Document de Recherche

Une première description du système d'information du modèle de micro-simulation de l'ACOSS

Cyrille Hagneré
François Legendre

n°2012-04

mars 2012

Agence centrale des organismes de Sécurité sociale
Direction des statistiques, des études et de la prévision
Mission recherche, études et publications

Une première description du système d'information du modèle de micro-simulation de l'ACOSS

Cyrille Hagneré et François Legendre †*

Première version : mai 2011

Résumé

Les employeurs sont tenus d'effectuer, auprès des organismes de sécurité sociale, un certain nombre de déclarations. D'une part, périodiquement, ils transmettent aux Unions de recouvrement des cotisations de sécurité sociale et d'allocations familiales (Urssaf) un Bordereau récapitulatif des cotisations (BRC) décomptant les cotisations dues. D'autre part, en début d'année, ils doivent faire parvenir à leur Caisse d'assurance retraite et de la santé au travail (Carsat) une Déclaration annuelle de données sociales (DADS) qui renseigne les assiettes des contributions et cotisations sociales de chacun de leurs salariés. L'originalité du modèle de micro-simulation, que l'Agence centrale des organismes de Sécurité sociale (l'ACOSS, la caisse nationale des URSSAF) est en train de développer, est de s'appuyer sur ces deux sources : les BRC qui livrent une information longitudinale infra-annuelle et les DADS qui apportent une connaissance transversale, propre à chaque salarié. Ce modèle est destiné à chiffrer des dispositifs ou des projets de mesure portant sur les cotisations et les contributions sociales. Le but de cette contribution est de détailler la démarche qui a présidé à la construction du système d'information sur lequel ce modèle est construit.

Abstract

Employers are required to make a number of declarations to Social security. On the one hand, periodically they transmit to the Urssaf network a summary Schedule of contributions (BRC) which summarizes the amounts of contributions to pay. On the other hand, once a year, they must submit an annual Declaration social data (DADS) that gives informations for each of their employees. The originality of the micro-simulation model, the Central Agency of Social Security (ACOSS, the national fund ot the URSSAF network) is currently developing, is to rely on these two sources: BRC which deliver an infra-annual longitudinal information and DADS that provide cross knowledge specific to each employee. This model is intended to evaluate measures on social contributions. The purpose of this paper is to detail the process that led to the construction of an information system on which the model is built.

Mots clefs : micro-simulation, données déclaratives, déclaration annuelle de données sociales (DADS), Bordereaux récapitulatifs de cotisations (BRC).

Code JEL : C63, C81

---

* Département des Statistiques, des Études et de la Prévision, Acoss – cyrille.hagnere@acoss.fr

† ÉRUDITE (Université PARIS-EST) et Acoss – f.legendre@u-pec.fr

Introduction

Chaque mois, ou chaque trimestre (pour les entreprises de moins de 10 salariés), les établissements employeurs sont tenus de transmettre aux Unions de recouvrement des cotisations de sécurité sociale et d'allocations familiales (URSSAF) un Bordereau récapitulatif des cotisations (BRC) indiquant le montant des assiettes salariales, les cotisations dues, les éventuels abattements ou exonérations, ainsi que les effectifs salariés en fin de mois. Par ailleurs, en début d'année, les établissements envoient aux URSSAF un Tableau récapitulatif (TR) permettant de précéder, le cas échéant, à des régularisations sur les montants déclarés sur les BRC de l'année écoulée.

L'exhaustivité de l'information sur la masse salariale et l'emploi du secteur privé contenue dans les BRC confère à l'Agence centrale des organismes de sécurité sociale (ACOSS), qui centralise les données, un rôle essentiel dans la production et l'expertise des statistiques relatives à l'évolution de l'emploi et des salaires, indispensables à l'analyse de la conjoncture.[1]

L'ACOSS fait cependant l'objet d'une demande d'expertise de plus en plus large. Celle-ci a trait, d'une part, à des catégories particulières d'employeurs, comme par exemple les indépendants ou les particuliers employeurs. Cette demande d'expertise sectorielle est satisfaite en mobilisant des sources *ad hoc*. D'autre part, il est demandé à l'ACOSS de développer une compétence spécifique des dispositifs qui impactent, de près ou de loin, le recouvrement des cotisations sociales comme, bien évidemment, les mesures de réduction ou d'exonération. Pour ce type d'expertise, les données des BRC s'avèrent en général insuffisantes. En effet, les BRC fournissant des informations agrégées au niveau de l'établissement, ils ne permettent pas de pratiquer des analyses fines de dispositifs applicables au niveau des salariés. Les données issues de la Déclaration annuelle de données sociales (DADS) constituent alors une source complémentaire aux BRC.

La DADS est une obligation déclarative que doivent faire parvenir en début d'année les établissements à leur Caisse d'assurance retraite et de la santé au travail (CARSAT, ex Caisse régionale d'assurance maladie, CRAM). Elle renseigne les assiettes des contributions et cotisations sociales de chacun des salariés qui ont été employés dans l'année qui vient de s'écouler. Les DADS permettent de garantir les droits des salariés, notamment les droits à la retraite. Elles sont aussi utilisées, depuis 2006, par l'administration fiscale pour pré-remplir les déclarations à l'impôt sur le revenu.

L'idée d'un système d'information pour mettre en cohérence des informations apportées par les BRC et les DADS s'impose ainsi progressivement. Ce système d'information permettrait

- d'harmoniser le retraitement des données chaque année pour disposer de séries chronologiques cohérentes ;

[1] L'ACOSS diffuse ses statistiques conjoncturelles dans les publications *Le baromètre économique* et *ACOSS Stat*, *cf.* www.acoss.fr.

- de microsimuler des dispositifs ou des projets de mesure portant sur les prélèvements sociaux, comme par exemple des réductions de cotisations, afin
    - de chiffrer leur impact financier et d'établir un premier bilan de leurs conséquences redistributives ;
    - d'évaluer *ex ante*, à l'aide d'une modélisation simple, leur impact notamment en termes d'emploi ;
- de constituer un réservoir de données permettant de réaliser des évaluations micro-économétrique *ex post* des politiques publiques.[2]

Le développement d'un modèle de microsimulation assis sur un tel système d'information a ainsi été inscrit parmi les objectifs de la Convention d'objectifs et de gestion (COG) conclue en 2010 entre l'État et l'ACOSS pour la période 2010-2013.

