Enquête SalSa 2009 – Note d'accompagnement

Martin Chevalier et Damien Cartron – 24 janvier 2014

Nom complet	Enquête sur les salaires auprès des salariés 2009
Acronyme	SalSa 2009
Période de collecte	Novembre 2008 - Janvier 2009
Organismes producteurs	ANR (financement), CNRS, ENS, INSEE, Universités Paris 1 et Paris 4 (conception du questionnaire), INSEE (tirage et collecte)
Coordination scientifique	Michel Gollac (INSEE)
Champ couvert	Salariés présents dans les DADS en 2006 et résidant dans les régions Picardie, Centre, Basse-Normandie, Lorraine, Alsace, Pays de la Loire, Midi-Pyrénées, Rhône-Alpes, Auvergne, Languedoc-Roussillon ou dans le département Essonne
Taille de la population	10 520 000
Taille de l'échantillon	3 117 individus (3 110 pondérés)
Taux de sondage moyen	$2,96 \times 10^{-4}$
Variable d'identification	IDENT
Variable de pondération	POND09
Documents diffusés	Note d'accompagnement, questionnaire, lettre-avis aux enquêtés, dictionnaire des variables, formats SAS, note sur le rapprochement des enquêtes SalSa 2009 et SalSa 2011

Présentation de l'enquête

L'enquête sur les salaires auprès des salariés (SalSa) 2009 est une des rares enquêtes quantitatives qui porte sur les salaires et les jugements et représentations qui y sont associés en France. Animée par un groupe de chercheurs [1] et d'étudiants issus de plusieurs disciplines (économie, sociologie, statistique) dans le cadre du séminaire « Salaires et justice » de l'École normale supérieure, cette enquête porte sur les modalités subjectives de perception des salaires ainsi que sur les critères de justice auxquels se réfèrent les individus pour les juger. Son originalité réside ainsi en ce qu'elle aborde, dans le cadre d'une enquête statistique sur un échantillon d'une taille conséquence (environ 5 000 salariés échantillonnés pour 3 000 enquêtés), certains aspects de la perception des salaires le plus souvent analysés par le biais d'enquêtes qualitatives : perception des écarts de salaire à différents niveaux, sentiment de discrimination, motivation associée au salaire, perspectives de salaire futur, interrogation sur le caractère juste de certaines rémunérations (PDG de grande entreprise, ministre, footballeur, etc.).

D'un point de vue matériel, l'enquête a été rendue possible par le financement de l'Agence nationale de la recherche (ANR Corpus 2007). Le questionnaire a été entièrement conçu par l'équipe d'étudiants et de chercheurs (CNRS, ENS, INSEE, Universités Paris 1 et Paris 4) ; le tirage de l'échantillon et la collecte ont été réalisés dans le cadre du réseau d'enquêteurs

1. Christian Baudelot (ENS), Jérôme Gautié (Université Paris 1), Michel Gollac (INSEE), Olivier Godechot (CNRS) et Claudia Senik (Université Paris 4).

de l'INSEE soit en face-à-face, soit par téléphone. Pour des raisons de disponibilité du réseau d'enquêteurs, le champ géographique de l'enquête est restreint à un ensemble de dix régions et d'un département (*cf. supra*). Le champ est ainsi constitué par les salariés du secteur privé, des entreprises publiques ou de la fonction publique territoriale ou hospitalière de cette zone géographique présents dans les Déclarations annuelles de données sociales (DADS) en 2006. Le calcul des pondérations initiales a été opéré par l'Unité de méthodologie statistique (UMS) de l'INSEE, leur redressement a été assuré par le groupe de chercheurs responsables de l'enquête et par Martin Chevalier (ENS Cachan, ENSAE).

Échantillonnage et poids de sondage

Une des spécificités de l'enquête SalSa 2009 tient au fait que la base de sondage utilisée pour réaliser son échantillonnage n'est pas une base de logements mais une base d'individus : elle est issue de l'exploitation 2006 du panel au 25$^{\text{ème}}$ des Déclarations annuelles de données sociales (DADS), restreint aux communes de l'échantillon-maître 1999 de dix régions et d'un département.

