examinons certains aspects pratiques du calcul des paramètres d'intérêt et de leurs erreurs quadratiques moyennes; à la section 4 enfin, nous exposons certains résultats empiriques concernant les estimations trimestrielles des salariés et des salaires et de leurs erreurs quadratiques moyennes (EQM) relatives.

2. PARAMÈTRES D'INTÉRÊT ET MODÈLES STATISTIQUES

Soit P_t la population (finie) d'entreprises « actives » au registre pour la période t en cours (mois ou trimestre). Les paramètres d'intérêt correspondent aux des totaux des variables *emploi* et *traitements et salaires* de la population visée P_t sous la forme suivante :

$$Y_t = \sum_{i \in P_t} y_{ti} \qquad (1)$$

où y_{ti} désigne la valeur de la variable d'intérêt *y* (par exemple le nombre de personnes ayant un emploi) pour l'entreprise *i* dans la période *t*. Nous voulons estimer Y_t à l'aide des données auxiliaires de la source administrative utilisée.

Chaque entreprise inscrite à la BSS doit, quand elle fait son versement du mois, remplir une formule avec des renseignements sur *(i)* le nombre de ses salariés, *(ii)* les salaires et traitements versés et *(iii)* les charges de sécurité sociale.

L'unité de base du registre ne correspond à aucune définition type comme celles qu'énonce le *règlement du Conseil européen* (CEE) *n° 696/93 relatif aux unités statistiques d'observation et d'analyse du système productif dans la Communauté*. Chaque formule appartient à une seule entreprise. Une entreprise peut toutefois en remplir plusieurs, et il est difficile de repérer toutes les formules qui émanent d'une même entreprise. Ajoutons que la liste de toutes les unités inscrites à la BSS pour chaque période de référence pose un certain problème d'excès de couverture, puisqu'il faut d'ordinaire des mois pour qu'une unité disparue soit rayée du registre de la sécurité sociale.

Puisque la procédure de transmission n'est pas la même pour toutes les unités, pour chaque période de référence, on dispose uniquement de renseignements sur un sous-ensemble d'unités pour la période en cours *t* (à l'heure actuelle, ce sous-ensemble comprend quelque 300 000 unités sur un total approximatif de 1 100 000 pour la population d'entreprises des secteurs de l'industrie et des services qui, en Italie, comptent moins de 500 salariés). Bien que plutôt nombreux, ce sous-ensemble ne constitue en rien un échantillon aléatoire selon une procédure de sélection déterminée. En fait, le caractère non aléatoire des unités observées peut créer un biais dans les estimateurs conventionnels de (1).

Soit A_t l'ensemble d'entreprises « *actives* » inscrites à la BSS. A_t peut être considéré comme la représentation disponible de la population visée P_t selon le registre administratif. Du fait de la présence d'erreurs de couverture par excès dans cette source administrative, on a $P_t \subseteq A_t$. Comme nous l'avons mentionné, si ces erreurs se produisent, c'est surtout qu'on doit parfois compter des mois pour qu'une unité disparue soit retirée du registre de la sécurité sociale. Ainsi, A_t peut comprendre des disparitions récentes (parmi les unités qui n'ont pas été sélectionnées) que nous devons prendre en compte pour dégager des estimations exemptes de biais de la variable d'intérêt; le total d'intérêt peut donc être représenté par

$$Y_t = \sum_{i \in A_t} y_{ti}\, \lambda_{ti} \qquad (2)$$

où λ_{ti} est une variable indicatrice égale à l'unité si l'entreprise $i \in P_t$ (si c'est une unité *active* qui figure donc à bon droit au registre administratif) et à 0 dans les autres cas.
Pour introduire la méthode d'estimation, on partitionne A_t de la manière suivante :

$$A_t = s_t \cup \bar{s}_t \qquad (3)$$

où

- s_t est le sous-ensemble d'unités sur lesquelles on observe la valeur de la variable d'intérêt pour l'intervalle en cours *t*; dans ce qui suit, nous appellerons s_t l'*échantillon*; à noter que, $\forall\, i \in s_t$, nous avons λ_{ti} =1, car s_t comprend les entreprises qui, ayant fait parvenir leurs états de rémunération au régime de sécurité sociale pour la période en cours par une procédure spéciale de transmission, sont jugées *actives* avec *certitude*;
- $\bar{s}_t$ est l'ensemble d'unités sur lesquelles nous n'observons pas la variable d'intérêt au moment *t*, ce que nous appellerons le *non-échantillon* dans ce qui suit.