L'intérêt de la construction de ce système d'information est en outre renforcé par le fait que les établissements remplissent maintenant de mieux en mieux leurs obligations déclaratives, améliorant ainsi la fiabilité des données. Plusieurs éléments contribuent à cette amélioration. Tout d'abord, les logiciels de paye sont aujourd'hui en mesure d'engendrer automatiquement ces déclarations. Ensuite, depuis 2010, le non respect de ces obligations déclaratives peut être assimilé à un cas de travail dissimulé. Par ailleurs, le régime d'exonération spécifique des heures supplémentaires ou complémentaires introduit par la loi de 2007 en faveur du travail, de l'emploi et du pouvoir d'achat (dite loi TEPA) a conduit à isoler, dans les DADS, la rémunération relative à ces heures. Nous disposons ainsi maintenant d'une information particulièrement fiable sur le nombre d'heures totales payées par période d'emploi ainsi que le nombre d'heures supplémentaires ou complémentaires.

Enfin, la dématérialisation des obligations déclaratives des établissements et les nouvelles capacités de calcul des ordinateurs modifient la donne. Il est maintenant tout-à-fait possible d'utiliser l'ensemble des données et non un échantillon comme cela était auparavant le cas. Les données restent toutefois extrêmement volumineuses et il reste utile de développer des outils adaptés pour réaliser le rapprochement des BRC et des DADS.

Cette contribution a pour objectif de présenter les principaux éléments de mise en cohérence des données issues des BRC d'une part et des DADS d'autre part. Dans une première section, nous discutons de la question de l'exploitation de l'exhaustivité des données et des solutions retenues. Dans une seconde section sont présentées les principales caractéristiques des données utilisées. Enfin, dans une troisième section, nous présentons la procédure développée dans le but de reconstituer des revenus individuels mensuels en tirant parti de la dimension temporelle des BRC et de la dimension individuelle des DADS. Ces deux dernières sections sont illustrées par les données d'un établissement « témoin ».

[2] Il est toutefois un peu illusoire de vouloir utiliser les DADS à des fins de comparaison fine dans le temps sur longue période. La note de ROUX (2001) détaille les difficultés que l'INSEE a rencontrées pour constituer un panel « long » (depuis 1976) à partir de cette source.

1. Utiliser les données exhaustives des BRC et des DADS

Les micro-ordinateurs les plus performants disposent d'une mémoire centrale de grande capacité et de plusieurs processeurs. Afin de tirer parti d'un tel matériel, il convient de réorganiser les traitements en tentant, d'une part, de placer l'ensemble des données dans la mémoire centrale de l'ordinateur pour limiter les entrées/sorties depuis ou vers le support externe et, d'autre part, de paralléliser les traitements pour utiliser concurremment tous les processeurs.

Pour rapprocher les BRC et les DADS, il est de toute façon utile de disposer en même temps de l'ensemble des données. Il se peut, par exemple, qu'il y ait une confusion entre les différents établissements d'une même entreprise et il est alors commode, pour retraiter les données, de pouvoir facilement itérer sur tous les établissements d'une entreprise. Nous avons conduit une étude préalable pour évaluer la faisabilité de cette première ambition : loger l'ensemble des données en mémoire centrale.

Pour l'année 2009, sur le champ « secteur privé », on dénombre près de 94 millions de lignes de décompte dans les BRC et près de 41 millions de périodes d'emploi dans les DADS. Supposons maintenant que la taille de la mémoire centrale soit de l'ordre de quatre giga-octets et que les observations des BRC soient de taille comparable aux observations des DADS. On dispose alors d'une trentaine d'octets pour coder les informations d'une observation.

Pour tenir notre ambition, il est nécessaire de recourir à des techniques sophistiquées pour représenter l'information de la manière la plus compacte possible. Nous avons développé une méthode systématique pour établir la représentation minimale (au sens de l'encombrement en mémoire centrale) des variables d'une table. Cette méthode comporte deux étapes. En premier lieu, en parcourant l'ensemble des observations, toutes les occurrences différentes de chaque variable sont consignées ; en second lieu, en fonction de ces occurrences, la représentation minimale est déterminée. Au cas d'une variable présentant un petit nombre de modalités, une table de ces dernières est engendrée et la variable est recodée comme l'indice dans cette table. Au cas d'une variable numérique entière, la variable est recodée comme son étendue. Dans ces deux cas, nous utilisons exactement le nombre de bits nécessaire.

Le gain est parfois très élevé. Par exemple, la variable de la DADS *« Code motif début de période »*, qui détaille la situation du salarié au début de la période d'emploi et qui est codée sur trois caractères, comporte les 27 modalités suivantes : « 001 », « 003 », « 005 », « 019 », « 021 », « 023 », « 025 », « 029 », « 031 », « 033 », « 035 », « 041 », « 057 », « 061 », « 069 », « 089 », « 095 », « 097 », « 111 », « 113 », « 119 », « 121 », « 123 », « 127 », « 143 », « 411 » et « 901 ». La modalité la plus fréquente (44 % des périodes d'emploi)[3] est celle qui est codée

[3] Une période d'emploi, au sens des DADS, est une période pendant laquelle le salarié a été payé et pendant laquelle le salarié au sein de l'établissement a bénéficié du même statut. Les périodes d'emploi au sens des DADS ne sont donc pas directement mobilisables pour évaluer la mobilité des salariés.

« 097 » et qui signifie « continuité d'activité en début de période ». La table utilisée pour recoder cette variable comporte 27 éléments (cette variable est toujours renseignée et il n'est donc pas nécessaire de prévoir une entrée dans cette table pour coder la valeur manquante). Il suffit ainsi d'utiliser 4 bits pour recoder cette variable contre 24 bits dans le codage initial (trois caractères qui occupent chacun un octet).

