Document d'études
Direction de l'animation de la recherche, des études et des statistiques

Numéro **221**

Juin 2018

# Reconstitution des mouvements de main-d'œuvre depuis 1993 : guide méthodologique

Kévin Milin
Dares

DARES
direction de l'animation de la recherche, des études et des statistiques

# Table des matières

# Résumé

Jusque mi-2015, les statistiques de mouvements de main-d'œuvre étaient élaborées à partir de deux sources : une déclaration mensuelle obligatoire pour les établissements de plus de 50 salariés (DMMO) et une enquête trimestrielle pour les établissements de moins de 50 salariés (EMMO). Depuis lors, elles sont établies à partir de la déclaration sociale nominative (DSN).

Le Document d'études présente les méthodes adoptées pour reconstruire les données sur les mouvements de main-d'œuvre à partir de la DSN et pour rétropoler l'ensemble des séries sur le passé. La première étape consiste à construire des bases de référence à partir des nouvelles données de la DSN, permettant de reconstituer le champ des séries historiques. Tous les établissements présents dans ces bases n'ont pas, tous les mois, déposé une DSN. Il s'agit donc de traiter la non-déclaration, afin de garantir la représentativité des statistiques. La méthode retenue pour y parvenir se fonde sur l'estimation économétrique d'une probabilité de réponse, suivie d'un calage sur marges. Cette méthode est appliquée au niveau des établissements, afin d'assurer la cohérence entre les microdonnées et les séries agrégées. Ensuite, une stratégie d'imputation est mise en œuvre pour corriger des anomalies de déclaration en DSN, en particulier sur les fins de contrat manquantes. Enfin, les données sont rétropolées, afin de corriger les ruptures de séries observées suite au changement de source, par le biais de repondérations sur les mouvements. *In fine*, ces différentes étapes permettent de fournir des données rétropolées depuis 1993 sur les établissements de plus de 50 salariés et depuis 1998 sur les établissements de plus de 10 salariés.

# Introduction

Les mouvements de main-d'œuvre (MMO) recensent l'ensemble des embauches et des fins de contrats de travail au niveau des établissements. Ils permettent de mesurer les entrées et les sorties selon le type de contrat (contrat à durée déterminée/contrat à durée indéterminée), la durée des contrats et les motifs de rupture.

Jusqu'en 2015, ces statistiques étaient élaborées à partir de deux sources : une déclaration mensuelle obligatoire pour les établissements de plus de 50 salariés (DMMO), et une enquête trimestrielle pour les établissements de moins de 50 salariés (EMMO).

À partir de 2013, la déclaration sociale nominative (DSN) a été progressivement mise en place. Elle est constituée de l'ensemble des paies versées par les établissements chaque mois. Elle vise à remplacer un grand nombre de déclarations administratives réalisées par les entreprises, dont celles portant sur les mouvements de main-d'œuvre. Dans ce contexte, au cours de l'année 2015, le taux de réponse à la DMMO/EMMO a nettement diminué car de plus en plus d'établissements sont entrés dans le dispositif de la DSN. C'est la raison pour laquelle la publication des données sur les MMO a été suspendue à partir du deuxième trimestre.

Depuis lors, des travaux ont été menés afin de diffuser à nouveau les statistiques sur les mouvements de main-d'œuvre. Il s'agit ici de détailler les méthodes adoptées pour reconstruire les données sur les MMO à partir de la DSN et pour rétropoler l'ensemble des séries sur le passé.

À cette fin, des éléments de cadrage sur le champ des MMO, les bases statistiques mobilisées et la montée en charge de la DSN sont tout d'abord évoqués (fiche 1). Le mode de correction de la non-déclaration en DSN est ensuite développé (fiche 2). Les traitements réalisés sur les données de la DSN, principalement sur les dates et motifs de fin de contrat, sont alors précisés (fiche 3). Enfin, le processus de rétropolation des séries est détaillé (fiche 4).

# Synthèse

## 1. Éléments de cadrage

Historiquement, les statistiques de mouvements de main-d'œuvre (MMO) couvrent les établissements de France métropolitaine, relevant du champ privé hors agriculture et hors intérimaires. Elles ne prennent en compte que partiellement les effets démographiques, liés aux changements de statut des établissements (de non-employeur à employeur) et à leur cessation d'activité. De fait, le dispositif MMO nécessite de faire appel au référentiel Sirus (Système d'identification au répertoire des unités statistiques), qui fournit des informations sur les établissements avec un certain délai. Dans ce contexte, les MMO portent majoritairement sur les établissements employeurs depuis plus de deux ans. Sur le champ des établissements de plus de 50 salariés soumis à la DMMO, ceux réalisant pour la première fois une déclaration sont toutefois pris en compte dès l'année suivante.

Ce champ a été reconstitué sur les données issues de la DSN, dans une base dite de référence, élaborée pour chacune des années 2015 à 2017. Ces bases comprennent tous les établissements réputés concernés par les MMO, suivant la définition historique rappelée ci-dessus.

En 2017, la quasi-totalité du champ précédemment couvert par le processus DMMO/EMMO l'est désormais *via* la DSN : seuls les établissements recourant au dispositif simplifié Tese (Titre Emploi-Service Entreprise) n'y figurent pas encore. La montée en charge a toutefois été progressive, avec différents paliers liés à l'instauration de plusieurs obligations légales, portant d'abord sur les plus gros employeurs, puis au fur et à mesure, sur les plus petits. Cette transition vers la DSN étalée dans le temps a une incidence sur les méthodes retenues pour traiter l'absence de déclaration.

## 2. Traitement de la non-déclaration des établissements

Tous les établissements présents dans la base de référence n'ont pas, tous les mois, déposé une DSN. L'enjeu du traitement de la non-déclaration est de s'assurer que ceux qui ont déposé une DSN sont représentatifs de la base de référence. Il s'agit donc d'affecter des poids aux établissements déclarants, afin de garantir la représentativité des statistiques.

Dans cette perspective, des probabilités de déclaration des établissements sont estimées à partir de modèles économétriques. En 2015 et 2016, la non-déclaration est très largement liée au fait de ne pas avoir basculé en DSN. C'est donc le comportement d'entrée en DSN qui est implicitement modélisé. Pour cette raison, plusieurs paliers d'entrée en DSN sont déterminés à partir des obligations légales et, à chaque palier, correspond une spécification de modèle. Les probabilités estimées à partir de ces modèles fournissent un premier jeu de pondérations.

Dans un second temps, un calage sur marges est effectué. Il vise à assurer une mise en cohérence avec les bases de référence et à gagner sur la variance des estimateurs. Il en ressort un second jeu de pondérations qui vient légèrement modifier le premier.

## 3. Traitement des anomalies de déclaration

Dans certaines DSN, il arrive que des contrats actifs au cours d'un mois ne soient pas retrouvés le mois suivant, sans qu'aucune fin de contrat ne soit déclarée. Dans la plupart des cas, il s'agit bien de fins de contrat mais qui sont mal déclarées, de sorte que ni le motif ni la date de fin de contrat ne sont connus.

Sur les dates de fin de contrat, le traitement consiste à utiliser la date de fin prévisionnelle lorsqu'elle est renseignée, sinon, à tirer aléatoirement une date durant le mois de disparition du contrat. De leur côté, les motifs de fin de contrat manquants sont déterminés par une imputation.

## 4. Rétropolation des données

Une fois corrigées de la non-déclaration et des anomalies de déclaration, les données de la DSN marquent une rupture par rapport aux séries issues du dispositif DMMO/EMMO. L'origine des écarts est expertisée en mobilisant les données détaillées des déclarations préalables à l'embauche (DPAE), qui sont connues sur la période du changement de source. Il en ressort que :

- Sur les contrats à durée déterminée (CDD) :
  - Le taux de fin de CDD de plus d'un mois ne présente aucune rupture.
  - Le taux de fin de CDD de moins d'un mois est nettement plus élevé en DSN, quelle que soit la taille de l'établissement. Ce phénomène est lié à une sous-déclaration de ces mouvements dans le processus DMMO/EMMO.
  - La tendance du taux de fin de CDD de moins d'un mois est toutefois faussée en DSN durant la période de transition car la phase de montée en charge s'est accompagnée d'un accroissement de la déclaration des CDD très courts, en particulier dans les établissements de moins de 50 salariés.
  - Symétriquement au taux de fin de CDD de moins d'un mois, une rupture de série est observée sur le taux d'entrée en CDD.
- Sur les contrats à durée indéterminée (CDI) :
  - Aucune rupture n'apparait sur les taux d'entrée ou taux de sortie dans les établissements de plus de 50 salariés.
  - À l'inverse, une rupture de série se dégage sur les taux d'entrée au moment du passage en DSN sur les séries des établissements de moins de 50 salariés.