La méthode proposée d'estimation est fondée sur un modèle prévisionnel où on suppose qu'une relation existe entre la variable d'intérêt et certaines variables auxiliaires. On estime ensuite les paramètres des modèles à partir du sous-ensemble d'unités contenant des données pour la période de référence. Lorsqu'on

modélise la relation entre l'*échantillon* et le *non-échantillon*, on doit bien tenir compte de deux sources possibles de biais :

- caractère non *aléatoire* de s_t : comme nous l'avons précisé, s_t comprend les unités qui choisissent elles-mêmes d'envoyer leurs états de rémunération par procédure spéciale de transmission informatique, alors que le moyen *normal* de communication avec le régime de sécurité sociale est la poste; parmi les méthodes proposées de traitement du *biais de sélection* (Royall, 1988), nous avons choisi comme solution dans cet exposé d'ajuster le modèle à l'intérieur de sous-populations (groupes de *régression*), en supposant que les coefficients du modèle seront les mêmes pour les unités d'un même groupe. Hidiroglou *et coll.* (1995) proposent pour leur part un bon critère de définition des groupes de régression;
- *erreurs de couverture* dans $\bar{s}_t$: les erreurs de listage par disparitions non inscrites au registre administratif peuvent faire que certains sous-ensembles de A_t différeront systématiquement de la population cible correspondante; on peut supposer qu'il existe une partition de A_t telle que, dans chacun des sous-ensembles en question, les probabilités de juste inclusion dans la source administrative seront à peu près égales pour toutes leurs unités respectives; dans ce qui suit, ces sous-ensembles seront appelés *groupes d'erreur au registre*.

Soit

- $\{A_{tw}\}_{w=1,..,W}$ une partition de A_t en *W groupes de régression*. Nous désignerons par $A_{tw} = s_{tw} \cup \bar{s}_{tw}$ le w^{e} groupe de *régression*, à savoir le sous-ensemble d'unités appartenant à A_t pour lequel le w^{e} modèle statistique est défini;
- $\{A_{tr}\}_{r=1,....,R}$ une partition de A_t en *R groupes d'erreur au registre*. Nous supposons que, dans chaque groupe *d'erreur au registre* A_{tr}, les probabilités d'erreur de couverture sont fixes et à peu près constantes sur l'intervalle temporel de l'estimation.

En notant que E(.) et C(.) représentent respectivement l'espérance et la variance du modèle, nous supposons aussi que, pour $w = 1,\ldots,W$, le modèle suivant de surperpopulation est valable :

$$\begin{aligned} &\mathrm{E}\left(y_{ti} \middle| \lambda_{ti} = 1\right) = \boldsymbol{\beta}'_{tw}\, \boldsymbol{x}_{ti} && \forall i \in A_{tw} \\ &\mathrm{E}\left(y_{ti}, y_{t'i'} \mid \lambda_{ti}, \lambda_{t'i'}\right) = \begin{cases} c_{ti}\sigma^2_{tw} & \text{pour } (t=t') \cap (i=i') \cap (\lambda_{ti}=1) \quad \forall (i,i') \in A_{tw} \\ 0 & \text{dans les autres cas} \end{cases} \end{aligned} \tag{4}$$

où $\boldsymbol{x}_{ti}$ est un vecteur de variables auxiliaires de dimension M_{tw} pour l'entreprise i au moment t, $\boldsymbol{\beta}_{tw}$ un vecteur-colonne de coefficients de régression de dimension M_{tw} et c_{ti} une constante connue ayant à voir avec la taille des entreprises inscrites à la source administrative.

Nous formons en outre l'hypothèse qu'une *juste inclusion* à la BSS est définie par une suite d'essais indépendants de Bernoulli indépendants de la valeur de y. Supposons que ces essais suivent une distribution identique dans chaque groupe d'erreur au registre A_{tr} ($r = 1,\ldots,R$). Nous postulons ainsi que, pour $r = 1,\ldots,R$, le modèle suivant est valable :

$$\begin{aligned} &\mathrm{E}(\lambda_{ti}) = p_{ti} = \theta_{tr} && \forall i \in A_{tr} \\ &\mathrm{E}(\lambda_{ti}, \lambda_{t'i'}) = \begin{cases} \theta_{tr}\,(1-\theta_{tr}) & \text{pour } (t=t') \cap (i=i') \\ 0 & \text{dans les autres cas} \end{cases} && \forall (i,i') \in A_{tr} \end{aligned} \tag{5}$$

