Enquête sur la structure des salaires (LSE): révision de la pondération

Clément Chevalier, Lionel Qualité
Office Fédéral de la Statistique

OFS / DSSM / METH

Colloque Francophone sur les Sondages
06 octobre 2021

Sommaire

L'enquête Suisse sur la structure des salaires

Enquête qui a lieu tous les deux ans :

- Environ 45'500 unités primaires (entreprises et administrations) interrogées dans l'enquête : échantillon brut
- Environ 33'500 unités répondantes, dans l'échantillon net
- Les unités livrent les salaires de leurs employés : environ 1'740'000 salaires utilisables pour les estimations.

L'enquête Suisse sur la structure des salaires

Enquête qui a lieu tous les deux ans :

- Environ 45'500 unités primaires (entreprises et administrations) interrogées dans l'enquête : échantillon brut
- Environ 33'500 unités répondantes, dans l'échantillon net
- Les unités livrent les salaires de leurs employés : environ 1'740'000 salaires utilisables pour les estimations.

Objectif : calculer une pondération pour ces salaires de sorte à optimiser la précision des résultats.

Plan de sondage

- Tirage de l'échantillon brut d'unités primaires selon un plan de Poisson avec taux de sondage homogène par strates.
- Stratification par NOGA x Grande Région x Classe de Taille

Plan de sondage

- Tirage de l'échantillon brut d'unités primaires selon un plan de Poisson avec taux de sondage homogène par strates.
- Stratification par NOGA x Grande Région x Classe de Taille

- Détails dans :
  Qualité, L. & Potterat, J. (2020). *Rapport de méthodes : Révision du plan de sondage pour l'enquête suisse sur la structure des salaires 2018*, Office Fédéral de la statistique OFS.

Poids d'un salaire $\leftrightarrow$ Poids d'une unité primaire

- Les unités primaires qui répondent livrent un certains pourcentage, f_i, des salaires de leurs employés.
- Le poids du salaire d'un employé de l'unité i devrait ainsi être relié au poids w_i de l'unité primaire :

$$w_i' = w_i / f_i$$

Poids d'un salaire $\leftrightarrow$ Poids d'une unité primaire

- Les unités primaires qui répondent livrent un certains pourcentage, f_i, des salaires de leurs employés.
- Le poids du salaire d'un employé de l'unité i devrait ainsi être relié au poids w_i de l'unité primaire :

$$w_i' = w_i / f_i$$

$\Rightarrow$ Notre travail consiste à calculer les poids w_i des unités primaires.

Étapes clé de la pondération

On distingue 2 étapes importantes dans la pondération :

1. Calcul de poids avant calage :

$$d_i := \frac{1}{\pi_i r_i}$$

où π_i est un taux de sondage (déjà connu) et r_i une probabilité de réponse (**à estimer**)

Étapes clé de la pondération

On distingue 2 étapes importantes dans la pondération :

1. Calcul de poids avant calage :

$$d_i := \frac{1}{\pi_i r_i}$$

où π_i est un taux de sondage (déjà connu) et r_i une probabilité de réponse (**à estimer**)

2. calage des poids d_i sur des variables auxiliaires pour obtenir les poids calés w_i.

De l'échantillon brut à l'échantillon net

Pour qu'une unité interrogée dans l'enquête (échantillon brut) fasse partie de l'échantillon net il faut que :

1. l'unité réponde à l'enquête,

De l'échantillon brut à l'échantillon net

Pour qu'une unité interrogée dans l'enquête (échantillon brut) fasse partie de l'échantillon net il faut que :

1. l'unité réponde à l'enquête,

2. les salaires livrés par l'unité ne soient pas tous éliminés par le Processus de Préparation Statistique des Données (PPSD).

De l'échantillon brut à l'échantillon net

Ainsi :

$$r_i = p_i^{\text{resp}} p_i^{\text{PPSD}}$$

où p_i^{resp} est la probabilité de réponse et p_i^{PPSD} est la probabilité de non-élimination par le PPSD.

