Construction des équivalents plein temps pour la Statistique Structurelle Suisse des Entreprises

Desislava Nedyalkova [1] & Daniel Assoulin [2]

[1] *Office fédéral de la statistique OFS, Espace de l'Europe 10, CH-2010 Neuchâtel*
desislava.nedyalkova@bfs.admin.ch
[2] *Office fédéral de la statistique OFS, Espace de l'Europe 10, CH-2010 Neuchâtel*
daniel.assoulin@bfs.admin.ch

Résumé. La construction des équivalents plein temps (EPT) pour la statistique structurelle suisse des entreprises (STATENT) se base sur des données d'enquêtes appariées aux données administratives de l'assurance-vieillesse et survivants (AVS) [1] et du registre des entreprises et établissements (REE). Les EPT ne sont pas recensés dans l'AVS. Ils doivent donc être construits sur la base d'un modèle estimé sur des données d'enquête appariées aux données de l'AVS. Cependant on observe des incohérences au niveau des emplois entre les différentes sources. Ces differences sont traitées dans le but d'obtenir une cohérence entre les EPT d'enquête et les données AVS. Les analyses mettent en cause l'utilisation d'une première approche traitant les incohérences par un simple ajustement par le ratio. Une autre approche a donc été développée. Elle permet de traiter ces différences en tenant mieux compte du type d'emplois.

Mots-clés. données administratives, harmonisation des données, intégration des données, modélisation, équivalents plein temps

1 Introduction

Le recensement des entreprises (RE) en Suisse a été effectué pour la dernière fois en 2008. Depuis 2011, le RE est modernisé par un nouveau système intégré appelé STATENT (statistique des entreprises) basé sur des données du registre des entreprises et établissements (REE), du registre des caisses de compensation (AVS) et d'enquêtes complémentaires comme l'enquête sur la statistique des emplois (STATEM). Les données de l'AVS sont appariées au niveau entreprise avec les données du REE.

Le passage du RE à la STATENT induit plusieurs changements de définition et de méthode. Les différences principales entre les deux univers portent sur les unités et les

[1] L'AVS est le principal pilier de la prévoyance vieillesse et survivants en Suisse. L'AVS est obligatoire et a pour but de couvrir les besoins vitaux d'une personne assurée en cas de retraite, ou de sa famille en cas de décès.

emplois [2] recensés ainsi que la périodicité. Par exemple, le RE avait lieu tous les 3-4 ans tandis que la STATENT est réalisée chaque année. Le RE se référait à une date précise tandis que la STATENT plutôt à un mois de référence (décembre).

Une autre différence majeure entre le RE et la nouvelle statistique structurelle des entreprises est la façon dont les EPT sont calculés. Dans le RE, le calcul des EPT était réalisé en tenant compte des informations relevées sur les taux d'occupation. Pour la STATENT, vu que ces informations sur les taux d'occupation ne sont plus disponibles, les EPT doivent être construits à l'aide d'un modèle (modèle EPT) utilisant des variables explicatives issues des registres (AVS, REE). Pour des entreprises pour lesquelles on dispose des EPT provenant d'une enquête complémentaire, ces derniers seront utilisés dans la STATENT. L'intégration des données d'enquête et des données AVS révèle des incohérences d'emplois entre les différentes sources. Ces différences doivent être traitées dans le but d'obtenir une cohérence entre les EPT d'enquête et les données AVS.

Nous commençons par une description du modèle EPT basé sur les données d'enquête appariées aux données AVS. Nous présentons ensuite deux manières de traiter les incohérences entre les différentes sources. Nous décrivons la première approche basée sur un ajustement par le ratio et la mettons en lien avec le modèle EPT. Nous exposons ensuite une deuxième approche basée sur l'idée de traiter les différences observées comme des échantillons PPS[3] de taille fixe et examinons sa mise en oeuvre.

2 Modèle EPT

Les données utilisées pour la construction du modèle pour l'année de référence 2011 sont principalement celles des mono-établissements de l'enquête STATEM, quatrième trimestre 2011, appariées aux données du registre AVS 2011 (mois de décembre). Le modèle est estimé séparément pour chacun des sous-ensembles suivants: hommes/secteur 2, hommes/secteur 3, femmes/secteur 2, femmes/secteur 3. Par souci de simplicité nous utilisons la même notation pour chaque sous-ensemble.