Pour corriger ces ruptures de séries, une rétropolation des données au niveau établissement est privilégiée. Cette approche a l'avantage de fournir des fichiers de données individuelles cohérents avec les agrégats publiés. Elle conduit à générer un jeu de poids : auparavant, chaque mouvement d'un établissement avait un poids implicite de 1 ; la rétropolation consiste à dilater ces poids pour pallier les ruptures de séries. Pour chaque strate (secteur-taille d'établissement), l'ampleur de la rupture est estimée, puis de nouvelles pondérations sont calculées de façon à s'assurer que la somme des mouvements sur la strate est bien égale à la somme attendue une fois la rupture de série prise en compte. Ces pondérations peuvent également être utilisées pour dilater les poids de

sondage de l'EMMO et de redressement de la DMMO. Au final, plusieurs jeux de poids sont créés sur :

- les entrées en CDI ;
- les sorties de CDI ;
- les fins de CDD très courts ;
- les entrées en CDD.

*In fine*, cette méthode permet de fournir des données rétropolées depuis 1993 (respectivement 1998) sur le champ des établissements de plus de 50 salariés (resp. 10 salariés). Sur les très petits établissements, les taux d'entrée et de sortie de CDI sont mis à disposition depuis 2007, date de mise en place de l'EMMO sur ce champ. En revanche, sur ces mêmes établissements de moins de 10 salariés, il est nécessaire d'avoir davantage de recul temporel sur les données de la DSN pour valider la correction de la sous-déclaration et produire les taux d'entrée et les taux de sortie de CDD.

# Fiche 1 – Eléments de cadrage

Le processus historique de production des statistiques des MMO reposait sur l'exploitation des déclarations de mouvements de main-d'œuvre (DMMO) pour les établissements de plus de 50 salariés (déclaration mensuelle), et de l'enquête sur les mouvements de main-d'œuvre (EMMO) pour les établissements de moins de 50 salariés (enquête trimestrielle). Ces déclarations historiques sont maintenant substituées par la déclaration sociale nominative (DSN).

Ce passage de l'ancien processus vers le nouveau a conduit à l'émergence de deux enjeux :

- parmi les DSN qui sont reçues, lesquelles doivent être prises en compte pour la production statistique des MMO, l'objectif étant de rester sur le champ le plus stable possible ?
- les DSN reçues, réputées exhaustives à un horizon cible, le sont-elles durant la phase de montée en charge et, si non, comment s'assurer de leur représentativité ?

L'objectif de cette première fiche est donc :

- de décrire la méthode adoptée afin de construire, pour chaque année, les bases de référence, c'est-à-dire la liste de l'ensemble des établissements français réputés concernés par ce dispositif (champ « MMO ») ;
- de décrire la montée en charge de la DSN, afin d'anticiper les besoins en matière de redressement de non-réponse.

## 1. Construction des bases de référence sur la période récente

Le champ MMO se restreint au secteur privé hors intérim et hors agriculture (industrie, construction et tertiaire) de France métropolitaine. Plus précisément, la base de sondage pour la construction de l'échantillon d'établissements de l'année N est construite à la fin de l'année N-1, à partir du référentiel Sirus[1]. Un établissement est inclus dans cette base de sondage s'il appartient au champ des MMO, c'est-à-dire :

- s'il ne relève pas de l'administration publique, des collectivités territoriales, de la défense nationale et il n'est pas un établissement de travail temporaire ;
- s'il apparait dans le dernier référentiel Sirus et si son effectif de référence est strictement positif.

Ainsi, dans l'ancien dispositif, l'échantillon d'établissements est composé :

- D'un côté des établissements de plus de 50 salariés devant effectuer une DMMO et appartenant au champ MMO. Le dénombrement était réputé exhaustif puisque la déclaration était obligatoire selon la législation. Les établissements ayant déposés spontanément une première DMMO dans le courant de l'année N-1 étaient intégrés à l'échantillon de l'année N.

[1] « Système d'identification au répertoire des unités statistiques ». Ce référentiel statistique de l'Insee est particulièrement utilisé pour le tirage d'échantillon d'établissements ou d'entreprises, pour les enquêtes de la statistique publique.

- De l'autre, des établissements de moins de 50 salariés tirés aléatoirement dans la base de sondage (EMMO). Compte tenu des délais d'actualisation de Sirus, de manière générale, seuls des établissements employeurs depuis plus de 2 ans pouvaient être intégrés à l'échantillon de l'EMMO.
- Les établissements cessés sont retirés de l'échantillon au fur et à mesure. Lors du trimestre de cessation, en cas de non-réponse de l'établissement, les effectifs sont considérés comme des mouvements de sortie (avec une imputation des motifs).

Les bases de référence étaient également complétées d'informations sur les établissements concernés, notamment en termes d'effectif de référence, de secteur d'activité, ou de localisation géographique. Pour la partie DSN, comme pour la partie historique, les données MMO les plus récentes sont privilégiées pour les variables concernant l'existence des établissements (c'est-à-dire le fait que les établissements ont une activité économique réelle ou non) et leur effectif salarié (cf. encadré 1 pour une discussion sur l'effectif de référence). À l'inverse, les informations telles que le secteur d'activité ou la région principale d'activité proviennent du répertoire Sirus.

Encadré 1 – L'effectif de référence

Pour les effectifs, la variable primordiale est l'effectif dit « de référence ». Cet effectif correspond à une mise à jour de celui présent dans Sirus de façon à intégrer de façon pertinente l'information issue des MMO lorsqu'elle est disponible ; il date ainsi majoritairement de l'année précédente (N-1). Plus précisément, l'effectif de référence peut dater :
(i) de l'année N-1 pour les établissements préalablement assujettis aux MMO (ce qui est notamment le cas pour une majeure partie des établissements de plus de 50 salariés) ;
(ii) de l'année N-2 pour les établissements non-assujettis aux MMO, ce qui est notamment la situation de nombreux établissements de moins de 50 salariés, puisque le taux de sondage de l'EMMO est d'environ 3 %[1].

Plus précisément, pour une année N donnée, l'effectif calculé depuis la source MMO est défini, pour un établissement répondant $i$, comme une moyenne des effectifs renseignés tout au long de l'année, avec une surpondération du dernier effectif connu :

$$eff_{(i),N}^{MMO} = \frac{eff_{\tau} + \frac{1}{\sum_{t=1}^{T} 1_{(i)\ \epsilon\ rep(t)}} . \sum_{t=1}^{T} effmoy_{t}^{(i)} . 1_{(i)\epsilon rep(t)}}{2}$$

Avec :

- $T$ valant 4 ou 12 selon la fréquence trimestrielle ou mensuelle ;
- $rep(t)$ l'ensemble des établissements répondants sur une période $t$ ;

[1] L'échantillon de l'EMMO était coordonné positivement avec celui de l'année précédente. Environ un quart de l'échantillon était renouvelé chaque année. L'effectif de référence des établissements de l'échantillon des MMO, datait majoritairement de l'année précédente (N-1).

- $effmoy_t^{(i)}$ la moyenne entre l'effectif en début de trimestre *t* (respectivement mois) et de l'effectif en fin de trimestre *t* (resp. mois) ;
- $eff_\tau$ le dernier effectif connu (c'est-à-dire : $\tau = \max\{t.1_{(i)\,\epsilon\, rep(t)}\}$).

## 2. Une montée en charge progressive de la déclaration sociale nominative

Le nombre de DSN réceptionnées a progressivement augmenté depuis la mise en place du dispositif en 2013 : d'environ 30 000 déclarations début 2015, à plus de 1 700 000 déclarations fin 2017. La montée en charge du dispositif s'est faite par paliers, conformément à la législation. Considéré globalement, le taux de déclaration s'établit, sur le champ MMO, à environ 10 % en avril 2015, et s'élève à plus de 75 % en 2017. Pour les établissements de plus de 50 salariés, il atteint 95 % en 2015. Les 5 % non-répondants sont soit des établissements non assujettis à la DSN, soit des établissements qui n'apparaissent pas dans les bases de référence en raison de l'actualisation retardée des répertoires Sirus. Enfin, la Dares n'a pas eu accès aux données relatives aux établissements adhérents au système Tese[2] avant 2018, ce qui explique en partie le plus faible taux de réponse pour les plus petits établissements.