Le recours à cette base de données présente des avantages et des inconvénients. La principale qualité de cette base de sondage est de fournir des informations de première importance dans la problématique de l'enquête, notamment le niveau de salaire et le domaine d'emploi (public ou privé) des individus. L'utilisation de cette base de sondage a ainsi permis de distinguer trois strates à échantillonner : les salariés de la fonction publique (strate 1), les salariés du privé dont le salaire est inférieur au neuvième décile (strate 2), les salariés du privé dont le salaire est supérieur au neuvième décile (strate 3). Les strates 1 et 3 sont surpondérées. Une seconde limite tient à la nécessité de procéder à une recherche d'adresse après le tirage des individus dans le panel DADS (l'adresse n'étant pas directement renseignée dans le fichier), ce qui implique un échantillonnage en deux phases distinctes pour corriger le biais introduit par les échecs de recherche. Enfin, la restriction du tirage à un sous-ensemble géographique limite la représentativité de l'enquête aux dix régions de gestion impliquées dans la collecte [2].

L'échantillonnage a donc été opéré en deux phases distinctes, avec dans chacune un traitement spécifique pour chacune des trois strates de salariés (Tableau 1).

- Première phase : tirage de 10 400 individus dans la base de sondage et détermination de leurs poids de sondage selon leur strate de tirage (Annexe A).
- Seconde phase : recherche d'adresse et correction au sein de chaque de strate des échecs de recherche.

L'échantillon final comporte ainsi 5 326 individus représentatifs d'une population de 10 675 800 individus environ.

Tableau 1 – Effectifs et poids par strate à chaque phase de l'échantillonnage

Strate		Phase 1		Phase 2	
		Effectif	Poids	Effectif	Poids
1	Salariés des collectivités territoriales et de la fonction publique hospitalière	2 000	709,78	1 019	1 393,09
2	Autres salariés du privé	7 400	1 134,43	3 789	2 215,57
3	Salariés du privé dont le salaire est supérieur au neuvième décile	1 000	861,47	518	1 663,07
	Ensemble	10 400	1 026,52	5 326	2 004,47

2. Cette limitation résulte de l'indisponibilité de l'intégralité du réseau d'enquêteurs de l'INSEE pour assurer la collecte dans les autres régions.

Correction de la non-réponse totale par repondération

3 117 des 5 326 personnes contactées ont pu être interrogées. Les causes d'échecs de collecte sont multiples : individus impossibles à joindre (918), sortis du champ (décédés ou qui ne sont plus salariés au moment de la collecte, 670), refus ou impossibilité de mener l'enquête (629). Par ailleurs, 9 individus échantillonnés dont 3 interrogés peuvent être considérés comme hors-champ, dans la mesure où leur département de résidence (Bouches-du-Rhône) n'appartient pas au champ de l'enquête. Enfin, 4 des 3 117 individus interrogés présentent des taux de réponse particulièrement bas (moins de 30 %), ce qui conduit à les considérer comme globalement non-répondants. Au total, l'échantillon des répondants comporte 3 110 individus sur les 5 317 individus interrogés appartenant au champ de l'enquête, soit un taux de réponse relativement élevé de 58,5 %.

Afin de garantir la représentativité des résultats, les données ont été repondérées pour tenir compte de cette non-réponse et de sa répartition potentiellement inégale au sein de l'échantillon. Les différentes formes de non-réponse mentionnées précédemment ont été traitées simultanément ; en particulier, le choix a été fait de considérer comme une forme de non-réponse le fait d'être sorti du champ, situation qui concerne principalement des individus salariés dans la base de sondage mais qui ne le sont plus au moment de l'enquête [3]. Plusieurs variables étaient disponibles pour mener à bien cette repondération : le sexe de l'individu, son salaire (codé en décile), l'appartenance à la fonction publique, la région de gestion et enfin (par le biais du code commune) la tranche d'unité urbaine de la commune de résidence. La régression logistique dichotomique utilisée pour évaluer la pertinence de ces variables et obtenir les probabilités de réponse prédites en fonction des caractéristiques de chaque individu est présentée en annexe (Annexe B).