De l'échantillon brut à l'échantillon net

Ainsi :

$$r_i = p_i^{\text{resp}} p_i^{\text{PPSD}}$$

où p_i^{resp} est la probabilité de réponse et p_i^{PPSD} est la probabilité de non-élimination par le PPSD.

Innovation pour la LSE 2018 : modèles logistiques pour estimer p_i^{resp} et p_i^{PPSD}.

Modèle logistique pour p_i^{resp}

Modèle de la forme :

$$\text{logit}(p_i^{\text{resp}}) = \beta_0 + \sum_{i=1}^{n_{\text{pred}}} \beta_j x_{i,j}$$

Modèle logistique pour p_i^{resp}

Modèle de la forme :

$$\text{logit}(p_i^{resp}) = \beta_0 + \sum_{i=1}^{n_{pred}} \beta_j x_{i,j}$$

Pour le secteur privé, quelques prédicteurs “classiques” ...

- NOGA (39 modalités)
- Grande Région x Classe de Taille ($7 * 3$ modalités)
- Classe de Taille à 8 modalités
- Nombre d’employés x NOGA
- Nombre d’employés x Grande Région

Modèle logistique pour p_i^{resp}

... et des prédicteurs originaux basés sur les données de l'AVS :

- Pourcentage, selon l'AVS, de salariés de l'unité primaire payés au dessus d'un revenu AVS médian dans la NOGA (variable catégorielle à 11 modalités).
  $\Rightarrow$ Les unités payant bien leur salariés ont tendance à répondre plus.
- Pourcentage, selon l'AVS, de salariés de sexe masculin (variable catégorielle à 11 modalités).
  $\Rightarrow$ Les unités qui emploient plus de femmes ont tendance à répondre plus.

Modèle logistique pour p_i^{PPSD}

L'élimination par le PPSD est modélisée séparément. C'est un processus différent de la non-réponse.

Modèle logistique pour p_i^{PPSD}

L'élimination par le PPSD est modélisée séparément. C'est un processus différent de la non-réponse.

Prédicteurs du modèle logistique :

- Un prédicteur lié à la classe de taille et à l'appartenance de l'unité au programme "profiling"
- Grande Région
- Nombre d'employés x Grande Région

Calage des poids

Objectif du calage : modifier les poids avant calage d_i en des poids w_i qui permettent de retrouver exactement les totaux de variables auxiliaires bien choisies.

Calage des poids

Objectif du calage : modifier les poids avant calage d_i en des poids w_i qui permettent de retrouver exactement les totaux de variables auxiliaires bien choisies.

Choix des variables de calage (pour l'estimateur d'un total)

$$\mathrm{Var}(\widehat{T}) = \sum_{i \in S} d_i(d_i - 1)\widehat{\varepsilon}_i^2 + \ldots$$

où $\widehat{\varepsilon}_i$ est un résidu dans un régression linéaire où l'on tente de prédire le total estimé de la variable d'intérêt (salaires) dans l'unité i :

$$\widehat{T}_i := \frac{\text{nombre d'employés que l'unité déclare avoir}}{\text{nombre de salaires livrés}} \sum_{k \in S_i} y_k$$

à l'aide des variables de calage.

Variables de calage

Quelques variables de calage “classiques” ...

- Nombre d'employés par NOGA ($\Rightarrow$ 39 variables de calage)
- Nombre d'employés par Grande Région x Classe de taille (21 variables)
- Idem avec le nombre d'unités primaires ($39 + 21$ variables)
- ...

