Direction des Statistiques Démographiques et Sociales

N° F1504

CONTRÔLES DES RÉMUNÉRATIONS DANS LES DÉCLARATIONS ANNUELLES DE DONNÉES SOCIALES (DADS)

Une analyse exploratoire pour améliorer la détection des points atypiques

Claire JACOD

Document de travail

Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des Documents de Travail
de la
DIRECTION DES STATISTIQUES DÉMOGRAPHIQUES ET SOCIALES

N°F1504

CONTRÔLES DES RÉMUNÉRATIONS DANS LES DÉCLARATIONS ANNUELLES DE DONNÉES SOCIALES (DADS)

Une analyse exploratoire pour améliorer la détection des points atypiques

AUTEUR : CLAIRE JACOD

(DIVISION EXPLOITATION DES FICHIERS ADMINISTRATIFS SUR L'EMPLOI ET LES REVENUS D'ACTIVITÉ)

Document de travail

novembre 2015

Ce document reprend un mémoire de Formation Continue Diplômante des Attachés, encadré par Emmanuel Gros (Département des méthodes statistiques).

Contrôles des rémunérations dans les DADS – Détection des points atypiques

Résumé

À l'Insee, les Déclarations Annuelles de Données Sociales (DADS) constituent la source annuelle de référence sur l'emploi et les salaires. La qualité des informations déclarées et leur adéquation aux besoins statistiques sont variables, notamment sur les rémunérations. Il faut donc les contrôler et éventuellement les corriger. Ces traitements sont actuellement réalisés au sein des applications DADS, sur le secteur privé, et Siasp (Système d'information sur les agents du service public), sur le secteur public, selon des méthodes proches, basées sur la modélisation du salaire horaire par les moindres carrés ordinaires, en utilisant des variables auxiliaires redressées.

Dans le cadre de cette étude, on propose une étape préliminaire de détection d'outliers, en amont des traitements effectués dans les applications DADS et Siasp, et donc avant le redressement des variables auxiliaires. Conformément aux attentes du futur système d'information sur l'emploi et les revenus, l'objectif est un repérage précoce et homogène sur l'ensemble du champ des valeurs atypiques, qui seront confirmées ou infirmées dans la suite du processus. Une fois ces données isolées en vue d'un traitement spécifique, il est possible de traiter sans délai les informations issues du cas nominal.

On montre que les moindres carrés ordinaires ne sont pas suffisamment robustes et qu'il est nécessaire d'utiliser une méthode d'estimation robuste non seulement aux points verticaux, mais également aux points leviers. On propose une méthode alternative, adaptée à la présence de variables explicatives discrètes et au volume important des données que l'on souhaite traiter.

Mots clés : Déclarations Annuelles de Données Sociales, source administrative, méthodes robustes, points atypiques, points leviers

Controls of compensation figures in Annual declaration of social data (DADS) - Outlier detection

Abstract

At Insee, Annual declaration of social data (DADS) is the annual reference source on employment and salaries. The quality of the declared information and its adequacy to statistical needs are variable, in particular where remunerations are concerned. It is thus necessary to control them and to correct them when appropriate. These treatments are now carried out by the DADS application, for the private sector, and the Siasp (Information system on the agents of the public service) application, for the public sector. Both applications use similar methods, based on the ordinary least squares modelisation of the hourly wage rate, using previously adjusted auxiliary variables.

In the course of this study we propose a preliminary stage of outliers detection, before the treatments carried out by the DADS and Siasp applications, and thus before adjustments to the auxiliary variables. In line with what is expected from the future information system on employment and salaries, the objective is an early and homogeneous detection of the outlier results on the whole field, outlier results which will be confirmed or corrected during the remaining process. Once these data have been isolated to be treated specifically, it is possible to handle the information pertaining to the general case without delay.

It is shown that ordinary least squares are not a robust method enough and that such a robust method must be used not only on vertical points, but also on leverage points. We propose an alternative method, adapted to the presence of discreet explanatory variables, as well as to the large volume of data we need to process.

Keywords : Annual declaration of social data (DADS), administrative source, robust methods, outlier, leverage points

SOMMAIRE

Introduction

A l'Insee, les Déclarations Annuelles de Données Sociales (DADS), complétées par les fichiers de paie de l'État, constituent la source annuelle de référence pour la production de données statistiques structurelles sur l'emploi salarié et les revenus d'activité salariale. Elles permettent de décrire les caractéristiques des emplois et les évolutions de salaires, et de réaliser des études longitudinales via le panel Tous salariés. Depuis la mise en place du nouveau dispositif de recensement de la population, elles sont également utilisées pour estimer le niveau annuel de l'emploi[1]. En outre, les DADS sont de plus en plus sollicitées par les chercheurs, pour des analyses à des niveaux géographique et sectoriel fins ou sur l'étude des carrières professionnelles.

La qualité des informations déclarées dans les DADS est variable, selon l'usage administratif et la nécessité pour le déclarant de bien les renseigner. En outre, les variables déclarées ne sont pas toujours adaptées aux besoins statistiques. Il est donc nécessaire de réaliser un ensemble de traitements sur les données collectées[2], afin d'en tirer une information statistique la plus fiable et la plus complète possible, notamment sur les concepts centraux que sont les rémunérations[3] et le temps de travail[4]. Cette exploitation est réalisée au sein du Département de l'emploi et des revenus d'activité (DERA), et s'appuie sur trois applications informatiques : l'application DADS, pour le secteur privé, l'application Siasp (Système d'information sur les agents des services publics) pour le secteur public[5] et le Frontal. Cette dernière application a été mise en place en 2012 pour prendre en charge le changement de format des DADS : une nouvelle norme de déclaration a en effet été mise en place pour les données sociales de 2011, appelée Norme de Déclaration Dématérialisée Des Données Sociales (N4DS). Le Frontal réceptionne les données au format N4DS et les transforme en ancienne norme, pour qu'elles puissent être prises en charge par les applications DADS et Siasp.

Actuellement, les contrôles de la qualité des données sur les rémunérations sont réalisés au sein des applications DADS et Siasp. Ils visent à diffuser des informations cohérentes sur les salaires et le temps de travail. Les méthodes utilisées s'appuient dans les deux cas sur des modèles de régression linéaire pour la détection de valeurs aberrantes et leur redressement, mais présentent des spécificités liées au mode de fonctionnement des applications.
Avec la mise en place de la N4DS, l'ensemble des données brutes issues des déclarations DADS et concernant les deux secteurs (privé et public) sont réunies en un seul applicatif, le Frontal, rendant possible une analyse plus homogène des rémunérations et du temps de travail, en amont des applications DADS et Siasp. Dès lors, contrôler les rémunérations au niveau du Frontal serait une opportunité pour poursuivre les efforts entrepris ces dernières années pour améliorer la comparabilité entre les secteurs public et privé, comme le préconisent le Conseil National de l'Information Statistique (CNIS), dans ses orientations de moyen terme 2014-2018, et le Conseil commun de la fonction publique, sachant que les données des deux secteurs sont d'ores et déjà diffusées au sein d'un même fichier, appelé fichier DADS-Grand format ou Tous salariés[6]. La nouvelle norme de déclaration comporte également de nouvelles variables sur le secteur public qui pourraient améliorer la détection des points aberrants. De plus, l'application Frontal étant en amont des dispositifs DADS et Siasp, elle dispose des données plus tôt, ce qui pourrait permettre de détecter plus rapidement les points atypiques.

L'objectif de cette étude est donc de déterminer une méthode, homogène sur la source DADS, de détection des valeurs atypiques de la rémunération horaire, qui permette d'identifier de façon mutualisée au niveau du Frontal des données à contrôler prioritairement par la suite dans les applications DADS et Siasp. On s'appuie sur des méthodes de régression robuste, en testant ces outils sur un gros volume de données, constituées par des déclarations administratives brutes dont la

1 Voir Bibliographie [3].

2 Voir Annexes §1.1.

3 Voir Définitions, p. 38.

4 Voir Définitions, p. 38.

5 Alimentée par les DADS des fonctions publiques hospitalière et territoriale et de certains établissements publics ainsi que par les fichiers de paie de l'État.

6 Les données issues du traitement des déclarations des particuliers employeurs permettent de compléter le champ salarié.

qualité n'a pas été contrôlée. La masse des informations à traiter a conduit à privilégier des solutions qui ne sont pas les plus optimales du point de vue statistique, mais qui restent réalistes en termes de temps de traitement.

L'étude est construite en cinq parties.
La première partie rappelle le contexte dans lequel l'étude a été réalisée.
La deuxième partie décrit les contrôles actuels et expose la problématique traitée dans le cadre de cette étude.
La troisième partie décrit les étapes préliminaires de contrôle des informations sur la rémunération et le temps de travail visant à exclure de l'analyse les observations incohérentes.
La quatrième partie expose les différentes méthodes possibles pour détecter les rémunérations horaires atypiques.
La cinquième et dernière partie explique la façon dont cette détection des points atypiques peut trouver une application concrète dans les systèmes applicatifs actuel et futur.

1. Contexte métier, méthodologie actuelle et problématique

1.1. Description de la source et norme des DADS

La DADS est une formalité déclarative que doit accomplir toute entreprise employant des salariés. Dans ce document destiné à la fois aux administrations sociales et fiscales, les employeurs, y compris les entreprises nationales, les administrations publiques et les collectivités locales, sont tenus, annuellement et pour chaque établissement, de communiquer aux organismes de Sécurité sociale d'une part[7], à l'administration fiscale d'autre part, la masse des traitements qu'ils ont versés, les effectifs employés et une liste nominative de leurs salariés indiquant pour chacun leurs caractéristiques d'emploi et le montant des rémunérations salariales perçues. Étant donné que les fichiers reçus par l'Insee comportent des informations sensibles (comme le NIR[8] des salariés, par exemple), l'ensemble du processus de traitement de ces données est soumis à des règles très strictes de confidentialité.

Le format et le contenu des informations transmises par les entreprises sont par ailleurs soumis à un ensemble de règles décrites dans une norme, qui évolue chaque année pour prendre en compte notamment les nouvelles réglementations en vigueur tant dans le domaine social que dans le domaine fiscal. Entre 2006 et 2011, les déclarations devaient être transmises dans la norme DADS-Unifiée (DADS-U). Depuis janvier 2012 et sur les données sociales de la validité 2011, les déclarations sont transmises dans une nouvelle norme, la Norme de Déclaration Dématérialisée Des Données Sociales (N4DS). Cette dernière constitue un profond changement par rapport à la DADS-U, en introduisant les modifications suivantes :

- introduction de structures modulaires ;
- variables adaptées au type de salarié, rendant les déclarations moins lourdes pour les entreprises. Le type de salarié, appelé code population[9], est défini de la façon suivante :

Tableau 1 : Type de salarié et code population

Type	Code population	Libellé	Libellé court
privé	10	salarié sous contrat de droit privé	salarié privé classique
	11	salarié artiste ou technicien sous contrat à durée déterminée dans le spectacle	salarié du spectacle
	13	salarié sous contrat de droit privé travaillant dans des organismes de droit public	salarié privé dans le public
	14	fonctionnaire détaché comme salarié sous contrat de droit privé	fonctionnaire détaché
public	40	fonctionnaire ou "ouvrier d'État"	fonctionnaire
	43	agent de droit public non fonctionnaire (y compris personnel médical hospitalier)	non-titulaire de la fonction publique
élu	42	élu	élu

- davantage d'informations disponibles sur le secteur public.

7 Pour les salariés relevant du régime général.

8 Numéro d'Identification au Répertoire.

9 Dans la suite du document, par souci de simplification, on utilisera les libellés courts du code population.

1.2. Utilisation de la source DADS au sein du système d'information sur l'emploi et les revenus d'activité

Les DADS constituent la source pivot du Système d'Information sur l'Emploi et les Revenus d'Activité (Siera), géré au sein du Département de l'emploi et des revenus d'activité (DERA) à l'Insee. Les DADS contribuent ainsi à remplir l'objectif du Siera de production d'une information structurelle annuelle sur l'emploi et les salaires.

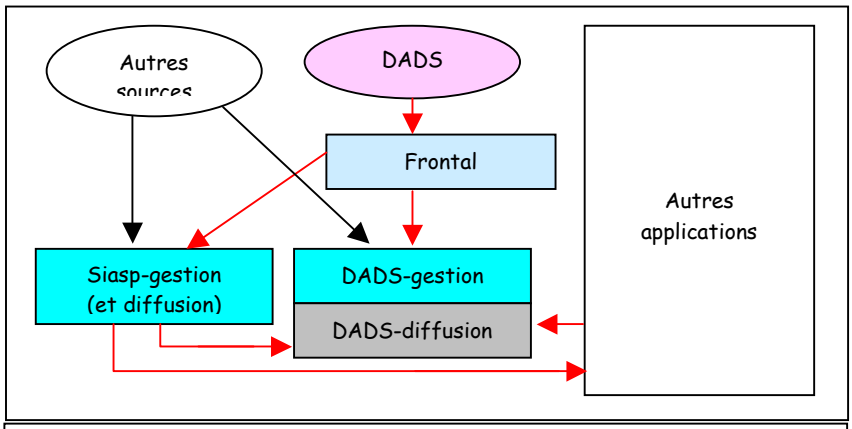
Depuis 2009 et avant la mise en place de la N4DS, les données étaient traitées via deux applications de gestion, chacune sur leur champ propre :

- L'application DADS, application pivot des estimations annuelles d'emploi et de salaires sur le secteur privé. La quasi-totalité des informations traitées par l'application provient des DADS, seules quelques informations parviennent à l'Insee directement de la Direction générale des finances publiques (DGFiP), pour des entreprises dont les salariés ne dépendent pas du régime général de la Sécurité sociale ;

- L'application Siasp (Système d'Information sur les Agents du Service Public), sur le secteur public. Les informations portant sur la fonction publique d'État (FPE) sont en partie présentes dans les DADS, mais l'application traite ce champ à partir d'une autre source : les fichiers de paie des agents de l'État. Ainsi, plus de trois millions de périodes d'emploi issues des DADS ne sont pas conservées dans Siasp. Les fichiers de paie sont une source mensuelle, les données sur la FPE sont annualisées pour être traitées conjointement avec les informations annuelles issues des DADS, sur les fonctions publiques territoriale et hospitalière (FPT et FPH).

Avec l'arrivée de la N4DS sur la validité 2011, le dispositif a été complété par une nouvelle application en 2012, appelée Frontal. Cette application a pour objectif principal d'assurer la continuité de la production des informations issues des DADS en transformant les données reçues en N4DS en ancienne norme DADS-U, que les applications DADS et Siasp sont capables de traiter. Elle stocke également les informations au format N4DS pour des besoins d'analyse des données administratives, en vue d'améliorer et compléter les produits et les services du système d'information actuel. Ce sont ces données qui sont exploitées dans le cadre de cette étude.

Schéma 1 : Dispositif applicatif d'exploitation de la source DADS



En rouge : contribution à l'objectif du Siera sur la production d'informations annuelles sur l'emploi et les salaires.

1.3. Précisions sur les unités statistiques en jeu

Dans toute la phase amont des traitements, que ce soit dans l'application Frontal ou dans les applications clientes du Frontal, l'unité de traitement de base est la **période** d'activité. Cette unité, très adhérente à la norme des DADS, que ce soit la N4DS ou la DADS-U, se caractérise par une fenêtre temporelle définie par une date de début et une date de fin, durant laquelle la situation du salarié reste inchangée au sein d'un établissement. Par exemple, pour un salarié changeant de contrat de travail en cours d'année au sein d'un même établissement, on trouvera deux périodes d'activité. Dès lors

qu'un salarié perçoit une rémunération ou que sa situation au sein d'un établissement lui ouvre des droits sociaux, l'établissement a obligation de déclarer une période d'activité pour ce salarié. Il existe donc des périodes d'activité qui ne correspondent pas à des périodes de travail, par exemple des salariés en congé parental.

1.4. Un gros volume d'informations traitées

Le Frontal a traité en 2011 plus de 2 000 fichiers, 1,8 million de déclarations, correspondant à 45,7 millions de « lignes salariés », correspondant à un salarié dans une déclaration, et 63,0 millions de périodes.
L'Insee reçoit certaines informations en double, certaines entreprises apportant des corrections dans leur déclaration dans un nouvel envoi. Les données sont donc filtrées au niveau du Frontal et 60,5 millions de périodes sont fournies aux applications DADS et Siasp.
Les informations sont également triées par application destinataire : DADS récupère 33,8 millions de lignes salariés et 48,0 millions de périodes, Siasp récupère 10,1 millions de lignes salariés et 12,5 millions de périodes[10].

2. Description des traitements actuels de cohérence des rémunérations

2.1. Les concepts de salaire

Du salaire de base au salaire net à payer, il existe plusieurs concepts de salaire selon les éléments pris en compte[11]. Avec les DADS, on cherche à mesurer, via le salaire brut et le salaire net, les rémunérations brute (versée par l'employeur) et nette (perçue par le salarié)[12]. Pour ce faire, les informations sources dont on dispose dans la déclaration ne sont pas toujours bien renseignées, ni toutes utilisables de la même façon selon le type de salariés.

En théorie, les différentes variables financières de rémunération doivent respecter l'équation suivante :

SALAIRE_NET = **SALAIRE_BRUT** - cotisations salariées - CSG[13] - CRDS[14]

On peut appréhender ces notions à partir des <u>données fiscales</u> :

SALAIRE_BRUT = NET_FISCAL[15] + cotisations salariées + CSG déductible
+ heures supplémentaires exonérées

SALAIRE_NET = NET_FISCAL - CSG non déductible - CRDS
+ heures supplémentaires exonérées

On peut appréhender ces notions à partir des <u>données sociales</u> :

SALAIRE_BRUT = (BASE_CSG[16] / part du salaire brut soumis à CSG)

SALAIRE_NET = (BASE _CSG / part du salaire brut soumis à CSG)
- cotisations salariées - CSG - CRDS

Dans les applications DADS et Siasp, le salaire brut est obtenu à partir de la base CSG (données sociales), tandis que le salaire net est obtenu à partir de la rémunération nette fiscale (données fiscales). Ces variables statistiques, issues d'informations différentes, sont donc susceptibles de ne pas être cohérentes entre elles et ne pas respecter les équations théoriques ci-dessus.

---

10 Dont plus de trois millions de périodes font partie du champ de la FPE et ne sont donc pas utilisées dans Siasp.

11 Voir Bibliographie [2].

12 Voir Annexes §1.2.

13 Contribution Sociale Généralisée.

14 Contribution à la Réduction de la Dette Sociale.

15 Voir Annexes §1.2.

16 Voir Annexes §1.2.

2.2. Un contrôle nécessaire des rémunérations

Les données déclarées dans les DADS ne sont pas toutes de la même qualité. Si les informations portant sur les rémunérations sont globalement remplies avec soin par les déclarants, car directement utiles à l'administration fiscale et aux organismes de protection sociale, pour l'ouverture des droits aux salariés, ce n'est pas le cas de l'ensemble des variables disponibles, notamment la PCS[17]. De plus les informations administratives ne sont pas toutes adaptées aux concepts statistiques que l'on souhaite mesurer[18].

Il est donc nécessaire de réaliser un ensemble de traitements sur les données collectées, afin d'en tirer une information statistique la plus fiable et la plus complète possible[19]. Ces traitements concernent notamment les concepts centraux que sont les rémunérations et le temps de travail. Dans cette optique, les applications DADS et Siasp ont mis au point des méthodes de contrôle et redressement des salaires et/ou temps de travail au niveau période, chacune sur leur champ respectif : le secteur privé pour l'application DADS et le secteur public pour Siasp.

2.3. Les contrôles de cohérence des rémunérations en amont de la chaîne : historique et méthode actuelle

2.3.1. Dans l'application DADS

Les salaires et le temps de travail sont à l'origine les variables centrales de l'application telle qu'elle a été mise en œuvre en 1977[20]. Ainsi, elles faisaient déjà l'objet de contrôles-redressements à cette époque. Ceux-ci ont été revus depuis, notamment les contrôles de cohérence des rémunérations, à l'occasion du projet de refonte des DADS[21] (1998-2002). Le travail des gestionnaires a été ciblé sur les grandes déclarations[22] et a été remplacé, pour les petites déclarations[23], par des redressements automatiques.

La détection des points atypiques a également été améliorée en deux temps :

- dans un premier temps, leur détection à partir de seuils fixes a été remplacée par un modèle de régression du salaire horaire, estimé sur l'échantillon au 1/12 de l'année N-1 : les 5 % des résidus standardisés de l'année N les plus élevés et les 5 % les plus faibles sont considérés comme atypiques ;
- dans un deuxième temps, le modèle de régression du salaire horaire[24] a été amélioré, en le réalisant sur l'ensemble de l'année N-1 (et non plus sur l'échantillon) et en isolant les points atypiques de l'année N de la même façon.

Le traitement actuel apporte séquentiellement des réponses à l'ensemble des problèmes posés sur les rémunérations et temps de travail, ordonnancement basé sur l'hypothèse empiriquement vérifiée que le salaire constitue l'information la plus fiable :

- la cohérence interne des variables financières : si les variables financières de rémunération ne vérifient pas des règles de cohérence de type « salaire net inférieur au salaire brut » ou si elles ne vérifient pas les équations de salaires définies plus haut[25], on corrige la ou les variables mise(s) en cause, en la(les) calculant à partir des autres variables financières ;

17 Profession et Catégorie Socioprofessionnelle.

18 Voir Annexes §1.1.

19 Voir Annexes §1.1.

20 Voir Bibliographie [12].

21 Le premier objectif de ce projet, appelé DADS2, était d'ouvrir l'utilisation des DADS à d'autres problématiques, notamment l'emploi, ses caractéristiques, ... Il a donc fallu mettre en place d'autres contrôles, notamment d'exhaustivité par rapport aux employeurs et lieux de travail, via Sirene (Système Informatique pour le Répertoire des Entreprises et de leurs Etablissements). Voir Bibliographie [13].

22 Voir Définitions, p. 38.

23 Voir Définitions, p. 38.

24 Voir Définitions, p. 38.

25 Voir §2.1.

- la vraisemblance du salaire horaire compte-tenu des caractéristiques du poste et du salarié (détection des salaires horaires hors normes) : on réalise une estimation du salaire net horaire de l'année N-1[26] par les moindres carrés ordinaires, avec les variables explicatives suivantes :
    - la catégorie socioprofessionnelle (CS) à deux positions ;
    - le sexe ;
    - l'âge quinquennal ;
    - l'activité de l'établissement (nomenclature agrégée en 38 postes) ;
    - la région de travail (distinction Paris/Province) ;
    - la tranche d'effectif de l'établissement (donnée par le répertoire Sirene[27]) ;
    - la catégorie juridique de l'entreprise.

  On détermine ainsi des coefficients et on calcule un salaire horaire théorique pour les observations de l'année N. On définit alors un écart $ecart\_abs$ entre le salaire horaire observé et salaire horaire théorique par :

  $$ecart\_abs = \left| \frac{\log(y_i) - \log(\hat{y}_i)}{\hat{\sigma}} \right|, \text{ où :}$$

  y_i est le salaire horaire observé,

  $\hat{y}_i$ est le salaire horaire théorique estimé à partir des coefficients calculés sur l'année N-1,

  $\hat{\sigma}$ est l'écart-type du modèle sur l'année N-1.

  Il y a soupçon de trop haut ou trop bas salaire lorsque $ecart\_abs > 4$.

  S'il y a soupçon de trop haut ou trop bas salaire ou bien si le salaire horaire observé est inférieur à 80 % du SMIC[28] horaire, alors la variable est redressée par Hot Deck[29] dans la classe d'imputation concernée, constituée par le croisement des variables explicatives du modèle et du type d'emploi[30]. Le nombre d'heures et les variables financières sont alors mis en cohérence avec le salaire horaire, en redressant prioritairement le nombre d'heures[31].

- la cohérence entre le nombre d'heures de travail et la durée d'emploi[32] : s'il y a des incohérences entre le nombre d'heures par jour et la condition d'emploi (temps plein, temps partiel), par rapport aux valeurs de référence calculées sur N-1, alors la quotité[33] et/ou la durée sont redressées.

Ces trois types d'anomalies peuvent faire intervenir les gestionnaires :

- systématiquement sur les **grandes déclarations** (500 périodes ou plus) ;
- au cas par cas sur les **déclarations moyennes**[34] (de 11 périodes à moins de 500 périodes), suivant la gravité et le nombre d'anomalies.

Les traitements sont totalement automatiques pour les **petites déclarations**, comprenant 10 périodes ou moins.

La détermination des classes d'imputation pour le Hot Deck prend en compte des variables issues des déclarations, qui peuvent être redressées par l'application DADS, comme la PCS, par exemple, mais également des variables récupérées de sources externes, comme la tranche d'effectif de l'établissement et l'activité de l'établissement, issues de la source Sirene.

---

26 L'estimation est réalisée sur les valeurs redressées. En fin de campagne, le modèle est ensuite ré-estimé sur l'année N et les redressements complétés.

27 Système Informatique pour le Répertoire des Entreprises et de leurs Etablissements. Voir également Annexes §1.1.

28 Salaire Minimum Interprofessionnel de Croissance.

29 Méthode stochastique d'imputation. Pour plus d'informations sur le redressement par Hot Deck, voir Bibliographie [3] et [11].

30 Le type d'emploi permet de distinguer les emplois dits « ordinaires » de quelques cas particuliers, comme les apprentis, les stagiaires ou les emplois aidés. Voir Bibliographie [20].

31 Les données financières sont jugées plus fiables

32 Voir Définitions, p. 38.

33 Voir Définitions, p. 38.

34 Voir Définitions, p. 38.

2.3.2. Dans l'application Siasp

Dans l'application Siasp, des premiers contrôles de cohérence des rémunérations ont été prévus dans les spécifications du projet de constitution du système Siasp. La méthode préconisée était basée sur celle de l'application DADS. Cependant, ces contrôles ont été spécifiés sur la base des informations reçues dans les fichiers de paie. Ils se sont révélés inadaptés dans la pratique à la source DADS et n'ont jamais été utilisés.
Des contrôles de cohérence des rémunérations ont alors été mis en place hors application en 2009, à partir de réflexions menées conjointement avec la Drees[35]. La méthode utilisée reprend les grands principes de la méthode de l'application DADS. Elle a été testée sur la FPH puis généralisée à l'ensemble des fonctions publiques, y compris sur la source des fichiers de paie portant sur la FPE[36].