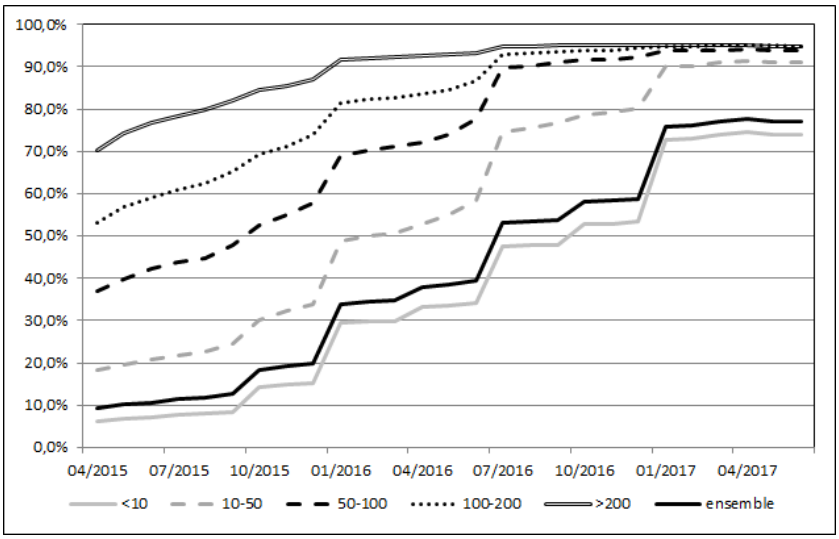
Plus précisément, le taux de réponse par date a évolué de la façon suivante (cf. encadré 2) :

- Plus de 70 % des établissements de plus de 200 salariés ont déposé une DSN dès avril 2015. Ce taux a augmenté progressivement tout au long de l'année 2015, pour atteindre plus de 90 % début 2016 ;
- Pour les établissements ayant entre 100 et 200 salariés, le taux de réponse est de 80 % en janvier 2016 ; celui des 50-100 salariés est de 70 % à la même date. Ces taux de réponse ont augmenté au cours du premier semestre 2016, pour dépasser 90 % en juillet 2016 (cf. graphique 1) ;
- Pour les établissements ayant entre 10 et 50 salariés, le taux de réponse en janvier 2016 est plus faible, de l'ordre de 50 %. Il a toutefois fortement progressé, jusqu'à 90 % en janvier 2017 ;
- Enfin, la hausse du taux de réponse des établissements de moins de 10 salariés s'est faite avec retard. Il s'élève à 74 % en juin 2017 (après 30 % en janvier 2016, et 6 % en avril 2015).

Globalement, le taux de réponse, qui était en-deçà des 10 % en avril 2015, s'élève à environ 77 % à la mi-2017 (95 % pour les plus de 50 salariés). Il n'est pas égal à 100 % alors que le dépôt de DSN est obligatoire pour tous les établissements du régime général depuis janvier 2017 :

- Pour les établissements de moins de 50 salariés, notamment de leur adhésion au système Tese (Titre emploi service entreprise) qui n'est pas encore intégré en DSN.
- Pour les établissements de plus de 50 salariés, notamment de l'absence des sociétés non commerciales en DSN ainsi que par d'une actualisation retardée des référentiels dans le cas de cessations.

[2] Le « Titre emploi service entreprise » est un système de déclaration simplifiée s'adressant aux petites entreprises (moins de vingt salariés) du régime général (http://www.letese.urssaf.fr/tesewebinfo/cms/presentation.html).




*Graphique 1 – taux de réponse à la DSN parmi les établissements concernés, selon la taille des établissements*

Encadré 2 – Détail de la montée en charge du dispositif DSN

Le dispositif DSN a été mis en place en 2013. Le nombre de DSN déposées est passé de 30 000 début 2015, à plus de 1 700 000 à la mi-2017. Cette montée en charge s'est faite par étapes correspondant à une séquence d'obligations législatives :

- Jusqu'en mars 2015 : la DSN se faisait sur la base du volontariat.

- D'avril à septembre 2015 : une première obligation intermédiaire a été fixée en avril 2015. En effet, à partir de cette date, « l*es employeurs redevables de cotisations et contributions sociales (au titre de l'année civile 2013) : d'un montant égal ou supérieur à 2 millions d'euros ou d'un montant égal ou supérieur à 1 million d'euros lorsqu'ils ont recours à un tiers déclarant et pour lesquels la somme totale des cotisations et contributions sociales déclarées par ce tiers au titre de l'année civile 2013 pour le compte de l'ensemble de ses clients est égale ou supérieure à 10 millions d'euros »*[1] doivent obligatoirement déposer des DSN. Lorsque cette DSN est déposée, la DMMO est réputée substituée.

- D'octobre à décembre 2015 : le mois d'octobre 2015 correspond au premier mois de déploiement de la deuxième phase de la DSN. Cette dernière correspond à un élargissement du périmètre des DSN, visant à substituer les déclarations unifiées des cotisations sociales[2], et les relevés mensuels de mission (RMM) pour les entreprises de travail temporaire.

- De janvier à mars 2016 : aucune obligation n'est à relever pour cette période. La hausse du taux de réponse est probablement liée au changement d'année, et à l'organisation interne des établissements.

- d'avril à juin 2016 : en avril 2016, débutait la période test de la troisième phase de la

DSN dont l'objectif était de substituer les déclarations de cotisations à la MSA (Mutuelle Sociale Agricole) et aux organismes complémentaires retraite, santé et prévoyance, ainsi que la déclaration annuelle des données sociales (DADS).

- de juillet à septembre 2016 : en juillet 2016, une seconde obligation intermédiaire entre en vigueur, basée sur les cotisations sociales versées en 2014 (cf. tableau A).

- d'octobre à décembre 2016 : en octobre 2016, la phase 3 de la DSN est étendue à tous les établissements.

- à partir de janvier 2017 : la DSN est rendue obligatoire pour tous les établissements du régime général, ainsi que pour les établissements relevant du régime agricole et ayant versés plus de 30 000 € de cotisations sociales en 2014.

- à partir d'avril 2017 : extension de l'obligation à tous les établissements du régime agricole.

| EMPLOYEURS OU TIERS MANDATÉS PAR L'EMPLOYEUR DONT LE PERSONNEL RELÈVE DU RÉGIME GÉNÉRAL ou d'un ou plusieurs régimes spéciaux mentionnés à l'article L. 711-1 du code de la sécurité sociale | | |
|---|---|---|
| **Déclarant** | **Montant de cotisations et contributions sociales dues au titre des périodes de paie de 2014** | **Obligation de transmettre une déclaration sociale nominative (DSN)** |
| Employeur sans tiers mandaté | Egal ou supérieur à 50 000 € | A compter de la paie du mois de juillet 2016 (DSN exigible le 5 ou le 15 août 2016 selon l'échéance applicable à l'employeur) |
| | Inférieur à 50 000 € | A compter de la paie du mois de janvier 2017 (DSN exigible le 5 ou le 15 février 2017 selon l'échéance applicable à l'employeur) |
| Tiers mandatés par l'employeur | Egal ou supérieur à 10 millions d'euros | A compter de la paie du mois de juillet 2016 (DSN exigible le 5 ou le 15 août 2016 selon l'échéance applicable à l'employeur) |
| | Inférieur à 10 millions d'euros | A compter de la paie du mois de janvier 2017 (DSN exigible le 5 ou le 15 février 2017 selon l'échéance applicable à l'employeur) |

*Tableau A – Dates limites pour la transmission obligatoire de DSN*
*source :* Legifrance, *décret n° 2016-611 du 18 mai 2016*

[1] Extrait du décret n° 2014-1082 du 24 septembre 2014.
[2] Elles correspondent aux bordereaux récapitulatifs de cotisations (BRC) pour l'Union de recouvrement des cotisations de sécurité sociale et d'allocations (Urssaf).

# Fiche 2 – Traitement de la non-déclaration des établissements

En 2015 et 2016, la DSN étant dans une phase de déploiement, tous les établissements n'en ont pas effectuées. En 2017, le taux de réponse à la DSN n'atteint pas encore les 100 % (cf. fiche 1). Une méthode de redressement de la non-réponse totale permet de corriger les effets de cette non-exhaustivité, en affectant un poids à chaque établissement déclarant pour garantir la représentativité des statistiques produites sur les mouvements de main-d'œuvre.

Plus précisément, plusieurs étapes jalonnent ce travail :

1. Les probabilités de déclaration de chaque établissement sont estimées pour chaque mois. Il convient de prendre en compte les effets de non-linéarités dans les estimations. Par exemple, la probabilité d'entrer dans le dispositif des DSN n'augmente pas linéairement en fonction de l'effectif de l'établissement.
2. Des groupes de réponse homogène (GRH) sont ensuite créés à partir des probabilités estimées de déclaration. Le poids affecté à chaque établissement vaut alors l'inverse du taux de réponse dans le GRH associé.
3. Un calage sur marge est enfin appliqué, notamment pour diminuer la variance des estimateurs.

Cette deuxième fiche a donc pour vocation de présenter à la fois les méthodes choisies pour redresser les données de la non-déclaration, mais aussi pour montrer la qualité de la correction effectuée.

## 1. Notations et rappel de la méthode de redressement de la non-réponse totale choisie

Tout au long de ce document, on note *y* une variable d'intérêt présente dans les données issues des DSN, comme par exemple le nombre d'entrées ou de sorties sur une période considérée. Par ailleurs, ayant à disposition un référentiel qui liste les établissements pour lesquels une DSN est attendue, on note $r_i$ la variable indicatrice de réponse pour l'unité *i* : $r_i = 1$ si l'établissement *i* a effectué une DSN, et $r_i = 0$ dans le cas contraire. Notons *bdr* la base de référence, et *R* l'ensemble des déclarants (ou répondants, c'est-à-dire les établissements pour lesquels $r_i = 1$).