Sur la base des informations concernant le taux d'occupation relevés dans l'enquête et la distribution des salaires en provenance de l'AVS on construit pour chaque agrégat NOGA section [4] quatres classes de salaire. Ces classes de salaire forment la base pour la construction de nos variables explicatives. La variable d'intérêt est le nombre d'EPT.

Le modèle que nous voulons estimer est:

$$y_i = \alpha_1 . V_{i1} + \sum_{j=2}^{4} \beta_{jkl} V_{ij} + \epsilon_i, \tag{1}$$

[2] Le recensement tenait compte des employées travaillant aux moins 6h par semaine, tandis que la STATENT 2011 contient les employées avec un salaire annuel $\geq$ 2300 CHF en 2011.

[3] Probability proportional to size.

[4] Nomenclature générale des activités économiques

où:

- y_i, le nombre d'EPT d'une entreprise i,
- V_{ij}, nombre d'employé(e)s de l'entreprise i dans la classe de salaire j ($j = 1, ..., 4$). $j = 1$ contient les plus petits et $j = 4$ les plus grands salaires. On note que $\sum V_{ij} = \text{EMPTOT\_AVS}$.
- α_1, coefficient de régression pour V_{i1},
- $\beta_{jk\ell}$, coefficient de régression pour V_{ij} dans la grande région k ($k = 1, ..., 7$) et la section NOGA ℓ,
- ϵ_i, résidu avec $E(\epsilon_i) = 0$ et $\text{Var}(\epsilon_i) = \sigma^2 \text{EMPTOT\_AVS}_i$.

3 Harmonisation des variables

Les données appariées sur lesquelles le modèle EPT est estimé présentent des différences entre le nombre d'employés (EMPTOT) de l'enquête et celui du registre AVS. Ces différences peuvent être p.ex. dues à des erreurs de mesure ou à la définition de l'emploi. Pour la STATENT la source de référence pour les variables de l'emploi est le registre AVS. D'où le besoin d'harmoniser les EPT d'enquête pour qu'ils soient en cohérence avec la source AVS. Pour traiter ces incohérences, il existe différentes méthodes d'ajustement de variables comme le *prorating* et le *generalized ratio adjustment* (Panekoek, 2011; Panekoek, 2014). Par exemple, la méthode de *prorating* est un ajustement multiplicatif simple appliqué sur des variables impliquées dans des règles de contrôle.

3.1 Première approche pour le traitement des différences

En s'inspirant des méthodes d'ajustement mentionnées ci-dessus, pour chaque entreprise i, nous définissons une nouvelle variable EPT_AVS (par sexe) comme suit:

$$\text{EPT\_AVS}_i = \eta_i \text{EPT\_S}_i, \tag{2}$$

où $\eta_i = \text{EMPTOT\_AVS}_i / \text{EMPTOT\_S}_i$.

Cette nouvelle variable *harmonisée avec la source AVS* va être utilisée pour la modélisation, de sorte que les EPT prédits seront cohérents avec les valeurs de EMPTOT_AVS. Si les incohérences sont traitées selon la définition (2), nous pouvons réécrire l'équation du modèle (1) de la manière suivante:

$$\eta_i \text{EPT\_S}_i = \alpha_1 . V_{i1} + \sum_{j=2}^{4} \beta_{jkl} V_{ij} + \epsilon_i, \tag{3}$$

ou encore

$$\text{EPT\_S}_i = \alpha_1 . \frac{V_{i1}}{\eta_i} + \sum_{j=2}^{4} \beta_{jkl} \frac{V_{ij}}{\eta_i} + \frac{\epsilon_i}{\eta_i}, \tag{4}$$

Si nous supposons que $\eta_i > 1$, nous pouvons interpréter l'équation (4) comme ceci: L'ajustement de EMPTOT_AVS se fait uniformément dans les quatre classes de salaire en réduisant le nombre d'employés par η_i. Ceci se justifie si les incohérences sont indépendantes des classes de salaire.

3.2 Nouvelle approche pour le traitement des différences

Nous présentons une alternative du modèle (2) dans laquelle les incohérences des variables ne sont plus traitées de manière uniforme. Nous étudions séparément les cas suivants:

- Cas 1: EMPTOT_AVS > EMPTOT_S.
- Cas 2: EMPTOT_S > EMPTOT_AVS.
- Cas 3: EMPTOT_S = EMPTOT_AVS.