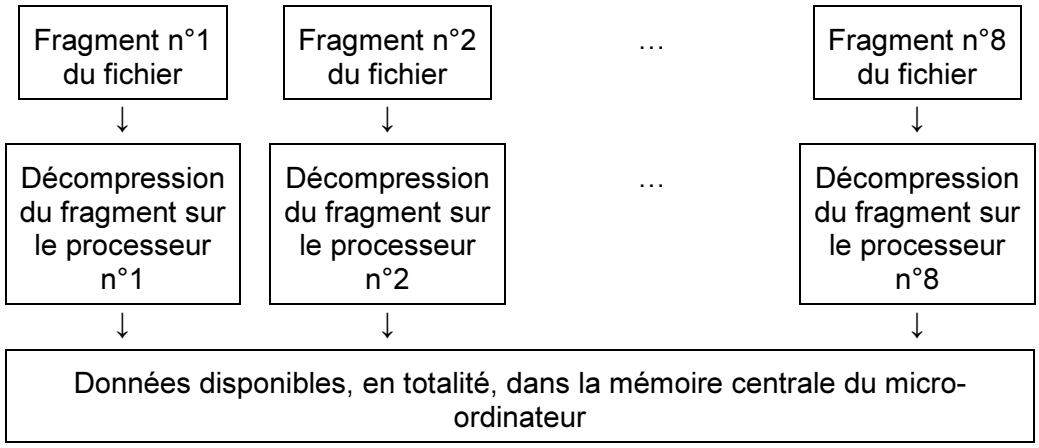
Il n'est bien sûr pas recommandé, en général, de coder l'information avec une granularité inférieure à l'octet. L'octet, un groupe de 8 bits, est la plus petite unité d'information disposant d'une adresse en mémoire centrale. Quand les données sont codées sur un champ de bits à l'intérieur d'un octet, deux instructions sont nécessaires pour accéder à l'information : il faut, d'une part, charger l'octet dans le registre du processeur et, d'autre part, extraire le champ de bits idoine. La situation est encore pire quand il faut mettre à jour l'information puisqu'il est alors nécessaire de préalablement lire l'octet dans sa totalité avant de mettre à jour le champ de bits correspondant et de réécrire l'octet en mémoire. Nous sommes conduits à cette solution uniquement parce qu'il nous est prioritaire d'obtenir la représentation la plus compacte des données.

Un grand nombre de détails ont été pris en compte dans cette entreprise de recodage des variables, détails accessoires qui conduisent cependant à grandement économiser la ressource rare que constitue la mémoire centrale de l'ordinateur. La chaîne de traitement comporte les étapes suivantes. En premier lieu, les identifiants des établissements et des entreprises (les SIREN et les SIRET qui nous ont été livrés cryptés pour assurer l'anonymat des données[4]) ont été remplacés par un numéro d'ordre pour réduire l'encombrement des données. En deuxième lieu, l'analyse systématique des variables, détaillée ci-avant, est mise en œuvre. Ensuite, les tables de travail sont construites, séparément pour les BRC et pour les DADS : elles ne comportent que les établissements dans le champ (le secteur privé) et que les variables utiles pour rapprocher les BRC et les DADS. En quatrième lieu, ces tables sont triées sur une clef dont le premier critère est l'identifiant de l'établissement. En dernier lieu, le rapprochement des BRC et des DADS est effectivement réalisé pour produire, pour chaque salarié au sein d'un établissement, une information cohérente comportant, à côté de grandeurs annuelles synthétiques, des profils mensuels obtenus par le truchement des BRC.

Par ailleurs, pour obtenir des temps de réponse qui ne soient pas rédhibitoires, il faut s'engager dans l'utilisation concurrente de tous les processeurs disponibles. Il n'est cependant pas trivial de coordonner les traitements dans un tel contexte. La première difficulté tient au fait que l'efficacité de la « parallélisation » des tâches relève des algorithmes mis en œuvre mais aussi des données sur lesquelles ces algorithmes s'appliquent ; la seconde au fait qu'il est difficile d'équilibrer la charge entre les différentes tâches quand ces dernières diffèrent : le temps total d'exécution est alors égal au temps de la tâche la plus lente. En outre, les gains obtenus peuvent être effacés par la nécessité de synchroniser l'accès en écriture aux variables.

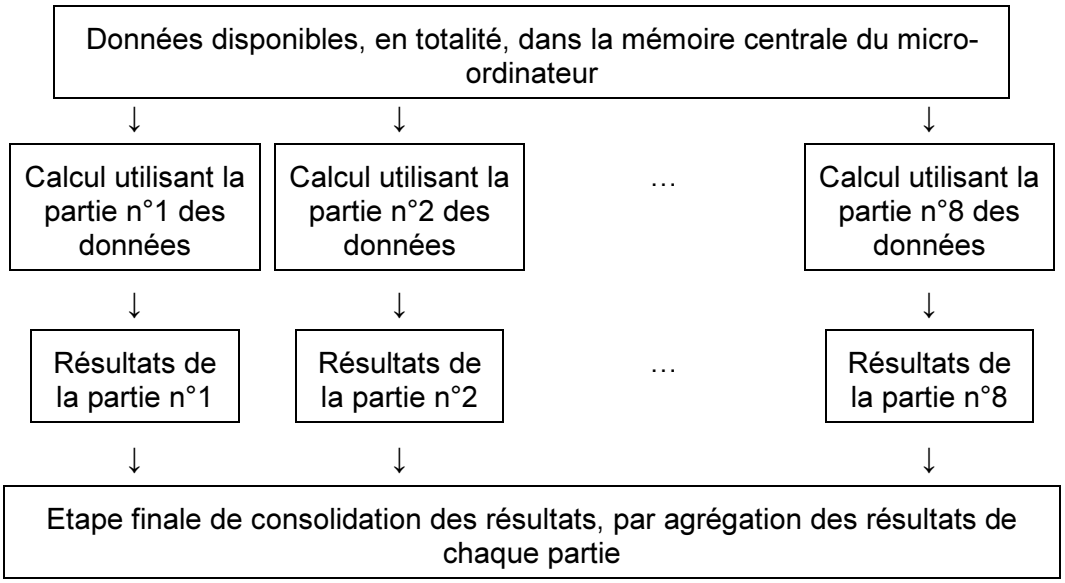
[4] Il en est de même, bien évidemment, pour les salariés du fichier DADS dont l'identifiant individuel est également crypté.