La méthode de correction proposée consiste à expliquer le mécanisme de réponse en attachant à chaque établissement une probabilité de répondre (de remettre une DSN) qui varie en fonction de ses caractéristiques, et de pondérer chaque établissement répondant par l'inverse de sa probabilité de réponse estimée.

Notons $p_i = P(r_i = 1),\ i \in bdr$ la probabilité que $i \in bdr$ ait remis une DSN. Cherchons à estimer le total de la variable d'intérêt (*i.e.* le nombre total d'entrées ou de sorties sur la période). En l'absence de non-réponse, ce total s'écrit :

$$t_y = \sum_{i \in bdr} y_i$$

Comme la base de référence est exhaustive, ce total est connu de façon exacte.

La correction de non-réponse consiste à remplacer l'expression précédente par une somme sur l'ensemble des répondants, avec prise en compte de la pondération :

$$\widetilde{t_y} = \sum_{i \in R} \frac{y_i}{p_i}$$

Étant donné que les probabilités de réponse ne sont pas connues, elles sont estimées *via* un modèle : $p_i = f(\boldsymbol{x}_i, \theta)$, où $\boldsymbol{x}_i$ représente l'information auxiliaire disponible dans la base de référence, et où $\theta$ est un vecteur de paramètres inconnus à estimer. En notant $\hat{p}_i = f(\boldsymbol{x}_i, \hat{\theta})$, la probabilité estimée de réponse de l'établissement *i,* l'estimateur du total de la variable d'intérêt *y* est donné par :

$$\widehat{t_y} = \sum_{i \in R} \frac{y_i}{\hat{p}_i}$$

La qualité de l'estimation est fortement liée à la qualité du modèle d'explication de la non-réponse[3]. De plus, il faut prendre garde à la dispersion éventuelle des probabilités estimées, qui peut entrainer une variance importante de l'estimateur $\widehat{t_y}$ (liée à de petites probabilités de réponse).

Afin d'apporter de la robustesse, on a recours à des classes de pondération, plutôt que d'utiliser directement les probabilités de déclaration (cf. Ardilly, 2006). On divise la base de référence en T classes ($C_1, C_2, \ldots C_T$), qui comportent à la fois des déclarants et des non-déclarants, et on affecte à chaque déclarant un poids égal à l'inverse du taux de réponse observé dans sa classe. En adoptant cette stratégie, l'estimateur prend la forme suivante :

$$\ddot{t_y} = \sum_{i \in R} \frac{y_i}{\widehat{p_i}} = \sum_{c=1}^{T} \sum_{i \in C_c} \frac{r_i}{\widehat{p_c}} y_i = \sum_{c=1}^{T} N_c \overline{y_{rc}}$$

où $N_c = \sum_{i \in C_c} 1$ ; $\widehat{p_c} = \frac{\sum_{i \in C_c} r_i}{N_c}$ ; $\overline{y_{rc}} = \frac{\sum_{i \in C_c} r_i y_i}{\sum_{i \in C_c} r_i}$

Et le biais de non-déclaration associé à l'estimateur du total de *y* prend la forme :

$$B(\ddot{t_y}) = E(\ddot{t_y} - t_y) \cong \sum_{c=1}^{T} \frac{1}{\bar{p}_c} \sum_{i \in C_c} (p_i - \bar{p}_c)(y_i - \bar{y}_c)$$

avec : $\bar{p}_c = \frac{\sum_{i \in C_c} p_i}{N_c}$ et $\bar{y}_c = \frac{\sum_{i \in C_c} y_i}{N_c}$

[3] Si le modèle de non-réponse n'est pas correctement spécifié, l'estimateur qui en découle peut être considérablement biaisé.

Cette écriture suggère que le biais dû à la non-déclaration est petit :

- si la probabilité moyenne de déclaration dans chaque classe est grande (*i.e.* que le taux de réponse dans la classe est grand) ;
- ou si la covariance entre la probabilité de faire une déclaration et la variable d'intérêt est faible.

Ainsi, il parait naturel de former des classes homogènes par rapport aux probabilités de réponse, puisque les variables d'intérêt sont nombreuses.

Plusieurs méthodes de formation des classes ont été testées sur le mois d'octobre 2015. Une méthode dite des scores s'est avérée la plus efficace, et a donc été retenue. Cette dernière se divise en trois étapes :

1. Modélisation du score, c'est-à-dire estimation de la probabilité de déposer une DSN pour chaque établissement apparaissant dans le référentiel ;
2. Formation des classes en se basant sur les probabilités de déclaration estimées selon la méthode décrite par Haziza et Beaumont, 2007 ;
3. Affectation d'un poids à chaque établissement (inverse du taux de réponse de la classe).

Comme la construction des classes se base principalement sur les probabilités estimées de déclaration, et donc sur le modèle utilisé, l'étape d'élaboration de modèles de déclaration/non-déclaration s'avère essentielle dans le processus d'affectation de poids aux établissements ayant fait des DSN. Les parties suivantes détaillent la manière de constrruire des modèles de non-réponse en DSN, ainsi que les éléments qui permettent leur validation.

## 2. Prise en compte des non-linéarités

Différentes variables sont intégrables dans la modélisation des probabilités de déclaration : le secteur d'activité, l'effectif salarié ou la zone géographique. Les variables retenues doivent caractériser les établissements et doivent être corrélées à leur comportement de déclaration. Par exemple, la taille des établissements ou leur appartenance à un groupe peut indirectement capter l'existence d'une structure interne, comme un service de ressources humaines ou de comptabilité, en charge des déclarations légales.

Un modèle paramétrique, tel que présenté en encadré 3, intègre deux variables quantitatives : l'effectif de l'établissement, ainsi que le nombre d'établissements de l'entreprise à laquelle appartient l'établissement considéré. Il est difficile d'imaginer que le lien entre la propension à effectuer une DSN et ces deux variables soit totalement linéaire : il peut, par exemple, exister des effets de seuil, c'est-à-dire qu'à partir d'une certaine taille de l'établissement et de l'entreprise, l'effet sur la probabilité de faire une DSN est marginal.

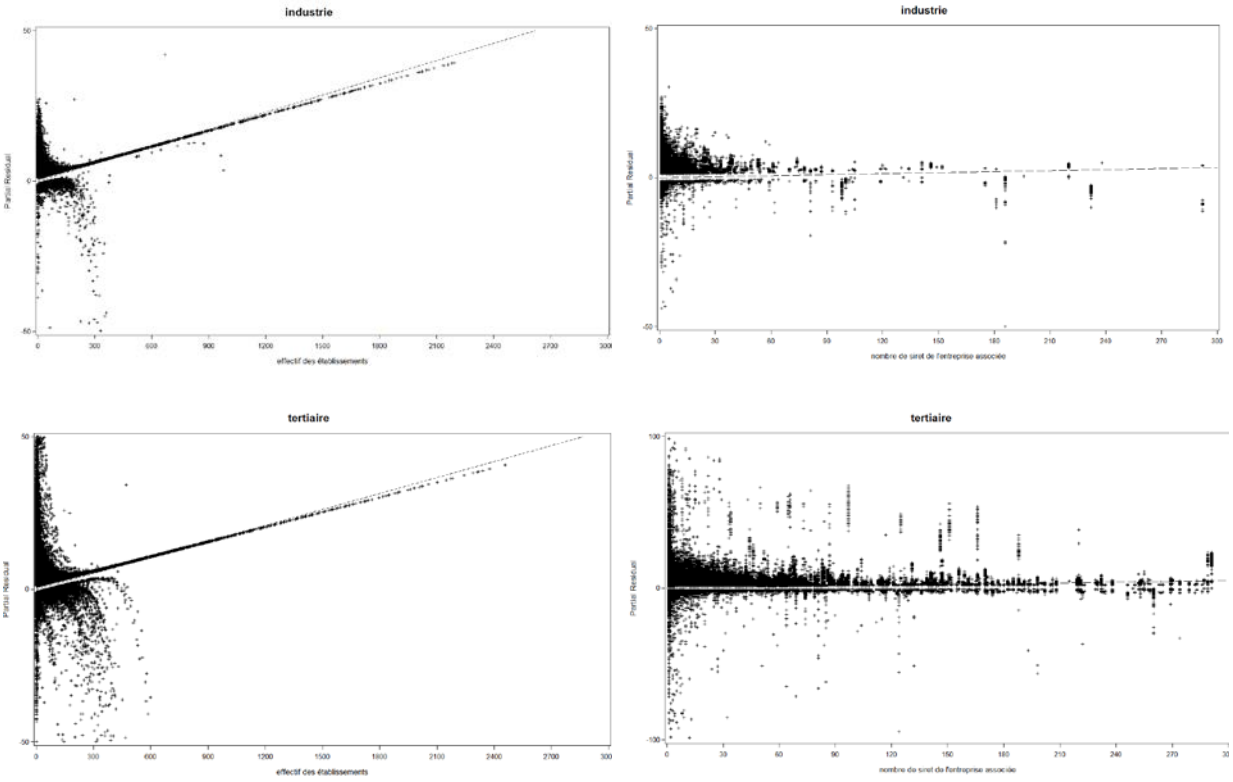
Afin de tester l'existence de non-linéarité, plusieurs méthodes peuvent être utilisées, notamment :

- Un modèle semi-paramétrique (*Generalized Additive Model*) intégrant les mêmes variables que le modèle dit « de base », du type : $logit(p_i) = ln(\frac{p_i}{1-p_i}) = f(\boldsymbol{X}_i) + \boldsymbol{\beta}.\boldsymbol{Z}_i$, avec $\boldsymbol{X}_i$ et $\boldsymbol{Z}_i$ des informations auxiliaires associées à l'établissement *i*, $\boldsymbol{\beta}$ un vecteur de paramètres

à estimer, *f* une fonction de forme inconnue. Cette méthode est similaire à celle présentée par Da Silva et Opsomer (2009), et permet dans le même temps d'estimer les probabilités de déclaration pour chaque établissement. Ces modèles ont été testés : ils nécessitent des temps de calcul trop importants.