Soient diff_ab = EMPTOT_AVS−EMPTOT_S (par sexe) et diff_ba = EMPTOT_S−EMPTOT_AVS (par sexe). Notons qu'une entreprise pour laquelle les EMPTOT sont trop différents, que ce soit chez les femmes ou chez les hommes, sera traitée comme si l'information d'enquête est manquante.

3.2.1 Traitement du premier cas

Nous supposons que EMPTOT_AVS > EMPTOT_S et estimons les coefficients du modèle suivant (par sexe et par secteur):

$$\text{diff\_ab}_i = \sum_{j=1}^{4} \tilde{\beta}_j V_{ij} + \epsilon_i,$$

sous l'hypothèse $\text{Var}(\epsilon_i) = \sigma^2 \text{EMPTOT\_AVS}_i$. Ceci n'est pas fait dans le but de modéliser (estimer) diff_ab_i, mais dans celui d'obtenir des valeurs $\tilde{\beta}_j$ estimés, $\widehat{\tilde{\beta}}_j$, qui peuvent être vues comme des mesures de la probabilité qu'une personne appartenant à l'ensemble des salariés dans la classe de salaire j ne soit pas prise en compte dans le nombre d'emplois de l'enquête.

Le tableau 1 contient les estimations, $\widehat{\tilde{\beta}_j}$, obtenues avec la procédure ROBUSTREG de SAS avec des poids proportionnels à $1/\text{EMPTOT\_AVS}_i$. Il nous indique par exemple que les coefficients pour la classe 1 (les plus petits salaires) sont plus grandes que ceux pour la classe 2. Ainsi, un traitement uniforme selon (4) ne semble pas être justifié.

	Secteur 2		Secteur 3	
$\widehat{\tilde{\beta}_j}$	Hommes	Femmes	Hommes	Femmes
$\widehat{\tilde{\beta}}_1$	0.6586	0.7239	0.8076	0.6925
$\widehat{\tilde{\beta}}_2$	0.5618	0.4952	0.6238	0.3904
$\widehat{\tilde{\beta}}_3$	0.1532	0.2406	0.3285	0.1251
$\widehat{\tilde{\beta}}_4$	0.0385	0.0507	0.0876	0.1293

Table 1: Tableau des valeurs $\widehat{\tilde{\beta}_j}$

Idée initiale - tirage d'un échantillon PPS à taille fixe

On suppose que, pour une entreprise i, la différence observée $n_i = \text{diff\_ab}_i$ correspond à un échantillon s_i

- de taille fixe (n_i),
- à probabilités proportionnelles aux $\widehat{\tilde{\beta}_j}$ (voir tableau 1)

de personnes qui sont dans l'AVS, mais pas dans l'enquête. Ainsi, pour une personne d de la classe de salaire j:

$$P(d \in s_i) \quad = \quad n_i \frac{\text{mos}_d}{\text{mos}_i} = \frac{n_i \widehat{\tilde{\beta}_j}}{\sum_{j=1}^4 \widehat{\tilde{\beta}_j} V_{ij}} = \widehat{\tilde{\beta}_j} \frac{n_i}{\widehat{n_i}} = \widehat{\tilde{\beta}}_j^*,$$

où $\text{mos}_d = \widehat{\tilde{\beta}_j}$ et $\text{mos}_i = \sum_{d \in i} \text{mos}_d = \sum_{j=1}^4 \widehat{\tilde{\beta}_j} V_{ij} = \widehat{n_i}$. Si $n_i \frac{\text{mos}_d}{\text{mos}_i} \geq 1$ alors la personne est automatiquement retirée (procédure selon Särndal et al. (1992, p.89)). On note que $\widehat{\tilde{\beta}}_j^*$ peut être interprété comme un $\widehat{\tilde{\beta}_j}$ ajusté de manière à ce que $\sum_{j=1}^4 \widehat{\tilde{\beta}}_j^* V_{ij} = n_i$.