Dans ce cadre, le modèle suivant est dérivé pour $w = 1,\ldots,W$ et $r = 1,\ldots,R$

$$\begin{aligned} &\mathrm{E}\left(y_{ti}\,\lambda_{ti}\right) = \boldsymbol{\beta}'_{tw}\,\mathbf{x}_{ti}\,\theta_{tr} && \forall i \in A_{t(w,r)} \\ &\mathrm{E}\left(y_{ti}\lambda_{ti}, y_{t'i'}\,\lambda_{t'i'}\right) = \begin{cases} c_{ti}\;\sigma^2_{t\,w}\;\theta_{tr} + \left(\boldsymbol{\beta}'_{tw}\,\mathbf{x}_{ti}\right)^2\,\theta_{tr}\left(1-\theta_{tr}\right) & \text{pour } \left(t=t'\right)\cap\left(i=i'\right) \quad \forall (i',i) \in A_{t(w,r)} \\ 0 \qquad\qquad \text{dans les autres cas} \end{cases} \end{aligned} \tag{6}$$

où $A_{t(w,r)} = A_{tw} \cap A_{tr}$.

Par traitement prévisionnel et après obtention des estimations $\tilde{\boldsymbol{\beta}}_{tw}$ ($w = 1,\ldots,W$) et $\tilde{\theta}_{tr}$ ($r = 1,\ldots,R$) des paramètres entrant dans les modèles (4) - (6), nous estimons le total Y_t par

$$\tilde{Y}_t = \sum_{w=1}^{W}\sum_{r=1}^{R}\left[\sum_{i\in s_{t,(w,r)}} y_{ti} + \sum_{i\in \bar{s}_{t,(w,r)}} \tilde{\boldsymbol{\beta}}'_{tw}\,\mathbf{x}_{ti}\,\tilde{\theta}_{tr}\right] \tag{7}$$

où $s_{t(w,r)} = s_t \cap A_{t(w,r)}$ et $\bar{s}_{t(w,r)} = \bar{s}_t \cap A_{t(w,r)}$.

Ainsi, pour une partition $A_{t(w,r)}$, l'estimation du total de population de la variable *y* est semblable à une *inférence à partir d'un modèle* pour les *totaux de domaines* (Chambers, 1997).

À la section suivante, nous examinerons en détail une forme explicite pour les estimations $\tilde{\boldsymbol{\beta}}_{tw}$ ($w = 1,\ldots,W$) et $\tilde{\theta}_{tr}$ ($r = 1,\ldots,R$), ainsi que certains aspects pratiques de la définition des partitions $\{A_{tw}\}_{w=1,\ldots,W}$ et $\{A_{tr}\}_{r=1,\ldots,R}$.

3. MÉTHODE D'ESTIMATION

Deux grandes questions se posent au sujet de la méthode d'estimation proposée :

(i) on a obtenu les estimations de $\boldsymbol{\beta}_{tw}$ et θ_{tr} séparément pour chaque ensemble de paramètres;

(ii) le principal problème à résoudre pour l'obtention d'estimations (en gros) exemptes de biais a été celui de la définition des partitions $\{A_{tw}\}_{w=1,\ldots,W}$ et $\{A_{tr}\}_{r=1,\ldots,R}$.

3.1 Estimation de $\boldsymbol{\beta}_{tw}$

Pour résoudre le problème du biais de sélection, nous avons obtenu séparément les estimations de $\boldsymbol{\beta}_{tw}$ en (7) dans chaque groupe de régression et tenté de définir une partition $\{A_{tw}\}_{w=1,\ldots,W}$ telle que, dans chacun de ces groupes, il y ait la plus grande homogénéité possible des paramètres $\boldsymbol{\beta}$ entre *échantillon* et *non-échantillon* constitutifs. La disponibilité à la période *t* de données sur les variables d'intérêt pour toutes les entreprises de la BSS visées l'année précédant la période de référence actuelle permet de vérifier l'homogénéité des $\boldsymbol{\beta}$ à l'aide des données réelles de cette période antérieure sous l'hypothèse de la stabilité des modèles dans le temps.