Figure 1 : Lecture concurrente d'un fichier compressé



Notre expérience reste en ce domaine limitée. Pour le moment, il nous est bien apparu que les transferts entre le support externe et la mémoire centrale constituent un goulet d'étranglement même quand les données font l'objet de la représentation compacte détaillée ci-avant. Un gain est obtenu quand les fichiers sont, d'une part, compressés et, d'autre part, enregistrés en plusieurs fragments de taille comparable – en autant de fragments que de processeurs disponibles sur le micro-ordinateur. La lecture du fichier est alors particulièrement efficace (*cf.* la figure 1 pour une représentation du processus). La compression des données réduit le volume à transférer ; les processeurs s'occupent à décompresser les données et à les placer en mémoire centrale ; en outre, la charge des différents processeurs est identique si bien que le micro-ordinateur est occupé à 100 % jusqu'à la fin de la lecture du fichier.

Figure 2 : L'algorithme *map and reduce*
(répartition des calculs et consolidation des résultats)



Pour chiffrer certaines mesures nouvelles, l'algorithme *map and reduce* est particulièrement efficace. En effet, il est compliqué de synchroniser l'accès en écriture aux variables qui représentent un résultat. Par contre, l'évaluation *ex ante* de la mesure nouvelle, si l'on se contente d'un chiffrage de « premier tour » ne

faisant pas intervenir d'interactions, peut être conduite établissement par établissement. Il est donc facile de diviser les observations en autant de sous ensembles qu'il y a de processeurs ; de réserver des variables locales pour stocker les résultats propres à la tâche et, dans une dernière étape, de consolider les résultats entre les différentes tâches. D'une façon générale, il est préférable de paralléliser le travail en dupliquant le même traitement sur des jeux de données différents. Il n'est alors nécessaire ni de synchroniser les accès en écriture aux variables ni d'équilibrer la charge entre les tâches.

Nous avons cherché systématiquement à paralléliser les traitements. Les gains sont au total assez variables. Pour les opérations de lecture des données, nous appliquons la tactique détaillée ci-avant ; le gain est sensible. Pour les opérations d'écriture, nous utilisons un algorithme similaire ; le gain est là encore plus sensible puisque le coût fixe élevé de la compression des données est réparti sur tous les processeurs. Pour valider les observations, nous exécutons en parallèle des tâches élémentaires, chacune spécialisée dans un aspect particulier de la validation. Il n'est pas nécessaire de verrouiller en écriture la variable booléenne qui spécifie que l'observation est invalide : une observation peut ne pas être conservée pour de multiples raisons. Il est apparu difficile d'équilibrer la charge entre ces différentes tâches et le gain de la « parallélisation » est dans ce cas faible. Pour trier les données, nous avons utilisé une bibliothèque spécialisée qui tire parti de l'architecture multi-processeur. Le bénéfice nous a semblé limité. Pour rapprocher BRC et DADS, nous appliquons une variante de l'algorithme *map and reduce* : le traitement est dupliqué en un grand nombre de tâches parce que le temps de traitement est finalement assez variable d'un jeu de données à l'autre même si les établissements sont initialement équi-répartis dans chaque jeu. Enfin, pour simuler une mesure nouvelle, la lecture des données est parallélisée et les calculs sont distribués en utilisant un *map and reduce.*

Notre ambition – exploiter les données de manière exhaustive – est ainsi satisfaite. Nous avons toutefois dû effectuer, pour cela, un détour de production significatif. Les temps de réponse, pour simuler par exemple une mesure simple de réduction de cotisations sociales, sont de l'ordre de quelques minutes.

Tout ce qui précède reste très abstrait. Nous retenons maintenant l'approche opposée en présentant les premiers développements du modèle de micro-simulation à partir d'un établissement « témoin ».

2. Les données issues des BRC et des DADS

Nous avons sélectionné un établissement particulier, que nous qualifions d'établissement « témoin », qui constitue un parangon et qui nous permet de rendre notre présentation moins abstraite. Cet établissement n'a pas été choisi au hasard. Il cumule les principales caractéristiques suivantes : ses salariés effectuent des heures supplémentaires ; des mouvements de main-d'œuvre sont observés dans l'année ; des primes sont versées certains mois ; le nombre de salariés reste limité.

Le bordereau récapitulatif des cotisations est une pièce comptable qui accompagne le versement des prélèvements sociaux collectés par les URSSAF. Initialement, sa

structure était simple : chaque ligne de décompte permettait d'isoler toutes les cotisations de tous les salariés pour lesquelles le barème est identique. Par exemple, il est fait masse de toutes les assiettes brutes déplafonnées de la Sécurité sociale de tous les salariés du régime général pour calculer le montant total de toutes les cotisations de Sécurité sociale dont le barème résulte de l'application d'un taux à l'assiette déplafonnée. Les lignes de décompte sont identifiées par un « code type personnel » ; le code le plus utilisé est le code « 100 », qui est associé aux cotisations de Sécurité sociale des salariés du régime général.

Tableau 1 : Le bordereau de cotisations du mois de janvier

Catégorie de salariés	Code type personnel	Nombre de salariés	Base	Salaires arrondis	Taux en %	Cotisations arrondies
REDUCTION SALARIALE HEURES SUP	3	3	—	275	—	–59
DEDUCTION PP HEURES SUP + 20 SAL	5	1	—	20	—	–10
REDUCTION HCR AVANTAGES EN NATURE	95	9	—	—	—	–246
RG CAS GENERAL	100	—	P	13 281	15,05	1 999
RG CAS GENERAL	100	—	T	13 281	21,94	2 914
TAXE SUR CONTRIBUTION PREVOYANCE	108	—	—	313	8,00	25
FNAL SUR TOTALITE DES SALAIRES	236	—	—	13 281	0,40	53
CSG CRDS REGIME GENERAL	260	—	—	13 171	8,00	1 054
REDUCTION FILLON	671	8	—	—	—	–1 694
APPRENTIS LOI 87 AVEC AT	705	—	P	618	0,10	1
APPRENTIS LOI 87 AVEC AT	705	—	T	618	1,69	10
TRANSPORT	900	—	—	13 899	1,35	188

Source : Etablissement « témoin », BRC de 2009.