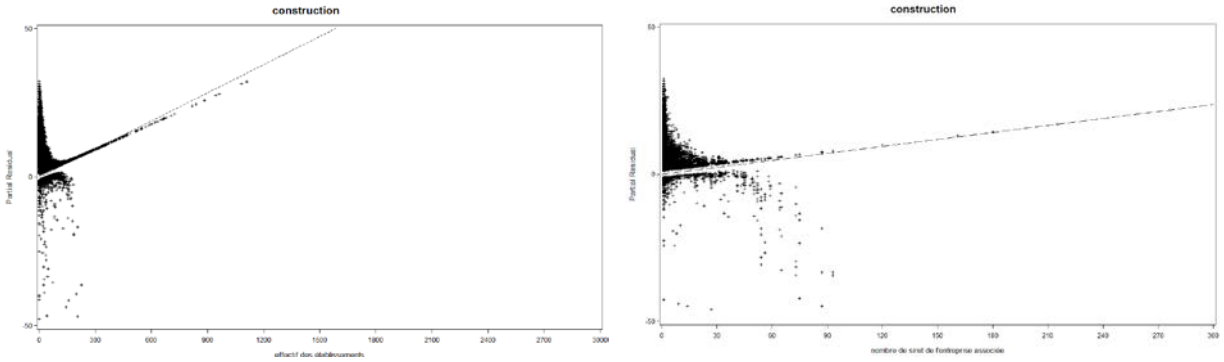
- L'analyse des résidus partiels[4]. Cela consiste à tracer les résidus partiels du modèle « de base » (cf. encadré 3), associés aux deux variables quantitatives, en fonction de ces mêmes variables quantitatives. Si le tracé est linéaire alors le modèle initial est accepté. *A contrario*, si une tendance non linéaire se dégage, le modèle initial doit être modifié. Par exemple, il est possible de :
  - remplacer la variable quantitative en question par une fonction de celle-ci donnant la même tendance que celle observée ;
  - ou encore de modifier cette variable en variable catégorielle.

Cette dernière méthode est donc appliquée au modèle « de base » présenté en encadré 3. Au regard des graphiques 2 à 7, une structure particulière se dégage des résidus partiels associés à l'effectif, et dans une moindre mesure, de ceux associés au nombre d'établissements par entreprise. En conséquence, il convient de prendre en compte cette non-linéarité dans les modèles. Pour cela, la création de classes pour rendre les variables quantitatives qualitatives est préférée à l'utilisation d'une fonction, puisqu'elle permet l'économie de conjectures sur la forme de la fonction la plus adaptée (cf. *infra*).




[4] Les résidus partiels associés à la variable $X$ se définissent de la manière suivante :

$$\hat{\varepsilon}_x = \frac{r_{i-}\hat{p}_i}{(1-\hat{p}_i).\hat{p}_i} + \beta_x.X$$




*Graphiques 2 à 7 – résidus partiels calculés à partir du modèle de base, pour le mois d'octobre 2015.*

Encadré 3 – Une première approche pour modéliser les probabilités de déclaration

Un modèle dit « de base » a été mobilisé en première approche, pour estimer les probabilités de déclaration des établissements pour un mois donné, sur chaque grand secteur (industrie, construction, tertiaire). Il s'écrit :

$$logit(p_i) = ln(\frac{p_i}{1-p_i})$$
$$= \gamma + \rho.eff_i + \sigma.n_i^{siret} + \sum_{1 \le j \le n_{apet}-1} \alpha_j.1_{apet_i=apet_j}$$
$$+ \sum_{1 \le j \le n_{region}-1} \beta_j.1_{region_i=region_j}$$

Avec :

- $eff_i$ l'effectif de l'établissement *i* (non catégorisé) ;
- $n_i^{siret}$ le nombre d'établissements de l'entreprise dans laquelle se situe l'établissement *i* appartient (non catégorisé) ;
- $apet_i$ le secteur d'activité de l'établissement *i*, au niveau le plus fin ; avant octobre 2015, des niveaux plus agrégés sont utilisés afin de garantir la convergence de l'algorithme itératif permettant l'obtention de l'estimateur du maximum de vraisemblance de $(\gamma, \rho, \sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ ;
- $region_i$ la zone géographique d'implantation de l'établissement *i*.

## 3. Recherche de modèle par période

Si les estimations sont effectuées sur les trois grands secteurs d'activité (industrie, construction, tertiaire), des différences subsistent entre les modèles retenus. En effet, pour un secteur donné, les variables sélectionnées dans le modèle ne sont pas forcément les mêmes d'un mois à l'autre, tout comme les classes retenues (nombres et bornes) pour transformer les variables quantitatives en variables qualitatives.

La recherche de modèle n'est néanmoins pas effectuée pour chaque mois, mais selon les paliers qui correspondent aux différentes phases de montée en charge de la DSN (cf. fiche 1).

## 4. Variables utilisées

Les modèles retiennent plusieurs variables :

- l'effectif de référence de l'établissement (catégorisée) ;
- le secteur d'activité ;
- la zone géographique d'implantation de l'établissement (région) ;
- le nombre d'établissements de l'entreprise à laquelle l'établissement appartient (catégorisé) ;
- la catégorie juridique de l'entreprise ;
- la date de création de l'établissement (constitution de plusieurs classes[5]) ;
- le chiffre d'affaire de l'entreprise ; la date d'entrée dans le dispositif des DSN dépend des montants passés des cotisations et contributions sociales des entreprises (cf. fiche 1). Le chiffre d'affaire de l'entreprise est alors la variable qui permet d'approcher le plus possible les règles d'entrée dans le processus des DSN.

De même, il est possible de rajouter des variables d'interactions pour augmenter la dimension explicative du modèle. Par exemple, dans le modèle purement additif (modèle de base), l'hypothèse sous-jacente est que, pour une région donnée, les coefficients associés à la variable effectif sont identiques pour les différents secteurs (cf. graphique 8). Dans un modèle avec des interactions (entre effectifs et secteurs), cette hypothèse est relâchée, les pentes peuvent être différentes (cf. graphique 9).

Le nombre de variables d'interactions de premier ordre à introduire est potentiellement important. Cependant, toutes ces variables ne peuvent pas être intégrées au modèle, au risque de nuire à l'estimation des paramètres. Par exemple, un trop grand nombre de coefficients à estimer peut amener :

- à une trop forte complexité calculatoire ;
- à de la séparabilité des données, impliquant la non-convergence de l'estimateur du maximum de vraisemblance ;
- ou encore à un sur-apprentissage.

En conséquence, plusieurs variables d'interactions ne sont pas testées dans la modélisation : d'une part celles qui intègrent la catégorie juridique et la région, d'autre part celles qui sont composées de variables dont le V de Cramer est inférieur à 0,15 (les interactions basées sur de fortes

[5] Pour les années de création des établissements, on considère six classes différentes :

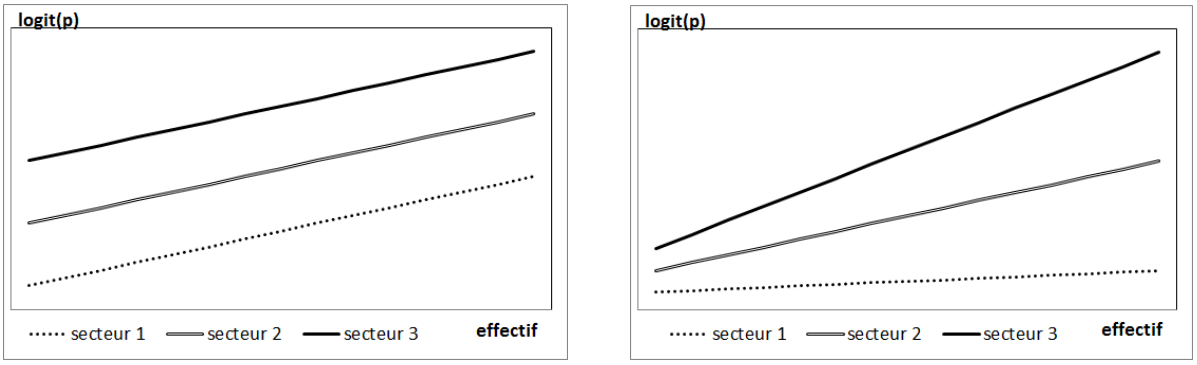
- les établissements créés avant 1998 ;
- entre 1998 et 2005 ;
- entre 2006 et 2010 ;
- entre 2011 et N-3, avec N l'année d'intérêt ;
- ceux crées après N-2 ;
- les établissements dont la date de création est inconnue ou incertaine.

dépendances sont donc rejetées). Ainsi, entre cinq et six interactions sont testées, les plus fréquentes étant :

- secteur d'activité x effectif ;
- secteur d'activité x chiffre d'affaire de l'entreprise (uniquement pour le secteur industriel) ;
- secteur d'activité x année de création de l'établissement ;
- effectif x nombre d'établissements dans l'entreprise ;
- effectif x année de création de l'établissement ;
- date de création de l'établissement x nombre d'établissements dans l'entreprise ;
- chiffre d'affaire x année de création de l'établissement.