Nous avons défini les groupes de régression pour trois sous-ensembles de A_t comportant une quantité différente de données auxiliaires. Il y a d'abord *a*) les unités établies pour plus d'une année sur lesquelles nous disposons d'un ensemble complet de variables auxiliaires. Ces variables sont le nombre de salariés, les salaires et traitements et les autres coûts de main-d'œuvre de l'année précédente, ainsi que le nombre total d'entreprises inscrites (environ 310 groupes ont été délimités pour ces unités). Il y a ensuite *b*) les unités établies pour plus d'une année sur lesquelles la *seule* variable auxiliaire disponible est le nombre d'entreprises inscrites (une centaine de groupes). Il y a enfin *c)* les unités établies pour moins d'une année

sur lesquelles les variables auxiliaires disponibles sont le nombre de salariés au moment de l'inscription initiale des entreprises à la BSS, tout comme le nombre total d'entreprises (60 groupes environ). Il faut ajouter que, pour chaque unité incluse dans A_t, nous disposons également de renseignements sur son lieu d'implantation et son activité économique (NACE, 1re révision).

Nous avons défini les partitions pour chacun des trois sous-ensembles a), b) et c) à l'aide de classes naturelles issues d'un croisement de l'*activité économique* (niveau à 2 chiffres de la NACE, 1re révision), de la *région* et de la *catégorie de taille* (pour les seuls sous-ensembles a) et c)), c'est-à-dire le nombre de salariés [observé l'année précédente (sous-ensemble a)) ou au moment de l'inscription initiale de l'entreprise à la BSS (sous-ensemble c))]. Si nous nous reportons aux critères précités pour la définition des groupes de régression, c'est que nous désirons obtenir des sous-ensembles le plus homogènes possible qui nous donnent des estimations (à peu près) exemptes de biais à de bas niveaux d'agrégation, et des groupes de régression sans chevauchement de *domaines d'intérêt* pour la publication des estimations définitives.
Dans certains cas, il a fallu réunir des groupes dont la taille d'échantillon était insuffisante pour dégager des estimations sûres des coefficients de régression. En revanche, dans certains groupes, nous avons *subdivisé* la partition (niveau à 3 chiffres de la NACE, 1re révision) pour une meilleure homogénéité des coefficients de régression estimés entre l'échantillon et le non-échantillon des groupes résultants.

Nous avons alors obtenu des estimations des coefficients du groupe général w ($w = 1,\ldots,W$) :

$$\tilde{\boldsymbol{\beta}}_{tw} = \left[\sum_{i \in s_{tw}} \boldsymbol{x}_{ti} \boldsymbol{x}_{ti}' / c_{ti} \right]^{-1} \sum_{i \in s_{tw}} \boldsymbol{x}_{ti}\, y_{ti} / c_{ti} \qquad (w = 1,\ldots,W). \qquad (8)$$

3.2 Estimation de θ_{tr}

Pour traiter la source possible de biais induit par les erreurs de couverture de la BSS, nous avons introduit les paramètres θ en (5) pour ainsi définir les probabilités qu'une unité au registre appartienne à la population visée. Nous avons tenu ce paramètre pour constant dans chaque sous-ensemble A_{tr} de la partition $\{A_{tr}\}_{r=1,\ldots,R}$ introduite à la section 2. La procédure de définition de la partition visait à l'obtention de sous-groupes d'une homogénéité interne optimale pour ce qui est des probabilités de juste inclusion au registre, mais pour une hétérogénéité maximale *entre les groupes*. Nous avons en outre défini la partition pour que les probabilités soient stables dans le temps, et ce, en essayant de réduire le biais éventuellement causé par l'aménagement de l'information de la BSS, le but étant que les données sur les erreurs de couverture au registre administratif ne soient disponibles que pour des périodes précédant d'un an la période *actuelle* de référence. Nous avons dégagé la partition $\{A_{tr}\}_{r=1,\ldots,R}$ en modélisant la relation entre la variable dichotomique λ introduite en (2) (et disponible à la période t en décalage d'un an par rapport à la période de référence actuelle) et un jeu de covariables de cette base. La partition est fondée sur des classes issues d'un croisement des valeurs des covariables les plus déterminantes (âge de l'entreprise, région, catégorie de taille, activité économique) choisies par régression non paramétrique (Breiman et coll., 1984).

Pour le r^{e} ($r = 1,\ldots,R$) groupe d'erreur au registre, nous estimons les probabilités de juste inclusion à la période t comme la *proportion* d'unités de la population visée observées un an avant la période de référence actuelle :

$$\tilde{\theta}_{tr} = \hat{\theta}_{t-12r} = \sum_{i=1}^{N_{t-12r}} \lambda_{t-12\,i} \Big/ N_{t-12r} \qquad (r = 1,\ldots,R) \qquad (9)$$

où N_{t-12r} désigne le nombre d'unités appartenant au r^{e} groupe d'erreur au registre un an avant la période de référence actuelle, où $\mathrm{E}\left(\hat{\theta}_{t-12r}\right) = \theta_{t-12r}$.