Plusieurs techniques de redressement de la non-réponse ont été mises en œuvre et comparées : redressement par groupes de réponse homogène d'une part, avec trois modes de constitution des groupes (manuel, automatisé avec un algorithme du type *Chi-Squared Automated Interaction Detection*, semi-automatisé à partir des probabilités prédites par un modèle de régression logistique), utilisation directe des probabilités prédites par un modèle de régression logistique d'autre part [4]. Cette comparaison a montré que le redressement utilisant les probabilités individuelles de réponse prédites par régression logistique présente un bon ajustement aux données et conduit à un vecteur de poids redressé avec de bonnes caractéristiques distributionnelles (faible dispersion notamment). C'est donc cette méthode qui a été privilégiée : une fois le comportement de réponse modélisé et le vecteur des probabilités de réponse prédites par le modèle obtenu, le poids de sondage de chaque répondant a été multiplié par l'inverse de sa probabilité de réponse. La somme totale des poids n'a été que peu modifiée par cette opération (Tableau 2).

Tableau 2 – Caractéristiques du vecteur de poids après redressement de la non-réponse

Nombre d'individus	3 110	Maximum	11 134,26
Somme des poids	10 648 719	D9	5 062,78
Moyenne	3 424,03	Q3	3 852,62
Ecart-type	1 276,44	Médiane	3 206,18
Rapport Q3/Q1	1,43	Q1	2 658,39
Rapport D9/D1	2,57	D1	1 971,89
Rapport Max/Min	6,94	Minimum	1 603,61

3. Une autre solution aurait consisté à les considérer comme « hors-champ » et à ne pas redresser l'échantillon en conséquence.

4. Le détail de ce travail de comparaison figure dans le rapport de stage de Martin Chevalier soutenu à l'ENSAE en juin 2013.

Calage sur les marges du panel DADS

L'accès aux données du panel DADS dans le cadre du Centre d'accès sécurisé distant (CASD) a enfin permis de mettre en œuvre un calage sur marges particulièrement fin. En effet, dans cette configuration la population de référence sur laquelle est calé l'échantillon coïncide presque exactement avec la base de sondage, dans la mesure où celle-ci est tirée de l'exploitation 2006 du panel DADS. L'utilisation dans le calage de variables communes à l'enquête et à la population de référence (sexe, année de naissance, domaine d'emploi et salaire) garantit ainsi la qualité du rapprochement entre les marges de ces deux sources de données.

Cependant, avant de pouvoir procéder au rapprochement en tant que tel il convient de redélimiter convenablement dans l'exploitation 2008 des DADS le champ de l'enquête SalSa 2009 [5]. Concrètement, cette opération consiste à identifier les observations à conserver pour constituer la population de référence, mais aussi à effectuer sur la base d'origine l'ensemble des opérations susceptibles de garantir l'intégrité des données utilisées (suppression des éventuels double-compte, intégration ou non des emplois annexes, etc.). Deux informations fournies dans les premiers éléments de documentation de l'enquête (fournis par UMS, responsable du tirage de l'échantillon) permettent d'évaluer la qualité de cette redélimitation : le nombre d'individus (à partir de la somme des poids de sondage) et la valeur du neuvième décile de revenus du privé (utilisé pour construire les strates). La comparaison de ces deux grandeurs au moment du tirage de l'enquête d'une part et dans l'exploitation du panel DADS 2006 d'autre part (Tableau 3) confirme la qualité de la redélimitation du champ.

Tableau 3 – Comparaison de la population de tirage et de sa redélimitation *ex post*

Caractéristique	Tirage	*Ex post*	Variation
Taille de la population	10 676 000	10 323 000	- 3,3 %
Neuvième décile de salaire du privé	28 893 €	28 591 €	- 1,0 %

Dès lors, des statistiques de calage ont pu être calculées à partir de la redélimitation du champ dans l'exploitation 2008 du panel DADS (Annexe C) et utilisées comme référence pour modifier à la marge la pondération de l'enquête après redressement de la non-réponse. La mise en œuvre concrète du calage sur marge s'est appuyée sur le macro-programme SAS CALMAR disponible sur le site de l'INSEE. Les quatre méthodes de calage sur marges disponibles (linéaire, du *Raking ratio*, logistique et linéaire tronquée) ont été mises en œuvre : les caractéristiques distributionnelles de la pondération obtenue ont conduit à privilégier la méthode logistique. La pondération finalement obtenue POND09 (Tableau 4) garantit ainsi la correction de la non-réponse à l'enquête ainsi que le calage des marges de cinq variables de l'échantillon sur leur contrepartie exacte dans l'exploitation 2008 du panel DADS.