Si l'application Siasp s'est inspirée de la méthode de l'application DADS pour le traitement des incohérences des rémunérations, la modélisation est ici réalisée sur les données de l'année N, et l'ordonnancement et les règles de décision utilisées sont différents de ceux de l'application DADS :

- Apurement et recalage préalables des variables :
  - Cohérence interne des variables financières, et recalage des variables incohérentes à partir des variables financières valides ;
  - Cohérence interne du temps de travail, en redressant d'abord la durée de travail puis la quotité.
- Cohérence entre traitement indiciaire et nombre d'heures :
  - Calcul du traitement théorique, à partir du nombre d'heures, de l'indice et de la valeur du point d'indice ;
  - Le nombre d'heures des périodes pour lesquelles l'écart entre le traitement théorique et le traitement observé est supérieur à 40 % est automatiquement redressé à partir de l'équation comptable du traitement théorique ;
- Cohérence entre le nombre d'heures et le salaire net :
  - On réalise une estimation du salaire net horaire de l'année N par les moindres carrés ordinaires, avec les variables explicatives suivantes[37] :
    - le statut ;
    - le sexe ;
    - l'âge ;
    - la région de travail (distinction métropole/départements d'outre-mer) ;
    - le type d'employeur, défini pour la source DADS à partir de la catégorie juridique.

    Les périodes atypiques (les premiers et derniers 5% de la distribution des résidus standardisés) de cette régression sont mises à part et seront redressées.
  - On réalise une deuxième estimation du salaire net horaire par les moindres carrés ordinaires, avec les mêmes variables explicatives, en excluant les périodes atypiques de la première régression. On détermine ainsi une deuxième série de périodes atypiques (les premiers et derniers 5% de la distribution des résidus standardisés) ;
  - Les valeurs atypiques sont redressées automatiquement sur le nombre d'heures, grâce aux valeurs prédites du salaire horaire du deuxième modèle[38] , le salaire étant considéré comme valide. La quotité puis, si nécessaire, la durée sont ensuite mises en cohérence avec le nombre d'heures corrigé.

2.4. Les autres contrôles réalisés dans la chaîne

Dans le cadre de ce document, on s'intéresse uniquement à la phase amont des traitements statistiques, sans aller jusqu'aux informations fournies dans les fichiers de diffusion.

---

35 Direction de la recherche, des études, de l'évaluation et des statistiques. La Drees exploite les produits de l'application Siasp sur la FPH.

36 Une modélisation par source est réalisée. Pour la source DADS, le champ détaillé par versant de la fonction publique est une variable explicative du modèle.

37 Les variables explicatives sont des variables déclarées ou bien recalculées par l'application Siasp, comme le statut, par exemple.

38 Pour plus d'informations sur l'imputation par la régression, voir Bibliographie [11].

Cependant, les contrôles et traitements de la chaîne actuelle ne se résument pas à cette phase amont. En effet, tout un ensemble de contrôles et de traitements supplémentaires sont mis en œuvre pour **garantir *in fine* la qualité statistique des informations statistiques produites**. Des vérifications accrues et traitements particuliers sont notamment réalisés sur les concepts d'emploi, d'activité et de localisation des salariés, ainsi que sur le secteur d'activité de l'établissement qui les emploie[39]. De plus, les données font l'objet de plusieurs validations, en interne à l'Insee comme en externe, en partenariat avec la Dares[40], la Drees et la DGAFP[41]. Ces validations portent à la fois sur le niveau et les évolutions des données diffusées.

2.5. Les objectifs d'une mutualisation de la détection des valeurs atypiques pour l'ensemble du champ

Les contrôles des rémunérations sont actuellement réalisés par les applications DADS et Siasp. Ils répondent au même besoin : diffuser des informations cohérentes sur les salaires et le temps de travail. Si les méthodes utilisées s'appuient toutes les deux sur des modèles de régression, les contrôles et redressements réalisés restent très liés au mode de fonctionnement des applications, notamment à la validité utilisée pour déterminer les points atypiques (N pour Siasp, N-1 pour DADS) et à l'étape de la chaîne de traitement à laquelle ils sont réalisés. Ils dépendent en particulier de l'étape préalable de redressement et de recodage des variables auxiliaires des modèles (cohérence interne des variables financières, codage de la PCS et du statut), étape réalisée de façon spécifique dans chacune des deux applications.

Avec la mise en place du Frontal, on dispose de l'ensemble des données brutes issues des DADS en une seule application. Il est donc possible d'étudier l'apport d'une détection mutualisée et davantage homogène sur la source DADS des observations atypiques en termes de rémunérations, en amont des applications DADS et Siasp. Cette mutualisation semble d'autant plus légitime que les données des deux secteurs sont diffusées au sein d'un même fichier, appelé fichier DADS-Grand format (Tous salariés). La nouvelle norme de déclaration comporte également de nouvelles variables sur le secteur public qui pourraient améliorer la détection des points aberrants. L'application Frontal étant en amont des dispositifs DADS et Siasp, elle dispose également des données plus tôt, ce qui pourrait permettre de détecter plus rapidement les points atypiques.

En contrepartie, ces nouvelles informations et les variables auxiliaires utilisées actuellement dans les redressements des rémunérations et/ou du temps de travail ne sont pas toutes de grande qualité, notamment les informations sur l'activité exercée par le salarié. Cette information étant essentielle pour redresser *in fine* les variables financières et de temps de travail, il n'est pas pertinent de réaliser la phase de redressement au niveau du Frontal. En revanche, il est utile de détecter le plus tôt possible les éventuelles incohérences sur les variables financières et de temps de travail, pour pouvoir guider les applications dans leurs contrôles.

L'objectif de cette étude est donc de déterminer une méthode de détection des points atypiques, au niveau du Frontal, qui puisse identifier des périodes à contrôler prioritairement dans les applications DADS et Siasp et parmi lesquelles il conviendra de vérifier s'il s'agit :

- de points atypiques mais normaux ;
- de périodes à redresser sur les variables financières et/ou sur les variables de temps de travail et/ou sur les variables auxiliaires.

On peut ainsi fournir aux applications une liste de périodes à confirmer ou à redresser. Cette information est directement utile aux applications lorsque les variables auxiliaires utilisées dans la modélisation du Frontal n'ont pas été redressées entre temps, spécialement dans le cas des petites et moyennes déclarations à destination de l'application DADS :

- pour les petites déclarations : les périodes atypiques sont redressées automatiquement dans l'application DADS, l'indicateur du Frontal peut permettre d'identifier des cas d'incohérence supplémentaires ;

39 Voir Bibliographie [20].

40 Direction de l'animation de la recherche, des études et des statistiques.

41 Direction générale de l'administration et de la fonction publique.

- pour les déclarations moyennes : l'identification des points atypiques au niveau du Frontal peut entrer en ligne de compte dans l'orientation de ces cas vers des traitements gestionnaires.

En revanche, lorsque les variables auxiliaires sont redressées entre temps dans les applications, l'information fournie par le Frontal est moins utile, car les observations atypiques détectées peuvent l'être non pas en raison de leur salaire horaire, mais en raison d'une mauvaise qualité des variables explicatives déclarées[42].

Sur les 60,5 millions de périodes transmises par le Frontal aux applications, 41,7 % sont issues de grandes déclarations à destination de l'application DADS[43] et 20,7 % sont à destination de Siasp.
Les trois quarts des périodes correspondent à des salariés privés classiques, cette proportion allant jusqu'à 96 % pour les déclarations à destination de l'application DADS. A contrario, les déclarations à destination de Siasp contiennent essentiellement des périodes de fonctionnaires et de non-titulaires de la fonction publique.

Pour les 18,3 millions de périodes issues de petites et moyennes déclarations à destination de DADS (soit 37,6 % des périodes), il peut être directement utile de détecter les individus aberrants. Pour les autres, le dispositif peut compléter l'information actuellement contrôlée au niveau des applications[44].

2.6. Avantages et inconvénients de la mutualisation

2.6.1. Avantages

- Une méthode homogène serait utilisée sur l'ensemble de la source DADS.
- Le Frontal-N4DS a traité la quasi-totalité des données fin mars N+1. Au-delà, les fichiers reçus sont peu nombreux et comportent peu d'informations. Les applications DADS et Siasp réalisent leurs contrôles de cohérence des rémunérations bien plus tard : ils sont réalisés en juillet et août N+1 côté Siasp, et s'étalent entre mars et décembre N+1 côté DADS. Ainsi, on dispose d'une période assez longue qui pourrait permettre de détecter au niveau de la source amont des périodes atypiques en termes de rémunérations, et de les lister pour qu'elles soient prises en charge prioritairement par les applications.
- On dispose en N4DS de nouvelles informations, qui n'existaient pas en DADS-U et ne sont donc pas transmises aux applications. Cela peut permettre de prendre en compte la diversité des situations présentes dans la source en amont. On dispose par exemple du code population de salariés, qui permet d'isoler certains cas particuliers et offre une approche plus pertinente que la distinction par catégorie juridique réalisée dans le partage des périodes entre les applications ;
- Les contrôles des rémunérations pourraient tous être basés sur les données de l'année N, indépendamment de la validité N-1, contrairement à ce que fait l'application DADS. Ainsi, on évite de perpétuer des éventuelles erreurs.

2.6.2. Inconvénients

- En utilisant la source en amont de tous les redressements pour réaliser les contrôles, on se base sur des informations déclarées, qui ne sont pas toujours fiables. On ne peut pas non plus utiliser directement en amont les méthodes des applications, puisqu'elles sont basées sur des informations redressées (notamment les PCS et le statut), et on doit reconstruire en amont des filtres utilisés dans les applications pour exclure certaines périodes de l'analyse (périodes inexploitées, etc.).

---

[42] La variable explicative la plus fréquemment redressée est la PCS : la PCS à quatre positions est modifiée pour un tiers des périodes dans l'application DADS. Mais elle l'est moins sur le premier chiffre (19 %), qui est le niveau que l'on utilise dans notre étude.

[43] Voir Annexes §1.3.

[44] Voir §5.

- Les interventions manuelles des gestionnaires sont réalisées via les applications DADS et Siasp. Au niveau du Frontal, on ne pourrait donc envisager que des redressements automatiques. Ceci n'est toutefois pas un inconvénient pour les déclarations qui ne sont de toute façon pas traitées par les gestionnaires (petites déclarations).
- En travaillant en amont, on utilise une méthode homogène sur l'ensemble de la source DADS, mais on ne peut pas prendre en compte les autres sources. Celles-ci pèsent peu dans l'application DADS, mais, côté Siasp, elles concernent la quasi-totalité de la FPE.

2.7. Points de vigilance

- La masse des données (60,5 millions de périodes initiales) est une forte contrainte.
- Pour qu'elle puisse trouver une application concrète, la méthode proposée doit pouvoir être compatible avec la future déclaration sociale nominative (DSN), qui doit remplacer la source DADS à l'horizon 2016.

3. Étape préliminaire : apurement et première détection de périodes aberrantes

3.1. Les éléments à contrôler

Point préalable et convention :
On ne s'intéresse pas, dans le cadre de cette étude, au cas des élus. D'une part, les rémunérations des élus ne correspondent pas à des salaires et d'autre part, leur temps de travail n'est pas renseigné. Leur rémunération est conservée pour diffusion dans l'application Siasp, mais sans être traitée.
Par convention, les points ou périodes détectés hors norme à l'issue de l'étape préliminaire (§3) seront appelés points ou périodes aberrant(e)s, les points ou périodes détectés hors norme à l'issue des différentes modélisations proposées (§4) seront appelés points ou périodes atypiques.

Dans les contrôles des rémunérations, la variable d'intérêt est, *in fine*, **le salaire horaire**. Il est calculé comme dans les applications, par le salaire net rapporté aux heures rémunérées :

SALAIRE_HORAIRE = SALAIRE_NET / HEURES_REMUNEREES

Il correspond donc au salaire horaire net. Pour pouvoir le calculer et s'assurer de sa cohérence, il faut vérifier les informations sur les **rémunérations**, au numérateur, et sur le **volume de travail**[45], au dénominateur.

Dans cette étape initiale d'apurement, on réalise donc **trois phases de contrôles** :
- contrôles de cohérence globale des variables financières et du volume de travail ;
- contrôles de cohérence interne des variables financières ;
- contrôles de cohérence globale du salaire horaire.

Les variables financières d'intérêt sont, *in fine*, dans les données statistiques, le salaire net et le salaire brut[46]. Elles sont construites différemment et de façon complexe dans les deux applications. Dans Siasp, par exemple, leur construction nécessite l'utilisation de 78 variables composites, utilisant des informations DADS-U ou bien des paramètres externes. Au niveau du Frontal, calculer ces variables n'a pas de sens, car il faudrait également contrôler toutes les variables entrant dans leur construction, et réaliser des traitements cohérents sur les champs DADS et Siasp. On contrôlera donc la base CSG et le net fiscal, qui sont les deux composantes principales des salaires net et brut.

Le volume de travail peut être appréhendé par deux notions dans les DADS :

[45] Voir Définitions, p. 38.

[46] Le salaire net et le salaire brut sont les variables de référence, voir §2.1 et Définitions, p. 38.

- Le nombre d'heures travaillées correspond *a priori* le mieux à la notion de volume de travail pour l'analyse du salaire horaire. Cependant, cette variable n'est en pratique pas renseignée pour certains types de salariés, comme les salariés au forfait jours, les journalistes pigistes, les travailleurs à domicile, les fonctionnaires… Ainsi, 20 % des périodes, dont quasiment toutes les périodes du code population des fonctionnaires, ont un nombre d'heures travaillées manquant[47]. On ne peut donc pas retenir cette notion pour appréhender le volume de travail, seulement pour le contrôler. En effet, le nombre d'heures travaillées est nul pour les personnes n'ayant pas travaillé pendant la période. On utilisera donc cette information pour exclure ces périodes de l'analyse ;
- Le nombre d'heures rémunérées[48] est obligatoire pour tous les types de salariés, en dehors des agents ou salariés au forfait jour[49], on l'utilisera donc à la fois pour les contrôles du volume de travail et pour le calcul du salaire horaire.

3.2. Des périodes particulières à traiter séparément

Les DADS étant une source administrative, les informations qui y sont déclarées ne sont pas toujours adaptées à nos besoins statistiques. Ainsi, **de nombreuses périodes décrivent des situations non-standard, qui font bien sens pour les organismes destinataires mais qui ne correspondent pas aux concepts statistiques que l'on souhaite mesurer. Elles doivent donc subir des traitements statistiques spécifiques.** Par exemple, l'Insee reçoit les déclarations des caisses de congés payés et de Pôle Emploi (en partie) pour les bénéficiaires de leurs allocations. **Ces périodes ne correspondent pas à des périodes de travail et ne doivent pas être comptabilisées dans l'emploi total : il est donc normal de les isoler.** Elles sont traitées à part dans l'application DADS. De même, pour certains apprentis et stagiaires, la base CSG peut être nulle. Un net fiscal nul est un cas plus courant, pouvant correspondre à des cas particuliers de rémunérations, d'épargne salariale ou bien des primes de licenciement. Certaines assistantes maternelles bénéficient, par exemple, d'exonérations fiscales particulières, ce qui génère des incohérences entre les variables financières, sans que cela soit une erreur. Ces cas particuliers doivent être traités à part, en tant que périodes non-standard.

Cette étape initiale d'apurement permet d'identifier, en plus des erreurs de déclarations qui doivent être corrigées par les applications, **une première partie de ces périodes particulières, qui sont exclues de la suite de notre étude**. **Une fois la liste des périodes identifiées comme particulières dans cette étape initiale d'apurement transmise aux applications DADS et Siasp, il reste à leur charge de distinguer :**

- les **périodes particulières à isoler** (associées à une prime de licenciement, par exemple) ;
- les **périodes particulières « normales »** (exonérations de certaines assistantes maternelles, par exemple) ;
- les **périodes particulières ou aberrantes à redresser**.

En effet, l'analyse et le traitement de ces cas particuliers nécessitent la récupération et/ou le redressement de variables explicatives (APET[50], PCS, statut…)[51] et doivent donc être réalisés au niveau des applications DADS et Siasp.

3.3. Filtre sur les heures et les variables financières

Les périodes ayant :

- des variables financières manquantes[52], négatives ou nulles,

47 Voir Annexes §2.1.

48 A noter qu'elle est bien souvent égale aux heures travaillées, sans que cela soit toujours à raison.

49 Elle ne concerne pas non plus les salariés travaillant à la pige ou les cachets isolés, par exemple. Elle sert à l'ouverture des droits à l'assurance maladie et au calcul de la prime pour l'emploi, contrairement au nombre d'heures travaillées, utilisé pour le calcul du temps d'exposition au risque d'accident du travail.

50 Activité Principale de l'Etablissement.

51 Voir Annexes §1.1.

52 Les variables financières ne peuvent être manquantes en théorie, puisqu'elles correspondent à des obligations sociales ou fiscales. En pratique, elles le sont pour 40 000 périodes sur la validité 2011, qui sont ici filtrées.

- et/ou des heures travaillées nulles,
- et/ou des heures rémunérées manquantes, négatives ou nulles

doivent être directement considérées comme aberrantes.

Il convient également de fixer une borne supérieure aux heures rémunérées :

- dans Siasp : les heures sont systématiquement plafonnées à 2 028 heures ;
- dans DADS : les heures sont systématiquement plafonnées à 2 200 heures ou 2 500 heures selon l'APET.

On se place donc ici dans le cadre général, en considérant comme aberrantes les périodes pour lesquelles les heures rémunérées dépassent 2 500 heures.

Heures : 12,2 millions de périodes sont détectées comme aberrantes sur les heures, sur 60,3 millions, soit 20,3 % des périodes[53].

- Le filtre sur les heures travaillées nulles conduit surtout à considérer comme aberrantes des périodes des salariés du privé « classiques », avec 12 % des périodes concernées.
- Le filtre sur les heures rémunérées manquantes conduit à considérer comme aberrantes 8,5 % de périodes. La plupart des heures rémunérées manquantes correspondent à des périodes de salariés du privé, et plus fréquemment les salariés du privé travaillant dans des organismes de droit public, et les fonctionnaires détachés.
- Les autres filtres sur les heures rémunérées détectent 10,2 % de périodes aberrantes, dont la plupart ont des heures rémunérées nulles.

Variables financières : 7,7 millions de périodes sont aberrantes sur les variables financières, soit 12,7 % des périodes[54].

Au total, 15,5 millions de périodes sont détectées comme aberrantes par ce filtre sur les heures et sur les variables financières, sur 60,3 millions, soit 25,8 % des périodes.
Ces périodes ne sont pas conservées dans la suite de notre étude, pour la détection des périodes atypiques, mais elles sont, bien sûr, conservées dans les applications DADS et Siasp pour y être traitées.

3.4. Cohérence interne des variables financières

Les variables financières doivent être cohérentes entre elles : il faut que le salaire net soit inférieur au salaire brut.

Une fois mises à part, dans la première phase d'apurement (filtre sur les heures et les variables financières), toutes les périodes avec un net fiscal et/ou une base CSG inexploitables, on se place dans le cadre général de la définition du salaire net et du salaire brut dans les applications, calculés à partir du net fiscal et de la base CSG. On réalise une approximation supplémentaire sur le salaire net en ne prenant pas en compte les rémunérations des heures supplémentaires exonérées, afin de ne pas réaliser davantage de contrôles. En effet, ces rémunérations sont issues d'informations à consolider, et ne concernent que peu de salariés. Cette mesure n'existe d'ailleurs plus en 2013. Ainsi, on calcule les salaires net et brut comme :

SALAIRE_NET = NET_FISCAL - CSG non déductible - CRDS

SALAIRE_BRUT = (BASE_CSG / part du salaire brut soumis à CSG)

La CSG non déductible s'obtient par :
CSG non déductible = BASE_CSG * TAUX_CSG * (1 - PART_CSG_DEDUCTIBLE)

La CRDS s'obtient par :
CRDS = BASE_CRDS * TAUX_CRDS

[53] Voir Annexes §2.2.
[54] Voir Annexes §2.2.

En 2011, le taux de CSG est de 7,5 %, la part de CSG déductible est de 68 %[55] et le taux de CRDS est de 0,5 %.

On peut alors vérifier la cohérence interne du salaire net et du salaire brut, en conservant une marge d'erreur à 20 %[56] : seront donc considérées directement comme aberrantes (et donc filtrées) les périodes pour lesquelles le salaire net est supérieur de plus de 20 % au salaire brut.

7,2 % des périodes ont un salaire net et un salaire brut incohérents. Cette proportion est plus importante pour les fonctionnaires, détachés ou non, et pour les salariés du spectacle.
La moitié des périodes filtrées parmi les fonctionnaires concernent des fonctionnaires d'État, pour lesquels les DADS ne sont actuellement pas utilisées et donc sur la qualité desquelles on ne dispose d'aucun élément. Pour les salariés du spectacle et les fonctionnaires détachés, également davantage concernés par des incohérences internes du salaire net et du salaire brut, de nombreuses périodes aberrantes ont une base CSG inférieure à 1 000 € et correspondent donc à de l'activité annexe, contribuant de manière négligeable à l'appareil productif, et traitées à part dans les applications. Il est donc important qu'elles soient également isolées dans l'analyse au niveau du Frontal. Parmi les périodes filtrées, il y a 16 000 cas de salaire net négatifs.

Tableau 2 : Nombre et pourcentage de périodes avec un salaire net supérieur au salaire brut ou supérieur de plus de 20 % au salaire brut, selon le code population

code population	total	salaire net > salaire brut	%	salaire net > 1,2* salaire brut	%
10 - salarié privé classique	35 582 201	3 177 260	8,9	2 170 915	6,1
11 - salarié du spectacle	851 375	361 795	42,5	131 759	15,5
13 - salarié privé dans le public	317 547	32 941	10,4	21 345	6,7
14 - fonctionnaire détaché	13 166	6 076	46,1	4 269	32,4
40 - fonctionnaire	4 845 599	2 048 035	42,3	639 110	13,2
43 - non-titulaire de la fonction publique	3 138 386	283 205	9,0	232 951	7,4
Total	44 748 274	5 909 312	13,2	3 200 349	7,2

3.5. Filtre sur le salaire horaire

Comme dans les applications DADS et Siasp, une borne inférieure est fixée, pour le salaire horaire, à 80 % du SMIC horaire. En 2011, le SMIC horaire net est de 9,19 €, ce qui correspond à une borne inférieure du salaire horaire net de 7,22 €. Ce dernier filtre conduit à qualifier 6,2 % des périodes d'aberrantes.

- Les agents non-titulaires de la fonction publique : les deux tiers des périodes ainsi filtrées parmi les non-titulaires de la fonction publique concernent la fonction publique d'État, pour laquelle les DADS ne sont actuellement pas utilisées et donc sur la qualité de laquelle on ne dispose d'aucun élément ;
- Les salariés privés dans le public : les périodes filtrées ont quasiment toute une base CSG inférieure à 1 000 € et correspondent donc à de l'activité annexe, traitée à part dans les applications.

Tableau 3 : Nombre et pourcentage de périodes avec un salaire net inférieur à 80% du SMIC, selon le code population

code population	total	Salaire horaire net < SMIC horaire net * 80 %	%
10 - salarié privé classique	33 411 286	1 967 658	5,9%
11 - salarié du spectacle	719 616	26 486	3,7%
13 - salarié privé dans le public	296 202	29 461	9,9%
14 - fonctionnaire détaché	8 897	247	2,8%
40 - fonctionnaire	4 206 489	21 454	0,5%
43 - non-titulaire de la fonction publique	2 905 435	517 962	17,8%
Total	41 547 925	2 563 268	6,2%

[55] Ainsi, le taux de CSG déductible est égal à : TAUX_CSG_ DEDUCTIBLE = TAUX_CSG * PART_CSG_DEDUCTIBLE = 5,1 %. Voir également Annexes §1.2.

[56] Une marge d'erreur de 5 % ou de 10 % conduit à éliminer beaucoup de périodes, particulièrement des fonctionnaires, détachés ou non, et des salariés du spectacle. Les approximations réalisées sur le calcul du salaire brut et du salaire net obligent à conserver une marge d'erreur importante.

3.6. Bilan

Sur l'ensemble des codes populations (hors élus), on conserve 64,7 % des périodes. Cette proportion est plus importante pour les salariés du privé et du spectacle.
Par contre, quasiment deux tiers des périodes des fonctionnaires détachés sont filtrées (il s'agit d'une population très particulière, pour laquelle les salariés sont souvent au forfait jour). 57,1 % des périodes des salariés de la fonction publique sont également filtrées, en particulier dans la fonction publique d'État, sur laquelle nous n'avons pas le recul nécessaire pour juger de la qualité ou de la cohérence des informations déclarées et qui n'est, de toutes façons, pas conservée dans Siasp.

Tableau 4 : Nombre de périodes avant et après étape préliminaire et pourcentage de périodes conservées dans la suite de l'étude, selon le code population

code population	Nombre initial de périodes	Nombre de périodes OK	%
10 - salarié privé classique	47 077 646	31 443 628	66,8%
11 - salarié du spectacle	1 033 440	693 130	67,1%
13 - salarié privé dans le public	462 682	266 741	57,7%
14 - fonctionnaire détaché	25 337	8 650	34,1%
40 - fonctionnaire	7 333 920	4 185 035	57,1%
43 - non-titulaire de la fonction publique	4 349 138	2 387 473	54,9%
total	60 282 163	38 984 657	64,7%

Si certains cas d'incohérences correspondent à des erreurs de déclaration, **la plupart des périodes filtrées dans cette première phase d'apurement correspond à des cas particuliers qui sont actuellement traités à part dans les applications**, qu'il s'agisse de périodes non travaillées à isoler (bénéficiaires de congés payés, par exemple), de cas d'incohérences « normales » à vérifier en fonction des caractéristiques des salariés avant d'être validées, de salariés au forfait jour pour lesquels on ne peut pas calculer de salaire horaire... **Il est donc normal que ces périodes soient également exclues de l'analyse au niveau du Frontal.**

46,4 % des périodes à conserver sont issues de petites et moyennes déclarations à destination de l'application DADS, contre une proportion initiale de 37,6 %. La phase initiale d'apurement conduit donc à se concentrer, pour la détection des rémunérations horaires atypiques, sur les périodes pour lesquelles l'apport d'une mutualisation au niveau du Frontal est le plus important, car non traitées par des gestionnaires.