Par substitution des expressions (8) et (9) en (7) et après un certain traitement algébrique, l'estimateur du total de population peut être représenté sous forme linéaire comme

$$\tilde{\tilde{Y}}_t = \sum_{i \in s_t} y_{ti}\, w_{ti} \qquad (10)$$

étant

$$w_{ti} = 1 + \left[\sum_{r=1}^{R} \sum_{l \in s_{t,(w,r)}} \boldsymbol{x}_{tl}\, \hat{\theta}_{t-12r} \right]' \left[\sum_{r=1}^{R} \sum_{l \in s_{t,(w,r)}} \boldsymbol{x}_{tl}\, \boldsymbol{x}_{tl}' / c_{tl} \right]^{-1} \boldsymbol{x}_{ti} / c_{ti} \quad \text{pour } i \in A_{tw} \qquad (11)$$

où l désigne la l^{e} entreprise dans $s_{t(w,r)}$.

3.3 Erreur quadratique moyenne de $\tilde{\tilde{Y}}_t$

L'erreur quadratique moyenne (EQM) de l'estimateur (7) est la somme de la variance du modèle et du carré du biais de modèle :

$$\mathrm{EQM}(\tilde{\tilde{Y}}_t) = \mathrm{E}\left(\tilde{\tilde{Y}}_t - Y_t\right)^2 = \mathrm{V}(\tilde{\tilde{Y}}_t - Y_t) + \left[\mathrm{E}(\tilde{\tilde{Y}}_t - Y_t)\right]^2 = \mathrm{V}(\tilde{\tilde{Y}}_t - \tilde{\tilde{Y}}_t) + \left[\mathrm{Biais}(\tilde{\tilde{Y}}_t)\right]^2 . \quad (12)$$

Il y a deux grandes sources de biais de modèle :

1. une piètre définition des groupes de régression, si bien que le vecteur des coefficients $\boldsymbol{\beta}$ peut systématiquement différer entre échantillon et non-échantillon : on peut alors supposer que, dans chaque groupe de régression, le vecteur des coefficients $\boldsymbol{\beta}_{\bar{s}_{tw}}$ ($w = 1,\ldots,W$) du non-échantillon peut être représenté comme la somme de $\boldsymbol{\beta}_{tw}$ et d'un vecteur $\boldsymbol{\alpha}_{tw}$ d'effets fixes, ce qui donne $\boldsymbol{\beta}_{\bar{s}_{tw}} = \boldsymbol{\beta}_{tw} + \boldsymbol{\alpha}_{tw}$;
2. l'hypothèse erronée de la constance des probabilités d'inclusion au registre entre deux années consécutives.

Compte tenu de ces points, une expression de l'approximation du premier ordre du carré du biais de $\tilde{\tilde{Y}}_t$ est donné par :

$$\left[\mathrm{Biais}(\tilde{\tilde{Y}}_t)\right]^2 \cong \left[\sum_{w=1}^{W} \sum_{r=1}^{R} \left[\boldsymbol{\beta}_{tw}\, \theta_{t-12r} - (\boldsymbol{\beta}_{tw} + \boldsymbol{\alpha}_{tw})\, \theta_{tr} \right]' \sum_{i \in \bar{s}_{t(w,r)}} \boldsymbol{x}_{ti} \right]^2 .$$

La variance de $\tilde{\tilde{Y}}_t$ dans (12) peut s'écrire de la façon suivante

$$\mathrm{V}(\tilde{\tilde{Y}}_t - Y_t) = \mathrm{V}_1 + \mathrm{V}_2 = \mathrm{V}\left[\sum_{w=1}^{W} \sum_{r=1}^{R} \sum_{i \in \bar{s}_t(r,w)} y_{ti}\, \lambda_{ti} \right] + \mathrm{V}\left[\sum_{w=1}^{W} \sum_{r=1}^{R} \tilde{\theta}_{tr}\, \tilde{\boldsymbol{\beta}}'_{tw} \sum_{i \in \bar{s}_{t(r,w)}} \boldsymbol{x}_{ti} \right]. \quad (13)$$