Calcul du nombre moyen de personnes devant être supprimées dans chaque classe

Les inconvénients d'utiliser un tirage PPS sont son aspect aléatoire et le fait que sa mise en oeuvre dans la production est plutôt compliquée. À la place d'utiliser un tirage aléatoire nous pouvons calculer le nombre moyen de personnes de la classe j sélectionnées dans l'échantillon s_i. Ce nombre moyen de personnes est donné par:

$$E(\sum_{d \in \mathcal{V}_{ij}} 1(d \in s_i)) \quad = \quad V_{ij} P(d \in s_i) = V_{ij} \widehat{\tilde{\beta}}_j^*, \tag{5}$$

où $\mathcal{V}_{ij}$ désigne l'ensemble des employés de l'entreprise i dans la classe de salaire j.

Comme dans le cas du tirage PPS, notre procédure commence par supprimer toutes les personnes pour lesquelles $\widehat{\widetilde{\beta}}_j^* \geq 1$. Ensuite, nous calculons le nombre moyen de personnes restantes qui doivent être éliminées selon l'équation (5). A la fin de la procédure, nous obtenons de nouvelles variables $\widetilde{V}_{ij} := V_{ij} - V_{ij}\widehat{\widetilde{\beta}}_j^*$ telles que $\sum_{j=1}^4 \widetilde{V}_{ij} = \text{EMPTOT\_S}_i$. Ces nouvelles variables vont remplacer les variables V_{ij} dans l'estimation du modèle (1) où y_i est donnée par EPT_S_i.

3.2.2 Traitement du deuxième cas

Nous supposons que EMPTOT_S > EMPTOT_AVS. Connaissant le nombre d'employés travaillant à temps partiel 3 (T_{i1}) , temps partiel 2 (T_{i2}) , temps partiel 1 (T_{i3}) et plein temps (T_{i4}) selon l'enquête, nous estimons par sexe et par secteur les coefficients du modèle suivant:

$$\text{diff\_ba}_i = \sum_{j=1}^{4} \gamma_j T_{ij} + \epsilon_i,$$

La procédure utilisée est PROC ROBUSTREG de SAS avec des poids proportionnels à $1/\text{EMPTOT\_S}_i$. Les estimations obtenues (tableau 2) peuvent être vues comme des mesures de la probabilité qu'une personne appartenant a l'ensemble des salariés travaillant à temps partiel 3, 2, 1 ou plein temps ne soit pas prise en compte dans le nombre d'employés de l' AVS. Le tableau 2 nous indique que les coefficients pour les personnes travaillant à temps partiel 3 ($\widehat{\gamma}_1$) sont plus grandes que ceux pour les personnes travaillant à plein temps ($\widehat{\gamma}_4$).

	Secteur 2		Secteur 3	
$\widehat{\gamma}_j$	Hommes	Femmes	Hommes	Femmes
$\widehat{\gamma}_1$	0.4687	0.5480	0.4484	0.4415
$\widehat{\gamma}_2$	0.6569	0.3285	0.2372	0.2513
$\widehat{\gamma}_3$	0.2558	0.0634	0.2031	0.0018
$\widehat{\gamma}_4$	0.0713	0.0654	0.1230	0.1399

Table 2: Tableau des valeurs préliminaires de $\widehat{\gamma}_j$

Adaptation des calculs pour le deuxième cas

Nous savons que pour éliminer les différences des EMPTOT et rendre les données cohérentes, nous devons supprimer pour chaque entreprise un nombre fixe de personnes, $n_i^* = \text{diff\_ba}_i$, contenues dans l'enquête. En utilisant les coefficients $\widehat{\gamma}_j$, nous appliquons la même procédure que pour le cas 1, avec les modifications requises, qui nous conduit aux variables $\widetilde{T}_{ij}$ telles que $\sum_{j=1}^4 \widetilde{T}_{ij} = \text{EMPTOT\_AVS}_i$.

Nous supposons que les équivalents plein temps de l'enquête peuvent être modélisés comme suit:

$$\text{EPT\_S}_i = \sum_{j=1}^{4} \delta_j T_{ij} + \epsilon_i, \tag{6}$$

où $\text{Var}(\epsilon_i) = \sigma^2 \text{EMPTOT\_S}_i$. Le modèle estimé, par sexe et par secteur, utilise la procédure GLM de SAS avec des poids proportionnels à $1/\text{EMPTOT\_S}_i$. Les résultats sont donnés dans le tableau 3.