Tableau 4 – Caractéristiques du vecteur de poids POND09

Nombre d'individus	3 110	Maximum	9 524,85
Somme des poids	10 520 000	D9	5 152,71
Moyenne	3 382,64	Q3	4 140,43
Ecart-type	1 308,86	Médiane	3 039,55
Rapport Q3/Q1	1,69	Q1	2 448,52
Rapport D9/D1	2,65	D1	1 945,62
Rapport Max/Min	7,37	Minimum	1 293,18

5. L'utilisation de l'exploitation 2008 du panel DADS pour le calage (et non de son exploitation 2006 qui a servi à l'échantillonnage) permet de rapprocher les informations de calage de la date de la collecte.

Apurement et redressement des données

Afin de garantir une exploitation la plus précise et la plus simple possible des données de l'enquête, un grand nombre de variables ont été apurées ou redressées avant diffusion.

La majeure partie des opérations d'apurement a cherché à assurer la cohérence des codages utilisés dans les enquêtes SalSa 2009 et 2011 (*cf.* la note sur le rapprochement des enquêtes SalSa 2009 et 2011). Ces opérations, qui n'induisent pas de perte d'information, prennent le plus souvent la forme d'un changement dans le codage des modalités de réponse des individus (non-réponse et refus notamment). Le codage des modalités des différentes variables est présenté dans le dictionnaire des variables de l'enquête ainsi que dans l'instruction de formatage SAS jointe aux fichiers de l'enquête.

Cependant, certaines variables ont également dû faire l'objet d'un redressement en tant que tel, dont les principaux sont présentés ci-après :

- **Variable synthétique de domaine d'emploi** : la variable FPPRIVE indique avec précision le domaine d'emploi du salarié, entre entreprise privée, entreprise publique, fonction publique d'État, territoriale et hospitalière. La variable PUBPRIVE du questionnaire ne permettant pas de distinguer les salariés de la fonction publique territoriale de ceux de la fonction publique hospitalière, le code APET de l'établissement d'emploi est utilisé pour ce faire : les salariés déclarant appartenir à la fonction publique territoriale ou hospitalière dont l'établissement a le code APET '86' de la NAF rev. 2 sont considérés comme appartenant à la fonction publique hospitalière, les autres sont affectés à la fonction publique territoriale.
- **Réconciliation de variables public / privé** : le questionnaire se divise très tôt en deux sous-questionnaires, selon que la personne interrogée déclare être salariée de la fonction publique ou d'une entreprise (privée ou publique). Si de manière générale les questions sont identiques dans les deux sous-questionnaires, certaines diffèrent légèrement et induisent donc dans le fichier des couples de variables proches sans être identiques, l'une pour la fonction publique et l'autre pour le privé. C'est en particulier le cas des variables QENT / QPUB et STATUENT / STATUPUB : afin de simplifier l'utilisation de ces informations, les variables Q et STATU sont créées dans des nomenclatures homogènes. Par ailleurs, le dictionnaire des variables de l'enquête rend compte des différences de formulation entre ces deux sous-questionnaires en utilisant le signe « / ».
- **Recodage et nettoyage de variables** : plusieurs variables catégorielles ou littérales du fichier ont fait l'objet d'un recodage ou d'un « nettoyage » spécifiques. Ainsi, les nationalités de la personne interrogée et de ses parents ont été uniformisées (une seule modalité de réponse par pays de nationalité), une centaine de diplômes codés en clair dans la variable DIPRECIS ont été réaffectés dans les modalités de la variable DIPLOME, les multiples variables relatives à la PCS de la personne interrogées ou à celle de ses parents ont été synthétisées dans les variables CS, PPERE et PMERE.