Avec les hypothèses du modèle (6), V_1 est donné par

$$\mathrm{V}_1 = \sum_{w=1}^{W} \sum_{r=1}^{R} \sum_{i \in \bar{s}_t(w,r)} c_{ti} \sigma_{tw}^2\, \theta_{tr} + \theta_{tr}(1 - \theta_{tr})\, \boldsymbol{\beta}_{tw}'\, \boldsymbol{x}_{ti}\, \boldsymbol{x}_{ti}'\, \boldsymbol{\beta}_{tw} \quad . \qquad (14)$$

On peut exprimer sous la forme suivante l'approximation du premier ordre de V_2 :

$$V_2 \cong V\left[\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in \bar{s}_{t(r,w)}}\left[\theta_{t-12r}\,\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}+\boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\tilde{\theta}_{t-12r}\right]\right]= \tag{15}$$

$$=\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in \bar{s}_{t(r,w)}}\left[\theta^2_{t-12r}\,\sigma^2_{tw}\,\boldsymbol{x}'_{ti}\left(\sum_{i\in s_{tw}}\boldsymbol{x}_{ti}\,\boldsymbol{x}'_{ti}/c_{ti}\right)^{-1}\boldsymbol{x}_{ti}+\left(\theta_{t-12r}\left(1-\theta_{t-12r}\right)\right)\boldsymbol{x}'_{ti}\boldsymbol{\beta}_{tw}\boldsymbol{\beta}'_{tw}\,\boldsymbol{x}_{ti}\right].$$

À l'aide des données réelles disponibles sur l'ensemble de la population et avec une période de référence précédant d'un an la période de référence actuelle (comme nous l'avons mentionné), il est possible d'établir une estimation de l'EQM par

$$\tilde{\mathrm{EQM}}(\tilde{Y}_t)=\left(\tilde{Y}_t-Y_t\right)^2 \tag{16}$$

où, par l'expression (12), on a $\mathrm{E}\left[\tilde{\mathrm{EQM}}(\tilde{Y}_t)\right]=\mathrm{E}\left(\tilde{Y}_t-Y_t\right)^2=\mathrm{EQM}(\tilde{Y}_t)$.

À partir de l'échantillon, on peut calculer l'estimation de la variance pour la période en cours en introduisant les estimations des paramètres inconnus σ^2_{tw}, $\boldsymbol{\beta}_{tw}$, θ^2_{t-12r}, θ_{t-12r} dans les expressions (14) et (15), ce qui donne :

$$\tilde{\mathrm{V}}(\tilde{Y}_t-Y_t)=\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in \bar{s}_{t\,(w,r)}}c_{ti}\tilde{\sigma}^2_{tw}\,\tilde{\theta}_{tr}+\theta_{tr}\left(1-\tilde{\theta}_{tr}\right)\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}\,\boldsymbol{x}'_{ti}\,\tilde{\boldsymbol{\beta}}_{tw}+ \tag{17}$$

$$+\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{i\in \bar{s}_{t(w,r)}}\left[\left(2\hat{\theta}^2_{t-12r}-\hat{\theta}_{t-12r}\right)\tilde{\sigma}^2_{tw}\,\boldsymbol{x}'_{ti}\left(\sum_{i\in s_{tw}}\boldsymbol{x}_{ti}\,\boldsymbol{x}_{ti}/c_{ti}\right)^{-1}\boldsymbol{x}_{ti}+\left(\hat{\theta}_{t-12r}\left(1-\hat{\theta}_{t-12r}\right)\right)\boldsymbol{x}'_{ti}\tilde{\boldsymbol{\beta}}_{tw}\,\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}\right]$$

avec

$$\tilde{\sigma}^2_{tw}=\frac{1}{n_{tw}-M_{tw}}\sum_{i=1}^{n_{tw}}\left(y_{ti}-\tilde{\boldsymbol{\beta}}'_{tw}\,\boldsymbol{x}_{ti}\right)^2/c_{ti}\quad; \tag{18}$$

où n_{tw} est le nombre d'unités de l'échantillon s_{tw}, alors que ($2\hat{\theta}^2_{t-12r}-\hat{\theta}_{t-12r}$) représente une estimation sans biais du carré de la probabilité θ^2_{t-12r}.