Malgré cette restriction à quelques variables d'interactions, la convergence des estimateurs n'est pas nécessairement assurée. Plusieurs solutions permettent de pallier ce problème :

- passer à des nomenclatures plus agrégées ;
- retirer les interactions qui offrent les plus grands degrés de liberté ;
- retirer les variables qui ne sont pas, *a priori*, significatives (tests basés sur l'estimation issue de la dernière itération de l'algorithme pour l'estimation du maximum de vraisemblance, variables d'interactions comprises).




*Graphiques 8 et 9 – Illustrations d'un modèle additif et d'un modèle avec interactions*

Ainsi, le modèle s'écrit sous la forme suivante :

$$
\begin{aligned}
&logit(p_i) \\
&= \beta^0 + \beta^1 . eff_i + \beta^2 . n_i^{siret} \\
&+ \sum_{1 \le j \le n_{secteur}-1} \beta_j^3 . 1_{secteur_{i=}secteur_j} + \sum_{1 \le j \le n_{region}-1} \beta_j^4 . 1_{region_{i=}region_j} \\
&+ \sum_{1 \le j \le n_{effc}-1} \beta_j^5 . 1_{effc_{i=}effc_j} + \sum_{1 \le j \le nc^{siret}-1} \beta_j^6 . 1_{nc_i^{siret}=nc_j^{siret}} + \sum_{1 \le j \le n_{cj}-1} \beta_j^7 . 1_{cj_{i=}cj_j} \\
&+ \sum_{1 \le j \le n_{cac}-1} \beta_j^8 . 1_{cac_{i=}cac_j} + \sum_{1 \le j \le n_{age}-1} \beta_j^9 . 1_{age_{i=}age_j} \\
&+ \sum_{1 \le j \le n_{secteur}.n_{age}-2} \beta_j^{10} . 1_{secteur_{i=}secteur_j} . 1_{age_{i=}age_j} \\
&+ \sum_{1 \le j \le nc^{siret}.n_{age}-2} \beta_j^{11} . 1_{nc_i^{siret}=nc_j^{siret}} . 1_{age_{i=}age_j} \qquad (*)
\end{aligned}
$$

Avec :

- $eff_i$ et $effc_i$ l'effectif et l'effectif catégorisé de l'établissement *i* ;
- $n_i^{siret}$ le nombre d'établissements de l'entreprise dans laquelle l'établissement *i* appartient ($nc_i^{siret}$ étant la même variable, mais catégorisée) ;
- $secteur_i$ le secteur d'activité de l'établissement *i* exprimé dans la nomenclature A88 ;
- $region_i$ la zone géographique d'implantation de l'établissement *i* ;
- $age_i$ l'année de création de l'établissement *i*, en cinq classes (avant 1999, 1999-2005, 2006-2010, 2011-2013, après 2014) ;
- $cj_i$ la catégorie juridique de l'entreprise associée à l'établissement *i* (les sociétés commerciales sont différenciées des autres sociétés) ;
- $cac_i$ le chiffre d'affaire catégorisé de l'entreprise intégrant l'établissement *i* ;
- $etatStat_i$ l'état statistique de l'établissement *i* tel qu'il apparaît dans le répertoire Sirus récent (actif, cessé économiquement, cessé juridiquement).

## 5. Non-linéarité : méthode retenue pour la création des classes

Au regard des résidus partiels associés à la variable effectif, l'existence d'une non-linéarité dans la relation entre la propension d'un établissement à effectuer une DSN et son effectif est vérifiée. Pour se prémunir de tout effet non linéaire entre variables d'intérêt et variables explicatives, les exogènes quantitatives sont catégorisées (effectif de l'établissement, chiffre d'affaire de l'entreprise, nombre d'établissements dans l'entreprise).

Toutefois, pour catégoriser une variable, le choix du nombre de classes, et de leurs bornes peut s'avérer délicat, et reste généralement arbitraires. En outre, il est souvent retenu des quartiles ou quintiles, sans que cela ne garantisse que ce soit le meilleur choix. Afin que les classes soient les plus pertinentes possibles, la méthodologie suivante est appliquée pour chaque secteur et pour chaque palier :

1- Découpage des distributions des effectifs et des chiffres d'affaire par le moyen d'un jeu de 20 quantiles. Étant donné que le chiffre d'affaire de l'entreprise est une variable continue, ce découpage permet d'obtenir 21 classes de taille égales (plus une pour les entreprises pour lesquelles le chiffre d'affaire n'est pas renseigné) ; au contraire, l'effectif étant une variable discrète, le nombre de classes n'est pas constant ;

2- Estimation du modèle (*), qui comprend toutes les variables préalablement citées, exceptées les variables d'interactions composées de l'effectif de l'établissement et du chiffre d'affaire de l'entreprise. Le retrait de ces quelques variables d'interactions permet notamment de faciliter l'écriture des tests de l'étape 3 ;

3- Application des tests d'égalité suivants[6]

$\forall j \in [1; n_{effc} - 2\,],\ \beta_j^5 = \beta_{j+1}^5$ et $\forall q \in [1\,; n_{cac} - 2\,],\ \beta_q^8 = \beta_{q+1}^8$

4- Fusion de la j[ième] et (j+1)[ième] classe d'effectif en lien avec le test qui présente la statistique de test la plus faible (en valeur absolue), et de la q[ième] et (q+1)[ième] classe du chiffre d'affaire ;

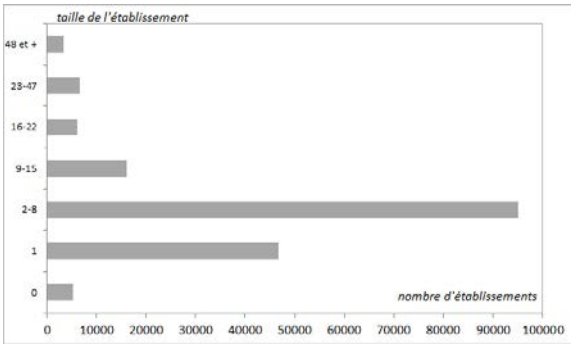
5- Réitération des étapes 2, 3 et 4 jusqu'à ce que tous les tests d'égalité soient rejetés au niveau 5 %.

Ainsi, l'application de cet algorithme débouche sur la création de classes pour la variable d'effectif et celle du chiffre d'affaire. Le nombre de classes, ainsi que leurs bornes diffèrent d'un palier à l'autre, et d'un secteur à l'autre (industrie, construction, tertiaire). Par exemple, dans le secteur de la construction, suite à la catégorisation de l'effectif, on obtient le même nombre de classes en janvier et juillet 2016, mais le découpage diffère (cf. graphiques 10 et 11). Par ailleurs, pour le chiffre d'affaire, douze classes ressortent de l'algorithme en janvier 2016, contre seize classes en juillet 2016 (cf. graphiques 12 et 13).

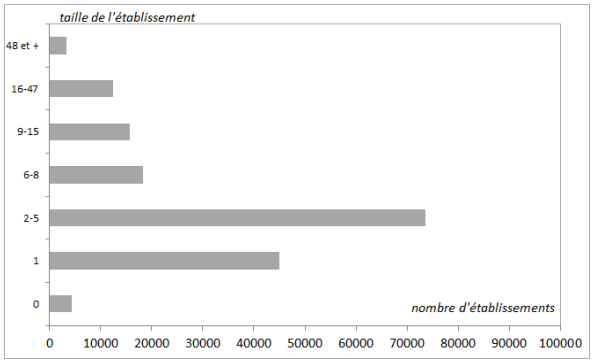
En outre, il est possible de vérifier facilement la présence de non-linéarité avec les modèles qui ont servi à la constitution des classes. Par exemple, en juillet 2016 dans le secteur de la construction :

- Le coefficient associé à l'effectif s'établit à -0,20 pour les établissements ayant seulement 1 salarié. Il s'élève à +0,27 pour les établissements ayant entre 16 et 47 salariés, alors que pour les établissements de plus de 48 salariés, le coefficient associé à la classe d'effectif est nul (cf. tableau 1) ;
- Le paramètre associé aux chiffres d'affaire importants (plus de 7978 k€) est plus bien fort que ceux des chiffres d'affaire plus modestes (1,69 contre un coefficient inférieur à 1,04 pour les autres classes ; cf. tableau 2), conformément aux règles d'entrée dans le dispositif des DSN.