Cette structure a été, au fil des années, réutilisée pour enregistrer des montants qui ne s'inscrivent pas exactement dans cette logique. Il en est ainsi des réductions de cotisation sur les bas salaires, intitulées dans le bordereau « Réduction Fillon ». Un « code type personnel » a été affecté à ce dispositif, la colonne « Salaires arrondis » normalement utilisée pour spécifier l'assiette du prélèvement n'est pas renseignée et la colonne « Cotisations arrondies » sert à indiquer le montant de la réduction. Pour la déduction forfaitaire patronale sur les heures supplémentaires, la colonne « Salaires arrondis » renseigne le *nombre d'heures* !

Tableau 2 : Les montants mensuels issus des BRC (en €)

Mois	Base brute de la Sécurité sociale	Rémunération des heures supplémentaires	Réduction de cotisations sur les bas salaires
Janvier	13 899	275	1 694
Février	12 481	211	1 521
Mars	16 988	158	1 056
Avril	12 699	353	1 445
Mai	14 306	272	1 320
Juin	13 712	247	1 328
Juillet	14 765	301	1 533
Août	13 659	199	1 652
Septembre	13 583	158	1 708
Octobre	14 902	198	1 484
Novembre	12 747	158	1 526
Décembre	23 739	861	82

Ensemble des bordereaux de l'établissement « témoin » pour 2009.

Au total, on peut disposer, par le biais des BRC, des montants agrégés portés dans le tableau 2. Notre attention se porte plus particulièrement sur l'assiette déplafonnée des cotisations, intitulée « Base brute de la Sécurité sociale ». Par ailleurs, du fait du régime social spécifique, la rémunération des heures supplémentaires est isolée ; on dispose en outre du montant des réductions de cotisations sur les bas salaires. On observe, pour cet établissement « témoin », que les rémunérations sont plus élevées les mois de mars et de décembre en raison, sans doute, du versement de primes ces deux mois. On voit aussi que les réductions sur les bas salaires font l'objet d'une forte variabilité au fil des mois : en septembre, la réduction est égale à 1 708 €, en décembre à 82 € seulement. Cette variabilité résulte, pour une faible part, de la variabilité de l'assiette et, pour une beaucoup plus grande part, du mode de calcul de la réduction qui fait intervenir le taux de salaire horaire tel qu'il peut être apprécié mois après mois. Aussi les primes mensuelles conduisent-elles à une augmentation du taux de salaire horaire mensuel et à une contraction du domaine d'éligibilité de la réduction sur les bas salaires.

Il faut ainsi retenir des BRC que ces derniers livrent une information infra-annuelle qui est *a priori* précieuse ; cette information est toutefois agrégée sur l'ensemble des salariés. C'est en cela que l'information apportée par les DADS est tout-à-fait complémentaire. Dans le tableau 3, nous portons les montants que nous pouvons obtenir grâce aux DADS. Ces montants sont cette fois-ci ventilés par salarié ; en revanche, ils ont trait à l'ensemble de l'activité annuelle du salarié. Plus précisément, dans le tableau 3, nous n'avons pas porté l'information brute : les chiffres de ce tableau sont déjà consolidés par salarié. Il se peut, en effet, que plusieurs périodes d'emploi soient détaillées pour certains salariés. Nous exploitons, le cas échéant, l'information infra-annuelle apportée par cette situation ; au cas général, toutefois, il y a une seule période d'emploi par salarié.

Tableau 3 : Les montants annuels issus des DADS (en €)

Salarié…	Base brute de la Sécurité sociale (en €)	Nombre total d'heures payées	Rémunération des heures supplémentaires (en €)	Nombre d'heures supplémentaires
…1	43 744	2 028	0	0
…2	27 945	2 099	2 757	279
…3	19 358	1 806	0	0
…4	1 708	108	0	0
…5	18 725	1 670	59	6
…6	19 779	1 820	0	0
…7	13 601	1 219	175	18
…8	13 481	1 310	186	20
…9	783	191	0	0
…10	14 575	1 455	192	21
…11	192	22	22	2
…12	3 589	325	0	0

DADS de l'établissement « témoin », consolidées par salarié, de l'année 2009.

Il convient de souligner que la forme de la DADS a récemment évolué avec la dématérialisation de la déclaration. La différence majeure a trait à l'unité statistique élémentaire. Actuellement, cette unité est la « période d'emploi » ; du temps des « DADS papier », seules les deux périodes d'emploi les plus longues dans l'année étaient renseignées. Il n'était donc pas possible de toujours en déduire, par exemple, la durée totale d'emploi dans l'année. Les montants étaient quant à eux consolidés sur l'ensemble de l'année. En outre, les heures payées étaient bien souvent mal renseignées. Les droits que les salariés peuvent acquérir du fait de leurs cotisations dépendent, dans la plupart des cas, du montant total des cotisations acquittées par l'employeur dans l'année. Aussi ces omissions ou ces inexactitudes étaient-elles sans conséquence. C'est notamment pour cette raison que l'INSEE se restreint souvent, dans les études réalisées à partir des DADS, aux seuls salariés employés à temps complet.

Nous portons, en premier lieu, une grande importance à la base brute de la Sécurité sociale. Il nous faut, quand les BRC sont rapprochés des DADS, comprendre les raisons qui conduisent à un écart entre l'agrégation temporelle de la base brute de la Sécurité sociale des BRC et l'agrégation individuelle des DADS. D'autres assiettes pourraient être utilisées comme l'assiette de la Contribution sociale généralisée (CSG). Dans de rares cas, l'établissement n'a pas à émettre une DADS, comme par exemple pour des salariés non résidents : les deux cumuls ont de bonnes raisons pour ne pas toujours coïncider. Les travaux préliminaires conduits à l'ACOSS montrent que plus de 99 % des établissements qui sont présents dans les DADS sont aussi présents dans les BRC. Ils montrent aussi que l'écart entre les deux assiettes agrégées est inférieur, en valeur absolue, à 1 % dans près de 92 % des cas. Nous sommes en train de développer l'outil qui va permettre de réduire ces anomalies. Celles-ci proviennent, dans la plupart des cas, d'une erreur sur l'identification de l'établissement. Nous cherchons alors soit à repérer correctement l'établissement soit à rapprocher les données en consolidant sur le périmètre de l'entreprise (et non de chacun de ses établissements).