Tableau 5 : Nombre et répartition des périodes après étape préliminaire, selon l'application destinataire, la taille de la déclaration et le code population

Code population	Siasp - ensemble		DADS						Total	
			grandes déclarations		déclarations moyennes		petites déclarations			
	nombre	%	nombre	%	nombre	%	nombre	%	nombre	%
10 - salarié privé classique	196 163	2,8%	13 700 543	97,9%	13 532 856	96,5%	4 014 066	99,1%	31 443 628	80,7%
11 - salarié du spectacle	11 155	0,2%	185 558	1,3%	460 537	3,3%	35 880	0,9%	693 130	1,8%
13 - salarié privé dans le public	260 657	3,8%	1 507	0,0%	3 832	0,0%	745	0,0%	266 741	0,7%
14 - fonctionnaire détaché	256	0,0%	5 129	0,0%	3 038	0,0%	227	0,0%	8 650	0,0%
40 - fonctionnaire	4 083 370	59,0%	86 797	0,6%	14 809	0,1%	59	0,0%	4 185 035	10,7%
43 - non-titulaire de la fonction publique	2 369 267	34,2%	15 059	0,1%	2 723	0,0%	424	0,0%	2 387 473	6,1%
Total	6 920 868	100,0%	13 994 593	100,0%	14 017 795	100,0%	4 051 401	100,0%	38 984 657	100,0%
% du total		17,7%		35,9%		36,0%		10,4%		100,0%

4. Méthodes de détermination des points atypiques

4.1. Méthode similaire aux méthodes utilisées dans les applications : la régression par les moindres carrés ordinaires

4.1.1. Notations

Soit $X = (x_{i,j})$ une matrice $n \times p$ de variables explicatives, $y = (y_1,..., y_n)'$ la variable d'intérêt (le salaire horaire dans notre analyse) avec n observations, et $\theta = (\theta_1,..., \theta_p)'$ un vecteur inconnu de p paramètres à estimer. Le modèle linéaire habituel est :

$y_i = \sum_{j=1}^{p} \theta_j x_{ij} + e_i$, pour $i = 1,..., n$, soit : $y = X\theta + e$, où $e = (e_1,..., e_n)'$ est le vecteur des erreurs.

L'hypothèse est faite que, pour une matrice X donnée, les e_i de e sont indépendants et identiquement distribués selon la loi de distribution $L(./\sigma)$, où σ est l'écart-type, paramètre habituellement inconnu. Souvent, $L(./\sigma) = \phi(.)$, la distribution de la loi normale centrée-réduite de densité $\phi(s) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-s^2}{2})$.

L'estimateur des moindres carrées ordinaires (MCO) ou ordinary least squares (OLS) $\hat{\theta}_{LS}$ de θ s'obtient comme la solution de : $\min_{\theta} \sum_{i=1}^{n} r_i^2$,

où $r = (r_1,..., r_n)'$ est le vecteur n des résidus pour une valeur donnée de θ et pour une matrice X .

Si X est de rang p, la solution est : $\hat{\theta}_{LS} = (X'X)^{-1} X'y$.

4.1.2. Spécification de la modélisation OLS du salaire horaire

On fait confiance à la codification du code population qui est, par construction, une information structurante pour l'ensemble de la déclaration des variables auxiliaires. On retiendra les variables explicatives suivantes :

- Pour tous les codes populations :
  - o des caractéristiques de temps de travail : durée de la période et quotité ;
  - o une caractéristique d'activité : le fait de travailler ou non en Île-de-France ;
  - o des caractéristiques du salarié : le sexe et l'âge.
- Pour chaque code population, on dispose également d'informations auxiliaires spécifiques pouvant expliquer le salaire horaire :

Tableau 6 : Informations auxiliaires utilisées selon le code population

Type		privé				public	
Code population concerné		10 - salarié privé classique	11 - salarié du spectacle	13 - salarié privé dans le public	14 - fonctionnaire détaché	40 - fonctionnaire	43 - non-titulaire de la fonction publique
Premier chiffre de la PCS	1 Agriculteurs exploitants	**oui**	non	non	non	non	non
	2 Artisans, commerçants et chefs d'entreprise	**oui**	non	non	non	non	non
	3 Cadres et professions intellectuelles supérieures	**oui**	**oui**	non	**oui**	**oui**	**oui**
	4 Professions Intermédiaires	**oui**	**oui**	**oui**	**oui**	**oui**	**oui**
	5 Employés	**oui**	**oui**	**oui**	**oui**	**oui**	**oui**
	6 Ouvriers	**oui**	non	non	**oui**	**oui**	**oui**
	9 Non renseigné	**oui**	**oui**	**oui**	**oui**	**oui**	**oui**

Type		privé				public	
Code population concerné		10 - salarié privé classique	11 - salarié du spectacle	13 - salarié privé dans le public	14 - fonctionnaire détaché	40 - fonctionnaire	43 - non-titulaire de la fonction publique
Fonction publique d'appartenance	FPT	non	non	**oui**	non	**oui**	**oui**
	FPH	non	non	**oui**	non	**oui**	**oui**
	FPE	non	non	**oui**	non	**oui**	**oui**
Catégorie	A	non	non	non	non	**oui**	**oui**
	B	non	non	non	non	**oui**	**oui**
	C	non	non	non	non	**oui**	**oui**
Type de contrat	contractuels	non	non	non	non	non	**oui**
	contrats aidés	non	non	non	non	non	**oui**
	autres	non	non	non	non	non	**oui**

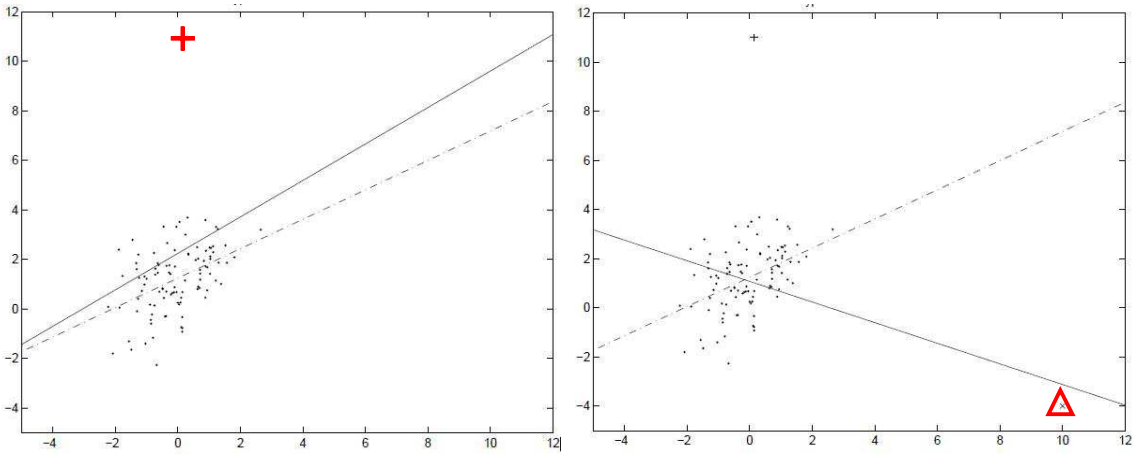
4.1.3. Problématique des points atypiques

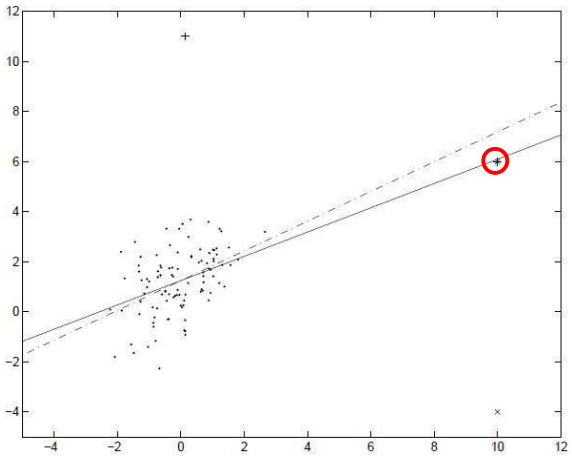
Les méthodes d'estimation classiques sont extrêmement vulnérables aux valeurs aberrantes (ou outliers). Il est donc nécessaire d'utiliser des méthodes d'estimation robustes dès lors que les données présentent des individus atypiques.

Il existe trois types d'individus atypiques :

- les points verticaux (points aberrants sur la variable y) : dans notre étude, les footballeurs et les traders en sont des exemples. Il s'agit pour ces exemples de points atypiques normaux, qui ne doivent pas être redressés par les applications ;
- les mauvais points leviers (points aberrants sur les X) : dans notre étude, les enfants salariés, qui reçoivent une rémunération pour des publicités, en sont un exemple. De la même façon, ces exemples correspondent à des points atypiques normaux, qui ne doivent pas être redressés par les applications ;
- les bons points leviers (points aberrants sur la variable y et sur les X).

Schéma 2 : Les trois types d'individus atypiques





Légende :

+ Point vertical

△ Mauvais point levier

○ Bon point levier

Les points verticaux et mauvais points leviers ont pour conséquence de biaiser les estimations. Les bons points leviers, par contre, améliorent la précision des paramètres de la régression (Rousseeuw et Van Zomeren (1990))[57]. Ils ne constituent pas le cœur de notre problème.

Dans notre analyse, étant données les particularités des données (points leviers, points verticaux…) et compte-tenu du peu d'éléments explicatifs disponibles et du biais déclaratif sur certains d'entre eux, on soupçonne le modèle classique OLS de ne pas être robuste. On cherche donc tout d'abord à vérifier cette hypothèse.

4.1.4. Influence des observations sur les paramètres du modèle OLS

4.1.4.1. Tests sur échantillons

On commence par s'intéresser au code population le plus fréquent et recouvrant peu de cas particuliers : le code population 10, des salariés privés classiques. Étant donné le nombre de périodes de ce code population, on teste l'hypothèse de robustesse du modèle OLS sur 50 échantillons de 100 000 périodes[58].

- Tests préalables

Quand on conserve toutes les observations des échantillons, les estimateurs des moindres carrés ont, en valeur absolue, des valeurs très élevées[59], ce qui suggère une faible robustesse du modèle. En supprimant du modèle les observations strictement supérieures au 99^{e} percentile du salaire horaire, les paramètres estimés sont nettement plus faibles, en valeur absolue[60]. Cependant, il faut vérifier la présence d'observations atypiques et influentes sur le modèle pour s'assurer que la méthode des OLS peut être utilisée.

- Statistique DFFITS d'influence des observations sur les paramètres du modèle OLS

On peut mesurer l'influence des observations sur les valeurs prédites du modèle grâce à la statistique suivante :

$$DFFITS(i) = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{(i)}}},$$

où $h_i = x_i (X'X)^{-1} x_i'$, $y_{(i)}$ correspond à la valeur de y en i estimée sans l'observation i , $s_{(i)}$ correspond à la valeur de s en i estimée sans l'observation i et $h_{(i)}$ correspond à la valeur de h en i estimée sans l'observation i .

---

[57] Voir Bibliographie [8], régression robuste.

[58] Il y a peu d'agriculteurs dans la population et donc dans les échantillons. On les agrège donc avec la catégorie « non renseignée ».

[59] Voir Annexes §3.1.

[60] Voir Annexes §3.1.

Cette statistique mesure la différence relative de valeur prédite pour la i ème observation avec toutes les observations et sans la i ème observation. De fortes valeurs indiquent la présence d'observations influentes et donc un risque de non-robustessee du modèle. Belsley, Kuh et Welsch (1980) préconisent alors l'utilisation d'un seuil ajusté défini par : $2\sqrt{p/n}$. Ainsi, si $DFFITS(i) > 2\sqrt{p/n}$, on considère l'observation i comme influente.

Sur les 50 échantillons[61], on compte en moyenne 5,2% d'observations influentes, chiffre très peu variable selon les échantillons. Compte-tenu de cette proportion d'observations influentes, il y a donc un risque important de non-robustesse du modèle OLS, et ce même en se restreignant aux observations inférieures au 99e percentile de salaire horaire.

4.1.4.2. Tests sur la population

Sur l'ensemble des salariés privés classiques, 4,9 % des observations sont influentes au sens de la statistique DFFITS. Pour les autres codes population, la part des observations influentes est variable : elle est du même ordre de grandeur que pour les salariés privés classiques pour les salariés du privé dans le public et pour les non-titulaires de la fonction publique. Par contre, seuls 2,6 % des périodes des salariés du spectacle sont détectées comme influentes.

Tableau 7 : 99e percentile du salaire horaire, nombre de périodes inférieures au 99e percentile du salaire horaire, nombre et pourcentage de périodes influentes, selon le code population

Code population	P99 du code population	Nombre d'observations utilisées dans le modèle (salaire horaire inférieur au 99e percentile du code population)	Nombre d'observations influentes	%
10 - salarié privé classique	28,0	31 129 192	1 525 434	4,9%
11 - salarié du spectacle	56,0	686 199	17 542	2,6%
13 - salarié privé dans le public	22,0	264 074	12 640	4,8%
14 - fonctionnaire détaché	86,7	8 564	312	3,6%
40 - fonctionnaire	23,1	4 143 185	157 289	3,8%
43 - non-titulaire de la fonction publique	40,1	2 363 599	111 220	4,7%
Total		38 594 813	1 824 437	4,7%

Au vu de ces critères, il convient donc de tester l'apport des méthodes robustes sur les OLS, pour chaque code population.

4.2. Théorie sur les méthodes d'estimations robustes

Les méthodes d'estimation classiques sont extrêmement vulnérables aux valeurs atypiques (outliers). Il est donc nécessaire d'utiliser des méthodes d'estimation robustes dès lors que les données présentent des individus atypiques. De nombreuses méthodes ont été développées pour traiter ce problème, les plus courantes étant la M-estimation, l'estimation « high breakdown value » ou de « point de rupture », ou encore la combinaison de ces deux méthodes[62] :

- La M-estimation, introduite par Huber en 1973, est l'approche la plus simple, sur les plans théorique et calculatoire. Bien qu'elle ne soit pas robuste aux points leviers, elle est largement utilisée lorsque les données sont supposées comporter des points aberrants sur la variable d'intérêt y ;

- La LTS-estimation (Least Trimmed Squares) est une méthode de « point de rupture », introduite par Rousseeuw (1984). Le point de rupture correspond à la proportion des données qui peuvent être arbitrairement changées sans changer arbitrairement la valeur de l'estimateur[63] ;

61 Voir Annexes §3.2.

62 Les méthodes citées ci-dessous correspondent aux méthodes proposées par le logiciel SAS et la procédure ROBUSTREG. Voir Bibilographie [4] et [8], procédure ROBUSTREG. L'algorithme de calcul via la procédure ROBUSTREG de SAS est décrit dans les Annexes §4.1.

63 Il existe également une autre méthode de « point de rupture », la S-estimation, introduite par Rousseeuw et Yohai (1984) et statistiquement plus efficiente que la LTS-estimation.

- La MM-estimation, introduite par Yohai (1987), combine M-estimation et méthode de « point de rupture ». Elle possède les propriétés des méthodes de « point de rupture », tout en étant statistiquement plus efficiente.

4.2.1. M-estimation

Au lieu de minimiser la somme des carrés des résidus, comme pour l'estimateur des moindres carrés, le M-estimateur[64] $\hat{\theta}_M$ de θ s'obtient comme la solution de[65] :

$$\min_{\theta} Q(\theta) \text{, où } Q(\theta) = \sum_{i=1}^{n} \rho(\frac{r_i}{\sigma}) ,$$

avec ρ une fonction positive, symétrique, et nulle en 0, appelée fonction objectif, et σ un paramètre d'échelle mesurant la dispersion des y_i.

L'estimateur est alors défini par sa fonction objectif ou sa fonction de score $\psi = \rho'$. Ainsi, les OLS sont un cas particulier de M-estimation, avec $\rho(x) = x^2/2$ et $\psi(x) = x$. Cependant, la forme carrée de sa fonction objectif le rend très sensible aux valeurs extrêmes : les OLS ne sont donc pas robustes à ce type de valeurs. Mais un choix adéquat de ρ peut permettre de limiter l'influence des valeurs extrêmes sur les estimations et ainsi les rendre robustes. La robustesse de la M-estimation découle donc directement de la forme de la fonction ρ.

Ainsi, le choix de la fonction ρ constitue le cœur du problème : la fonction objectif doit permettre des estimations à la fois robustes et proches de l'efficacité sur les distributions étudiées. Il existe dans la littérature de nombreuses fonctions objectif. Pour les distributions à queues épaisses, il convient de choisir une fonction objectif permettant une estimation des plus robustes : on choisira donc une fonction de score dite « redescendante », c'est-à-dire qui tend vers zéro à l'infini ou s'annule à partir d'une certaine distance à l'origine.

4.2.2. LTS-estimation

Cette méthode, proposée par Rousseeuw (1984) est une méthode très robuste. Au lieu de minimiser la somme des carrés des résidus, comme pour l'estimateur des moindres carrés, le LTS-estimateur $\hat{\theta}_{LTS}$ de θ consiste à minimiser la somme des carrés des h [66] plus petits résidus[67].

Ainsi, $\hat{\theta}_{LTS}$ s'obtient comme la solution de :

$$\min_{\theta} Q_{LTS}(\theta) \text{, où } Q_{LTS}(\theta) = \sum_{i=1}^{h} r_{(i)}^2 ,$$

avec $r_{(1)}^2 \le r_{(2)}^2 \le ... \le r_{(n)}^2$ les résidus au carré ordonnés.

L'estimation qui en découle est donc plus robuste que les OLS, puisque les observations ayant les résidus les plus élevés n'entrent pas en jeu dans la minimisation et donc n'affectent pas les estimations.

Le paramètre h permet de définir le point de rupture[68] (ou breakdown value) de l'estimateur (égal à $(n-h)/n$) et de réaliser ainsi le compromis entre robustesse (point de rupture maximal de 50 %, obtenu pour $h = n/2+1$) et efficience (pour $h = n$, $\hat{\theta}_{LTS} = \hat{\theta}_{LS}$, l'estimateur est efficace mais non robuste aux observations atypiques).

---

64 Voir Bibliographie [1].

65 Voir Bibliographie [9].

66 En pratique, $n/2+1 \le h \le (3n+p+1)/4$.

67 Voir Bibliographie [1].

68 Voir introduction du §4.2.

Le nombre total de sous-ensembles de taille h parmi n observations étant de C_n^h, une recherche exhaustive est en général impossible, et la résolution de l'équation $\min_{\theta} Q_{LTS}(\theta)$ s'effectue alors de façon numérique via des algorithmes itératifs. L'algorithme de calcul de LTS-estimation utilisé par la procédure ROBUSTREG de SAS est l'algorithme FAST-LTS (Rousseeuw et Van Drissen (2000))[69].

4.2.3. MM-estimation

La MM-estimation[70] est une combinaison entre méthode de point de rupture et M-estimation, introduite par Yohai (1987), l'idée étant de bénéficier simultanément de la robustesse des méthodes de point de rupture et de l'efficacité de la M-estimation. Elle se déroule en trois étapes[71] :

- dans un premier temps, un estimateur de point de rupture est utilisé pour déterminer une première estimation robuste du paramètre d'intérêt θ et en déduire des résidus associés ;
- dans un second temps, ces résidus servent de base au calcul d'une estimation robuste par M-estimation du paramètre d'échelle σ mesurant la dispersion des y_i ;
- enfin, une seconde M-estimation, utilisant le paramètre d'échelle estimé à l'étape précédente et une nouvelle fonction objectif, inférieure en tout point à celle de l'étape précédente, permet d'obtenir l'estimation finale du paramètre d'intérêt θ.

Les deux premières étapes assurent un point de rupture élevé à l'estimateur final, ce qui le rend particulièrement robuste aux points verticaux et points leviers. La troisième étape et le choix de la fonction objectif associée garantit quant à elle l'efficience de l'estimateur. Ainsi, la MM-estimation permet de choisir l'efficience de l'estimateur indépendamment de sa robustesse, combinant les avantages de la méthode de point de rupture et de la M-estimation.

4.2.4. Type d'observations atypiques et méthodes préconisées

La méthode d'estimation robuste à privilégier dépend du type d'individus atypiques présents dans les données :

Tableau 8 : Type d'individus présents dans les données et méthodes d'estimation adaptées

Type d'individus présents dans les données	Méthode d'estimation adaptée		
	OLS	M	MM
standards	**oui**	**oui**	**oui**
points verticaux	**non**	**oui**	**oui**
mauvais points leviers	**non**	**non**	**oui**

Dans notre étude, les variables explicatives quantitatives sont bornées : l'âge (de 0 à 100 ans), la quotité (de 0 % à 100 %) et la durée (de 1 à 365 jours). L'impact de valeurs atypiques des valeurs explicatives sur l'estimation des paramètres est donc plus limité, ce qui devrait également limiter l'impact des mauvais points leviers. On commence donc par tester la M-estimation.

4.3. M-estimation et test de Hausman : la M-estimation est-elle préférable aux OLS ?

L'analyse des statistiques d'influence DFFITS suggère d'utiliser une méthode robuste plutôt que les OLS. La méthode robuste la plus courante est la M-estimation.

Le test de Hausman[72] permet de choisir la méthode à utiliser en fonction des données que l'on souhaite modéliser, et faire ainsi un compromis entre efficience et robustesse.

Le test se présente de la façon suivante :

---

[69] Voir Bibliographie [8], procédure QUANTREG.

[70] L'algorithme de calcul via la procédure ROBUSTREG de SAS est décrit en Annexes §4.2.

[71] Voir Bibliographie [19].

[72] Voir Bibliographie [7].

- hypothèse H0 : il n'y a pas de problème d'outliers ;
- hypothèse H1 : il y a un problème d'outliers.

Sous H0, l'estimateur des OLS et le M-estimateur sont convergents mais seul l'estimateur des OLS est efficient. Sous H1, le M-estimateur est convergent mais pas l'estimateur des OLS.

On pose $q = \hat{\beta}_M - \hat{\beta}_{LS}$.

Sous H0, comme $\hat{\beta}_{LS}$ est efficient, q et $\hat{\beta}_{LS}$ sont orthogonaux et donc $V(q)+V(\hat{\beta}_{LS}) = V(\hat{\beta}_M)$.

Soit $m = \hat{q}'(V(\hat{\beta}_M) - V(\hat{\beta}_{LS}))^{-1}\hat{q}$. Sous l'hypothèse nulle, m suit asymptotiquement une loi χ^2_p.

Si $m > \chi^2_{p,0.95}$, alors on rejette H0 et donc l'estimateur des OLS est biaisé par la présence d'outliers et la M-estimation est préférable. Il faut donc utiliser une méthode robuste.

Sur les 50 échantillons de 100 000 observations des salariés privés classiques[73], en restreignant l'analyse aux périodes pour lesquelles le salaire horaire est inférieur au 99e percentile[74], on a :

$$\min_{i=1,...,50} m_i = 10695 \text{ et } \chi^2_{11,0.95} = 19{,}7 .$$

Donc $m > \chi^2_{p,0.95}$: H0 est donc très largement rejetée. Le modèle OLS n'est pas adapté aux données, il faut utiliser une méthode robuste.

C'est également le cas pour les modèles complets sur les salariés privés classiques et les autres codes population :

Tableau 9 : Résultats du test de Hausmann, selon le code population

Code population	m	nombre de paramètres p du modèle	statistique du $\chi^2_{p,0.95}$ correspondante	Modèle préférable
10 - salarié privé classique	3 510 632	12	21,0	M
11 - salarié du spectacle	58 515	9	16,9	M
13 - salarié privé dans le public	99 165	10	18,3	M
14 - fonctionnaire détaché	1 052	10	18,3	M
40 - fonctionnaire	375 641	14	23,7	M
43 - non-titulaire de la fonction publique	488 131	16	26,3	M

Cependant, bien que nos variables explicatives quantitatives soient bornées, le logiciel SAS détecte des points leviers lors de la M-estimation, et suggère d'utiliser des méthodes robustes aux points leviers, comme les méthodes LTS et MM.

4.4. Prise en compte des points verticaux et points leviers

4.4.1. MM-estimation et variables qualitatives

Lorsqu'il y a des variables explicatives quantitatives et qualitatives, le modèle peut s'écrire de la façon suivante :

$$y_i = \sum_{j=1}^{p-q} \theta_j x_{ij} + \sum_{k=1}^{q} \gamma_k I_{ik} + e_i \text{, pour } i = 1,...,n \text{, soit : } y = \theta X + \gamma I + e ,$$

où il y a $p-q$ variables quantitatives X et q indicatrices I équivalentes aux variables catégorielles.

Le modèle OLS gère très bien les variables qualitatives, en traitant les indicatrices issues de la décomposition des variables catégorielles de la même façon que les variables quantitatives (Draper et

[73] Comme précédemment, les agriculteurs, peu nombreux dans les échantillons, sont agrégés avec la catégorie « non renseignée ».

[74] Voir Annexes §3.3.

Smith (1981), Hardy (1993))[75]. La méthode de M-estimation gère aussi les variables catégorielles, comme l'a montré l'analyse de Birch et Myers (1982), en résolvant le système d'équations avec indicatrices par l'algorithme itératif des MCO pondérés[76]. Cependant, les M-estimateurs ne sont pas robustes aux points leviers.

Les méthodes robustes aux points leviers que sont les LTS et MM-estimateurs, par exemple, ne peuvent cependant pas être appliquées aux variables explicatives qualitatives. En effet, ces méthodes sont basées sur un algorithme d'échantillonnage itératif de h observations[77], où $\frac{n+p+1}{2} \leq h \leq n$. Dans le cas de variables explicatives dont certaines sont des indicatrices, la plupart des échantillons de h observations ne seront pas de plein rang, ce qui empêche la résolution du système d'équations. Il n'est donc pas possible d'utiliser ces méthodes avec des variables qualitatives.

4.4.2. Méthodes préconisées en présence de variables explicatives quantitatives et qualitatives

4.4.2.1. Méthode RDL

Cette méthode, proposée par Hubert et Rousseeuw (1999) [78], fonctionne en deux étapes :

- dans un premier temps, on détecte les points leviers, sur les régresseurs quantitatifs ;
- dans un deuxième temps, on estime les paramètres du modèle en utilisant les régresseurs quantitatifs et qualitatifs et en sous-pondérant ces points leviers.

Cette deuxième étape doit permettre la détection des outliers sur la variable d'intérêt, une fois pris en compte les points leviers détectés dans la première étape.

Pour la première étape, on utilise l'estimateur MVE (minimum volume ellipsoid) (Rousseeuw 1985), dont l'objectif est la minimisation du volume d'une ellipse contenant un nombre donné (par défaut $\frac{n+p+1}{2}$) d'observations dans l'espace des X. Cet estimateur fournit une distance robuste RD définie par :

$$RD(x_i) = \sqrt{(x_i - T(X))C(X)^{-1}(x_i - T(X))^t},$$

où les composants de $x_i = (x_{i1},...,x_{ip})$ sont les régresseurs quantitatifs, $T(X)$ le centre de la plus petite ellipse contenant un nombre donné d'observations X (par défaut, $\frac{n+p+1}{2}$) et $C(X)$ la forme de l'ellipse.

Pour un grand nombre d'observations, cette distance robuste suit une loi du χ^2_p. Ainsi, les observations pour lesquelles la distance robuste est élevée par rapport à la distribution du χ^2_p correspondent à des points leviers.