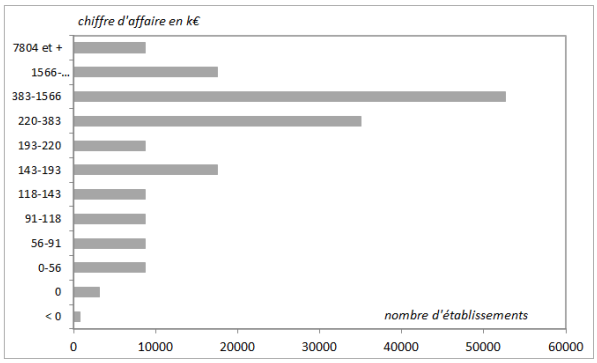
janvier 2016

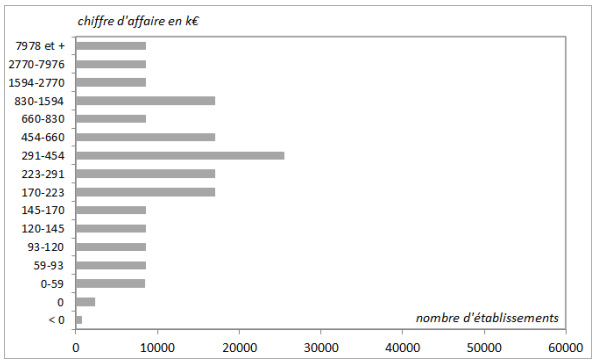



juillet 2016




[6] Pour rendre les tests interprétables, les matrices de variance-covariance sont préalablement modifiées, afin d'adapter les tests aux effets de sur-dispersion (respectivement de sous-dispersion). Si la sur-dispersion (resp. sous-dispersion) n'intervient pas dans l'estimation des paramètres, elle rend les tests trop significatifs (resp. pas assez significatifs).




*Graphiques 10 à 13 – Exemples de classes retenues pour les effectifs et les chiffres d'affaires, pour les mois de janvier et juillet 2016, dans le secteur de la construction.*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| mai-15 | effectif | 0 | 1 | 2-3 | 4-5 | 6-10 | 11-22 | 23-47 | 48 et + | |
| | coefficient | -1,77 | -0,46 | -0,21 | 0,13 | 0,47 | 0,68 | 0,47 | 0,00 | |
| oct-15 | effectif | 1 | 2-7 | 8-48 | 49-1516 | 0,00 | | | | |
| | coefficient | -0,02 | 0,08 | 0,18 | 0,39 | 0,00 | | | | |
| janv-16 | effectif | 0 | 1 | 2-8 | 9-15 | 16-22 | 23-47 | 48 et + | | |
| | coefficient | -0,96 | 0,11 | 0,17 | 0,26 | 0,19 | 0,07 | 0,00 | | |
| avr-16 | effectif | 0 | 1 | 2-2 | 3-8 | 9-22 | 23-47 | 48 et + | | |
| | coefficient | -0,66 | 0,07 | 0,17 | 0,12 | 0,18 | 0,00 | 0,00 | | |
| juil-16 | effectif | 0 | 1 | 2-5 | 6-8 | 9-15 | 16-47 | 48 et + | | |
| | coefficient | -0,82 | -0,20 | -0,07 | -0,01 | 0,15 | 0,27 | 0,00 | | |
| oct-16 | effectif | 0 | 1 | 2-5 | 6-8 | 9-10 | 11-15 | 16-47 | 48 et + | |
| | coefficient | -0,75 | -0,31 | -0,13 | -0,06 | 0,08 | 0,15 | 0,29 | 0,00 | |
| janv-17 | effectif | 0 | 1 | 2-2 | 3-4 | 5-10 | 11-13 | 14-20 | 21-43 | 44 et + |
| | coefficient | -1,02 | -0,63 | -0,16 | -0,07 | 0,06 | -0,09 | 0,34 | 0,55 | 0,00 |

*Tableau 1 – estimation des paramètres associés aux classes des effectifs (secteur de la construction)*

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mai-15 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-58 | 58-287 | 287-824 | 824-2725 | 2725-7738 | 7743 et + | | | | | | | | | | |
| | coefficient | 0,00 | 1,01 | -1,73 | -0,82 | -0,65 | -0,26 | 0,45 | 2,86 | | | | | | | | | | |
| oct-15 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-61 | 61-93 | 93-120 | 120-223 | 223-292 | 292-389 | 389-2784 | 2785-7925 | 7928 et + | | | | | | | |
| | coefficient | 0,00 | -0,83 | -1,04 | -0,27 | 0,03 | 0,26 | 0,44 | 0,35 | 0,43 | 0,59 | 1,93 | | | | | | | |
| janv-16 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-56 | 56-91 | 91-118 | 118-143 | 143-193 | 193-220 | 220-383 | 383-1566 | 1566-7802 | 7804 et + | | | | | | |
| | coefficient | 0,00 | -1,19 | -1,38 | -0,38 | 0,01 | 0,14 | 0,23 | 0,29 | 0,38 | 0,45 | 0,54 | 1,31 | | | | | | |
| avr-16 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-57 | 57-92 | 92-119 | 119-144 | 144-169 | 169-222 | 222-386 | 386-2750 | 2750-7936 | 7937 et + | | | | | | |
| | coefficient | 0,00 | -1,10 | -1,33 | -0,36 | 0,04 | 0,20 | 0,28 | 0,34 | 0,48 | 0,57 | 0,66 | 1,33 | | | | | | |
| juil-16 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-59 | 59-93 | 93-120 | 120-145 | 145-170 | 170-223 | 223-291 | 291-454 | 454-660 | 660-830 | 830-1594 | 1594-2770 | 2770-7976 | 7978 et + | | |
| | coefficient | 0,00 | -1,47 | -1,72 | -0,71 | -0,27 | -0,07 | 0,03 | 0,13 | 0,28 | 0,37 | 0,46 | 0,57 | 0,64 | 0,80 | 1,04 | 1,69 | | |
| oct-16 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-60 | 60-94 | 94-121 | 121-146 | 146-171 | 171-197 | 197-225 | 225-293 | 293-391 | 391-458 | 458-666 | 666-839 | 839-1611 | 1612-2805 | 2805-8096 | 8097 et + |
| | coefficient | 0,00 | -1,56 | -1,80 | -0,78 | -0,33 | -0,11 | 0,01 | 0,10 | 0,18 | 0,30 | 0,35 | 0,43 | 0,48 | 0,61 | 0,67 | 0,81 | 1,06 | 1,63 |
| janv-17 | chiffre d'affaire en k€ | non-renseigné | < 0 | 0-52 | 52-80 | 80-107 | 107-132 | 132-156 | 156-207 | 207-237 | 237-272 | 272-509 | 509-622 | 622-1040 | 1040 et + | | | | |
| | coefficient | 0,00 | -1,55 | -1,89 | -0,88 | -0,43 | -0,03 | 0,16 | 0,38 | 0,61 | 0,72 | 0,84 | 0,95 | 1,06 | 1,17 | | | | |

*Tableau 2 – estimation des paramètres associés aux classes des chiffres d'affaires (secteur de la construction)*

## 6. Les nouveaux modèles corrigent-t-il totalement le biais de non-déclaration ?

L'objectif principal de toutes les méthodes de redressement est de réduire le biais de non-réponse, qui est dû au fait que les déclarants présentent des caractéristiques différentes des non-déclarants. La correction et l'ampleur du biais de non-réponse dépend du type de non-réponse présent dans les données :

- un mécanisme de non-réponse est uniforme (« Missing Completely At Random ») si la probabilité de non-réponse de dépend pas des variables d'intérêt ($\forall i,\ p_i = p$) . Ce type de mécanisme au niveau de la population n'est généralement pas réaliste, mais s'avère particulièrement efficace dans notre contexte de classes de repondération (cf. *supra*) ;
- un mécanisme de non-réponse est ignorable (« Missing At Random »), si après avoir pris en compte toute l'information auxiliaire appropriée *x,* la probabilité de réponse ne dépend pas des variables d'intérêt ($P(r_i = 1/y, x) = P(r_i = 1 \,/\, x)$).
- si un mécanisme de non-réponse n'est pas ignorable, il est qualifié de non-ignorable (« Non Missing At Random » ; lorsqu'il existe un lien de causalité entre la variable d'intérêt et la probabilité de non-réponse). Dans ce cas, les statistiques sur la variable d'intérêt sont biaisées.