Il faut observer que les DADS livrent une information relativement fiable, depuis peu, sur les heures payées : le nombre total d'heures payées et le nombre d'heures

supplémentaires (salarié à temps complet) ou complémentaires (salarié à temps partiel). Les DADS permettront maintenant d'obtenir de la connaissance sur les taux de salaire horaires.

Tableau 4 : Les prorata du nombre de jours payés issus des périodes d'emploi des DADS (en %)

	Mois…											
Salarié…	**…1**	**…2**	**…3**	**…4**	**…5**	**…6**	**…7**	**…8**	**…9**	**…10**	**…11**	**…12**
…1	100	100	100	100	100	100	100	100	100	100	100	100
…2	100	100	100	100	100	100	100	100	100	100	100	100
…3	100	100	100	100	100	100	100	100	100	100	100	100
…4	100	100	100	100	100	100	100	100	100	100	100	100
…5	100	100	100	100	100	100	100	100	100	100	100	100
…6	100	100	100	100	100	100	100	100	100	100	100	100
…7	100	100	100	100	100	100	100	100	100	100	100	100
…8	100	100	100	100	100	100	100	100	100	100	100	100
…9	100	100	36	0	0	0	0	0	0	0	0	0
…10	100	100	100	100	100	100	100	100	100	100	100	100
…11	0	0	0	0	0	0	16	3	0	0	0	0
…12	0	0	0	0	0	0	0	100	100	97	0	0

Enfin, les périodes d'emploi des DADS permettent d'établir un calendrier, au cours de l'année, pour chaque salarié. Nous synthétisons l'information apportée par les dates de début et de fin des périodes d'emploi en calculant des prorata de nombre de jours payés par mois (*cf.* le tableau 4). Pour notre établissement « témoin », nous observons une assez faible mobilité : 9 salariés sur 12 sont présents toute l'année ; un salarié quitte l'établissement au cours du mois de mars ; enfin, deux salariés ont été employés dans l'année sans rester dans l'établissement.

C'est à partir de ces intrants, détaillés dans les tableaux 2, 3 et 4, que nous cherchons à mensualiser les montants individuels obtenus à partir des DADS en nous calant sur l'information infra-annuelle apportée par les BRC.

3. L'imputation de montants mensuels pour chaque salarié

Nous reprenons ici l'établissement « témoin » décrit précédemment afin de présenter la procédure que nous sommes en train de perfectionner pour mensualiser les montants annuels des DADS qui sont propres à chaque salarié.

Il est bien sûr tentant, à partir des données disponibles, de chercher à imputer une information mensuelle pour chaque salarié qui figure dans les DADS. Il s'agit, en quelque sorte, de reconstituer les bulletins mensuels de paye des salariés. Cette tentative est motivée par deux raisons assez différentes. D'un côté, nous voudrions apporter de la connaissance sur les pratiques salariales et sur notamment l'ampleur des primes qui sont versées périodiquement. De l'autre côté, nous voudrions pouvoir chiffrer des mesures pour lesquelles la dimension temporelle est primordiale, telle que celle qui va conduire à l'annualisation de la réduction des cotisations sur les bas salaires. En effet, la loi de financement de la Sécurité sociale

pour 2011 prévoit que l'éligibilité au dispositif sera déterminée à partir d'un taux de salaire horaire calculé sur l'année et non plus mois après mois. Aussi les entreprises qui versaient par exemple un treizième mois verront-elles leurs réductions diminuer.

Notre tactique pour imputer des montants mensuels repose sur les six étapes suivantes.

1. Détermination du nombre mensuel d'heures payées de base ;
2. Imputation préliminaire de la rémunération mensuelle de base ;
3. Détermination du nombre mensuel d'heures supplémentaires ;
4. Imputation de la rémunération mensuelle des heures supplémentaires ;
5. Imputation sur barème des réductions de cotisations sur les bas salaires ;
6. Imputation définitive de la rémunération mensuelle de base en limitant les écarts qui apparaissent sur les réductions de cotisations.

Nous avons systématiquement distingué, pour la rémunération et le nombre d'heures, une grandeur de base et une grandeur liée aux heures supplémentaires pour les trois raisons suivantes. En premier lieu, nous disposons de cette ventilation dans les intrants, dans les BRC et dans les DADS. En deuxième lieu, cette distinction est importante pour le calcul du taux de salaire horaire qui détermine la réduction des cotisations sur les bas salaires. Enfin, la rémunération des heures supplémentaire n'est *a priori* pas affectée par les primes salariales.

Tableau 5 : L'imputation des heures de base par mois

	Mois…											
Salarié…	**…1**	**…2**	**…3**	**…4**	**…5**	**…6**	**…7**	**…8**	**…9**	**…10**	**…11**	**…12**
…1	169	169	169	169	169	169	169	169	169	169	169	169
…2	152	152	152	152	152	152	152	152	152	152	152	152
…3	150	150	150	150	150	150	150	150	150	150	150	150
…4	9	9	9	9	9	9	9	9	9	9	9	9
…5	139	139	139	139	139	139	139	139	139	139	139	139
…6	152	152	152	152	152	152	152	152	152	152	152	152
…7	100	100	100	100	100	100	100	100	100	100	100	100
…8	108	108	108	108	108	108	108	108	108	108	108	108
…9	81	81	29	0	0	0	0	0	0	0	0	0
…10	107	107	107	107	107	128	128	128	128	128	128	128
…11	0	0	0	0	0	0	17	3	0	0	0	0
…12	0	0	0	0	0	0	0	110	110	106	0	0

L'imputation des heures mensuelles de base s'appuie sur le nombre total d'heures payées, le nombre total d'heures supplémentaires et les prorata du nombre de jours payés par mois, en supposant une répartition en fonction de ces prorata. Cette imputation est portée dans le tableau 5 pour notre établissement « témoin ». Prenons le cas du deuxième salarié : son nombre total d'heures de base est égale à 1 820, obtenu à partir du calcul 2 099 – 279. Comme ce salarié est employé toute l'année, le nombre d'heures par mois est égal à 152, obtenu à partir du calcul

1 820 / 12. Notons que les différentes périodes d'emploi sont là utilisées. Par exemple, le salarié numéro 10 figure dans les DADS pour deux périodes d'emploi, la première s'étend sur les cinq premiers mois de l'année, la seconde sur les sept derniers mois. Comme le nombre d'heures est connu pour chaque période d'emploi, on voit ainsi qu'il est imputé à ce salarié, dans le tableau 5, 107 heures par mois puis 128 heures par mois.