Pour la 2e étape, après avoir identifié ces points leviers, on leur affecte un poids w_i défini par :

$$w_i = \min\left\{1, \frac{p}{RD(x_i)^2}\right\}, \text{ pour } i = 1,...,n$$

---

[75] Voir Bibliographie [16].

[76] L'algorithme est décrit en Annexes §4.1.

[77] Voir Bibliographie [17].

[78] Voir Bibliographie [16].

On utilise ensuite la méthode L1 (ou LAV pour Least absolute value)[79] pondérée par ces poids w_i pour estimer les paramètres du modèle, consistant à résoudre :

$$\min_{\theta,\gamma}\left\{\sum_{i=1}^{n} w_i \left| r_i(\theta,\gamma)\right|\right\},$$

où les r_i sont les résidus du modèle.

4.4.2.2. Méthode SLTS

Lorsqu'il y a des variables qualitatives parmi les variables explicatives, la méthode SLTS (Smoothed Least Trimmed Squares), proposée par Čížek[80] est une alternative à la méthode RDL. Elle peut être utilisée quel que soit le type des variables explicatives, et le SLTS-estimateur est censé être peu sensible aux problèmes de classement des variables qualitatives, contrairement à la méthode RDL, ce qui est intéressant dans notre étude. Elle permet également d'optimiser la robustesse du modèle testé. En revanche, l'algorithme (décrit ci-dessous) nécessite de nombreuses itérations pour réaliser cette optimisation, et donc demande à la fois du temps et d'importantes ressources en mémoire, ce qui n'est pas envisageable compte-tenu du nombre de données que l'on souhaite traiter.

Le SLTS-estimateur $\hat{\theta}_{SLTS}$ de θ s'obtient comme la solution de :

$$\min_{\theta,w} Q_{SLTS}(\theta,w) \text{ , où } Q_{SLTS}(\theta,w)=\sum_{i=1}^{n} w_i r_{(i)}^2 \text{ ,}$$

avec $r_{(1)}^2 \le r_{(2)}^2 \le ... \le r_{(n)}^2$ les résidus au carré ordonnés et $w=(w_1,...,w_n)'$ un vecteur de poids tel que $w_1 \ge w_2 \ge ... \ge w_n \ge 0$.

La forme du SLTS-estimateur est proche de celle des moindres carrés pondérés (WLS), à la différence que les poids sont affectés aux carrés ordonnés des résidus et non pas aux carrés non ordonnés des résidus. Le comportement et les propriétés de l'estimateur dépendent donc entièrement du choix des poids. Par exemple :

- Si $w_1 = ... = w_n = 1$, alors le SLTS-estimateur est équivalent à l'estimateur OLS ;
- Si $w_1 = ... = w_h = \frac{n}{h}$ et $w_{h+1} = ... = w_n = 0$, alors le SLTS-estimateur est équivalent au LTS-estimateur.

L'enjeu est donc de choisir un jeu de poids ayant des propriétés intéressantes et permettant de réaliser un arbitrage optimal entre robustesse et efficacité. Čížek propose de retenir les poids définis par :

$$w_i = f_\lambda\left(\frac{2i-1}{2n},\omega\right) \text{ pour } i=1,...,n \text{ ,}$$

avec $f_\lambda(x,\omega)=\frac{1}{1+e^{\omega(x-\lambda)}}\Big/\int_0^1 \frac{1}{1+e^{\omega(x-\lambda)}}\,dx$ et $\lambda \in \left[\frac{1}{2};1\right]$ une constante fixée.

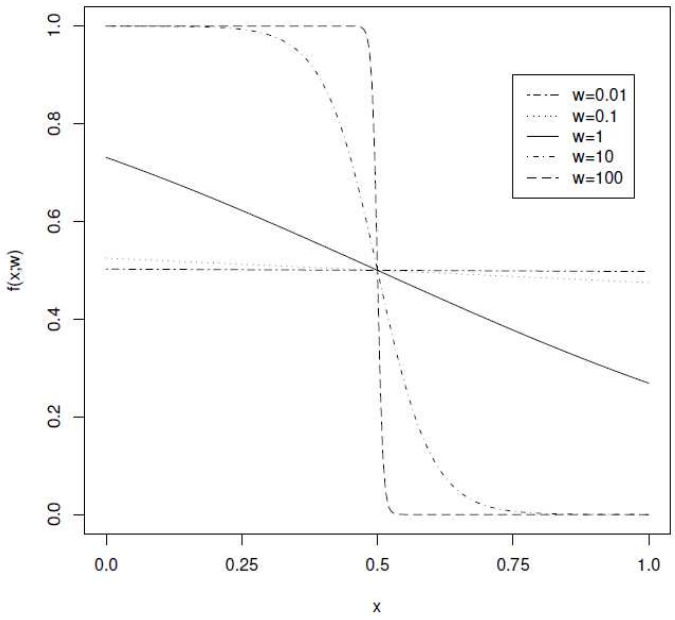
La forme de la fonction est la suivante pour $\lambda = \frac{1}{2}$, en fonction du paramètre ω :

---

[79] Cette méthode peut être utilisée à partir de la procédure QUANTREG de SAS, voir Bibliographie [5], [8] (procédure QUANTREG) et [15].

[80] Voir Bibliographie [6].

Graphique 1 : Forme de la fonction de poids proposée par Čížek



Ainsi :

- Lorsque $\omega \to \infty$, alors le SLTS-estimateur est équivalent au LTS-estimateur ;
- Lorsque $\omega \to 0$, alors le SLTS-estimateur est équivalent à l'estimateur OLS.

Le SLTS-estimateur est d'autant plus robuste que ω est grand. Le paramètre ω permet donc de jouer sur la robustesse et l'efficience du modèle.

Reste à effectuer l'estimation du SLTS-estimateur. Pour ce faire, **Čížek propose une procédure en deux étapes** (voir schéma p. 35) :

- dans un premier temps, un algorithme itératif – procédant par recherches successives, selon une stratégie proche de celle utilisée pour les estimations LTS – permet **d'estimer, pour un ω et une fonction de poids donnés, le SLTS-estimateur $\hat{\theta}_{SLTS}(\omega)$** ;
- dans un second temps, un second algorithme permet de **déterminer le paramètre ω de façon à réaliser le meilleur compromis possible entre robustesse et efficacité de l'estimation**[81].

Le premier algorithme d'estimation du SLTS-estimateur $\hat{\theta}_{SLTS}(\omega)$, à ω et fonction de poids fixés, se déroule en 8 étapes, pour un seuil K_S donné. L'objectif des 7 premières étapes est de déterminer un minimum local avec des poids définis par une permutation aléatoire Π donnée, la dernière étape vise à s'assurer, par itérations successives sur les premières étapes, qu'il s'agit bien d'un minimum global :

- étape 1 : on génère une permutation aléatoire $\Pi = (\pi_1, ..., \pi_n)'$ de $\{1, ..., n\}$;
- étape 2 : on définit des poids : $v = (v_1, ..., v_n)'$ tels que $v_i = w_{\pi_i}$;

[81] On peut également optimiser à la fois sur ω et λ, ce qui revient à utiliser un deuxième algorithme différent.

- étape 3 : on calcule l'estimateur des moindres carrés pondérés $\hat{\theta}^0$ avec les poids v et on fixe $k=0$, où k est le nombre d'itérations. On calcule $S_S(X, y, w; \hat{\theta}^0) = Q_{WLS}(\theta) = \sum_{i=1}^{n} v_i r_i^2(\hat{\theta}^0)$;
- étape 4 : on trie les résidus $r_i(\hat{\theta}^k)$, ce qui aboutit à une nouvelle permutation $\Pi = (\pi_1, ..., \pi_n)'$ telle que : $\left|r_{\pi_1}(\hat{\theta}^k)\right| \leq ... \leq \left|r_{\pi_{n1}}(\hat{\theta}^k)\right|$;
- étape 5 : on définit des poids : $v = (v_1, ..., v_n)'$ tels que $v_i = w_{\pi_i}$;
- étape 6 : on calcule l'estimateur des moindres carrés pondérés $\hat{\theta}^{k+1}$ avec les poids v. On calcule également : $S_S(X, y, w; \hat{\theta}^{k+1}) = \sum_{i=1}^{n} v_i r_i^2(\hat{\theta}^{k+1}) = Q_{SLTS}(\hat{\theta}^{k+1}, w)$;
- étape 7 : si $S_S(X, y, w; \hat{\theta}^k) > S_S(X, y, w; \hat{\theta}^{k+1})$, alors on fixe $k = k+1$ et on recommence à l'étape 4. Sinon, on va à l'étape 8 ;
- étape 8 : $S_S(X, y, w; \hat{\theta}^k) \leq S_S(X, y, w; \hat{\theta}^{k+1})$, alors on compte le nombre K de séquences de l'algorithme qui ont été réalisées sans améliorer le minimum global $S_S(X, y, w; \theta)$. On compare $S_{S,k,K} = S_S(X, y, w; \hat{\theta}^k)$ et les $S_{S,J}$, avec $J = 1, ..., K-1$, correspondant aux $S_{S,k_J,J} = S_S(X, y, w; \hat{\theta}^{k_J})$ finaux, obtenus lors des séquences d'estimation précédentes.
    - S'il est le plus petit, alors on recommence à l'étape 1 avec une nouvelle permutation aléatoire de départ ;
    - Sinon :
        - Si $K < K_S$, alors on recommence à l'étape 1 ;
        - Sinon, l'algorithme s'arrête et alors : $\hat{\theta}_{SLTS,\omega} = \hat{\theta}^k$ et $S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega}) = S_S(X, y, w; \hat{\theta}^k) = Q_{SLTS}(\hat{\theta}^k, w)$.

Pour déterminer la valeur de ω à retenir, Čížek propose de fixer $\lambda = 1/2$ et $K_S = 100$, et de commencer par tester le premier algorithme sur une valeur initiale ω_0 raisonnablement élevée (par exemple 50), puis de faire décroître graduellement la valeur de ω ($\omega_{l+1} = 0{,}8\omega_l$) dans un deuxième algorithme. Il faut alors définir un critère d'arrêt, indiquant dans quelle mesure le ω testé est acceptable, c'est-à-dire dans quelle mesure il ne dégrade pas trop la robustesse de l'estimateur. Čížek définit deux critères d'arrêt : le premier basé sur la fonction objectif et le second basé sur les résidus.

- Critère sur la fonction objectif : la fonction objectif est décroissante, atteignant son maximum $S_{S,LS}$ (fonction objectif de l'estimateur OLS) quand $\omega \to 0$, et son minimum $S_{S,LTS}$ (fonction objectif du LTS-estimateur) quand $\omega \to \infty$. Le ratio $R = S_{S,LS} / S_{S,LTS}$, testé par Čížek, fournit un bon indicateur de la contamination des données, ou bien, dit autrement, de la probabilité que les OLS ne fonctionnent pas. Comme on souhaite choisir un ω adapté au « besoin de robustesse » lié au degré de contamination des données, on retient un ω tel que $S_S^*(\omega) / S_S^*(\omega_0)$ ne dépasse pas trop R. On s'arrête donc quand la fonction optimale $S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega})$ est plus grande qu'un critère $c \times R_N \times S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega_0})$, avec

$R_N = \dfrac{\lambda}{F_{\chi^2_3}(F^{-1}_{\chi^2_3}(\lambda))}$, qui correspond à la limite asymptotique de R, et $c \in [1;2]$ un paramètre, qu'on fixera à $c = 1$. Pour $\lambda = 1/2$, $R_N = 7{,}01$.

- Critère sur les résidus : dans le cas de données particulièrement contaminées, le critère d'arrêt basé sur la fonction objectif n'est pas suffisant, car il est possible de choisir un ω pour lequel les estimations sont déjà trop affectées par les points atypiques. On complète donc ce critère par une nouvelle règle de décision portant sur les résidus. En effet, les résidus pondérés permettent de décrire l'influence des observations sur la fonction objectif, ils sont donc un bon indicateur de robustesse du modèle. Le principe de la règle de décision sur les résidus est le suivant : l'estimation réalisée pour ω_0 est la plus robuste, car ω_0 est la valeur de ω la plus élevée. Elle est donc l'estimation pour laquelle on a réduit l'influence d'un maximum d'observations ayant des résidus élevés. En faisant décroître ω, on augmente donc l'influence de certaines observations ayant des résidus élevés. On cherche à vérifier qu'on ne l'augmente pas trop par rapport à l'estimation la plus robuste, en ω_0, en définissant une sorte d'intervalle de confiance « de robustesse de l'estimateur » pour les résidus pondérés, défini comme $[m_0 \pm C \times mad_0]$, où :

  $m_0 = med_i(r_i(\hat{\theta}_{SLTS,\omega_0}))$ est la médiane des résidus de $\hat{\theta}_{SLTS,\omega_0}$,

  $mad_0 = med_i(\left| r_i(\hat{\theta}_{SLTS,\omega_0}) - med_j(r_j(\hat{\theta}_{SLTS,\omega_0})) \right|)$ est la déviation absolue à la médiane (median absolute deviation) de $\hat{\theta}_{SLTS,\omega_0}$, et C un paramètre choisi de sorte que $[m_0 \pm C \times mad_0]$ soit un intervalle de confiance pour les λn observations ayant les plus petits résidus, les autres observations devant alors être sous-pondérées de manière à ce que leurs résidus pondérés appartiennent à l'intervalle de confiance. Pour $\lambda = 1/2$, on peut choisir $C \in [0{,}48;0{,}72]$.

  Si, pour un ω et donc une estimation donnés, il existe des résidus pondérés en dehors de cet intervalle de confiance de robustesse, alors on considère que les observations associées ont une influence trop importante sur la fonction objectif et les estimations. On arrête alors les itérations sur ω et le SLTS-estimateur est l'estimateur du précédent ω, considéré comme le juste compromis entre robustesse et efficience.

Ce deuxième algorithme s'effectue en 6 étapes :

- étape 1 : on fixe $\omega_0 = 50$. On fixe également un $\omega_{\min} = 1$, par exemple (ce qui revient à faire 17 itérations au maximum sur ω) ;
- étape 2 : on calcule le SLTS-estimateur $\hat{\theta}_{SLTS,\omega_0}$ avec la boucle décrite précédemment, et on obtient $S_S{}^*(X, y, w; \hat{\theta}_{SLTS,\omega_0})$. On calcule également m_0, la médiane des résidus de $\hat{\theta}_{SLTS,\omega_0}$, et mad_0, la déviation absolue à la médiane de $\hat{\theta}_{SLTS,\omega_0}$;
- étape 3 : on calcule $\omega_{l+1} = 0{,}8\omega_l$. Si $\omega_{l+1} < \omega_{\min}$, alors on s'arrête et $\hat{\theta}_{SLTS} = \hat{\theta}_{SLTS,\omega_l}$;
- étape 4 : sinon, on calcule le SLTS-estimateur $\hat{\theta}_{SLTS,\omega_{l+1}}$ avec la boucle décrite précédemment, et on obtient $S_S{}^*(X, y, w; \hat{\theta}_{SLTS,\omega_{l+1}})$;

- étape 5 : si $S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega_{l+1}}) > c \times R_N \times S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega_0})$, alors on s'arrête et on garde alors $\hat{\theta}_{SLTS} = \hat{\theta}_{SLTS,\omega_l}$;
- étape 6 : sinon, si $S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega_{l+1}}) \leq c \times R_N \times S_S^*(X, y, w; \hat{\theta}_{SLTS,\omega_0})$, on calcule les résidus pondérés $\frac{\sqrt{w_i}}{\max_i \sqrt{w_i}} r_i(\hat{\theta}_{SLTS,\omega_{l+1}})$.

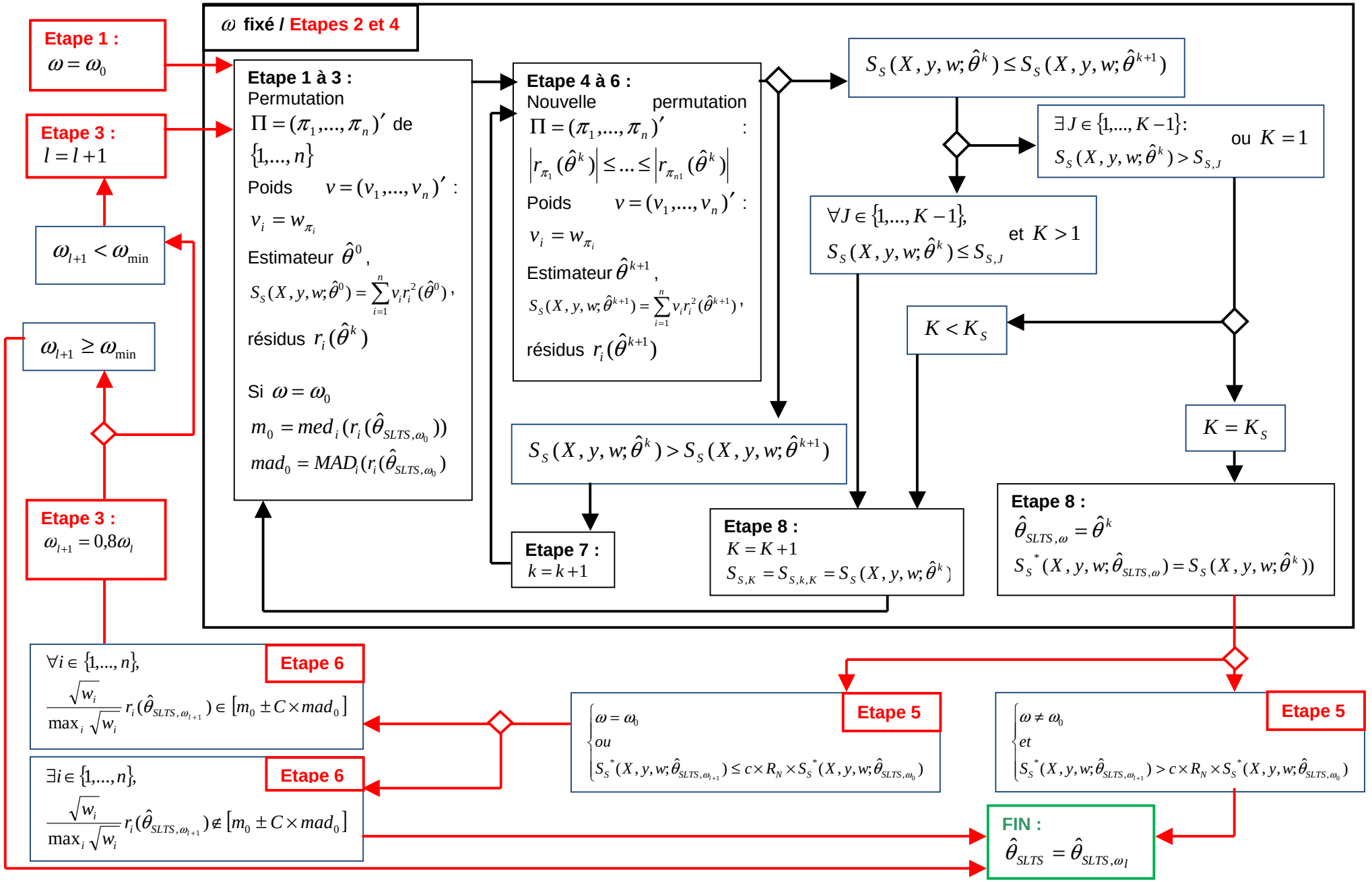
  Si $\frac{\sqrt{w_i}}{\max_i \sqrt{w_i}} r_i(\hat{\theta}_{SLTS,\omega_{l+1}}) \in [m_0 \pm C \times mad_0]$, pour tout $i \in \{1,...,n\}$, alors on recommence à l'étape 3 et on fixe $l = l+1$. Sinon, on s'arrête et alors $\hat{\theta}_{SLTS} = \hat{\theta}_{SLTS,\omega_l}$.

Autrement dit, à λ et K_S fixés, il faut boucler sur ω en plus des itérations pour un ω fixé. Les tests réalisés par Čížek montrent que l'algorithme converge relativement rapidement pour les échantillons de plusieurs milliers d'observations. Mais, dans notre cas d'étude et les 31 millions d'observations à modéliser sur les salariés privés classiques, son utilisation semble peu adaptée.

On peut cependant tester, pour quelques échantillons de 100 000 observations des salariés privés classiques, les résultats obtenus en utilisant le premier algorithme de la méthode sur des ω choisis, afin d'observer les paramètres estimés et les temps de calcul[82].

[82] Voir §4.4.5.

Schéma récapitulatif de l'algorithme de la SLTS-estimation



4.4.3. Une méthode proche de la méthode RDL : la méthode RDM

Le principe est identique à celui de la méthode RDL :

- en substituant, dans la première étape, la distance robuste calculée par le MM-estimateur grâce à l'algorithme MCD (minimum covariance determinant)[83], proposé par Rousseeuw (1984) à la distance robuste calculée par l'estimateur MVE. L'estimateur MCD a pour objectif de trouver h observations parmi n dont la matrice de variance-covariance a le plus petit déterminant. Le point de rupture est identique à celui de l'estimateur MVE, mais le MCD possède plusieurs avantages sur le MVE, dont le fait de calculer des distances robustes plus précises que celles du MVE[84].
- en substituant, dans la deuxième étape, un M-estimateur pondéré à l'estimateur L1 pondéré pour la détection des outliers. En effet, le M-estimateur est statistiquement plus efficient que l'estimateur L1[85].

4.4.4. Tests des méthodes RDL et RDM

4.4.4.1. Spécification des modèles

Les variables quantitatives pour la modélisation du salaire horaire sont les mêmes pour tous les codes populations :

- des caractéristiques de temps de travail : durée de la période et quotité ;
- l'âge du salarié.

Ce sont ces trois variables qui sont utilisées dans la première étape de la méthode RDM : la modélisation MM.

Pour la deuxième étape (M-estimation pondérée), on rajoute les variables qualitatives :

- Certaines variables qualitatives sont communes à tous les codes population :
  - le sexe du salarié ;
  - le fait de travailler ou non en Île-de-France ;
- Les autres sont spécifiques à chaque code population et sont les mêmes que celles utilisées pour la modélisation OLS du salaire horaire[86].

4.4.4.2. Spécificités des modèles sur les salariés privés classiques et les fonctionnaires

Les périodes des salariés privés classiques et des fonctionnaires sont très nombreuses. Afin de faciliter la modélisation, on raisonne par sous-population pour les deux étapes de chaque méthode. Pour les salariés privés classiques, on réalise une analyse croisée par sexe et premier chiffre de la PCS (CS1)[87]. Il reste des sous-populations grandes, au sein desquelles il existe une certaine diversité du salaire horaire : parmi les ouvriers, chez les hommes, et parmi les employés, chez les femmes. Pour les hommes ouvriers, on distingue les ouvriers qualifiés des ouvriers non qualifiés, obtenus à partir de la CS à deux positions (CS2). Pour les femmes employées, on utilise directement la CS2[88], car ces catégories contiennent des types de salariés très disparates en termes d'activité et de salaire. Pour les fonctionnaires, une analyse par sexe est suffisante pour permettre l'aboutissement des calculs.

---

83 Voir Bibliographie [17].

84 Voir Bibliographie [14].

85 Voir Bibliographie [18].

86 Voir §4.1.2.

87 On regroupe les agriculteurs exploitants et les artisans, commerçants et chefs d'entreprise, en introduisant la CS1 comme variable explicative du modèle.

88 Voir Annexe §5.2.

4.4.4.3. Résultats des méthodes RDL et RDM

- 1e étape :

Sur les 50 échantillons de 100 000 observations des salariés privés classiques, en tronquant le salaire horaire au 99e percentile, la MM-estimation sur les variables quantitatives[89] détecte en moyenne 16 470 points leviers, soit 16,6 % des observations. Cette proportion est très stable sur les 50 échantillons. En moyenne, les points leviers ont un poids de 0,74 pour la 2e étape.

Sur la population totale, en tronquant le salaire horaire au 99e percentile du code population concerné, la MM-estimation sur les variables quantitatives[90] détecte en moyenne 12,6 millions de points leviers, soit 32,5 % des observations. Cette proportion est variable, allant de 19 % pour les périodes des fonctionnaires détachés à 42 % pour les femmes salariées privées classiques, employées des services directs aux particuliers. En moyenne, les points leviers ont un poids de 0,5 pour la 2e étape.

- 2e étape :

Sur les 50 échantillons de 100 000 observations des salariés privés classiques, en tronquant le salaire horaire au 99e percentile, la méthode RDL détecte systématiquement plus d'individus atypiques que la méthode RDM[91] : en moyenne 7 100 individus atypiques, contre 6 700 pour la méthode RDM, soit respectivement 7,2 % et 6,8 % des observations. 99 % des individus détectés comme atypiques par la méthode RDM le sont également par la méthode RDL.

Sur l'ensemble de la population, en tronquant le salaire horaire au 99e percentile du code population concerné, la méthode RDL détecte plus d'individus atypiques que la méthode RDM[92], sauf pour les populations suivantes :

- les femmes salariées privées classiques, agents de surveillance ou employées de commerce ;
- les périodes des salariés privés dans le public : pour ce code population, les deux méthodes détectent un fort pourcentage (19,6 % et 15,9 %) d'individus atypiques, les variables explicatives étant peu nombreuses et peu discriminantes du salaire horaire pour cette population.

Globalement, les deux méthodes détectent un pourcentage proche de points atypiques : 2,1 millions de périodes détectées comme atypiques par la méthode RDM, contre 2,3 millions de périodes détectées comme atypiques par la méthode RDL, soit respectivement 5,5 % et 6,1 % des périodes. La plupart des périodes atypiques sont détectées en commun : les périodes détectées par les deux méthodes représentent 88,1 % des outliers de la méthode RDL et 96,4 % des outliers de la méthode RDM.

4.4.4.4. Comparaison des résultats des méthodes M et RDM

- Les périodes atypiques :

La méthode RDM est censée être meilleure que la M-estimation simple pour la détection de points atypiques, du fait de l'existence de points leviers dans les données. On peut comparer les diagnostics obtenus à partir de ces deux méthodes.

Sur les 50 échantillons de 100 000 observations des salariés privés classiques, en tronquant le salaire horaire au 99e percentile, la méthode M détecte systématiquement plus d'individus atypiques que la méthode RDM[93] : en moyenne 6 900 individus atypiques, contre 6 700 pour la méthode RDM, soit respectivement 6,9 % et 6,8 % des observations. La quasi-totalité des individus détectés comme atypiques par la méthode RDM le sont également par la méthode M.

Sur la population totale, en tronquant le salaire horaire au 99e percentile du code population concerné, les deux méthodes détectent un pourcentage proche de points atypiques[94] : 2,1 millions de périodes détectées comme atypiques par la méthode RDM, contre 2,2 millions de périodes détectées comme atypiques par la méthode M, soit respectivement 5,5 % et 5,7 % des périodes. La plupart des périodes atypiques sont détectées en commun : les périodes détectées par les deux méthodes représentent 96,5 % des outliers de la méthode RDM.