Ainsi, un an après la période en cours, on peut obtenir une estimation du carré du biais qui est elle-même exempte de biais :

$$\left[\tilde{\mathrm{Biais}}(\tilde{Y}_t)\right]^2=\left(\tilde{Y}_t-Y_t\right)^2-\tilde{\mathrm{V}}(\tilde{Y}_t-Y_t)\,. \tag{19}$$

Pour obtenir une estimation *actuelle* du biais, il est peut-être bon de procéder comme on le fait normalement avec les fonctions *généralisées de variance* (Wolter, 1985) en supposant que le biais quadratique relatif est une fonction décroissante de la fraction de sondage (*n/N*). Le biais quadratique relatif peut alors se modéliser de la manière suivante :

$$\ln\left|\left(\tilde{\mathrm{Biais}}(\tilde{Y}_t)\right)^2/\tilde{Y}_t^{\,2}\right|=a_t+b_t\,\ln\left(n_t/N_t\right)+\varepsilon_t \tag{20}$$

où a_t et b_t sont des coefficients fixes et ε_t, un résidu aléatoire. Si on pose que les coefficients a_t et b_t sont à peu près constants dans le temps, on peut estimer le biais quadratique *actuel* :

$$\left(\tilde{\tilde{\mathrm{Biais}}}(\tilde{Y}_t)\right)^2=\tilde{Y}_t^{\,2}\exp\left(\tilde{a}_{t-1}+\tilde{b}_{t-1}\ln\left(n_t/N_t\right)\right) \tag{21}$$

où $\tilde{a}_{t-1}$ et $\tilde{b}_{t-1}$ sont des estimations des paramètres *a* et *b* obtenues un an avant la période en cours.

4. QUELQUES INDICATIONS EMPIRIQUES

La méthode présentée aux sections 2 et 3 sert maintenant à l'estimation de l'emploi total et des salaires et traitements en valeur brute en Italie au premier trimestre de 2000 pour la population d'entreprises des secteurs de l'industrie et des services qui comptent moins de 500 salariés.

Pour chaque variable d'intérêt, nous avons établi des estimations de variance et de biais quadratique en nous reportant respectivement aux expressions (17) et (19). Pour juger du rendement de cette opération d'estimation et de l'incidence relative du biais qui entre dans l'erreur quadratique moyenne par rapport à la variance, nous avons dégagé les mesures suivantes : (*i*) la racine carrée de l'erreur quadratique moyenne relative en pourcentage : $\mathrm{REQMR}\,(\tilde{Y}) = 100\sqrt{\mathrm{E\tilde{Q}M}\,(\tilde{Y})}\,/\tilde{Y}$, qui représente l'ordre de grandeur de la racine de l'EQM pour les estimations; (*ii*) biais relatif en pourcentage : $\mathrm{BR}\,(\tilde{Y}) = 100\,\tilde{\mathrm{B}}\mathrm{ias}\,(\tilde{Y})/\tilde{Y}$; pourcentage de la composante biais : $\mathrm{CBR}\,(\tilde{Y}) = 100\left[\tilde{\mathrm{B}}\mathrm{iais}(\tilde{Y})\right]^2 / \mathrm{E\tilde{Q}M}\,(\tilde{Y})$.

Au tableau 1 à la page suivante, nous présentons les indices qui précèdent pour l'emploi et les salaires et traitements en valeur brute dans l'ensemble de l'économie et dans chaque secteur de l'industrie.

Il suffit d'observer les tableaux pour juger que les estimations sont d'une bonne précision pour l'EQMR : les estimations du nombre de salariés et des salaires et traitements en valeur brute dans l'ensemble de la population présentent respectivement une EQMR de 0,45 % et 0,03 %. Comme on pouvait s'y attendre, la variabilité des estimations tient principalement de la composante biais.

5. EXAMEN ET OBSERVATIONS EN CONCLUSION

Dans le présent document, nous avons présenté une méthode d'estimation des indicateurs de facteur travail à l'aide de données administratives. Dans une foule d'ISN, on s'intéresse de plus en plus aux données tirées de sources administratives comme solution d'appoint ou de rechange aux enquêtes directes, ce qui devrait permettre de répondre à une demande croissante de données statistiques actuelles et détaillées, surtout dans le cas des enquêtes auprès des entreprises, des exploitations agricoles et des établissements.

On doit préciser que la base de données sur la sécurité sociale recèle un riche jeu de variables sur le marché du travail et que, dans ce cas, on ne se trouve pas à imposer un plus grand fardeau aux entreprises. Il reste que le recours à la BSS pour la statistique à court terme de l'emploi et des salaires présente certains inconvénients auxquels nous avons tenté de nous attaquer dans notre exposé. Plus précisément,

1. l'unité de base de la BSS ne correspond pas à une définition de l'unité statistique qui se prête à l'analyse d'un système économique;
2. le sous-ensemble de données s'appliquant à la période en cours pour l'estimation des coefficients de modèle n'est pas un échantillon aléatoire;
3. les procédures de collecte de données et les problèmes de mise à jour de la BSS (unités introduites, retirées ou transformées) causent des problèmes de couverture de cette source administrative, si bien que la liste des unités actives définies par cette base peut systématiquement différer de la population visée.