Variables de calage

... et d'autres plus originales :

- nb_sup40, nb_sup50, nb_sup60 : nombre de revenus, selon l'AVS, supérieurs aux percentiles à $40\%, 50\%, 60\%$ des revenus AVS sur toute la Suisse.
- nb_sup50NOGA x NOGA : pour chaque NOGA, on cale le nombre de revenus, selon l'AVS, supérieurs au revenu AVS médian de la NOGA.
- nb_sup50GR x Grande Région : idem avec la grande région
- IncomeTotalAVS x Section NOGA et IncomeTotalAVS x Classe de Taille

Algorithme de calage : garantir $w_i \geq 1$

Calage classique : modifier les poids d_i en des poids w_i de sorte à satisfaire :

$$\sum_{i \in S} w_i x_{i.j}^{\text{cal}} = \sum_{\text{cadre de sondage}} x_{i.j}^{\text{cal}} := t_j^{\text{cadre}}$$

Algorithme de calage : garantir $w_i \geq 1$

Calage classique : modifier les poids d_i en des poids w_i de sorte à satisfaire :

$$\sum_{i \in S} w_i x_{i.j}^{\text{cal}} = \sum_{\text{cadre de sondage}} x_{i.j}^{\text{cal}} := t_j^{\text{cadre}}$$

Calage en appliquant un facteur multiplicatif positif sur les poids d_i : programme Calmar de l'INSEE.

Algorithme de calage : garantir $w_i \geq 1$

Calage classique : modifier les poids d_i en des poids w_i de sorte à satisfaire :

$$\sum_{i \in S} w_i x_{i.j}^{\text{cal}} = \sum_{\text{cadre de sondage}} x_{i.j}^{\text{cal}} := t_j^{\text{cadre}}$$

Calage en appliquant un facteur multiplicatif positif sur les poids d_i : programme Calmar de l'INSEE.

Problème : il est fréquent d'avoir $w_i < 1$.

Algorithme de calage : garantir $w_i \geq 1$

L'équation classique de calage est modifiée de la façon suivante :

$$\sum_{i \in S} w_i x_{i,j}^{\text{cal}} - \sum_{i \in S} x_{i,j}^{\text{cal}} = t_j^{\text{cadre}} - \sum_{i \in S} x_{i,j}^{\text{cal}}$$

ce qui donne

$$\sum_{i \in S} (w_i - 1) x_{i,j}^{\text{cal}} = t_j^{\star \text{ cadre}}$$

où $t_j^{\star \text{ cadre}} = t_j^{\text{cadre}} - \sum_{i \in S} x_{i,j}^{\text{cal}}$.

Algorithme de calage : garantir $w_i \geq 1$

L'équation classique de calage est modifiée de la façon suivante :

$$\sum_{i \in S} w_i x_{i,j}^{\text{cal}} - \sum_{i \in S} x_{i,j}^{\text{cal}} = t_j^{\text{cadre}} - \sum_{i \in S} x_{i,j}^{\text{cal}}$$

ce qui donne

$$\sum_{i \in S} (w_i - 1) x_{i,j}^{\text{cal}} = t_j^{\star \text{ cadre}}$$

où $t_j^{\star \text{ cadre}} = t_j^{\text{cadre}} - \sum_{i \in S} x_{i,j}^{\text{cal}}$.

$\Rightarrow$ Calage des "poids moins 1", $d_i - 1$, similaire au calage de départ.

Algorithme de calage : garantir $w_i \geq 1$

Avantage du nouvel algorithme de calage : les poids sont tous supérieurs ou égaux à 1.

Algorithme de calage : garantir $w_i \geq 1$

Avantage du nouvel algorithme de calage : les poids sont tous supérieurs ou égaux à 1.

Inconvénients :

- Pas de marge de manœuvre pour modifier des poids d_i très proches de 1
- Calage parfois plus difficile à atteindre

Une précision améliorée pour estimer des salaires médians

Noga	CV visé	CV attendu	CV noAVS	CV
05-09	3.0	1.2	0.7	0.6
10-11	3.0	1.3	0.6	0.5
12	10.3	8.9	2.7	2.7
13-15	4.0	3.3	1.3	1.0
16-18	3.0	1.1	0.8	0.7
19-20	3.0	2.0	1.3	0.7
21	3.0	1.8	0.5	0.3
22-23	3.0	1.0	0.6	0.4
24-25	3.0	0.9	0.6	0.5
26	3.0	1.5	0.5	0.3
27	3.0	2.8	0.7	0.4
28	3.0	1.1	0.6	0.4
29-30	3.0	1.8	0.6	0.5
31-33	3.0	1.1	1.2	0.9
35	3.0	2.0	0.6	0.5
36-39	3.0	1.2	1.1	0.9
41-43	3.0	0.8	0.5	0.5
45-46	3.0	0.9	0.7	0.5
47	3.0	1.0	0.3	0.3
49-52	3.0	1.9	0.7	0.5