89 Voir Annexes §5.1.
90 Voir Annexes §5.2.
91 Voir Annexes §5.1.
92 Voir Annexes §5.2.
93 Voir Annexes §6.1.
94 Voir Annexes §6.2.

Les deux méthodes détectent un fort pourcentage (19,6 % et 34,8 %) d'individus atypiques pour le code population 13, les variables explicatives étant peu nombreuses et peu discriminantes du salaire horaire pour cette population.

- Le salaire horaire selon le type de périodes :

Sur les 50 échantillons, les points atypiques détectés par l'une ou l'autre méthode ont un salaire horaire significativement supérieur à celui des périodes non détectées comme aberrantes[95].
Les points atypiques détectés par la M-estimation mais non détectés comme atypiques par la méthode RDM ont en moyenne un salaire horaire significativement inférieur aux autres points atypiques.

Sur la population totale, les points atypiques détectés par l'une ou l'autre méthode ont également un salaire horaire significativement supérieur à celui des périodes non détectées comme aberrantes[96].
Les points atypiques détectés par la M-estimation mais non détectés comme atypiques par la méthode RDM ont en moyenne un salaire horaire significativement inférieur aux autres points atypiques, tout comme les points atypiques détectés par la méthode RDM mais non détectés comme atypiques par la M-estimation, sauf pour trois sous-populations :

- les femmes salariées privées classiques, employées civiles et agents de service de la fonction publique ;
- les hommes salariés privés classiques, agriculteurs exploitants ou artisans, commerçants et chefs d'entreprise ;
- les hommes fonctionnaires.

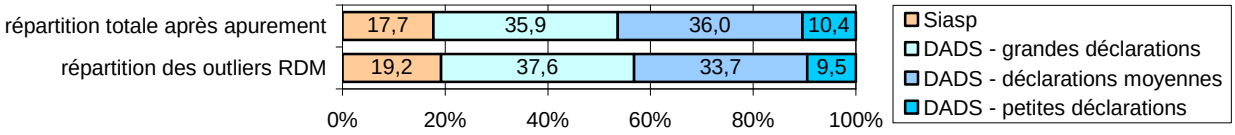
Ainsi, **les périodes atypiques détectées par les méthodes M et RDM n'ont pas le même profil de salaire horaire, même si elles détectent une grosse majorité de points aberrants en commun.** Si les temps de calculs sont largement inférieurs avec la méthode M, **il reste préférable d'utiliser la méthode RDM pour détecter les périodes atypiques et fournir aux applications une information de meilleure qualité**, la méthode RDM étant supérieure à la seule M-estimation du fait des points leviers. **Elle est donc la solution proposée pour la détection des périodes atypiques en termes de rémunérations au niveau du Frontal.**

4.4.4.5. Bilan des points atypiques de la méthode RDM selon le type de déclaration

La répartition des outliers détectés par la méthode RDM par taille de déclaration et application destinataire[97] est similaire à celle du total des périodes conservées à l'issue de l'étape initiale d'apurement[98] :

- 37,6 % sont issus de grandes déclarations à destination de l'application DADS et 19,2 % sont à destination de Siasp. L'information sur ces points atypiques peut être utilisée **en complément** de l'information contrôlée au niveau des applications ;
- 33,7 % et 9,5 % sont issus de moyennes et petites déclarations à destination de l'application DADS. L'information sur ces points atypiques peut être utilisée **directement** par l'application.

Graphique 2 : Répartition des périodes après apurement et des outliers RDM, selon l'application destinataire et la taille des déclarations



4.4.5. Tests sur échantillons de la méthode STLS

On teste, pour 10 échantillons de 100 000 observations des salariés du privé classique, les résultats obtenus en utilisant le premier algorithme de la méthode sur des ω choisis, afin d'observer les paramètres estimés et les temps de calcul.

---

95 Voir Annexes §6.1.
96 Voir Annexes §6.2.
97 Voir Annexes §7.
98 Voir §3.6.

La première phase de l'algorithme, pour un ω fixé, nécessite déjà entre 10 minutes et 8 heures[99]. Étant donné que la méthode SLTS nécessite, dans un deuxième algorithme, de boucler sur ω, elle n'est pas envisageable compte-tenu de la masse de données que l'on souhaite traiter.

Quand on s'intéresse aux paramètres estimés, ceux-ci sont très variables selon le ω fixé, ce qui est logique. En effet, un ω élevé implique une estimation très robuste, mais pas nécessairement efficiente. A contrario, un ω faible est associée à une estimation non robuste et donc non convergente en présence d'outliers. En faisant décroître progressivement ω, on augmente l'efficience de l'estimation. En vérifiant parallèlement qu'on ne dégrade pas trop sa robustesse, on s'assure d'un ω optimal correspondant au juste compromis entre robustesse et efficience.

En prenant en compte le critère sur la fonction objectif testé dans la deuxième partie de l'algorithme SLTS, le ω optimal se situerait entre 30 et 40. Les paramètres estimés pour ces valeurs de ω sont plus ou moins proches selon la variable explicative concernée, on peut donc penser qu'ils sont différents des paramètres finaux qui seraient obtenus avec le ω optimal. Ils restent également éloignés des paramètres obtenus avec la méthode RDM, sans que cela puisse permettre d'en conclure quoi que ce soit sur l'une ou l'autre méthode.

4.5. Temps de calculs et ressources nécessaires

4.5.1. Temps de calculs

Sur les 50 échantillons de 100 000 observations, le temps moyen passé pour la M-estimation est de 8 secondes, contre 3 minutes 46 pour la méthode RDM et 3 minutes 28 pour la méthode RDL.
Sur l'ensemble de la population, les M-estimations prennent un temps cumulé de 19 minutes 36, tandis que les méthodes RDM et RDL sont beaucoup plus gourmandes en temps : 6h42 pour la méthode RDM et 9h25 pour la méthode RDL. La méthode RDM nécessite plus de 20 fois plus de temps que la méthode M. Cependant, elle reste préférable à la M-estimation, compte-tenu des différences d'outliers détectés, mais également de sa prise en compte des points leviers, et le temps passé reste raisonnable, sachant que le traitement ne doit être réalisé qu'une fois par validité. La méthode RDM reste également, en temps comme en efficacité statistique, supérieure à la méthode RDL.

4.5.2. Ressources nécessaires

Les ressources en mémoire temporaire varient selon les algorithmes utilisés par les méthodes d'estimations[100] que nous avons vues :

- La M-estimation nécessite $3n + 2p^2 + 30p$ octets ;
- La LTS-estimation comme la MM-estimation nécessitent $np + 12n + 4p^2 + 60p$ octets, essentiellement utilisés pour stocker les informations issues des différents échantillons utilisés pour l'estimation ;
- La méthode L1 nécessite $np + 6n + p^2 + 4p$ octets.

Etant donné le nombre d'observations de notre étude, c'est le paramètre n qui est déterminant. Ainsi, sous l'hypothèse de réalisation des estimations sur les sous-populations dans leur ensemble, les besoins en mémoire sont très importants. C'est la méthode RDL la plus gourmande en ressources, du fait de la combinaison de la MM-estimation et de la L-estimation : au total, 1,2 Go sont nécessaires pour réaliser cette estimation. La RDM-estimation nécessite 1,8 fois moins de ressources, avec 700 Mo. Enfin, la M-estimation a des besoins 11,5 fois inférieurs à la méthode RDL, avec 110 Mo.

Tableau 10 : Ressources nécessaires pour les différentes méthodes d'estimation, selon le code population

Code population	nombre de périodes concernées	ressources nécessaires (Ko)		
		M	RDM	RDL
10 - salarié privé classique ; échantillon	90 000	294	1 856	3 223
10 - salarié privé classique	31 129 192	91 199	577 593	1 033 587
11 - salarié du spectacle	686 199	2 011	12 733	20 774
13 - salarié privé dans le public	264 074	774	4 901	8 253
14 - fonctionnaire détaché	8 564	26	160	268
40 - fonctionnaire	4 143 185	12 139	76 877	145 659
43 - non-titulaire de la fonction publique	2 363 599	6 926	43 857	87 712
Total	38 594 813	113 075	716 120	1 296 253

99 Voir Annexes §8.

100 Voir Bibliographie [8], procédure ROBUSTREG.

5. Utilisations possibles des résultats et pistes d'amélioration

5.1. Utilisation directe des résultats

La méthode RDM a été testée ici sur la validité 2011, première année en norme N4DS. Un nouveau test sur la validité 2012, une fois l'ensemble des données reçues par le Frontal, fin novembre 2013, permettrait de valider la méthode, sachant qu'une meilleure qualité des données déclarées est attendue. Ce test permettrait également de comparer les résultats obtenus par la méthode sur les données reçues fin mars N+1, une fois la quasi-totalité des informations reçues et en amont de leur traitement dans les applications DADS et Siasp, et fin novembre N+1, une fois la campagne terminée et l'intégralité des informations reçues par le Frontal. Ainsi, la liste des périodes atypiques détectées par le Frontal pourrait, dès la validité 2013, être fournie aux applications afin de compléter leur dispositif de contrôle des rémunérations.

5.2. Utilisation des méthodes testées dans le système d'information actuel

5.2.1. Pour les contrôles des rémunérations réalisés au niveau des applications

L'étude montre ici l'inadéquation des OLS pour détecter les points atypiques. Il conviendrait de vérifier que ce n'est pas le cas au niveau des applications, avec des informations auxiliaires redressées. Car dans ce cas, les coefficients du ou des modèles sont biaisés, du fait de la non-robustesse de l'estimation par les moindres carrés ordinaires, et l'imputation par la valeur prédite réalisée dans Siasp n'est également pas satisfaisante. Dans ce cas, l'étude réalisée pourrait être complétée par des tests de détection et de redressement des observations atypiques au niveau des applications DADS et Siasp.

5.2.2. Pour la détection des observations atypiques en diffusion de l'application DADS

Il est nécessaire de vérifier et éventuellement corriger des observations atypiques ou aberrantes sur les données diffusées, avant leur mise à disposition, sur des variables comme le salaire mensuel net par EQTP (équivalent temps plein). Actuellement, seules les distributions sont étudiées[101]. Les contrôles pourraient donc être complétés et automatisés en testant et en utilisant des méthodes robustes étudiées dans le cadre de cette étude.

5.3. Utilisation dans le cadre du futur système d'information

L'arrivée de la DSN en 2016 induira une large refonte du Siera. Une des limites du système actuel est qu'il est conçu pour traiter les cas particuliers dans le même temps que le cas nominal. Il ne permet donc pas d'avoir un retour rapide sur les données, puisqu'il faut attendre d'avoir traité la quasi-totalité des cas pour disposer d'informations macros à analyser. Dans l'objectif de mieux suivre la production et de pouvoir expertiser plus rapidement les données reçues, on souhaite, dans le cadre du futur système, pouvoir isoler les cas particuliers (dont les observations atypiques) pour traiter rapidement et prioritairement le cas nominal. Ainsi, l'analyse réalisée est adaptée aux besoins du futur système d'information et pourra y trouver une application concrète.

[101] Analyse des évolutions annuelles de quantiles et des rapports interquantiles.

Conclusion

Les DADS constituent la source annuelle de référence sur l'emploi et les salaires. La qualité des informations déclarées y est variable et les données renseignées, adaptées aux besoins des organismes destinataires, ne correspondent pas toujours aux concepts statistiques que l'on souhaite mesurer. Ainsi, de nombreux traitements statistiques sont nécessaires, dont des contrôles et redressements des rémunérations. Ces derniers sont actuellement réalisés au niveau des applications DADS et Siasp, selon des méthodes proches mais très liées au mode de fonctionnement des applications, basées sur la modélisation du salaire horaire par les moindres carrés ordinaires.

En cherchant à utiliser une méthode homogène sur l'ensemble de la source, en amont des applications, pour la détection des observations atypiques en termes de rémunérations, on montre que les moindres carrés ordinaires ne sont pas suffisamment robustes et qu'il est nécessaire d'utiliser une méthode robuste non seulement aux points verticaux, mais également aux points leviers.

Certaines variables explicatives du modèle étant discrètes, les méthodes robustes classiques ne peuvent pas être utilisées. On propose donc une méthode alternative, la méthode RDM, adaptée à la fois aux points leviers et aux points verticaux, et fonctionnant avec des variables explicatives discrètes. Elle est plus gourmande en temps et ressources que la M-estimation, notamment car elle fonctionne en deux étapes, mais elle reste néanmoins adaptée au gros volume des données que l'on souhaite traiter.

La méthode testée permet ainsi d'isoler les observations atypiques en termes de rémunérations en amont des applications DADS et Siasp, et de leur fournir une information complémentaire. Elle rentre également dans le cadre du futur système d'information sur l'emploi et les revenus d'activité, en permettant d'isoler le cas nominal des cas particuliers pour les traiter séparément afin de pouvoir expertiser sans délai les informations issues du cas nominal.

Glossaire

APET	Activité Principale de l'Établissement
CNAV	Caisse Nationale d'Assurance Vieillesse
CNIS	Conseil National de l'Information Statistique
CRDS	Contribution à la Réduction de la Dette Sociale
CS	Catégorie Socioprofessionnelle
CS1	Catégorie Socioprofessionnelle à une position
CS2	Catégorie Socioprofessionnelle à deux positions
CSG	Contribution Sociale Généralisée
CV	Coefficient de Variation (écart-type rapporté à la moyenne)
DADS	Déclarations Annuelles de Données Sociales (source ; application ; produits)
DADS-U	Déclarations Annuelles de Données Sociales - Unifiée (norme)
DARES	Direction de l'Animation de la Recherche, des Études et des Statistiques
DERA	Département de l'Emploi et les Revenus d'Activité
DGAFP	Direction Générale de l'Administration et de la Fonction Publique
DGFIP	Direction Générale des FInances Publiques
DREES	Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
DSN	Déclaration Sociale Nominative
EFA	Division Exploitation des Fichiers Administratifs sur l'emploi et les revenus
EQTP	Équivalent Temps Plein
FPE	Fonction Publique d'État
FPH	Fonction Publique Hospitalière
FPT	Fonction Publique Territoriale
LAV ou L1	Least Absolute Value estimate
LTS	Least Trimmed Squares
MCO	Moindres Carrés Ordinaires
MVE	Minimum Variation Estimator
N4DS	Norme de Déclaration Dématérialisée De Données Sociales
NIR	Numéro d'Identification au Répertoire
OLS ou LS	Ordinary Least Squares
PCS	Professions et Catégories Socioprofessionnelles
RDM	Robust Distance M-estimator
RDL	Robust Distance LAV estimate
SIASP	Système d'Information sur les Agents du Service Public
SIERA	Système d'Information sur l'Emploi et les Revenus d'Activité
SIRENE	Système Informatique pour le Répertoire des ENtreprises et de leurs Établissements
SMIC	Salaire Minimum Interprofessionnel de Croissance
SLTS	Smooth Trimmed Least Squares estimator

Définitions

Contrôle des rémunérations :

Le **contrôle des rémunérations** dans la source DADS s'intéresse à la fois aux **informations financières** et de **volume de travail**. En effet, on cherche à s'assurer de la cohérence du **salaire horaire** et donc :

- de la cohérence interne des informations financières ;
- de la cohérence interne des informations sur le volume, la **durée** et la **quotité** de travail ;
- de la cohérence des informations financières relativement au volume de travail.

La variable d'intérêt est donc, in fine, le **salaire horaire**. Il est calculé, dans les applications, par le salaire net rapporté au nombre d'heures rémunérées :

SALAIRE_HORAIRE = SALAIRE_NET / HEURES_REMUNEREES

Il correspond donc au **salaire horaire net**. Pour pouvoir le calculer et s'assurer de sa cohérence, il faut vérifier les informations sur les **rémunérations**, au numérateur, et sur le **volume de travail**, au dénominateur.

Quand on parle de **contrôle des rémunérations dans les applications DADS et Siasp**, on **inclut**, par extension, le **redressement des variables financières et de volume, durée et quotité de travail**.

Rémunérations, informations financières, variables financières :

La rémunération d'un salarié est mesurée, dans les applications actuelles, par le **salaire brut** et le **salaire net**, visant à mesurer les **rémunérations versée par l'employeur et perçue par le salarié**.

Dans les applications DADS et Siasp, le **salaire brut** est obtenu à partir de la **base CSG** (donnée sociale), tandis que le **salaire net** est obtenu à partir de la **rémunération nette fiscale** ou **net fiscal** (donnée fiscale). Ces variables statistiques, issues d'informations répondant à des logiques administratives différentes, sont donc susceptibles de ne pas être cohérentes entre elles et ne pas respecter les équations théoriques de salaire.

Temps de travail :

L'information centrale du **temps de travail** est le **volume de travail**, utilisé pour le calcul du **salaire horaire** et qui correspond, pour les applications, au **nombre d'heures rémunérées**.

D'autres **variables relatives au temps de travail** sont également mises en cohérence lors du contrôle-redressement des rémunérations :

- la **durée de travail** : elle correspond au nombre de jours entre la date de début et la date de fin de la période considérée ;
- la **quotité de travail** : elle correspond, pour une période, à la part de temps de travail par rapport au temps plein ou complet.

Taille des déclarations :

Dans l'application DADS, la taille des déclarations impacte les traitements réalisés et est un critère d'intervention des gestionnaires. Elle n'a aucun rôle dans l'application Siasp. On définit, dans l'application DADS, trois catégories de déclarations :

- les **grandes déclarations**, comprenant 500 périodes ou plus. En cas d'anomalie, les grandes déclarations sont traitées par des gestionnaires ;
- les **déclarations moyennes**, comprenant de 11 périodes à moins de 500 périodes. Ces déclarations peuvent être traitées automatiquement ou par des gestionnaires, suivant la gravité et le nombre d'anomalies ;
- les **petites déclarations**, comprenant 10 périodes ou moins. Les traitements des petites déclarations sont totalement automatiques.

Bibliographie

[1] ALMA Ö. G., « Comparison of Robust Regression Methods in Linear Regression », Int. J. Contemp. Math. Sciences, vol. 6, n°9, 2011, http://www.m-hikari.com/ijcms-2011/9-12-2011/almaIJCMS9-12-2011.pdf

[2] BRIZARD A., « Le salaire et ses évolutions dans le secteur marchand non agricole : éléments méthodologiques », Premières Synthèses, Dares, n°41.1, octobre 2006, http://travail-emploi.gouv.fr/IMG/pdf/2006-10-41-1-3.pdf

[3] CARON N., « La correction de la non-réponse par repondération et par imputation », Document de travail, Insee, n°M0502, novembre 2005, http://www.insee.fr/fr/publications-et-services/docs_doc_travail/m0502.pdf

[4] CHEN C., « Robust Regression and Outlier Detection with the ROBUSTREG Procedure », SAS Statistics and Data Analysis, Paper 265-27, 2002, http://www2.sas.com/proceedings/sugi27/p265-27.pdf

[5] CHEN C., « An Introduction to Quantile Regression and the QUANTREG Procedure », SAS Statistics and Data Analysis, Paper 213-30, 2005, http://www2.sas.com/proceedings/sugi30/213-30.pdf

[6] ČÍŽEK P., An Robust Estimation with Discrete Explanatory Variables », Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, n°2002, 76, http://www.econstor.eu/bitstream/10419/65326/1/727071114.pdf

[7] DEHON C., « Régression linéaire et robustesse : théorie et applications », août 2011, https://www.yumpu.com/fr/document/view/16863700/regression-lineaire-et-robustesse-theorie-et-applications/

[8] DOCUMENTATION SAS :
- Procédure QUANTREG : http://support.sas.com/rnd/app/papers/quantreg.pdf
- Procédure ROBUSTREG : http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#rreg_toc.htm
- Régression robuste : http://support.sas.com/documentation/cdl/en/imlug/59656/HTML/default/viewer.htm#robustregexpls_sect10.htm

[9] FERRARI N., « Prévoir l'investissement des entreprises – Un indicateur des révisions dans l'enquête de conjoncture sur les investissements dans l'industrie », Document de travail, 2005, http://www.insee.fr/fr/publications-et-services/docs_doc_travail/g2005-09.pdf.

[10] GODINOT A., « Pour comprendre le recensement de la population », Insee méthodes, Insee, mai 2005, http://www.insee.fr/fr/publications-et-services/sommaire.asp?codesage=imeths01

[11] HAZIZA D., « Traitement de la non-réponse », polycopié FCDA.

[12] HERNU P., « La nouvelle chaîne d'exploitation des DAS », Courrier des statistiques n°24, pp.39-43, octobre 1982, https://www.epsilon.insee.fr/jspui/bitstream/1/14108/1/cs24.pdf

[13] LAGARDE S., « La nouvelle exploitation exhaustive des DADS », Courrier des statistiques n°85-86, pp.68-70, juin 1998, https://www.epsilon.insee.fr/jspui/bitstream/1/14352/1/cs85-86.pdf

[14] RITSCHARD G., ANTILLE G., « A Robust Look at the Use of Regression Diagnostics », The Statistician, vol.41, n°1, p.41-53, 1992, http://www.jstor.org/discover/10.2307/2348635?uid=3738016&uid=2129&uid=2&uid=70&uid=4&sid=21102576008973.

[15] ROUSSEEUW P. J., « Least Median of Squares Regression », 1984, http://web.ipac.caltech.edu/staff/fmasci/home/statistics_refs/LeastMedianOfSquares.pdf.

[16] ROUSSEEUW P., HUBERT M., « Robust regression with both continuous and binary regressors », 1997, http://wis.kuleuven.be/stat/robust/papers/1997/rdl1.pdf.

[17] ROUSSEEUW P., VAN DRIESSEN K., « A Fast Algorithm for the Minimum Covariance Determinant Estimator », Technometrics, vol.41, n°3, août 1999, http://www.jstor.org/stable/1270566?seq=2.

[18] SIMPSON J. R., « A Combined Biased-Robust Estimator for Dealing with Influence and Colinearity in Regression », 1994, http://www.dtic.mil/dtic/tr/fulltext/u2/a281793.pdf.

[19] STUART C., « Robust Regression », avril 2011, http://www.maths.dur.ac.uk/Ug/projects/highlights/CM3/Stuart_Robust_Regression_report.pdf.

[20] « Guide méthodologique des DADS-Grand format », validité 2010, Insee, septembre 2012.

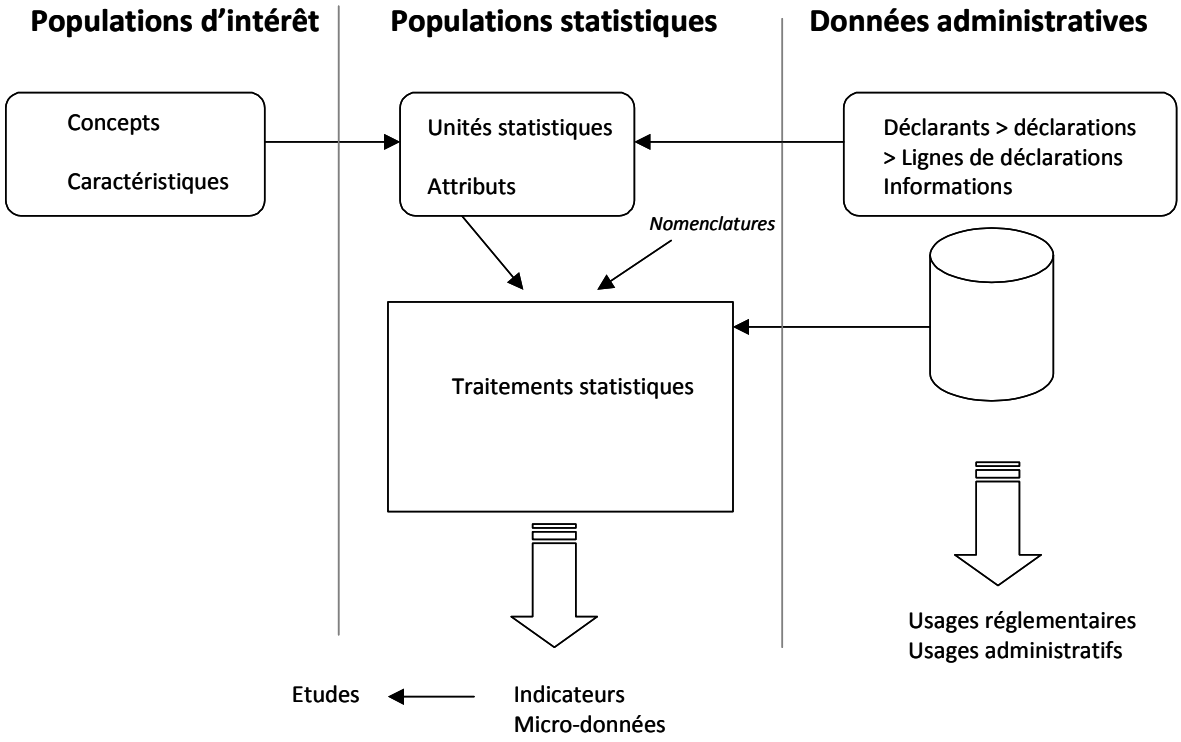
[21] Spécifications de l'application Siasp, Insee.

Annexes

1. Contexte et problématique

1.1. Statistiques et données administratives

La définition d'un système d'information statistique[102] requiert de définir d'une part les concepts statistiques ou économiques et d'autre part la façon dont on va mesurer ces concepts à partir des sources d'information existantes (données administratives ou enquêtes). Dans le cas des sources administratives, les principales difficultés pour le statisticien concernent les questions de champ (couverture complète ou non), de qualité (*a priori*, fiabilité uniquement des données servant à la gestion), leur stabilité dans le temps (stabilité des concepts, leur mesure peut éventuellement évoluer dans le temps, ce qui nécessite une adaptation du système statistique et éventuellement des rétropolations).



Pour cela, le statisticien met en place plusieurs étapes de traitement, avec les enjeux suivants :

- Définition des unités statistiques :
  - Concilier les impératifs de gestion et les besoins de la statistique, documenter les choix ;
  - Comment prendre en compte les évolutions d'environnement ;
- Validation des données d'identification avec les répertoires :
  par exemple, dans le cas des DADS :
  - les établissements sont identifiés grâce au répertoire Sirene[103] ;
  - l'identification des salariés permet de redresser leur date de naissance, fréquemment déclarée au premier jour du mois par l'employeur ;
- Définition des caractéristiques :
  - Mode de calcul ;
  - Codifications (automatiques, manuelles)

  par exemple, dans le cas des DADS : la PCS[104] ;

[102] Ce paragraphe reprend des éléments décrits dans le cadre du programme d'évolution du Siera.