	Secteur 2		Secteur 3	
$\widehat{\delta}_j$	Hommes	Femmes	Hommes	Femmes
$\widehat{\delta}_1$	0.1101	0.0946	0.0783	0.0908
$\widehat{\delta}_2$	0.2995	0.2929	0.2775	0.2807
$\widehat{\delta}_3$	0.6319	0.6543	0.6451	0.6603
$\widehat{\delta}_4$	0.9977	0.9938	0.9976	0.9879

Table 3: Tableau des estimations $\widehat{\delta}_j$

En utilisant les coefficients estimés du modèle (6), nous calculons une variable EPT_S ajustée et cohérente avec la variable EMPTOT_AVS, que nous désignons par $\widetilde{\text{EPT\_S}}$ et qui sera la variable modélisée dans le deuxième cas:

$$\widetilde{\text{EPT\_S}}_i = \text{EPT\_S}_i \frac{\sum \widehat{\delta}_j \widetilde{T}_{ij}}{\sum \widehat{\delta}_j T_{ij}}.$$

3.3 Implications pour l'estimation du modèle EPT

Le tableau 4 présente les variables qui seront utilisées pour l'estimation du modèle EPT dans le cas 1, le cas 2 et le cas d'égalité entre EMPTOT_S et EMPTOT_AVS.

Cas	Variable d'intérêt	Variables explicatives
1	EPT_S	$\widetilde{V}_{ij}$
2	$\widetilde{\text{EPT\_S}}$	V_{ij}
3	EPT_S	V_{ij}

Table 4: Variables utilisées dans le modèle

L'estimation du modèle se base donc sur une harmonisation du nombre d'employés de l' AVS et de l'enquête. On notera que l'on fait ainsi l'hypothèse implicite que ces deux ensembles de même taille correspondent aux mêmes employés.

3.4 Calcul des EPT dans la STATENT

Pour toutes les entreprises pour lesquelles on a les informations d'enquête (EPT et EMPTOT), la variable EPT_STATENT sera calculée ainsi:

Cas 1: $\text{EPT\_STATENT}_i = \text{EPT\_ENQUETE}_i \frac{\sum_{j=1}^{4} \widehat{\beta}_j V_{ij}}{\sum_{j=1}^{4} \widehat{\beta}_j \widetilde{V}_{ij}}$, où $\widehat{\beta}_j$ sont les coefficients estimés du modèle EPT.

Cas 2: $\text{EPT\_STATENT}_i = \text{EPT\_ENQUETE}_i \frac{\sum_{j=1}^{4} \widehat{\delta}_j \widetilde{T}_{ij}}{\sum_{j=1}^{4} \widehat{\delta}_j T_{ij}}$, où $\widehat{\delta}_j$ sont les coefficients estimés du modèle (6).

Cas 3: $\text{EPT\_STATENT}_i = \text{EPT\_ENQUETE}_i$.

Pour toutes les autres entreprises les EPT se basent sur le modèle EPT (1).

4 Conclusion

Les données provenant des différentes sources contiennent des incohérences par rapport aux nombre d'employés. Les données doivent être harmonisées pour obtenir une cohérence entre les EPT d'enquête et les données AVS, source de référence pour les variables d'emploi dans STATENT. Les analyses remettent en question l'application d'un ajustement simple des EPT par le ratio suivant la définition (2). Les divergences semblent plutôt être dues à de bas revenus ou à de faibles taux d'occupation. Un tirage de taille fixe à probabilités proportionnelles à la taille paraît plus approprié pour traiter ce type d'incohérences, mais a l'inconvenient d'être aléatoire et difficile à mettre en oeuvre. L'utilisation de tailles d'échantillon espérées à la place d'un échantillonnage aléatoire pour ajuster le nombre d'employés dans les différentes classes permet de surmonter ce problème.

Bibliographie

[1] Pannekoek, J. (2011), Models and algorithms for micro-integration, dans Final Report on WP2: Methodological developments, ESSNET on Data Integration, disponible sur http://www.cros-portal.eu/content/wp2-development-methods

[2] Panekoek, J. (2014) Method: Reconciling Conflicting Microdata, dans *Memobust Handbook on Methodology of Modern Business Statistics*, disponible sur http://www.cros-portal.eu/content/reconciling-conflicting-microdata-method

[3] Särndal, C.E., Swensson, B. et Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.