Noga	CV visé	CV attendu	CV noAVS	CV
53	4.0	3.8	0.2	0.2
55-56	3.0	1.1	0.2	0.2
58-60	3.0	3.0	1.3	1.2
61	3.0	2.1	1.6	0.9
62-63	3.0	1.2	0.7	0.4
64, 66	3.0	1.8	1.0	0.6
65	3.0	2.4	0.4	0.3
68	4.0	2.6	0.4	0.4
69-71	3.0	1.0	0.6	0.5
72	3.0	2.2	0.8	0.6
73-75	4.0	3.2	0.7	0.5
77, 79-82	5.0	3.8	0.6	0.5
78	3.0	1.2	0.8	0.7
84	4.0	3.0	0.9	0.8
85	3.0	2.6	0.4	0.4
86-88	3.0	1.2	0.2	0.2
90-93	5.0	3.3	1.3	1.0
94-95	4.0	2.7	0.4	0.4
96	3.0	1.4	0.5	0.5

… et des niveaux de salaires parfois différents

NOGA	2016 old method	2016 new method	2018 new method
05-09	6'190	6'255	6'213
10-11	5'296	5'191	5'272
12	9'784	9'484	8'863
13-15	5'208	5'035	5'095
16-18	5'973	5'948	6'037
19-20	7'608	7'473	7'712
21	9'835	9'467	9'747
22-23	5'984	5'969	6'040
24-25	6'000	6'013	6'094
26	6'875	6'817	6'829
27	6'710	6'857	6'829
28	6'882	6'901	7'010
29-30	6'779	6'805	6'814
31-33	6'117	6'118	6'266
35	8'181	8'149	8'363
36-39	5'778	5'667	5'783
41-43	6'106	6'121	6'200
45-46	6'529	6'519	6'628
47	4'797	4'820	4'875
49-52	6'214	6'148	5'782

NOGA	2016 old method	2016 new method	2018 new method
53	5'896	5'854	5'840
55-56	4'337	4'353	4'413
58-60	7'622	7'682	7'786
61	8'869	8'854	8'798
62-63	8'887	8'922	8'990
64, 66	9'502	9'557	9'546
65	8'806	8'761	8'900
68	6'729	6'747	6'754
69-71	7'690	7'785	7'945
72	9'157	9'292	9'042
73-75	6'634	6'512	6'522
77,79-82	5'160	5'096	5'101
78	5'520	5'493	5'402
85	7'238	7'136	7'450
86-88	6'178	6'126	6'178
90-93	6'000	6'005	6'260
94-95	7'200	7'228	7'250
96	4'043	4'063	4'120

Conclusion et perspectives – LSE 2020

- La LSE est une enquête importante avec beaucoup d'outils et de publications liés (Salarium, indicateurs d'inégalités de salaires H/F, ...)
- La nouvelle pondération améliore la précision des publications, grâce notamment à une nouvelle source de données : les revenus AVS

Conclusion et perspectives – LSE 2020

- La LSE est une enquête importante avec beaucoup d'outils et de publications liés (Salarium, indicateurs d'inégalités de salaires H/F, ...)
- La nouvelle pondération améliore la précision des publications, grâce notamment à une nouvelle source de données : les revenus AVS

Perspectives :

- Rapport de méthodes en préparation
- Pondération unique secteur privé / administrations publiques
- Utilisation *pour les publications de l'OFS* d'estimateurs jugés plus fiables (mais plus complexes) pour calculer la précision des résultats
- Utilisation de davantage d'information auxiliaire
- Calage direct des poids des unités secondaires (salaires)