[103] Système Informatique pour le Répertoire des Entreprises et de leurs Etablissements.

[104] Profession et Catégorie Socioprofessionnelle.

- Enrichissement avec des données externes (répertoires …) :

par exemple, dans le cas des DADS : la tranche d'effectif des établissements, fournie par le répertoire Sirene ;

- Modélisation ;

- Apurement et imputation :
  - Traitement des incohérences internes ;
  - Traitement de la non-réponse partielle, des échecs de codage ;
  - Redressements éventuels sur des données externes ;
- Le cycle de vie des données :
  - L'évolution des définitions, normes, consignes de recueil, cinématique ;
  - Dates d'effet/dates de gestion :

par exemple, dans le cas des DADS : les différences entre dates d'effet et de gestion peuvent expliquer des salaires négatifs ;

  - Ruptures de séries ;
- Indicateurs de qualité, redressements et diffusion des données.

1.2. Constitution des salaires à partir d'une fiche de paie

Exemple d'un salarié ayant travaillé 100 heures à 10 € bruts de l'heure, sans heures supplémentaires :

Informations	Nombre d'heures ou base de calcul	Taux (taux légal applicable aux revenus d'activité salariée[105])	Gains	Retenues
Salaire brut : **SALAIRE_BRUT**	100 (heures)	10,00 (€/heure)	1000,00 €	
Cotisations salariées	1000,00 € (SALAIRE_BRUT)	13,41 %		134,10 €
CSG déductible	970,00 € (**BASE_CSG**) (BASE_CSG = SALAIRE_BRUT * part du salaire brut soumis à CSG[106])	5,10 %		49,47 €
Net imposable : NET_FISCAL (NET_FISCAL = SALAIRE_BRUT - Cotisations salariées - CSG déductible)			816,43 €	
CSG non déductible	970,00 €	2,40 %		23,28 €
Cotisation Remboursement Dette Sociale	970,00 €	0,50 %		4,85 €
NET A PAYER : **SALAIRE_NET** (SALAIRE_NET = NET_FISCAL - CSG déductible - CRDS)			788,30 €	

1.3. Répartition initiale des périodes selon le type de déclarations

Code population	Siasp - ensemble		DADS						Total	
			grandes déclarations		déclarations moyennes		petites déclarations			
	nombre	%	nombre	%	nombre	%	nombre	%	nombre	%
10 - salarié privé classique	779 110	6,2%	24 245 500	96,1%	17 675 207	96,4%	4 377 829	99,0%	47 077 646	77,8%
11 - salarié du spectacle	15 488	0,1%	363 184	1,4%	613 991	3,3%	40 777	0,9%	1 033 440	1,7%
13 - salarié privé dans le public	403 774	3,2%	51 411	0,2%	6 639	0,0%	858	0,0%	462 682	0,8%
14 - fonctionnaire détaché	2 489	0,0%	14 949	0,1%	7 481	0,0%	418	0,0%	25 337	0,0%
40 - fonctionnaire	6 801 817	54,3%	511 797	2,0%	20 189	0,1%	117	0,0%	7 333 920	12,1%
42 - élu	244 265	1,9%	14	0,0%	169	0,0%	47	0,0%	244 495	0,4%
43 - non-titulaire de la fonction publique	4 288 397	34,2%	55 444	0,2%	4 749	0,0%	548	0,0%	4 349 138	7,2%
Total	12 535 340	100,0%	25 242 299	100,0%	18 328 425	100,0%	4 420 594	100,0%	60 526 658	100,0%
		20,7%		41,7%		30,3%		7,3%		100,0%
Nombre de déclarations associées	88 035		10 074		431 467		1 243 351		1 772 927	

105 Voir http://www.service-public.fr.

106 La part du salaire brut soumis à CSG est de 97 %.

2. Étape préliminaire[107]

2.1. Les heures travaillées manquantes

code population	nombre total de périodes	nombre de périodes sans heures travaillées	%
10 - salarié privé classique	47 077 646	3 869 889	8,2
11 - salarié du spectacle	1 033 440	88 309	8,5
13 - salarié privé dans le public	462 682	102 740	22,2
14 - fonctionnaire détaché	25 337	3 693	14,6
40 - fonctionnaire	7 333 920	7 160 344	97,6
43 - non-titulaire de la fonction publique	4 349 138	899 821	20,7
total	60 282 163	12 124 796	20,1

2.2. Filtre sur les heures et les variables financières

- Filtre sur les heures :

code population	nombre total de périodes	nombre de périodes atypiques sur les heures	%
10 - salarié privé classique	47 077 646	10 652 442	22,6%
11 - salarié du spectacle	1 033 440	177 818	17,2%
13 - salarié privé dans le public	462 682	96 240	20,8%
14 - fonctionnaire détaché	25 337	7 908	31,2%
40 - fonctionnaire	7 333 920	654 388	8,9%
43 - non-titulaire de la fonction publique	4 349 138	660 773	15,2%
total	60 282 163	12 249 569	20,3%

- heures travaillées nulles :

code population	nombre total de périodes	nombre de périodes avec heures travaillées nulles	%
10 - salarié privé classique	47 077 646	5 577 168	11,8%
11 - salarié du spectacle	1 033 440	53 106	5,1%
13 - salarié privé dans le public	462 682	21 509	4,6%
14 - fonctionnaire détaché	25 337	2 164	8,5%
40 - fonctionnaire	7 333 920	8 679	0,1%
43 - non-titulaire de la fonction publique	4 349 138	316 963	7,3%
total	60 282 163	5 979 589	9,9%

- heures rémunérées manquantes :

code population	nombre total de périodes	nombre de périodes sans heures rémunérées	%
10 - salarié privé classique	47 077 646	4 396 348	9,3%
11 - salarié du spectacle	1 033 440	117 806	11,4%
13 - salarié privé dans le public	462 682	59 207	12,8%
14 - fonctionnaire détaché	25 337	3 918	15,5%
40 - fonctionnaire	7 333 920	440 799	6,0%
43 - non-titulaire de la fonction publique	4 349 138	120 256	2,8%
total	60 282 163	5 138 334	8,5%

- heures rémunérées aberrantes :

code population	nombre total de périodes	nombre de périodes avec heures rémunérées négatives ou nulles	nombre de périodes avec heures rémunérées >2500	nombre de périodes avec heures rémunérées nulles ou >2500	%
10 - salarié privé classique	47 077 646	5 352 028	53 871	5 405 899	11,5%
11 - salarié du spectacle	1 033 440	38 095	9	38 104	3,7%
13 - salarié privé dans le public	462 682	31 117	271	31 388	6,8%
14 - fonctionnaire détaché	25 337	3 799	6	3 805	15,0%
40 - fonctionnaire	7 333 920	206 566	2 333	208 899	2,8%
43 - non-titulaire de la fonction publique	4 349 138	469 966	3 832	473 798	10,9%
total	60 282 163	6 101 571	60 322	6 161 893	10,2%

[107] Rappel : de nombreuses périodes des salariés de la fonction publique sont aberrantes, en particulier dans la fonction publique d'État, qui n'est pas conservée dans Siasp.

- Filtre sur les variables financières :

code population	nombre total de périodes	Base CSG <=0	%	net fiscal <=0	%	Base brute sécurité sociale <=0	%	Nombre total de périodes atypiques pour les variables financières	%
10 - salarié privé classique	47 077 646	3 551 486	7,5%	2 323 977	4,9%	3 370 762	7,2%	4 283 223	9,1%
11 - salarié du spectacle	1 033 440	11 935	1,2%	5 930	0,6%	8 055	0,8%	13 592	1,3%
13 - salarié privé dans le public	462 682	73 125	15,8%	16 967	3,7%	71 809	15,5%	78 245	16,9%
14 - fonctionnaire détaché	25 337	6 970	27,5%	1 186	4,7%	5 401	21,3%	7 155	28,2%
40 - fonctionnaire	7 333 920	2 364 147	32,2%	1 100 811	15,0%	2 004 882	27,3%	2 376 521	32,4%
43 - non-titulaire de la fonction publique	4 349 138	856 927	19,7%	515 438	11,9%	788 333	18,1%	903 343	20,8%
total	60 282 163	6 864 590	11,4%	3 964 309	6,6%	6 249 242	10,4%	7 662 079	12,7%

- Bilan du filtre sur les variables financières et le volume de travail :

code population	nombre total de périodes	Nombre total de périodes atypiques (variables financières et heures)	%	nombre de périodes atypiques sur les heures	%	Nombre total de périodes atypiques pour les variables financières	%
10 - salarié privé classique	47 077 646	11 495 445	24,4%	10 652 442	22,6%	4 283 223	9,1%
11 - salarié du spectacle	1 033 440	182 065	17,6%	177 818	17,2%	13 592	1,3%
13 - salarié privé dans le public	462 682	145 135	31,4%	96 240	20,8%	78 245	16,9%
14 - fonctionnaire détaché	25 337	12 171	48,0%	7 908	31,2%	7 155	28,2%
40 - fonctionnaire	7 333 920	2 488 321	33,9%	654 388	8,9%	2 376 521	32,4%
43 - non-titulaire de la fonction publique	4 349 138	1 210 752	27,8%	660 773	15,2%	903 343	20,8%
total	60 282 163	15 533 889	25,8%	12 249 569	20,3%	7 662 079	12,7%

3. Tests de la régression robuste M et des moindres carrés

3.1. Paramètres OLS

- Définition des paramètres :

  - Duree : durée de la période ;
  - Cs_i : indicatrice d'appartenance à la CS1 i, pour i de 2 à 6 ;
  - Sexe_f : indicatrice du sexe, vaut 1 si le salarié rattaché à la période est une femme, vaut 0 sinon ;
  - Age : âge du salarié ;
  - Quotite : quotité travaillée sur la période ;
  - Idf_etab : indicatrice du lieu de travail, vaut 1 pour l'Île-de-France, vaut 0 sinon.

- Estimation des moindres carrés, ensemble de l'échantillon :

échantillon	Constante	duree	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
Moyenne	-81,78	0,21	-72,16	607,74	137,28	10,87	-33,94	-70,69	2,70	0,40	-92,59
Variance	19680,46	0,04	6956,01	327258,67	30280,40	2327,18	4410,86	9443,68	10,54	0,26	10101,04
CV	-171,5%	91,0%	-115,6%	94,1%	126,8%	443,7%	-195,7%	-137,5%	120,3%	125,8%	-108,6%
1	-171,41	0,11	-70,48	726,43	40,82	46,75	-25,92	-81,04	5,34	0,36	-108,83
2	-72,76	0,26	-96,13	-17,90	407,61	33,77	-53,84	-129,51	2,33	0,64	-107,05
3	511,78	0,41	-393,06	468,36	-292,96	-270,25	-353,20	-168,12	1,13	-2,11	-136,21
4	-83,97	0,42	-69,03	782,42	451,05	50,68	-87,84	-234,16	3,82	0,86	-229,78
5	-244,98	0,16	-43,34	1182,37	40,17	21,91	41,15	30,80	4,19	0,96	-153,65
6	36,54	0,25	-73,55	681,27	156,99	3,24	-50,89	-85,88	0,11	0,09	-61,67
7	-185,74	0,18	-62,67	1456,70	0,72	24,10	-20,84	-34,66	4,16	0,87	-195,31
8	-220,87	0,10	-26,41	501,11	0,93	3,56	22,82	73,22	3,78	0,57	-56,14
9	-80,41	0,31	-26,97	-1,36	304,09	6,72	11,65	27,10	-0,46	0,83	-77,82
10	41,21	0,14	-39,75	-16,19	128,02	-4,47	-34,26	-51,14	-0,46	0,18	-29,94
11	-42,39	0,00	-20,61	3,69	69,97	3,86	-13,05	-19,71	1,58	0,18	-15,58
12	-80,03	0,20	-69,28	2,07	193,20	24,29	4,42	-88,82	2,39	0,45	-66,83
13	-38,40	0,71	-200,99	115,38	768,64	23,41	-133,53	-252,52	1,61	1,11	-208,02
14	104,80	0,39	-67,13	904,89	-15,94	-1,78	1,49	-47,53	-1,17	-0,25	-156,22
15	-64,13	0,34	-77,01	-44,34	503,02	-10,99	-53,19	-48,35	0,33	1,10	-134,12

échantillon	Constante	duree	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
16	-20,15	0,06	-20,83	699,52	35,98	9,83	-14,22	-39,30	1,49	0,14	-100,26
17	-323,44	0,33	-188,99	2203,64	162,19	96,53	-114,38	-320,59	11,90	1,02	-349,95
18	-14,19	0,34	-99,29	558,29	220,68	45,50	-74,97	-151,21	1,97	0,34	-155,29
19	-139,38	0,24	-58,72	-19,05	366,25	3,67	-18,47	5,93	2,64	0,56	-95,81
20	-140,30	-0,02	-68,80	1153,24	48,03	51,41	-29,24	-139,50	5,81	0,51	-158,56
21	-37,93	0,00	-4,21	49,93	31,98	4,10	4,13	4,75	0,86	0,18	-8,33
22	25,37	-0,07	23,41	283,78	19,16	10,42	2,77	-23,40	-0,13	0,20	-28,19
23	-64,62	0,12	12,24	45,06	128,97	-13,44	30,31	78,66	-0,94	0,66	-28,96
24	-214,40	0,04	4,27	468,24	72,48	5,57	65,81	84,56	2,49	0,98	-65,43
25	-38,88	0,42	-88,88	473,46	375,80	-33,65	-60,95	-13,12	0,79	0,26	-68,17
26	-200,62	0,03	6,15	41,61	7,20	4,20	113,95	90,94	2,55	0,67	-18,93
27	-270,77	0,04	-114,46	1148,69	13,40	47,66	-34,12	-144,21	8,50	0,59	-63,07
28	-83,54	0,19	-61,94	563,15	95,64	3,39	-30,77	-31,31	1,84	0,57	-102,50
29	-8,72	0,01	-18,46	24,64	107,79	2,90	-2,86	-30,63	0,64	0,25	-28,71
30	10,21	0,46	-103,53	564,16	261,97	4,07	-59,62	-66,39	-0,34	0,39	-152,70
31	-186,20	0,12	-115,68	897,78	110,61	35,74	-53,66	-157,94	6,24	0,38	44,50
32	44,14	0,18	-34,12	794,46	31,79	2,61	-25,78	-54,37	-0,54	0,26	-110,79
33	46,58	0,18	-72,51	80,69	228,97	0,62	-50,87	-104,32	-0,28	0,29	-20,86
34	-141,85	0,03	38,53	879,44	22,78	47,74	-12,57	-89,56	4,57	0,58	-120,28
35	-269,62	0,55	-255,12	1231,50	351,77	90,56	-114,48	-283,13	9,65	0,75	-282,69
36	-168,26	0,13	-22,50	645,41	144,51	44,14	-35,91	-136,98	6,35	0,23	-121,32
37	-158,82	0,33	-22,56	399,60	321,77	12,31	-27,44	-49,02	5,70	-0,23	-140,97
38	-410,99	0,39	-210,96	1123,75	267,01	23,60	-31,97	-114,86	8,37	0,95	203,52
39	-18,43	-0,03	1,45	14,75	12,00	2,21	12,06	1,31	0,58	0,14	-0,26
40	-14,87	0,29	-72,96	465,32	136,29	22,53	-45,35	-95,14	0,68	0,51	-91,11
41	-37,84	-0,03	4,99	8,09	27,83	1,63	2,64	-7,29	1,22	0,20	-1,33
42	-68,41	0,11	1,80	565,97	-8,12	-18,85	36,68	87,24	-1,41	0,65	74,23
43	-261,14	0,42	-286,29	1225,12	137,91	-4,22	-145,95	-249,21	9,61	0,75	-31,87
44	-199,91	0,01	-97,60	1689,61	121,46	54,53	-68,93	-195,47	7,81	0,82	-196,43
45	-61,86	0,03	-41,94	447,78	-10,66	9,37	-13,61	-51,32	2,19	0,04	79,74
46	2,89	0,68	-122,00	2373,84	84,35	-2,17	-83,21	-78,17	0,70	0,36	-345,90
47	-64,52	0,40	-94,24	1167,14	-8,32	10,39	-7,80	-81,15	1,23	0,72	-177,51
48	108,65	0,15	7,61	780,00	-14,97	-31,40	-7,77	64,35	-2,00	-0,47	-92,28
49	-17,71	-0,02	14,36	20,29	19,40	13,87	1,61	-2,90	0,55	0,12	-2,39
50	-98,63	0,25	-109,90	572,96	183,85	31,43	-56,78	-131,56	4,92	0,03	-133,30

- Estimation des moindres carrés, échantillon filtré sur les valeurs du salaire horaire inférieures au P99 :

échantillon	Constante	duree	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
Moyenne	6,78	0,00082	2,46	5,94	2,45	-0,27	-0,47	-0,39	0,05	0,00846	0,76
Variance	0,00405	0,00000	0,21937	0,00433	0,00230	0,00170	0,00175	0,00061	0,00000	0,00000	0,00073
CV	0,9%	7,8%	19,0%	1,1%	2,0%	-15,0%	-8,9%	-6,3%	1,5%	4,3%	3,6%
1	6,77	0,00076	2,10	5,87	2,36	-0,34	-0,53	-0,39	0,05	0,00874	0,75
2	6,78	0,00081	2,35	5,89	2,39	-0,31	-0,46	-0,37	0,05	0,00850	0,82
3	6,77	0,00076	1,82	5,96	2,45	-0,25	-0,47	-0,44	0,05	0,00842	0,77
4	6,76	0,00090	2,19	6,00	2,45	-0,25	-0,46	-0,43	0,05	0,00846	0,75
5	6,77	0,00073	2,12	5,91	2,48	-0,26	-0,47	-0,40	0,05	0,00907	0,73
6	6,83	0,00080	2,33	5,90	2,39	-0,28	-0,48	-0,42	0,05	0,00823	0,74
7	6,69	0,00077	3,06	5,93	2,49	-0,25	-0,43	-0,40	0,05	0,00879	0,75
8	6,91	0,00091	2,54	5,90	2,42	-0,29	-0,49	-0,40	0,05	0,00770	0,73
9	6,69	0,00075	2,15	6,02	2,48	-0,27	-0,42	-0,34	0,05	0,00883	0,77
10	6,85	0,00076	2,06	5,89	2,37	-0,35	-0,54	-0,39	0,04	0,00901	0,81
11	6,86	0,00071	3,13	5,96	2,43	-0,31	-0,48	-0,41	0,05	0,00833	0,77
12	6,75	0,00084	3,83	5,92	2,48	-0,21	-0,40	-0,38	0,05	0,00819	0,75
13	6,73	0,00083	1,78	5,91	2,46	-0,26	-0,42	-0,39	0,05	0,00901	0,71
14	6,79	0,00082	2,27	5,97	2,43	-0,34	-0,50	-0,33	0,05	0,00883	0,76
15	6,81	0,00082	2,18	5,99	2,46	-0,30	-0,49	-0,43	0,05	0,00861	0,73
16	6,80	0,00087	2,39	5,92	2,42	-0,25	-0,46	-0,41	0,05	0,00842	0,78
17	6,84	0,00081	2,04	5,86	2,49	-0,27	-0,47	-0,38	0,05	0,00788	0,76
18	6,76	0,00073	2,25	5,94	2,46	-0,27	-0,45	-0,40	0,05	0,00863	0,73
19	6,68	0,00087	2,27	5,90	2,55	-0,20	-0,42	-0,37	0,05	0,00868	0,79
20	6,80	0,00082	2,40	5,91	2,41	-0,29	-0,46	-0,35	0,05	0,00824	0,82
21	6,80	0,00083	3,03	5,86	2,43	-0,29	-0,48	-0,39	0,05	0,00850	0,78
22	6,72	0,00084	2,86	5,98	2,43	-0,26	-0,45	-0,38	0,05	0,00857	0,72
23	6,79	0,00088	2,74	5,85	2,44	-0,29	-0,50	-0,39	0,04	0,00902	0,79
24	6,85	0,00083	2,41	5,91	2,43	-0,27	-0,48	-0,42	0,05	0,00790	0,76
25	6,78	0,00080	2,08	5,95	2,40	-0,24	-0,46	-0,42	0,05	0,00888	0,77

échantillon	Constante	duree	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
26	6,74	0,00082	2,67	6,01	2,48	-0,23	-0,43	-0,40	0,05	0,00838	0,77
27	6,81	0,00070	2,82	5,88	2,45	-0,31	-0,52	-0,37	0,05	0,00868	0,73
28	6,89	0,00080	2,57	5,83	2,41	-0,32	-0,53	-0,41	0,05	0,00802	0,77
29	6,74	0,00084	2,52	5,99	2,49	-0,23	-0,45	-0,41	0,05	0,00859	0,75
30	6,69	0,00085	2,70	6,17	2,55	-0,18	-0,39	-0,40	0,05	0,00829	0,75
31	6,77	0,00084	2,50	5,97	2,44	-0,31	-0,45	-0,35	0,05	0,00832	0,76
32	6,79	0,00081	2,91	5,82	2,47	-0,27	-0,45	-0,37	0,05	0,00861	0,77
33	6,78	0,00088	3,28	6,03	2,41	-0,27	-0,47	-0,41	0,05	0,00825	0,79
34	6,77	0,00087	2,30	5,91	2,43	-0,26	-0,50	-0,40	0,05	0,00888	0,76
35	6,80	0,00086	2,37	5,97	2,50	-0,29	-0,45	-0,37	0,05	0,00817	0,76
36	6,70	0,00105	1,78	5,96	2,38	-0,29	-0,47	-0,36	0,05	0,00901	0,80
37	6,74	0,00077	2,50	6,03	2,51	-0,24	-0,42	-0,41	0,05	0,00822	0,76
38	6,94	0,00086	2,30	5,92	2,36	-0,39	-0,58	-0,39	0,05	0,00781	0,73
39	6,87	0,00080	2,09	5,81	2,38	-0,32	-0,55	-0,43	0,05	0,00822	0,81
40	6,72	0,00085	2,15	5,91	2,43	-0,26	-0,49	-0,39	0,05	0,00862	0,77
41	6,78	0,00077	1,75	5,97	2,43	-0,29	-0,51	-0,40	0,05	0,00873	0,74
42	6,76	0,00083	2,55	6,00	2,46	-0,28	-0,45	-0,38	0,05	0,00845	0,75
43	6,80	0,00083	2,06	5,98	2,47	-0,25	-0,46	-0,40	0,05	0,00827	0,81
44	6,81	0,00081	2,51	6,05	2,51	-0,25	-0,44	-0,38	0,05	0,00799	0,75
45	6,85	0,00088	3,94	5,94	2,38	-0,27	-0,48	-0,41	0,05	0,00815	0,73
46	6,88	0,00073	2,34	5,98	2,37	-0,34	-0,52	-0,38	0,05	0,00770	0,78
47	6,75	0,00073	2,56	6,02	2,52	-0,24	-0,46	-0,46	0,05	0,00861	0,70
48	6,83	0,00076	2,90	5,96	2,45	-0,26	-0,49	-0,40	0,05	0,00816	0,74
49	6,61	0,00094	2,75	5,94	2,45	-0,19	-0,37	-0,40	0,05	0,00907	0,76
50	6,71	0,00090	1,79	5,94	2,53	-0,25	-0,41	-0,39	0,05	0,00846	0,76

3.2. DFFITS

Échantillon	Nombre d'observations influentes	Pourcentage d'observations influentes
1	5 245	5,3
2	5 184	5,2
3	5 231	5,3
4	5 201	5,3
5	5 248	5,3
6	5 236	5,3
7	5 156	5,2
8	5 181	5,2
9	5 197	5,2
10	5 273	5,3
11	5 147	5,2
12	5 255	5,3
13	5 277	5,3
14	5 148	5,2
15	5 138	5,2
16	5 325	5,4
17	5 237	5,3
18	5 121	5,2
19	5 308	5,4
20	5 234	5,3
21	5 161	5,2
22	5 168	5,2
23	5 201	5,3
24	5 246	5,3
25	5 226	5,3
26	5 194	5,2
27	5 204	5,3
28	5 250	5,3
29	5 252	5,3
30	5 213	5,3
31	5 148	5,2
32	5 239	5,3
33	5 202	5,3
34	5 244	5,3
35	5 216	5,3
36	5 254	5,3
37	5 169	5,2
38	5 180	5,2
39	5 283	5,3

Échantillon	Nombre d'observations influentes	Pourcentage d'observations influentes
40	5 192	5,2
41	5 140	5,2
42	5 235	5,3
43	5 157	5,2
44	5 163	5,2
45	5 257	5,3
46	5 129	5,2
47	5 243	5,3
48	5 178	5,2
49	5 246	5,3
50	5 131	5,2

3.3. Tests de Hausman

Test sur l'échantillon filtré sur les valeurs du salaire horaire inférieures à P99 :

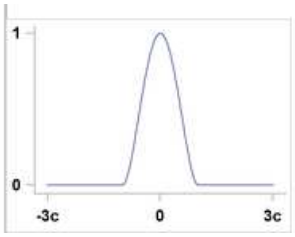
échantillon	valeur de la statistique d'Hausman : m	Modèle le plus efficient (valeur du Chi2 : 19,7)
1	11320,4	M
2	11204,3	M
3	11239,8	M
4	11123,4	M
5	11352,3	M
6	11234,4	M
7	11235,9	M
8	10962,0	M
9	10934,8	M
10	11388,8	M
11	11097,1	M
12	11582,8	M
13	11056,3	M
14	11275,0	M
15	11079,8	M
16	11235,0	M
17	10981,7	M
18	10880,7	M
19	11228,0	M
20	10878,3	M
21	11322,5	M
22	11381,6	M
23	11234,0	M
24	11202,5	M
25	11052,5	M
26	11231,6	M
27	10986,7	M
28	11861,4	M
29	11490,0	M
30	10820,9	M
31	10781,5	M
32	11169,4	M
33	11496,4	M
34	11201,8	M
35	11351,1	M
36	11201,5	M
37	11103,3	M
38	11176,2	M
39	11706,2	M
40	11257,4	M
41	10836,3	M
42	11476,7	M
43	11129,3	M
44	11034,8	M
45	11337,5	M
46	11348,0	M
47	11398,8	M
48	10695,1	M
49	11267,9	M
50	11277,1	M

4. Algorithmes SAS des méthodes robustes

4.1. M-estimation

L'algorithme de calcul de la M-estimation utilisé par la procédure ROBUSTREG de SAS est l'algorithme IRLS (iteratively reweighted least squares), qui consiste à réaliser une estimation des moindres carrés pondérés par itération. Les poids sont calculés, à chaque itération, en appliquant la fonction de poids choisie aux résidus obtenus lors de l'itération précédente.