Annexes

A Poids de sondage à l'issue de la première phase d'échantillonnage

La première phase de l'échantillonnage correspond à la sélection des individus dans la base de sondage avant la recherche d'adresse et la correction des éventuels échecs de recherche (seconde phase). De façon synthétique, il est possible de représenter la première phase de l'échantillonnage par un sondage à trois degrés :

1. **Tirage dans l'échantillon-maître** : le premier degré correspond à la sélection des unités primaires (UP) de l'échantillon-maître, que l'on assimile [6] à un sondage aléatoire à probabilités inégales. Pour chaque UP i, la probabilité d'inclusion des individus est connue et est notée π_i.
2. **Tirage dans le panel DADS** : le deuxième degré correspond à la sélection, au sein de chaque UP i, des seuls individus appartenant au panel DADS (sous-ensemble des déclarations annuelles de données sociales correspondant aux seuls individus nés en octobre d'une année paire). Le nombre d'individus sélectionnés dans chaque UP i est noté N_i, chacun d'entre eux ayant une probabilité d'inclusion (au niveau de l'UP i) de $\frac{1}{25}$.
3. **Sondage aléatoire simple par strate** : le troisième degré correspond à la sélection, par sondage aléatoire simple au sein de chaque strate h de chaque UP i, des $n_{i,h}$ individus finalement tirés. Avec $N_{i,h}$ le nombre total d'individus de la strate h parmi les N_i individus appartenant au panel DADS dans l'UP i, alors la probabilité d'inclusion des individus pour ce troisième degré de tirage est $\frac{n_{i,h}}{N_{i,h}}$.

Les trois degrés de tirage étant indépendants, la probabilité $\pi_{i,h}$ d'inclusion d'un individu de l'UP i appartenant au panel DADS et à la strate h s'obtient en multipliant les probabilités d'inclusion de chaque degré :

$$\pi_{i,h} = \pi_i \times \frac{1}{25} \times \frac{n_{i,h}}{N_{i,h}}$$

Dans cette expression, π_i et $N_{i,h}$ sont connus, tandis que $n_{i,h}$ et $\pi_{i,h}$ sont déterminés conjointement par les contraintes imposées au tirage. La première de ces contraintes est celle d'équipondération par strate : l'échantillon est construit de sorte à ce qu'au sein de chaque strate h tous les individus aient la même probabilité π_h d'être tirés :

$$\forall i \in EM,\ \pi_h = \pi_{i,h} = \pi_i \times \frac{1}{25} \times \frac{n_{i,h}}{N_{i,h}} \qquad \text{soit} \qquad \forall i \in EM,\ n_{i,h} = \frac{25 \times \pi_h \times N_{i,h}}{\pi_i}$$

En outre, le nombre total d'individus n_h à tirer pour chaque strate h est fixé et connu :

$$n_h = \sum_{i \in EM} n_{i,h} = 25 \times \pi_h \times \sum_{i \in EM} \frac{N_{i,h}}{\pi_i}$$

Dès lors, il est possible de déterminer la valeur de π_h et de donner une expression directement calculable de $n_{i,h}$:

$$\pi_h = \frac{n_h}{25 \times \sum_{i \in EM} \frac{N_{i,h}}{\pi_i}} \qquad \text{et } \forall i \in EM,\ n_{i,h} = n_h \times \frac{N_{i,h}}{\sum_{i \in EM} \frac{N_{i,h}}{\pi_i}}$$

On vérifie ainsi que la probabilité d'inclusion de première vague est bien constante au sein de chaque strate h et que la somme sur i des $n_{i,h}$ donne bien n_h. Le vecteur des poids de première phase s'obtient alors simplement en associant à chaque individu, en fonction de sa strate (1, 2 ou 3), l'inverse de sa probabilité d'inclusion π_h.

6. De façon plus rigoureuse, le tirage des UP de l'échantillon-maître résulte lui aussi d'un sondage à plusieurs degrés (deux ou trois selon la taille de l'unité urbaine de la commune) : les probabilités d'inclusion π_i utilisées ici sont les probabilités synthétiques associées à cet échantillonnage.