La détermination des unités a comporté un premier traitement de la BSS en vue du regroupement en un même enregistrement de tous les états de rémunération que l'on jugeait appartenir à la même entité

économique; il a fallu utiliser différentes variables clés pour apparier les divers enregistrements de la base de données.

Nous avons traité la question du biais de sélection créé par le caractère non aléatoire du sous-ensemble d'unités observées pour la période en cours en ajustant le modèle à l'intérieur de sous-populations homogènes (groupes de *régression*).

Nous avons pris en compte les erreurs de couverture par excès – qui s'expliquent en grande partie par les disparitions non inscrites à la BSS – par une modélisation des probabilités qu'une unité du registre administratif appartienne à la population visée à l'intérieur de sous-ensembles (*groupes d'erreur au registre*); nous avons supposé que, dans chaque sous-ensemble, les probabilités en question étaient à peu près égales.

Quelques problèmes demeurent sans solution dans la méthode d'estimation que nous proposons. Il y a principalement la question de la modélisation explicite des erreurs de couverture à l'aide, par exemple, de modèles par espace d'états (Tam, 1987). Avec une telle modélisation, on tiendrait compte de la variabilité dans le temps des probabilités de juste inclusion au registre administratif en fonction des variations du cycle économique. On pourrait retenir une démarche semblable de modélisation des paramètres $\boldsymbol{\beta}$ par un modèle récursif d'actualisation des estimations où on partirait des valeurs observées de la période précédente.

TABLEAU 1. – EQM et biais pour les variables « nombre de salariés » et « salaires et traitements en valeur brute » par secteur d'activité économique pour la population d'entreprises comptant moins de 500 salariés en Italie au 1er trimestre 2000

Secteur de l'industrie	Nombre de salariés				Salaires et traitements en valeur brute		
	Valeur observée	REQMR	BR	CBR	REQMR	BR	CBR
Fabrication	3 236 249	0,87	0,85	95,95	0,03	0,028	87,12
Construction	746 553	1,03	1,00	93,95	1,88	1,86	98,85
Commerce de gros et de détail	1 157 101	1,28	1,27	98,10	0,85	0,82	94,66
Hébergement, aliments et boissons	326 975	2,31	2,28	97,49	4,83	4,82	99,38
Transports et communications	396 060	1,20	1,07	80,39	0,34	0,30	78,24
Finances, assurances et affaires immobilières	148 052	1,33	1,00	56,61	4,35	3,19	53,83
Services aux entreprises	850 568	0,47	0,33	49,77	0,24	0,20	70,18
Total	6 861 556	0,45	0,44	94,29	0,03	0,029	93,65

BIBLIOGRAPHIE

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984), *Classification and Regression Trees*, Wadsworth International, Belmont, CA

.

Chambers, R. L. (1997), "Weighting and calibration in sample survey estimation", dans: Malaguerra, C., Morgenthaler, S., Ronchetti, E. (éds), *Proceedings of the Conference on Statistical Science honouring the Bicentennial of Stefano Franscini's Birth*, Basel: Birkhäuser Verlag.

Falorsi P. D., Pallara A., Succi R., et Russo A. (2000), "Estimating indicators of labour input from administrative records having coverage and measurement errors", *ICES II – Proceedings of the Second International Conference on Establishment Surveys*, (available at web site *http://www.eia.doe.gov/ices2/#errata*).

Hidiroglou, M.A., Latouche, M., Armstrong, B., et Gossen, M. (1995), "Improving Survey Information Using Administrative Records: the Case of the Canadian Employment Survey," *Proceedings of the Annual Research Conference, U.S Bureau of the Census*, pp. 171-197.

Royall, R.M. (1988), "The Prediction Approach to Sampling Theory", dans: Krishnaiah, P.R., Rao, C.R. (éds.), *Handbook of Statistics, vol. 6*, Elsevier Science Publishers, pp. 399-413.

Tam, S.M. (1987), "Analysis of Repeated Surveys Using a Dynamic Linear Model," *International Statistical Review*, 55, pp. 63-73.

Wolter, K.M. (1985), *Introduction to variance estimation*, New-York; Springer-Verlag.