La fonction de poids par défaut est la fonction Bisquare, définie par :



$$W(x,c)=\begin{cases}(1-(\frac{x}{c})^2)^2 & si\ |x|<c\\ 0 & si\ |x|\geq c\end{cases}$$, avec c une constante.

La valeur de c définie par défaut est : $c=4{,}685$. Elle est définie de façon à ce que le M-estimateur correspondant converge asymptotiquement à 95 % vers une loi gaussienne.

A la première itération, l'algorithme réalise une estimation OLS non pondérée. L'algorithme s'arrête lorsqu'un critère de convergence est satisfait, ou bien lorsque le nombre d'itérations atteint 1 000. Par défaut, le critère de convergence utilisé est le changement relatif des coefficients.

4.2. MM-estimation

La MM-estimation réalisée par la proc ROBUSTREG de SAS consiste en la procédure suivante :

1/ Calcul d'un estimateur de point de rupture initial convergent $\hat{\theta}_1$ de θ. Cet estimateur est calculé par défaut par la procédure ROBUSTREG de SAS avec la méthode LTS, en raison de sa rapidité ;

2/ Calcul de $\hat{\sigma}_1$ tel que :

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi(\frac{y_i-x_i'\hat{\theta}_1}{\hat{\sigma}_1})=\beta,$$

où $\beta=\int\chi(s)d\Phi(s)$ et χ est, par défaut, la fonction bisquare de Tukey définie par :

$$\chi_{k_0}(s)=\begin{cases}3(\frac{s}{k_0})^2-3(\frac{s}{k_0})^4+(\frac{s}{k_0})^6 & si\ |s|\leq k_0\\ 1 & si\ |s|>k_0\end{cases}$$

$k_0=2{,}9366$ par défaut, de façon à ce que l'estimateur asymptotiquement convergent $\hat{\sigma}_1$ ait un point de rupture de 25 %.

3/ Détermination du minimum local $\hat{\theta}_{MM}$:

$\min_{\theta} Q_{MM}(\theta)$, où $Q_{MM}(\theta)=\sum_{i=1}^{n}\rho(\frac{r_i}{\hat{\sigma}_1})$, tel que : $Q_{MM}(\hat{\theta}_{MM})\leq Q_{MM}(\hat{\theta}_1)$

Pour cette étape, la procédure ROBUSTREG utilise l'algorithme de la M-estimation.
La fonction ρ utilisée par défaut est la fonction de Tukey :

$$\rho(s)=\chi_{k_1}(s)=\begin{cases}3(\frac{s}{k_1})^2-3(\frac{s}{k_1})^4+(\frac{s}{k_1})^6 & si\ |s|\leq k_1\\ 1 & si\ |s|>k_1\end{cases}$$

$k_1=3{,}440$ par défaut, de façon à ce que le MM-estimateur correspondant converge asymptotiquement à 85 % vers une loi gaussienne.

5. Tests RDM et RDL

5.1. Tests sur les 50 échantillons de 100 000 périodes des salariés privés classiques

L'échantillon est filtré sur les valeurs du salaire horaire inférieures à P99.

- Résultats de la 1e étape :

échantillon	Points leviers		Poids associés	
	nombre	%	moyenne	CV
1	16 938	17,1%	0,74	24,5%
2	16 565	16,7%	0,75	24,2%
3	16 494	16,7%	0,74	25,1%
4	16 461	16,6%	0,76	24,6%
5	17 034	17,2%	0,74	24,5%
6	15 922	16,1%	0,74	23,9%
7	16 429	16,6%	0,73	29,8%
8	16 625	16,8%	0,75	24,2%
9	16 610	16,8%	0,75	24,6%
10	16 397	16,6%	0,75	24,1%
11	16 365	16,5%	0,73	24,2%
12	16 610	16,8%	0,74	24,3%
13	16 333	16,5%	0,73	24,0%
14	16 096	16,3%	0,74	23,8%
15	16 301	16,5%	0,75	24,0%
16	16 987	17,2%	0,74	24,7%
17	15 847	16,0%	0,73	24,1%
18	16 948	17,1%	0,74	24,5%
19	16 513	16,7%	0,75	24,0%
20	17 069	17,2%	0,74	24,5%
21	16 353	16,5%	0,74	24,2%
22	16 650	16,8%	0,75	24,2%
23	16 955	17,1%	0,74	24,4%
24	16 269	16,4%	0,74	24,1%
25	16 205	16,4%	0,73	23,7%
26	16 487	16,7%	0,75	24,0%
27	16 786	17,0%	0,74	24,6%
28	16 022	16,2%	0,74	24,2%
29	16 578	16,8%	0,74	24,3%
30	16 885	17,1%	0,74	24,8%
31	16 491	16,7%	0,74	24,3%
32	16 847	17,0%	0,74	24,8%
33	15 926	16,1%	0,74	23,8%
34	17 044	17,2%	0,74	24,6%
35	16 745	16,9%	0,75	24,1%
36	16 301	16,5%	0,74	24,1%
37	16 237	16,4%	0,73	25,2%
38	15 992	16,2%	0,74	23,8%
39	16 615	16,8%	0,75	23,8%
40	16 620	16,8%	0,74	24,5%
41	16 451	16,6%	0,75	24,0%
42	16 173	16,3%	0,73	23,9%
43	16 140	16,3%	0,75	24,7%
44	16 108	16,3%	0,75	24,2%
45	15 854	16,0%	0,74	23,9%
46	16 914	17,1%	0,74	24,4%
47	16 117	16,3%	0,74	24,1%
48	16 567	16,7%	0,75	24,0%
49	16 081	16,2%	0,74	24,2%
50	16 598	16,8%	0,74	24,5%
Moyenne	16 471	16,6%	0,74	
CV	2,06%		0,85%	

- Résultats de la 2e étape :

	nombre d'outliers			% d'outliers		
échantillon	RDM	RDL	communs	RDM	RDL	communs
1	6 718	7 233	6 659	6,8%	6,8%	6,7%
2	6 685	7 072	6 619	6,8%	7,1%	6,7%
3	6 545	6 837	6 486	6,6%	6,9%	6,6%
4	6 662	7 045	6 599	6,7%	7,1%	6,7%
5	6 761	7 197	6 671	6,8%	7,3%	6,7%
6	6 728	7 092	6 639	6,8%	7,2%	6,7%
7	6 786	7 093	6 688	6,9%	7,2%	6,8%
8	6 759	7 235	6 678	6,8%	7,3%	6,7%
9	6 781	7 174	6 691	6,8%	7,2%	6,8%
10	6 636	7 071	6 570	6,7%	7,1%	6,6%
11	6 760	7 110	6 683	6,8%	7,2%	6,8%
12	6 753	7 138	6 656	6,8%	7,2%	6,7%
13	6 640	7 023	6 570	6,7%	7,1%	6,6%
14	6 627	7 000	6 536	6,7%	7,1%	6,6%
15	6 561	6 898	6 450	6,6%	7,0%	6,5%
16	6 748	7 198	6 680	6,8%	7,3%	6,7%
17	6 688	7 133	6 612	6,8%	7,2%	6,7%
18	6 615	7 072	6 563	6,7%	7,1%	6,6%
19	6 652	7 099	6 577	6,7%	7,2%	6,6%
20	6 596	7 047	6 538	6,7%	7,1%	6,6%
21	6 691	7 108	6 630	6,8%	7,2%	6,7%
22	6 731	7 184	6 657	6,8%	7,3%	6,7%
23	6 755	7 206	6 693	6,8%	7,3%	6,8%
24	6 699	7 073	6 623	6,8%	7,1%	6,7%
25	6 692	7 066	6 620	6,8%	7,1%	6,7%
26	6 874	7 295	6 804	6,9%	7,4%	6,9%
27	6 658	7 068	6 564	6,7%	7,1%	6,6%
28	6 648	7 063	6 557	6,7%	7,1%	6,6%
29	6 873	7 322	6 801	6,9%	7,4%	6,9%
30	6 706	7 162	6 650	6,8%	7,2%	6,7%
31	6 684	7 075	6 631	6,8%	7,1%	6,7%
32	6 590	7 009	6 506	6,7%	7,1%	6,6%
33	6 763	7 131	6 686	6,8%	7,2%	6,8%
34	6 761	7 138	6 683	6,8%	7,2%	6,8%
35	6 751	7 137	6 644	6,8%	7,2%	6,7%
36	6 748	7 244	6 706	6,8%	7,3%	6,8%
37	6 631	6 901	6 506	6,7%	7,0%	6,6%
38	6 786	7 234	6 722	6,9%	7,3%	6,8%
39	6 838	7 222	6 764	6,9%	7,3%	6,8%
40	6 683	7 055	6 610	6,8%	7,1%	6,7%
41	6 775	7 136	6 688	6,8%	7,2%	6,8%
42	6 600	6 956	6 553	6,7%	7,0%	6,6%
43	6 820	7 166	6 754	6,9%	7,2%	6,8%
44	6 664	6 967	6 566	6,7%	7,0%	6,6%
45	6 720	7 194	6 667	6,8%	7,3%	6,7%
46	6 672	7 100	6 621	6,7%	7,2%	6,7%
47	6 755	7 162	6 683	6,8%	7,2%	6,8%
48	6 502	6 891	6 415	6,6%	7,0%	6,5%
49	6 764	7 164	6 683	6,8%	7,2%	6,8%
50	6 772	7 074	6 678	6,8%	7,1%	6,7%
Moyenne	6 706	7 105	6 631	6,8%	7,2%	6,7%

5.2. Tests sur les codes population

- Classes utilisées pour la modélisation des salariés privés classiques :

  - Hommes :

CS1	Regroupement	CS2	Nombre de périodes
1 Agriculteurs exploitants / 2 Artisans, commerçants et chefs d'entreprise			50 659
3 Cadres et professions intellectuelles supérieures			964 058
4 Professions Intermédiaires			2 268 113
5 Employés			3 414 696
6 Ouvriers	ensemble		7 548 184
	Ouvriers qualifiés	ensemble	5 064 075
		62 - Ouvriers qualifiés de type industriel	1 872 464
		63 - Ouvriers qualifiés de type artisanal	1 687 382
		64 - Chauffeurs	822 007
		65 - Ouvriers qualifiés de la manutention, du magasinage et du transport	682 222
	Ouvriers non qualifiés	ensemble	2 484 109
		66 - Ouvriers non qualifiés de type industriel	1 483 571
		67 - Ouvriers non qualifiés de type artisanal	938 826
		68 - Ouvriers agricoles et assimilés	61 712
9 Autres (non renseigné)			1 265 911
Total			15 511 621

  - Femmes :

CS1	CS2	Nombre de périodes
1 Agriculteurs exploitants / 2 Artisans, commerçants et chefs d'entreprise		20 320
3 Cadres et professions intellectuelles supérieures		734 584
4 Professions Intermédiaires		2 419 255
5 Employés	ensemble	8 308 844
	52 - Employés civils et agents de service de la fonction publique	1 288 380
	53 - Agents de surveillance	56 775
	54 - Employés administratifs d'entreprise	2 387 718
	55 - Employés de commerce	1 959 908
	56 - Personnels des services directs aux particuliers	2 616 063
6 Ouvriers		2 461 119
9 Autres (non renseigné)		1 673 449
Total		15 617 571

- Résultats de la 1e étape :

code population	sexe	CS1	CS2 ou regroupement	nombre de périodes	nombre de points leviers	%	Poids moyen des points leviers
10 - salarié privé classique	**ensemble**			**31 129 192**	**10 236 592**	**32,9%**	**0,49**
	femmes	ensemble		15 617 571	5 448 479	34,9%	0,41
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	20 320	7 038	34,6%	0,27
		3 - Cadres et professions intellectuelles supérieures	ensemble	734 584	260 546	35,5%	0,61
		4 - Professions Intermédiaires	ensemble	2 419 255	901 876	37,3%	0,23
		5 - Employés	ensemble	8 308 844	2 978 320	35,8%	0,47
			52 - Employés civils et agents de service de la fonction publique	1 288 380	432 372	33,6%	0,30
			53 - Agents de surveillance	56 775	12 681	22,3%	0,22
			54 - Employés administratifs d'entreprise	2 387 718	721 357	30,2%	0,62
			55 - Employés de commerce	1 959 908	723 137	36,9%	0,39
			56 - Personnels des services directs aux particuliers	2 616 063	1 088 773	41,6%	0,48
		6 - Ouvriers	ensemble	2 461 119	815 075	33,1%	0,36
		9 - Autres (non renseigné)	ensemble	1 673 449	485 624	29,0%	0,37
	hommes	ensemble		15 511 621	4 788 113	30,9%	0,57
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	50 659	14 262	28,2%	0,36
		3 - Cadres et professions intellectuelles supérieures	ensemble	964 058	366 948	38,1%	0,50
		4 - Professions Intermédiaires	ensemble	2 268 113	733 492	32,3%	0,63
		5 - Employés	ensemble	3 414 696	1 151 364	33,7%	0,58
		6 - Ouvriers	ensemble	7 548 184	2 014 382	26,7%	0,59
			ouvriers qualifiés	5 064 075	1 213 420	24,0%	0,68
			ouvriers non qualifiés	2 484 109	800 962	32,2%	0,40
		9 - Autres (non renseigné)	ensemble	1 265 911	507 665	40,1%	0,30
11 - salarié du spectacle	**ensemble**			**686 199**	**203 869**	**29,7%**	**0,35**
13 - salarié privé dans le public	**ensemble**			**264 074**	**105 796**	**40,1%**	**0,68**
14 - fonctionnaire détaché	**ensemble**			**8 564**	**1 655**	**19,3%**	**0,69**
40 - fonctionnaire	**ensemble**			**4 143 185**	**1 540 551**	**37,2%**	**0,51**
	femmes	ensemble		2 771 199	958 827	34,6%	0,59
	hommes	ensemble		1 371 986	481 724	42,4%	0,32
43 - non-titulaire de la fonction publique	**ensemble**			**2 363 599**	**573 363**	**24,3%**	**0,67**
Total				**38 594 813**	**12 561 826**	**32,5%**	**0,50**

- Résultats de la 2e étape :

code population	sexe	CS1	CS2 ou regroupement	nombre de périodes	nombre d'outliers			% d'outliers		
					RDM	RDL	communs	RDM	RDL	communs
10 - salarié privé classique	**ensemble**			**31 129 192**	**1 718 565**	**1 906 986**	**1 679 488**	**5,5%**	**6,1%**	**5,4%**
	femmes	ensemble		911 021	1 052 514	872 495	5,8%	6,7%	5,6%	5,7%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	20 320	153	318	152	0,8%	1,6%	0,7%
		3 - Cadres et professions intellectuelles supérieures	ensemble	734 584	907	1 579	907	0,1%	0,2%	0,1%
		4 - Professions Intermédiaires	ensemble	2 419 255	55 410	56 824	54 386	2,3%	2,3%	2,2%
		5 - Employés	ensemble	8 308 844	509 020	635 207	471 523	6,1%	7,6%	5,7%
			52 - Employés civils et agents de service de la fonction publique	1 288 380	51 184	172 063	21 160	4,0%	13,4%	1,6%
			53 - Agents de surveillance	56 775	2 567	2 549	2 517	4,5%	4,5%	4,4%
			54 - Employés administratifs d'entreprise	2 387 718	111 719	118 940	111 719	4,7%	5,0%	4,7%
			55 - Employés de commerce	1 959 908	142 177	139 783	139 008	7,3%	7,1%	7,1%
			56 - Personnels des services directs aux particuliers	2 616 063	201 373	201 872	197 119	7,7%	7,7%	7,5%
		6 - Ouvriers	ensemble	2 461 119	160 536	166 673	160 536	6,5%	6,8%	6,5%
		9 - Autres (non renseigné)	ensemble	1 673 449	184 995	191 913	184 991	11,1%	11,5%	11,1%
	hommes	ensemble		15 511 621	807 544	854 472	806 993	5,2%	5,5%	5,2%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	50 659	55	169	55	0,1%	0,3%	0,1%
		3 - Cadres et professions intellectuelles supérieures	ensemble	964 058	1 490	2 865	1 490	0,2%	0,3%	0,2%
		4 - Professions Intermédiaires	ensemble	2 268 113	41 714	53 461	41 714	1,8%	2,4%	1,8%
		5 - Employés	ensemble	3 414 696	254 912	261 167	254 732	7,5%	7,6%	7,5%
		6 - Ouvriers	ensemble	7 548 184	374 299	393 473	373 929	5,0%	5,2%	5,0%
			ouvriers qualifiés	5 064 075	232 691	247 253	232 688	4,6%	4,9%	4,6%
			ouvriers non qualifiés	2 484 109	141 608	146 220	141 241	5,7%	5,9%	5,7%
		9 - Autres (non renseigné)	ensemble	1 265 911	135 074	143 337	135 073	10,7%	11,3%	10,7%
11 - salarié du spectacle	**ensemble**			**686 199**	**8 056**	**12 343**	**8 056**	**1,2%**	**1,8%**	**1,2%**
13 - salarié privé dans le public	**ensemble**			**264 074**	**51 818**	**42 002**	**42 002**	**19,6%**	**15,9%**	**15,9%**
14 - fonctionnaire détaché	**ensemble**			**8 564**	**395**	**399**	**384**	**4,6%**	**4,7%**	**4,5%**
40 - fonctionnaire	**ensemble**			**4 143 185**	**73 822**	**91 327**	**73 202**	**1,8%**	**2,2%**	**1,8%**
	femmes	ensemble		2 771 199	54 152	68 096	53 778	2,0%	2,5%	1,9%
	hommes	ensemble		1 371 986	19 670	23 231	19 424	1,4%	1,7%	1,4%
43 - non-titulaire de la fonction publique	**ensemble**			**2 363 599**	**279 896**	**282 649**	**254 530**	**11,8%**	**12,0%**	**10,8%**
Total				**38 594 813**	**2 132 552**	**2 335 706**	**2 057 662**	**5,5%**	**6,1%**	**5,3%**

6. Comparaison des diagnostics des méthodes RDM et M

6.1. Tests sur les 50 échantillons de 100 000 périodes des salariés privés classiques

L'échantillon est filtré sur les valeurs du salaire horaire inférieures à P99.

- Nombre d'outliers :

	nombre d'outliers			% d'outliers		
échantillon	RDM	M	communs	RDM	M	communs
1	6 718	6 906	6 718	6,8%	6,8%	6,8%
2	6 685	6 849	6 685	6,8%	6,9%	6,8%
3	6 545	6 753	6 531	6,6%	6,8%	6,6%
4	6 662	6 870	6 649	6,7%	6,9%	6,7%
5	6 761	6 917	6 761	6,8%	7,0%	6,8%
6	6 728	6 883	6 727	6,8%	7,0%	6,8%
7	6 786	6 996	6 771	6,9%	7,1%	6,8%
8	6 759	6 886	6 759	6,8%	7,0%	6,8%
9	6 781	6 958	6 778	6,8%	7,0%	6,8%
10	6 636	6 846	6 618	6,7%	6,9%	6,7%
11	6 760	6 975	6 745	6,8%	7,0%	6,8%
12	6 753	6 872	6 753	6,8%	6,9%	6,8%
13	6 640	6 876	6 627	6,7%	6,9%	6,7%
14	6 627	6 775	6 627	6,7%	6,8%	6,7%
15	6 561	6 766	6 539	6,6%	6,8%	6,6%
16	6 748	6 888	6 746	6,8%	7,0%	6,8%
17	6 688	6 836	6 688	6,8%	6,9%	6,8%
18	6 615	6 776	6 615	6,7%	6,8%	6,7%
19	6 652	6 784	6 652	6,7%	6,9%	6,7%
20	6 596	6 768	6 596	6,7%	6,8%	6,7%
21	6 691	6 858	6 691	6,8%	6,9%	6,8%
22	6 731	6 906	6 730	6,8%	7,0%	6,8%
23	6 755	6 925	6 754	6,8%	7,0%	6,8%
24	6 699	6 919	6 687	6,8%	7,0%	6,8%
25	6 692	6 899	6 672	6,8%	7,0%	6,7%
26	6 874	7 071	6 874	6,9%	7,1%	6,9%
27	6 658	6 824	6 656	6,7%	6,9%	6,7%
28	6 648	6 806	6 648	6,7%	6,9%	6,7%
29	6 873	6 994	6 871	6,9%	7,1%	6,9%
30	6 706	6 847	6 705	6,8%	6,9%	6,8%
31	6 684	6 854	6 683	6,8%	6,9%	6,8%
32	6 590	6 738	6 588	6,7%	6,8%	6,7%
33	6 763	6 909	6 763	6,8%	7,0%	6,8%
34	6 761	6 937	6 761	6,8%	7,0%	6,8%
35	6 751	6 887	6 751	6,8%	7,0%	6,8%
36	6 748	6 925	6 748	6,8%	7,0%	6,8%
37	6 631	6 862	6 618	6,7%	6,9%	6,7%
38	6 786	6 946	6 786	6,9%	7,0%	6,9%
39	6 838	6 971	6 830	6,9%	7,0%	6,9%
40	6 683	6 819	6 683	6,8%	6,9%	6,8%
41	6 775	6 893	6 773	6,8%	7,0%	6,8%
42	6 600	6 798	6 583	6,7%	6,9%	6,6%
43	6 820	7 027	6 803	6,9%	7,1%	6,9%
44	6 664	6 887	6 653	6,7%	7,0%	6,7%
45	6 720	6 911	6 720	6,8%	7,0%	6,8%
46	6 672	6 829	6 672	6,7%	6,9%	6,7%
47	6 755	6 872	6 754	6,8%	6,9%	6,8%
48	6 502	6 620	6 502	6,6%	6,7%	6,6%
49	6 764	6 900	6 762	6,8%	7,0%	6,8%
50	6 772	6 939	6 770	6,8%	7,0%	6,8%
Moyenne	6 706	6 875	6 702	6,8%	6,9%	6,8%

- Salaire horaire moyen par type de périodes :

échantillon	RDM		M	
	outliers	périodes OK	outliers	Dont outliers non communs avec RDM
1	18	9,1	17,9	15
2	17,9	9,2	17,9	14,9
3	18,1	9,2	18	15,9
4	18	9,2	17,9	15,3
5	17,9	9,2	17,9	15,1
6	17,8	9,2	17,8	14,6
7	18	9,2	17,9	15,8
8	18,1	9,2	18	14,5
9	18,1	9,2	18	14,2
10	18	9,2	17,9	15,2
11	18,1	9,2	18	15,3
12	17,9	9,2	17,9	14,5
13	17,9	9,2	17,8	15,5
14	18,1	9,2	18	14,8
15	18,2	9,2	18,1	14,9
16	17,9	9,2	17,8	14,6
17	18	9,2	18	14,6
18	18	9,2	17,9	14,9
19	18,1	9,2	18	14,5
20	18,1	9,2	18	14,6
21	17,9	9,2	17,8	14,4
22	18	9,2	17,9	14,9
23	17,9	9,2	17,9	15,2
24	18	9,2	17,9	15
25	17,9	9,2	17,8	15
26	18	9,2	17,9	15,4
27	18,1	9,2	18	14,3
28	18,1	9,2	18	14,4
29	18	9,2	17,9	14,6
30	17,9	9,2	17,8	14,2
31	18	9,2	17,9	14,4
32	17,9	9,2	17,8	14,4
33	17,9	9,2	17,8	14,7
34	18,1	9,2	18	14,9
35	18,1	9,2	18	14,5
36	18	9,2	17,9	14,7
37	17,9	9,2	17,8	14,9
38	18	9,2	18	14,6
39	18	9,2	17,9	14
40	18	9,2	17,9	14,6
41	18	9,2	17,9	14,5
42	18	9,2	17,9	15,3
43	18,2	9,2	18,1	15,5
44	18	9,2	17,9	15,3
45	17,9	9,2	17,9	15
46	18,1	9,2	18	14,9
47	17,9	9,2	17,9	14,4
48	18	9,2	17,9	15
49	17,9	9,2	17,8	14,4
50	18	9,2	18	15,5
moyenne	18	9,2	17,9	14,8
CV	0,40%	0,20%	0,40%	2,90%

- Intervalles de confiance de la moyenne du salaire horaire par type de période :

échantillon	Type de périodes					
	périodes OK		Outliers M et OK RDM		Outliers communs	
	borne inférieure de l'IC	borne supérieure de l'IC	borne inférieure de l'IC	borne supérieure de l'IC	borne inférieure de l'IC	borne supérieure de l'IC
1	9,11	9,15	14,63	15,38	17,91	18,11
2	9,16	9,19	14,48	15,22	17,85	18,05
3	9,15	9,18	15,56	16,31	17,99	18,18
4	9,14	9,17	14,88	15,72	17,89	18,09
5	9,14	9,17	14,64	15,48	17,81	18,01
6	9,15	9,19	14,19	14,95	17,75	17,95
7	9,13	9,17	15,35	16,16	17,89	18,09
8	9,15	9,19	14,11	14,97	17,99	18,20
9	9,17	9,20	13,76	14,55	18,02	18,22
10	9,16	9,19	14,82	15,57	17,89	18,09
11	9,14	9,17	14,91	15,63	18,02	18,22
12	9,16	9,19	14,13	14,87	17,81	18,01
13	9,14	9,18	15,09	15,85	17,80	18,00
14	9,18	9,21	14,37	15,16	18,00	18,21
15	9,17	9,20	14,52	15,30	18,09	18,30
16	9,14	9,17	14,12	15,03	17,78	17,98
17	9,15	9,19	14,20	14,98	17,93	18,13
18	9,16	9,19	14,50	15,29	17,86	18,07
19	9,14	9,18	14,01	15,01	17,95	18,15
20	9,17	9,20	14,19	15,09	17,96	18,16
21	9,15	9,18	13,99	14,86	17,83	18,04
22	9,13	9,16	14,42	15,32	17,93	18,13
23	9,14	9,17	14,75	15,65	17,83	18,03
24	9,14	9,17	14,65	15,40	17,92	18,12
25	9,15	9,18	14,62	15,39	17,84	18,05
26	9,13	9,16	15,04	15,77	17,90	18,10
27	9,17	9,20	13,96	14,68	17,95	18,16
28	9,14	9,18	13,97	14,77	17,97	18,17
29	9,14	9,17	14,20	15,01	17,86	18,06
30	9,16	9,19	13,78	14,69	17,76	17,96
31	9,16	9,19	14,00	14,83	17,88	18,09
32	9,16	9,20	13,95	14,82	17,76	17,96
33	9,15	9,18	14,23	15,10	17,82	18,02
34	9,13	9,16	14,56	15,31	17,96	18,16
35	9,14	9,17	14,02	14,93	17,97	18,16
36	9,16	9,19	14,30	15,14	17,88	18,08
37	9,16	9,20	14,58	15,20	17,85	18,06
38	9,18	9,21	14,16	14,97	17,94	18,15
39	9,13	9,16	13,58	14,43	17,85	18,05
40	9,13	9,17	14,18	14,93	17,90	18,10
41	9,16	9,19	14,03	14,90	17,90	18,11
42	9,16	9,19	14,92	15,64	17,91	18,11
43	9,13	9,16	15,18	15,87	18,13	18,33
44	9,14	9,18	14,92	15,62	17,90	18,11
45	9,13	9,16	14,67	15,35	17,84	18,04
46	9,15	9,19	14,42	15,29	17,97	18,18
47	9,13	9,16	13,87	14,84	17,83	18,04
48	9,18	9,22	14,42	15,63	17,89	18,10
49	9,13	9,17	14,01	14,72	17,82	18,02
50	9,15	9,18	15,12	15,97	17,93	18,13