Tableau 6 : L'imputation préliminaire des rémunérations mensuelles de base (en €)

	Mois…												
Salarié…	**…1**	**…2**	**…3**	**…4**	**…5**	**…6**	**…7**	**…8**	**…9**	**…10**	**…11**	**…12**	μ_i
…1	3 459	3 115	4342	3 214	3 653	3 453	3 672	3 145	3 143	3 452	3 229	5 867	21,5
…2	1 992	1 794	2 500	1 850	2 103	1 988	2 114	1 811	1 810	1 987	1 859	3 378	13,8
…3	1 531	1 379	1 921	1 422	1 617	1 528	1 625	1 392	1 391	1 528	1 429	2 596	10,7
…4	1 35	1 22	1 70	1 25	1 43	1 35	1 43	1 23	1 23	1 35	1 26	2 29	15,8
…5	1 476	1 329	1 853	1 371	1 559	1 474	1 567	1 342	1 341	1 473	1 378	2 504	11,2
…6	1 564	1 409	1 963	1 453	1 652	1 561	1 660	1 422	1 421	1 561	1 460	2 653	10,8
…7	1 062	956	1 333	986	1 121	1 060	1 127	965	965	1 059	991	1 801	11,2
…8	1 051	947	1 320	977	1 110	1 050	1 116	956	955	1 049	981	1 783	10,3
…9	334	301	149	0	0	0	0	0	0	0	0	0	4,3
…10	1 020	918	1 280	947	1 077	1 216	1 293	1 107	1 107	1 215	1 137	2 066	10,0
…11	0	0	0	0	0	0	1 45	2 5	0	0	0	0	8,6
…12	0	0	0	0	0	0	0	1 172	1 171	1 245	0	0	12,4
π_t	0,95	0,86	1,19	0,88	1,00	0,95	1,01	0,86	0,86	0,95	0,89	1,61	

Ce premier calcul est ensuite utilisé pour établir une imputation préliminaire des rémunérations mensuelles (*cf.* le tableau 6). Notons $\overline{h}_{it}$ le nombre d'heures de base du salarié *i* le mois *t*. Nous supposons, pour imputer la rémunération de base mensuelle notée $\overline{w}_{it}$, la décomposition multiplicative suivante :

$$\overline{w}_{it} = \overline{h}_{it} \times \mu_i \times \pi_t$$

où μ_i est un effet individuel propre au salarié *i* et où π_t est un effet temporel propre au mois *t*. Dans cette écriture, les primes mensuelles sont exactement indexées sur le salaire de base : un salarié qui dispose d'un salaire deux fois plus élevé bénéficiera d'une prime deux fois plus grande.

Nous disposons par ailleurs des deux marges suivantes. D'une part, à partir des DADS, nous avons la rémunération annuelle de base du salarié *i,* notée $\overline{w}_i$. D'autre part, à partir des BRC, nous détenons la masse salariale mensuelle de base du mois *t*, notée $\overline{w}_t$. Aussi, les effets individuels et temporels doivent-ils respecter les deux contraintes suivantes :

$$\overline{w}_i = \sum_t \overline{w}_{it} = \sum_t \overline{h}_{it} \times \mu_i \times \pi_t = \overline{h}_i \times \mu_i \times \sum_t \pi_t$$

et

$$\overline{w}_t = \sum_i \overline{w}_{it} = \sum_i \overline{h}_{it} \times \mu_i \times \pi_t = \overline{h}_t \times \pi_t \times \sum_i \mu_i$$

où $\overline{h}_i$ est le total des heures de base du salarié *i* (qui figure dans les DADS) et où $\overline{h}_t$ est le total des heures de base du mois *t* (qui figure dans les BRC). Il faudrait en outre introduire une contrainte identifiante pour définir de manière unique les effets individuels et temporels.

Nous mettons en œuvre un algorithme itératif pour résoudre numériquement ce problème non linéaire. Les résultats figurent dans le tableau 6. On voit bien, dans ce tableau, que les effets individuels (les μ_i) retracent la hiérarchie des salaires : comme le produit des effets temporels est presque égal à 1, ces effets individuels s'interprètent comme le taux de salaire horaire moyen du salarié *i*. Ainsi le taux de salaire horaire brut dans cet établissement varierait de 8,6 € à 21,5 €. Le cas du salarié numéro 9 est particulier. Il s'agit sans doute d'un apprenti en alternance (*cf.* le tableau 1 où l'on apprend la présence d'au moins un apprenti dans l'établissement au mois de janvier) pour lequel le nombre d'heures renseigné n'est pas égal au nombre d'heures effectuées en entreprise. C'est pour cette raison que son taux de salaire horaire est manifestement sous-estimé.

Les effets temporels (les π_t) résument bien la politique de l'établissement en matière de primes mensuelles. On voit, en particulier, que les mois de mars et de décembre sont caractérisés par un effet temporel égal, respectivement, à 1,19 et à 1,61. L'établissement distribuerait ainsi deux primes par an, la seconde étant trois fois plus importante que la première. Notons que ces effets temporels relèvent bien *a priori* d'un phénomène structurel. Ces effets sont calculés à partir des rémunérations de base, à l'exclusion donc des rémunérations des heures supplémentaires qui peuvent retracer des phénomènes conjoncturels. En outre, ces effets sont établis en contrôlant du nombre d'heures de base : ce n'est pas par exemple parce que de nombreux salariés quittent l'établissement un certain mois que l'effet temporel ce mois là sera plus faible.