B Modélisation logistique de la probabilité de répondre à l'enquête

Nombre d'observations utilisées	5 317
Nombre de répondants	3 110
Pourcentage de concordance	66,4 %

Variable	DL	Statistique de Wald	P-valeur
Sexe	1	31,6842	0,0001
Salaire (déciles)	9	317,5472	0,0001
Fonction publique	1	3,8404	0,0500
Région de gestion	9	47,1060	0,0001
Tranche d'unité urbaine	8	45,6962	0,0001

Variable	Modalité	Paramètre	Erreur standard	P-valeur
Constante		0,2524	0,1849	0,1721
Sexe	Femme	0,3501***	0,0622	0,0001
	Homme	*Ref*		
Salaire (déciles)	D1	-1,1984***	0,1314	0,0001
	D2	-0,9368***	0,1299	0,0001
	D3	-0,6801***	0,1291	0,0001
	D4	-0,4185***	0,12,87	0,0011
	D5	*Ref*		
	D6	0,3083**	0,1349	0,0223
	D7	0,1543	0,1337	0,2482
	D8	0,2446*	0,1346	0,0683
	D9	0,3622***	0,1370	0,0082
	D10	0,1896	0,1342	0,1576
Fonction publique	Oui	0,1542*	0,0787	0,0500
	Non	*Ref*		
Région de gestion	Picardie (22)	0,4208**	0,1648	0,0107
	Centre (24)	-0,0105	0,1545	0,9456
	Basse-Normandie (25)	*Ref*		
	Lorraine (41)	0,5173***	0,1630	0,0015
	Alsace (42)	0,3475**	0,1720	0,0434
	Pays de la Loire (52)	0,6618***	0,1496	0,0001
	Midi-Pyrénées (73)	0,2218	0,1565	0,1565
	Rhône-Alpes (82)	0,3147**	0,1395	0,0240
	Auvergne (83)	0,6873***	0,1966	0,0005
	Languedoc-Roussillon (91)	0,2030	0,1588	0,2012
Tranche d'unité urbaine	Commune rurale	0,0835	0,1327	0,5291
	2 000 - 4 999	-0,0201	0,1497	0,8931
	5 000 - 9 999	*Ref*		
	10 000 - 19 999	-0,4160**	0,1804	0,0211
	20 000 - 49 999	-0,3546**	0,1580	0,0248
	50 000 - 99 999	-0,3021**	0,1508	0,0451
	100 000 - 199 999	-0,2434	0,1667	0,1443
	200 000 - 1 999 999	-0,4362***	0,1302	0,0008
	Unité urbaine de Paris	-0,1105	0,2031	0,5863

*Note : Coefficients significativement différents de 0 à * 10 %, ** 5 %, *** 1%.*

C Distributions comparées des variables de calage

Variables	Modalités	Répondants SalSa 2009		DADS
		Poids de tirage	Correction non-réponse	
Sexe	Homme	51,8 %	53,8 %	54,1 %
	Femme	48,2 %	46,2 %	45,9 %
Année de naissance	1976-1992	27,5 %	30,5 %	36,4 %
	1960-1975	46,8 %	44,7 %	40,8 %
	1944-1959	25,8 %	24,7 %	22,8 %
Domaine d'emploi	Entreprises	84,3 %	86,0 %	85,6 %
	Fonction publique territoriale et hospitalière	9,7 %	9,0 %	9,0 %
	Fonction publique d'Etat	5,9 %	5,1 %	5,4 %
Salaire	Q1 ($\leq$ 8 413 €)	20,9 %	29,5 %	25,0 %
	Q2 (8 414 € – 15 672 €)	20,9 %	20,8 %	25,0 %
	Q3 (15 673 € – 21 655 €)	29,4 %	24,9 %	25,0 %
	Q3D10 (21 656 € – 30 192 €)	18,5 %	15,5 %	15,0 %
	D10 (> 30 192 €)	10,43 %	9,3 %	10,0 %
Lieu de résidence	Picardie (22)	8,0%	7,5%	7,0 %
	Centre (24)	8,5%	9,6%	9,2 %
	Basse-Normandie (25)	5,9%	6,24%	5,0 %
	Lorraine (41)	8,4%	8,0 %	7,4 %
	Alsace (42)	6,3%	6,1 %	6,5 %
	Pays de la Loire (52)	14,9%	12,9 %	13,2 %
	Midi-Pyrénées (73)	8,5%	8,9 %	10,0 %
	Rhône-Alpes (82)	22,8%	23,6 %	23,4 %
	Auvergne (83)	4,8%	4,2 %	4,9 %
	Languedoc-Roussillon (91)	7,8%	8,4 %	8,5 %
	Essonne (91)	4,2%	4,5 %	5,2 %