6.2. Tests sur l'ensemble de la population

- Nombre d'outliers :

code population	sexe	CS1	CS2 ou regroupement	nombre de périodes	nombre d'outliers			% d'outliers		
					RDM	M	communs	RDM	M	communs
10 - salarié privé classique	**ensemble**			**31 129 192**	**1 718 565**	**1 717 740**	**1 693 120**	**5,5%**	**5,5%**	**5,4%**
	femmes	ensemble		15 617 571	911 021	906 257	894 465	5,8%	5,8%	5,7%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	20 320	153	172	131	0,8%	0,8%	0,6%
		3 - Cadres et professions intellectuelles supérieures	ensemble	734 584	907	862	824	0,1%	0,1%	0,1%
		4 - Professions Intermédiaires	ensemble	2 419 255	55 410	54 901	54 386	2,3%	2,3%	2,2%
		5 - Employés	ensemble	8 308 844	509 020	508 323	499 841	6,1%	6,1%	6,0%
			52 - Employés civils et agents de service de la fonction publique	1 288 380	51 184	49 499	49 292	4,0%	3,8%	3,8%
			53 - Agents de surveillance	56 775	2 567	2 899	2 567	4,5%	5,1%	4,5%
			54 - Employés administratifs d'entreprise	2 387 718	111 719	115 283	111 468	4,7%	4,8%	4,7%
			55 - Employés de commerce	1 959 908	142 177	143 737	141 025	7,3%	7,3%	7,2%
			56 - Personnels des services directs aux particuliers	2 616 063	201 373	196 905	195 489	7,7%	7,5%	7,5%
		6 - Ouvriers	ensemble	2 461 119	160 536	158 213	155 523	6,5%	6,4%	6,3%
		9 - Autres (non renseigné)	ensemble	1 673 449	184 995	183 786	183 760	11,1%	11,0%	11,0%
	hommes	ensemble		15 511 621	807 544	811 483	798 655	5,2%	5,2%	5,1%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	50 659	55	24	19	0,1%	0,0%	0,0%
		3 - Cadres et professions intellectuelles supérieures	ensemble	964 058	1 490	1 592	1 444	0,2%	0,2%	0,1%
		4 - Professions Intermédiaires	ensemble	2 268 113	41 714	41 441	40 080	1,8%	1,8%	1,8%
		5 - Employés	ensemble	3 414 696	254 912	259 558	253 824	7,5%	7,6%	7,4%
		6 - Ouvriers	ensemble	7 548 184	374 299	379 879	374 299	5,0%	5,0%	5,0%
			ouvriers qualifiés	5 064 075	232 691	237 642	232 691	4,6%	4,7%	4,6%
			ouvriers non qualifiés	2 484 109	141 608	142 237	141 608	5,7%	5,7%	5,7%
		9 - Autres (non renseigné)	ensemble	1 265 911	135 074	128 989	128 989	10,7%	10,2%	10,2%
11 - salarié du spectacle	**ensemble**			**686 199**	**8 056**	**8 735**	**8 056**	**1,2%**	**1,3%**	**1,2%**
13 - salarié privé dans le public	**ensemble**			**264 074**	**51 818**	**91 779**	**51 807**	**19,6%**	**34,8%**	**19,6%**
14 - fonctionnaire détaché	**ensemble**			**8 564**	**395**	**397**	**395**	**4,6%**	**4,6%**	**4,6%**
40 - fonctionnaire	**ensemble**			**4 143 185**	**73 822**	**80 276**	**71 971**	**1,8%**	**1,9%**	**1,7%**
	femmes	ensemble		2 771 199	54 152	59 497	52 885	2,0%	2,1%	1,9%
	hommes	ensemble		1 371 986	19 670	20 779	19 086	1,4%	1,5%	1,4%
43 - non-titulaire de la fonction publique	**ensemble**			**2 363 599**	**279 896**	**282 299**	**279 375**	**11,8%**	**11,9%**	**11,8%**
Total				**38 594 813**	**2 132 552**	**2 181 226**	**2 104 724**	**5,5%**	**5,7%**	**5,5%**

- Salaire horaire moyen par type de périodes :

code population	sexe	CS1	CS2 ou regroupement	outliers RDM	périodes OK RDM	outliers M	dont outliers M et OK RDM
10 - salarié privé classique	femmes	1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	24,7	11,2	24,8	11,2
		3 - Cadres et professions intellectuelles supérieures	ensemble	23,6	15,2	24,9	15,2
		4 - Professions Intermédiaires	ensemble	25,8	11,6	23,0	11,4
		5 - Employés	52 - Employés civils et agents de service de la fonction publique	16,2	9,0	16,3	9,1
			53 - Agents de surveillance	16,9	8,8	16,5	8,8
			54 - Employés administratifs d'entreprise	18,7	9,3	18,6	9,3
			55 - Employés de commerce	14,0	7,9	14,0	7,9
			56 - Personnels des services directs aux particuliers	12,9	7,7	13,0	7,7
		6 - Ouvriers	ensemble	14,3	7,9	14,4	8,0
		9 - Autres (non renseigné)	ensemble	17,7	8,3	17,8	8,3
	hommes	1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	25,9	12,4	26,7	12,4
		3 - Cadres et professions intellectuelles supérieures	ensemble	17,3	16,0	17,9	16,0
		4 - Professions Intermédiaires	ensemble	23,4	12,0	23,5	12,0
		5 - Employés	ensemble	16,5	8,4	16,4	8,4
		6 - Ouvriers	ouvriers qualifiés	17,9	9,0	17,8	9,0
			ouvriers non qualifiés	14,5	8,1	14,4	8,1
		9 - Autres (non renseigné)	ensemble	17,9	8,4	18,1	8,4
11 - salarié du spectacle	ensemble	ensemble	ensemble	37,3	12,8	36,6	12,8
13 - salarié privé dans le public	ensemble	ensemble	ensemble	8,0	7,6	8,8	7,0
14 - fonctionnaire détaché	ensemble	ensemble	ensemble	45,5	16,4	45,5	16,4
40 - fonctionnaire	femmes	ensemble	ensemble	12,7	12,7	12,7	12,7
	hommes	ensemble	ensemble	18,4	12,7	18,6	12,7
43 - non-titulaire de la fonction publique	ensemble	ensemble	ensemble	19,5	9,1	19,4	9,1
moyenne				19,8	10,5	19,8	10,5
CV				42,0%	27,0%	41,3%	27,3%

- Intervalles de confiance de la moyenne du salaire horaire par type de période :

code population	sexe	CS1	CS2 ou regroupement	type							
				périodes OK		M hors RDM		RDM hors M		communs	
				borne inférieure de l'IC	borne supérieure de l'IC	borne inférieure de l'IC	borne supérieure de l'IC	borne inférieure de l'IC	borne supérieure de l'IC	borne inférieure de l'IC	borne supérieure de l'IC
10 - salarié privé classique	femmes	1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	11,1	11,2	23,3	24,1	21,5	22,6	24,8	25,4
10 - salarié privé classique	femmes	3 - Cadres et professions intellectuelles supérieures	ensemble	15,2	15,2	26,7	27,2	9,3	12,6	24,4	25,3
10 - salarié privé classique	femmes	4 - Professions Intermédiaires	ensemble	11,4	11,4	22,7	22,7	11,5	13,0	26,0	26,1
10 - salarié privé classique	femmes	5 - Employés	52 - Employés civils et agents de service de la fonction publique	9,0	9,0	**13,1**	**13,2**	**13,0**	**13,1**	16,2	16,3
10 - salarié privé classique	femmes	5 - Employés	53 - Agents de surveillance	8,7	8,8	13,6	13,6			16,8	17,0
10 - salarié privé classique	femmes	5 - Employés	54 - Employés administratifs d'entreprise	9,3	9,3	16,2	16,2	13,7	14,1	18,7	18,7
10 - salarié privé classique	femmes	5 - Employés	55 - Employés de commerce	7,8	7,9	11,1	11,1	10,6	10,7	14,0	14,1
10 - salarié privé classique	femmes	5 - Employés	56 - Personnels des services directs aux particuliers	7,7	7,7	10,3	10,4	10,2	10,2	13,0	13,0
10 - salarié privé classique	femmes	6 - Ouvriers	ensemble	7,9	7,9	11,5	11,5	11,0	11,0	14,4	14,4
10 - salarié privé classique	femmes	9 - Autres (non renseigné)	ensemble	8,2	8,3	13,3	13,5	13,2	13,3	17,7	17,8
10 - salarié privé classique	hommes	1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	12,4	12,5	**25,5**	**28,2**	**24,9**	**26,1**	26,0	27,3
10 - salarié privé classique	hommes	3 - Cadres et professions intellectuelles supérieures	ensemble	15,9	16,0	23,3	25,5	16,0	21,7	16,7	17,8
10 - salarié privé classique	hommes	4 - Professions Intermédiaires	ensemble	12,0	12,0	23,0	23,1	20,0	20,3	23,5	23,5
10 - salarié privé classique	hommes	5 - Employés	ensemble	8,4	8,4	13,4	13,4	12,5	12,5	16,5	16,5
10 - salarié privé classique	hommes	6 - Ouvriers	ouvriers qualifiés	9,0	9,0	15,0	15,0			17,8	17,9
10 - salarié privé classique	hommes	6 - Ouvriers	ouvriers non qualifiés	8,1	8,1	11,7	11,7			14,4	14,5
10 - salarié privé classique	hommes	9 - Autres (non renseigné)	ensemble	8,4	8,4			14,0	14,0	18,1	18,1
11 - salarié du spectacle	ensemble	ensemble	ensemble	12,8	12,8	28,1	28,3			37,1	37,4
13 - salarié privé dans le public	ensemble	ensemble	ensemble	7,0	7,0	9,9	9,9	7,1	7,2	8,0	8,0
14 - fonctionnaire détaché	ensemble	ensemble	ensemble	16,3	16,5	26,1	36,3			44,4	46,7
40 - fonctionnaire	femmes	ensemble	ensemble	12,0	12,0	20,2	20,3	8,5	8,7	19,2	19,3
40 - fonctionnaire	hommes	ensemble	ensemble	12,2	12,2	**12,0**	**12,3**	**11,7**	**12,2**	12,0	12,1
43 - non-titulaire de la fonction publique	ensemble	ensemble	ensemble	9,1	9,1	11,1	11,3	15,8	16,5	19,4	19,5

7. Répartition des outliers détectés par la méthode RDM par type de déclaration

code population	sexe	CS1	CS2 ou regroupement	nombre d'outliers RDM	Type de déclaration			
					déclarations Siasp	grandes déclarations DADS	moyennes déclarations DADS	petites déclarations DADS
10 - salarié privé classique	**ensemble**			**1 718 565**	**1,2%**	**45,6%**	**41,4%**	**11,7%**
	femmes	ensemble		911 021	1,4%	42,3%	42,1%	14,1%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	153	1,8%	3,8%	47,2%	47,2%
		3 - Cadres et professions intellectuelles supérieures	ensemble	907	2,0%	42,2%	42,4%	13,5%
		4 - Professions Intermédiaires	ensemble	55 410	1,7%	43,7%	44,7%	9,8%
		5 - Employés	ensemble	509 020	2,0%	44,9%	39,1%	14,1%
			52 - Employés civils et agents de service de la fonction publique	51 184	1,3%	63,3%	31,3%	4,2%
			53 - Agents de surveillance	2 567	1,8%	61,0%	31,6%	5,7%
			54 - Employés administratifs d'entreprise	111 719	7,2%	43,4%	38,1%	11,3%
			55 - Employés de commerce	142 177	0,1%	61,7%	29,5%	8,8%
			56 - Personnels des services directs aux particuliers	201 373	0,6%	29,0%	48,5%	21,9%
		6 - Ouvriers	ensemble	160 536	0,2%	70,6%	24,8%	4,4%
		9 - Autres (non renseigné)	ensemble	184 995	0,8%	10,4%	64,9%	24,0%
	hommes	ensemble		807 544	1,0%	49,3%	40,6%	9,0%
		1 - Agriculteurs exploitants / 2 - Artisans, commerçants et chefs d'entreprise	ensemble	55	2,3%	3,1%	58,2%	36,4%
		3 - Cadres et professions intellectuelles supérieures	ensemble	1 490	1,3%	23,2%	44,7%	30,7%
		4 - Professions Intermédiaires	ensemble	41 714	3,0%	48,1%	44,3%	4,6%
		5 - Employés	ensemble	254 912	2,1%	43,7%	44,5%	9,6%
		6 - Ouvriers	ensemble	374 299	0,2%	65,6%	29,8%	4,4%
			ouvriers qualifiés	232 691	0,3%	60,5%	34,3%	4,9%
			ouvriers non qualifiés	141 608	0,1%	74,1%	22,2%	3,6%
		9 - Autres (non renseigné)	ensemble	135 074	0,7%	15,4%	62,0%	21,9%
11 - salarié du spectacle	**ensemble**			**8 056**	**2,4%**	**41,6%**	**53,2%**	**2,7%**
13 - salarié privé dans le public	**ensemble**			**51 818**	**94,0%**	**1,6%**	**3,9%**	**0,6%**
14 - fonctionnaire détaché	**ensemble**			**395**	**0,5%**	**22,0%**	**76,2%**	**1,3%**
40 - fonctionnaire	**ensemble**			**73 822**	**97,4%**	**2,4%**	**0,1%**	**0,0%**
	femmes	ensemble		54 152	97,4%	2,6%	0,1%	0,0%
	hommes	ensemble		19 670	97,7%	2,1%	0,3%	0,0%
43 - non-titulaire de la fonction publique	**ensemble**			**279 896**	**95,8%**	**4,0%**	**0,2%**	**0,0%**
Total				**2 132 552**	**19,2%**	**37,6%**	**33,7%**	**9,5%**

8. Tests SLTS

échan-tillon	méthode	ω	temps passé (en mi-nutes)	critère d'arrêt sur S	paramètres										
					cons-tante	durée	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
1	SLTS	50	130	104 613	4,986	0,000	1,360	145,352	2,718	0,610	-0,851	-1,680	0,117	0,011	-0,367
1	SLTS	40	291	117 336	6,001	0,000	3,260	200,856	3,563	0,479	0,284	-0,359	0,034	0,008	0,088
1	SLTS	30	143	**STOP**	5,912	0,000	2,871	287,232	3,357	0,469	0,207	-0,435	0,040	0,009	0,089
1	SLTS	20	19	**STOP**	5,714	0,000	2,706	402,060	3,450	0,496	0,134	-0,518	0,048	0,009	0,055
1	SLTS	10	5	**STOP**	-3,156	0,004	-2,205	546,738	5,799	3,049	-2,262	-6,001	0,390	0,026	-6,023
1	Synthèse SLTS	moyenne			3,891	0,001	1,598	316,448	3,777	1,021	-0,498	-1,798	0,126	0,013	-1,232
1	Synthèse SLTS	CV			102%	292%	140%	51%	31%	111%	-219%	-134%	120%	59%	-218%
1	RDM				6,774	0,001	2,101	5,865	2,362	-0,341	-0,534	-0,387	0,047	0,009	0,745
2	SLTS	50	125	59 937	7,653	0,003	3,224	5,810	481,196	-0,138	-2,265	-2,679	0,055	0,021	-1,146
2	SLTS	40	79	94 127	6,667	0,000	4,073	8,872	18,049	0,383	0,399	-0,236	0,012	0,007	0,157
2	SLTS	30	92	**STOP**	6,743	0,002	2,469	7,451	32,098	0,167	-0,668	-1,221	0,046	0,010	0,016
2	SLTS	20	61	**STOP**	6,412	0,000	2,606	10,017	10,232	0,468	0,423	-0,192	0,017	0,007	0,120
2	SLTS	10	5	**STOP**	-9,194	0,013	-2,156	6,075	139,353	1,553	-0,723	-1,668	0,301	0,085	-4,540
2	Synthèse SLTS	moyenne			4,176	0,003	2,053	7,348	113,882	0,349	-0,561	-1,064	0,080	0,023	-0,775
2	Synthèse SLTS	CV			172%	172%	118%	25%	175%	184%	-196%	-98%	153%	146%	-259%
2	RDM				6,781	0,001	2,347	5,894	2,390	-0,313	-0,465	-0,368	0,046	0,009	0,821
3	SLTS	50	63	58 220	12,356	0,002	-2,285	157,348	-1,997	-4,639	-5,318	-1,192	0,046	-0,001	-0,013
3	SLTS	40	90	92 367	7,183	-0,001	2,142	216,517	2,626	0,123	0,256	-0,102	0,003	0,005	0,112
3	SLTS	30	59	**STOP**	11,912	0,001	-1,425	305,025	-1,788	-4,784	-4,921	-0,503	0,022	0,005	0,179
3	SLTS	20	10	**STOP**	95,398	0,000	-86,642	-73,427	-85,930	-88,185	-88,109	-0,144	0,007	0,005	0,090
3	SLTS	10	4	**STOP**	26,872	0,012	-14,028	582,293	-8,561	-9,302	-17,519	-11,600	0,181	-0,066	-9,271
3	Synthèse SLTS	moyenne			26,750	0,002	-16,648	198,942	-15,543	-17,850	-19,346	-2,318	0,051	-0,007	-1,347
3	Synthèse SLTS	CV			138%	219%	-225%	120%	-242%	-210%	-191%	-215%	145%	-424%	-311%
3	RDM				6,766	0,001	1,822	5,959	2,445	-0,252	-0,474	-0,442	0,047	0,008	0,767
4	SLTS	50	161	80 579	5,999	0,001	1,581	1774,115	2,453	0,560	-0,971	-2,010	0,095	0,014	-1,230
4	SLTS	40	98	153 817	26,872	0,012	-14,028	585,290	-8,561	-9,302	-17,519	-11,600	-0,181	-0,066	-9,270
4	SLTS	30	78	**STOP**	6,957	-0,001	0,867	8,356	36,500	0,366	0,477	-0,051	0,003	0,005	0,091
4	SLTS	20	10	**STOP**	6,380	0,001	3,849	455,989	3,673	0,435	-0,121	-0,878	0,039	0,009	0,149
4	SLTS	10	4	**STOP**	9,312	0,018	-0,149	621,569	92,339	1,455	-8,913	-13,836	0,176	0,035	-14,036
4	Synthèse SLTS	moyenne			10,381	0,005	-1,010	575,213	21,475	-1,123	-4,587	-4,803	0,030	0,001	-3,922
4	Synthèse SLTS	CV			86%	156%	-705%	114%	191%	-400%	-169%	-136%	445%	4245%	-164%
4	RDM				6,756	0,001	2,185	5,998	2,454	-0,247	-0,456	-0,434	0,047	0,008	0,748
5	SLTS	50	112	87 208	5,997	0,000	4,287	225,848	3,660	0,235	0,341	0,058	0,031	0,007	0,187
5	SLTS	40	66	130 915	5,085	0,001	2,057	1423,468	2,658	0,429	-1,162	-2,362	0,127	0,013	-0,758
5	SLTS	30	75	**STOP**	4,980	0,001	3,523	723,875	3,037	0,055	-0,130	-0,001	0,070	0,013	-0,120
5	SLTS	20	14	**STOP**	6,843	0,000	4,372	12,113	6,346	0,391	0,505	-0,076	0,004	0,006	0,070
5	SLTS	10	368	**STOP**	-1,122	0,002	-0,801	876,663	6,819	2,155	-0,435	-5,102	0,278	0,042	-6,737
5	Synthèse SLTS	moyenne			4,757	0,001	2,604	544,661	4,162	0,503	-0,223	-1,320	0,093	0,015	-1,102
5	Synthèse SLTS	CV			66%	104%	83%	102%	47%	169%	-299%	-171%	117%	99%	-269%
5	RDM				6,767	0,001	2,120	5,908	2,485	-0,262	-0,474	-0,396	0,045	0,009	0,734
6	SLTS	50	79	88 503	9,396	0,004	1,472	998,842	291,790	-0,711	-1,583	-1,335	0,006	0,008	0,205
6	SLTS	40	86	191 872	6,908	-0,001	-0,089	10,591	2,457	0,411	0,557	-0,060	0,002	0,005	0,161
6	SLTS	30	190	**STOP**	6,876	-0,001	6,861	275,185	2,266	0,409	0,544	-0,064	0,002	0,005	0,159
6	SLTS	20	49	**STOP**	6,882	0,000	-0,115	388,504	3,072	0,407	0,522	-0,100	0,004	0,005	0,159
6	SLTS	10	232	**STOP**	11,483	0,012	-0,227	532,794	48,919	-0,588	-2,869	-3,518	-0,045	0,011	-1,523
6	Synthèse SLTS	moyenne			8,052	0,002	1,670	368,638	58,498	-0,056	-0,551	-0,912	0,002	0,007	-0,018
6	Synthèse SLTS	CV			26%	226%	182%	99%	215%	-1040%	-287%	-165%	934%	35%	-4304%
6	RDM				6,830	0,001	2,330	5,897	2,385	-0,284	-0,479	-0,424	0,046	0,008	0,740

échan-tillon	méthode	ω	temps passé (en mi-nutes)	critère d'arrêt sur S	paramètres										
					cons-tante	durée	cs_2	cs_3	cs_4	cs_5	cs_6	sexe_f	age	quotite	idf_etab
7	SLTS	50	25	82 373	6,933	-0,006	3,202	253,708	3,144	0,117	-0,551	-1,880	0,088	0,014	-1,128
7	SLTS	40	36	104 069	6,995	-0,001	7,747	1364,862	3,445	0,395	0,505	-0,075	0,002	0,004	0,105
7	SLTS	30	22	**STOP**	5,353	0,000	5,467	556,720	3,354	0,404	0,206	-0,171	0,057	0,010	-0,074
7	SLTS	20	347	**STOP**	4,861	0,004	2,536	951,592	2,307	-0,577	-0,452	0,701	0,071	0,018	-0,771
7	SLTS	10	252	**STOP**	-1,350	0,004	2,470	1089,224	3,731	2,799	-1,831	-5,106	0,275	0,038	-11,062
7	Synthèse SLTS	moyenne			4,937	0,000	3,959	703,667	3,061	0,476	-0,434	-1,159	0,090	0,015	-2,032
7	Synthèse SLTS	CV			69,6%	1262,5%	57,8%	62,5%	17,6%	268,7%	-208,1%	-200,5%	115,9%	84,1%	-234,5%
7	RDM				6,693	0,001	3,056	5,926	2,485	-0,250	-0,428	-0,399	0,047	0,009	0,752
8	SLTS	50	123	54 233	7,042	-0,001	4,084	9,143	1,129	0,383	0,515	-0,089	0,000	0,004	0,065
8	SLTS	40	143	69 950	5,073	0,001	4,367	137,867	3,309	0,201	0,187	0,132	0,068	0,009	0,188
8	SLTS	30	185	**STOP**	5,093	0,000	4,307	197,705	3,231	0,158	0,154	0,124	0,067	0,009	0,222
8	SLTS	20	20	**STOP**	6,660	0,000	3,661	10,996	3,134	0,366	0,503	-0,126	0,011	0,005	0,216
8	SLTS	10	26	**STOP**	-3,439	0,003	0,651	378,979	2,115	0,270	0,980	3,002	0,222	0,034	-2,390
8	Synthèse SLTS	moyenne			4,520	0,001	3,354	123,436	2,567	0,188	0,318	0,441	0,069	0,012	-0,158
8	Synthèse SLTS	CV			95,1%	190,0%	46,8%	124,2%	36,9%	52,6%	104,5%	304,8%	128,0%	103,7%	-726,4%
8	RDM				6,912	0,001	2,536	5,896	2,418	-0,291	-0,494	-0,400	0,045	0,008	0,728
9	SLTS	50	411	113 787	7,067	-0,001	2,448	7,353	135,577	0,549	-0,203	-2,035	0,050	0,016	-0,461
9	SLTS	40	461	132 910	6,744	0,000	5,483	9,139	20,574	0,418	0,537	-0,130	0,004	0,007	0,119
9	SLTS	30	198	**STOP**	6,922	-0,001	7,662	9,468	32,640	0,389	0,496	-0,079	0,003	0,005	0,096
9	SLTS	20	85	**STOP**	6,738	0,000	2,329	8,421	56,494	0,399	0,472	-0,154	0,007	0,007	0,124
9	SLTS	10	10	**STOP**	3,659	0,016	1,758	6,908	146,334	0,407	0,369	0,818	-0,012	0,040	-3,055
9	Synthèse SLTS	moyenne			6,340	0,003	3,703	7,864	65,673	0,312	0,196	-0,330	0,016	0,014	-0,408
9	Synthèse SLTS	CV			22,7%	298,7%	68,6%	14,1%	89,4%	21,2%	156,3%	-316,1%	142,8%	106,8%	-337,0%
9	RDM				6,692	0,001	2,147	6,016	2,480	-0,266	-0,421	-0,344	0,047	0,009	0,772
10	SLTS	50	198	66 681	6,953	-0,001	-0,056	10,886	1,209	0,383	0,535	-0,049	0,001	0,005	0,115
10	SLTS	40	248	77 621	6,941	-0,001	7,097	7,650	1,219	0,385	0,536	-0,049	0,001	0,005	0,118
10	SLTS	30	265	**STOP**	6,925	-0,001	6,853	6,590	2,547	0,316	0,640	-0,046	0,001	0,006	0,131
10	SLTS	20	491	**STOP**	6,715	0,001	6,609	10,967	2,842	0,380	0,502	-0,032	0,002	0,008	0,214
10	SLTS	10	190	**STOP**	14,839	0,003	-0,235	66,921	35,919	0,545	0,672	-5,696	0,039	0,042	1,481
10	Synthèse SLTS	moyenne			8,178	0,000	3,736	18,172	7,703	0,291	0,411	-1,036	0,015	0,012	0,472
10	Synthèse SLTS	CV			44%	355%	103%	143%	197%	29%	18%	-244%	113%	131%	127%
10	RDM				6,850	0,001	2,058	5,893	2,375	-0,350	-0,544	-0,388	0,045	0,009	0,806