Tableau 7 : L'imputation de la rémunération mensuelle des heures supplémentaires (en €)

	Mois…											
Salarié…	**…1**	**…2**	**…3**	**…4**	**…5**	**…6**	**…7**	**…8**	**…9**	**…10**	**…11**	**…12**
…1	0	0	0	0	0	0	0	0	0	0	0	0
…2	221	170	158	284	219	199	225	160	158	159	158	646
…3	0	0	0	0	0	0	0	0	0	0	0	0
…4	0	0	0	0	0	0	0	0	0	0	0	0
…5	0	0	0	0	0	0	0	0	0	0	0	59
…6	0	0	0	0	0	0	0	0	0	0	0	0
…7	17	13	0	22	17	15	17	12	0	12	0	50
…8	18	14	0	23	18	16	18	13	0	13	0	53
…9	0	0	0	0	0	0	0	0	0	0	0	0
…10	19	14	0	24	18	17	19	13	0	13	0	54
…11	0	0	0	0	0	0	22	0	0	0	0	0
…12	0	0	0	0	0	0	0	0	0	0	0	0

Nous imputons la rémunération mensuelle des heures supplémentaires en utilisant une modélisation analogue. Au cas de notre établissement « témoin », nous obtenons les chiffres portés dans le tableau 7. Il y a moins d'enjeux à imputer cet élément de rémunération. Les montants sont plus faibles ; tous les salariés ne sont pas nécessairement concernés. Pour autant, il nous est nécessaire de procéder à cette imputation. En effet, le calcul des réductions de cotisations sur les bas salaires exige de disposer, d'un côté, d'un taux de salaire horaire mensuel et, de l'autre côté, d'une assiette mensuelle. Le taux de salaire doit toutefois être calculé à partir d'éléments qui excluent les heures supplémentaires. Par contre, l'assiette mensuelle est une assiette totale, y compris la rémunération des heures supplémentaires.

Tableau 8 : L'imputation sur barème des réductions de cotisations sur les bas salaires (en €)

	Mois…												
	…1	**…2**	**…3**	**…4**	**…5**	**…6**	**…7**	**…8**	**…9**	**…10**	**…11**	**…12**	**Total**
Déclarées	1 694	1 521	1 056	1 445	1 320	1 328	1 533	1 652	1 708	1 484	1 526	82	16 349
Imputées	1 381	1 787	413	1 610	1 052	1 340	1 178	1 954	1 941	1 520	1 686	0	15 863
Différence	313	–266	643	–165	268	–12	355	–302	–233	–36	–160	82	486

Nous sommes ainsi en mesure d'imputer, sur barème, les réductions de cotisations sur les bas salaires pour chaque salarié. Il nous donc loisible de calculer le montant mensuel de ces réductions pour, éventuellement, infirmer notre méthode d'imputation des rémunérations. Pour l'établissement « témoin », les résultats de cette imputation sur barème sont portés dans le tableau 8. Il en ressort que les réductions sont fortement sous-estimées au mois de mars. En revanche, en décembre, elles sont légèrement sous-estimées. L'hypothèse principale de notre décomposition multiplicative – à savoir un régime de prime exactement hiérarchisé – serait ainsi remise en cause pour le mois de mars. Il faut sans doute envisager, dans notre établissement « témoin », la présence d'un système dual de prime où la prime du moi de mars aurait la nature d'une prime de performance, dont la part serait croissante avec la rémunération, et où la prime du mois de décembre aurait la nature d'une prime générale, liée par exemple à l'ancienneté ou au poste occupé (*cf.* BIGNON et FOLQUES (2009) qui mettent en évidence, à partir de l'enquête sur le coût de la main-d'œuvre et la structure des salaires, l'importance respective de ces deux types de prime).

C'est ainsi que nous nous engageons dans une modélisation un peu plus sophistiquée de l'imputation des rémunérations mensuelles où seraient distingués différents régimes de prime en fonction notamment du taux de salaire horaire moyen de chaque salarié. Notre démarche s'apparente à un ajustement économétrique dans le but de minimiser les erreurs mensuelles qui apparaissent, *ex post*, entre le montant mensuel des réductions de cotisations déclaré par l'établissement et le montant mensuel qui résulte de l'imputation sur barème à partir des rémunérations mensuelles estimées.

4. Conclusion

Nous avons jeté les bases d'un système d'information de mise en cohérence des deux principales obligations déclaratives des employeurs auprès des organismes de recouvrement des cotisations sociales : d'une part, les BRC, une déclaration qui accompagne le versement mensuel ou trimestriel des prélèvements sociaux et, d'autre part, les DADS, une déclaration annuelle par salarié utilisée pour établir les droits sociaux et les obligations fiscales des salariés.

Il nous a fallu développer des technologies particulières pour satisfaire une ambition particulière : celle d'utiliser les données exhaustivement (près de 94 millions de lignes de décompte dans les BRC et près de 41 millions de périodes d'emploi dans les DADS). Nous avons mis au point, d'un côté, une façon de représenter l'information de manière particulièrement compacte et, de l'autre côté, une utilisation concurrente des processeurs du micro-ordinateur dont nous disposons.

Le système d'information ainsi construit doit permettre de reconstituer les revenus perçus mensuellement par les salariés. Ce travail d'imputation implique la mise en œuvre de différents algorithmes numériques qui restent à perfectionner pour décrire au mieux la chronique du versement des salaires.

5. Références bibliographiques

Commission de la norme DADS-U (2009), *Cahier technique DADS-U V08R09*, note technique qui était disponible sur le site http://www.net-entreprises.fr.

BIGNON N. et FOLQUES D. (2009), « La structure des rémunérations en 2006 : les primes représentent en moyenne 12,4 % de la rémunération dans le secteur concurrentiel », *Premières Informations*, n° 31.4, juillet.

DEPIL S. et KERJOSSE R. (2010), « Les salaires dans les entreprises en 2008 : une hausse conséquente contrebalancée par l'inflation », *INSEE Première*, n°1 300, juin.

ROUX S. (2001), *Refonte du panel DADS : principes et premières estimations d'emploi et de salaire*, note INSEE.