Il est difficile de connaître le type de mécanisme de non-réponse qui se trouve dans les données, et de déterminer l'ampleur du biais de non-réponse induit. Toutefois, introduire les DPAE (Déclarations préalables à l'embauche) dans les modèles fournit des informations (limitées) sur le biais.

S'il s'avère difficile de mesurer l'ampleur d'un éventuel biais dû à la non-déclaration, il est néanmoins possible de juger de l'ampleur de l'erreur totale, et par conséquent d'évaluer l'efficacité de la correction apportée. Les estimations obtenues à partir de l'échantillon de déclarants peuvent, en effet, être comparées aux agrégats connus, qui sont eux-mêmes calculés à partir des bases de référence.

### *La Déclaration préalable à l'embauche*

Généralement, il est impossible de savoir si l'on se trouve dans une situation d'un mécanisme de non-réponse ignorable, ou d'un mécanisme de non-réponse non-ignorable. Néanmoins, un taux d'entrée dans les établissements est calculable *via* les DPAE. Cette variable est *a priori* une bonne approximation du taux d'entrée issu des DSN. Par conséquent, en testant le pouvoir explicatif de ces variables dans les modélisations de la non-déclaration, il est possible d'avoir une idée du mécanisme de non-réponse :

- L'hypothèse nulle des tests de nullité appliqués aux paramètres associés aux taux d'entrée en CDI dans les établissements entre les mois de mai 2015 et décembre 2016 est acceptée dans plus de 80 % des cas au niveau 1 % (cf. tableau 3) ;
- Quant aux tests de nullité des paramètres associés aux taux d'entrée en CDD, l'hypothèse nulle est acceptée dans un peu plus de 70 % des cas.

Ainsi, le comportement de réponse à la DSN ne dépendrait que faiblement des flux de main-d'œuvre des établissements. Cette absence d'effet de causalité signifierait que l'on se trouve dans une situation d'un mécanisme de non-réponse ignorable, c'est-à-dire une situation où le biais de non-déclaration est corrigé par le traitement par pondération décrit précédemment.

| | construction | industrie | tertiaire | Ensemble |
|---|---|---|---|---|
| taux d'entrée en cdd | 25 | 30 | 30 | **28** |
| | 75 | 70 | 70 | **72** |
| taux d'entrée en cdi | 15 | 5 | 30 | **17** |
| | 85 | 95 | 70 | **83** |

*Tableau 3 – bilan des tests de nullité des coefficients associés aux taux d'entrée dans les établissements*
*Note de lecture : les variables sont issues des DPAE ; les tests ont été appliqués pour chaque mois entre mai 2015 et décembre 2016 ; part de test dont l'hypothèse nulle a été acceptée au niveau 1 % (en %).*

### *L'erreur totale*

S'il n'est pas possible de calculer le biais causé par la non-déclaration de certains établissements, il est toutefois possible de connaître l'erreur totale, c'est-à-dire l'erreur résiduelle après redressement des données. Pour cela, on compare les estimations obtenues à partir de l'échantillon de déclarants (notées $\widehat{X_\pi}$), aux agrégats connus calculés sur les bases de référence (répondants et non-répondants confondus ; notés $X$). L'erreur totale s'écrit alors sous la forme $\theta = \widehat{X_\pi} - X = \sum_{i\,\epsilon\,R} p_i . X_i - \sum_{i\,\epsilon\,bdr} X_i$ et l'erreur relative $\varepsilon = \frac{(\widehat{X_\pi} - X)}{X}$. Pour que le calcul soit rendu possible, il est nécessaire que la variable soit renseignée pour tous les établissements des bases de référence. En conséquence, trois variables sont retenues :

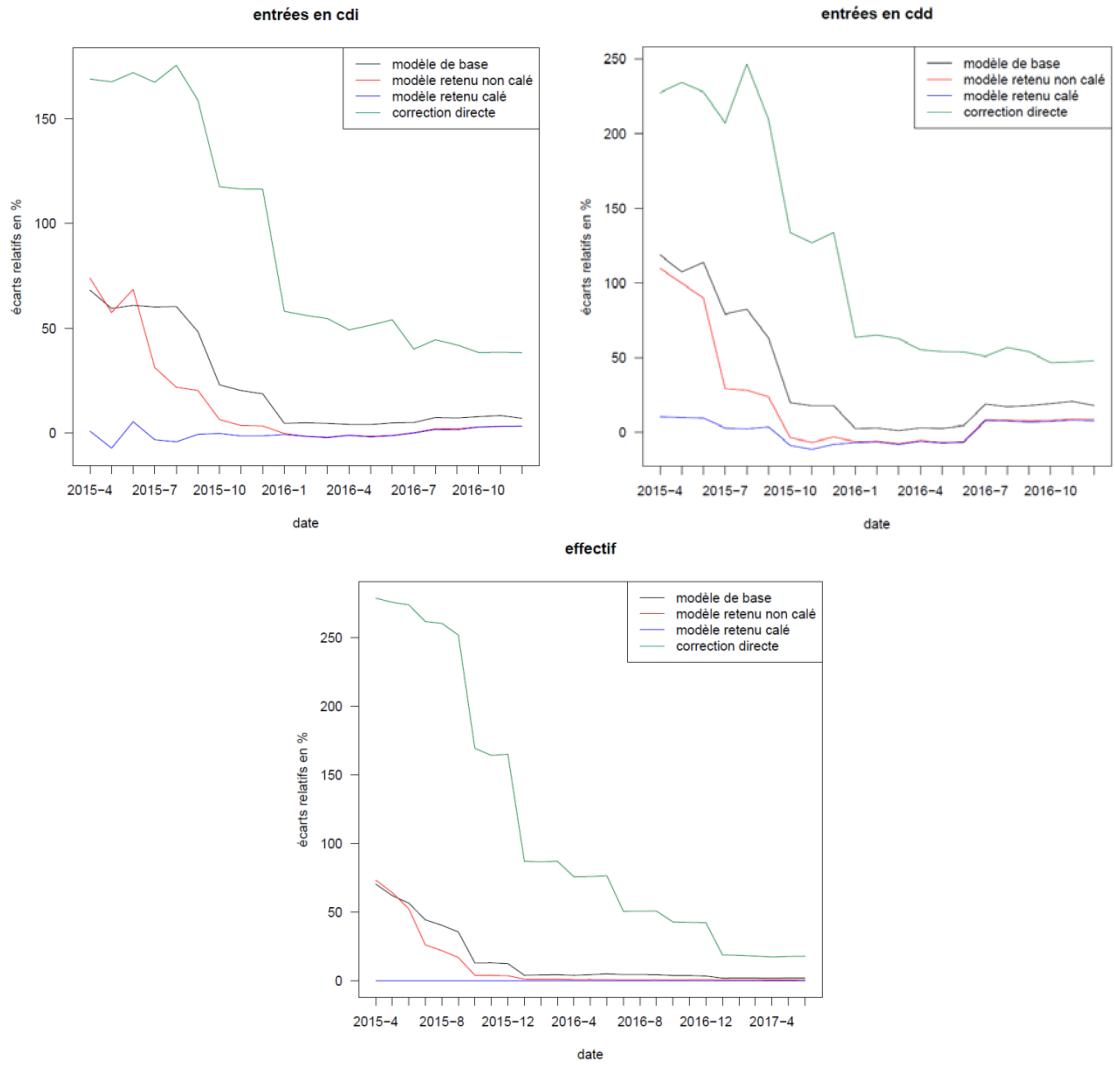
- l'effectif de référence ;
- le nombre d'entrées en CDI (issu des DPAE) ;
- le nombre d'entrées en CDD (des DPAE).

En outre, l'obtention des estimations nécessite l'utilisation de jeux de pondérations. Ici, différentes estimations sont confrontées. Elles sont calculées à partir :

- des poids constants (égaux à l'inverse du taux de déclaration ; « correction directe » efficace et suffisante dans le cas d'un mécanisme de non-réponse uniforme) ;
- des poids non calés issus du modèle « de base », après création des groupes de réponse homogène (GRH) ; « correction modèle de base » ;
- des poids non calés issus des modèles retenus, après création des GRH ; correction « modèles retenus » ;
- des poids calés issus des modèles retenus (pour le calage, cf. fiche 5).

Globalement, les estimations obtenues *via* les poids issus des modèles retenus minimisent l'erreur totale, comparativement à celles qui sont obtenues avec les poids du modèle de base (cf. graphiques 14 à 16). Plus particulièrement, si la correction « modèle de base » surestime, en moyenne, de 18,7 % le nombre d'entrées en CDD (respectivement 7,1 % en CDI) au second semestre 2016, la correction « modèles retenus » le surestime de 8,4 % (resp. 2,3 %). À noter

qu'en utilisant les poids calés issus des modèles retenus, la surestimation est réduite à 7,4 % pour le nombre d'entrées en CDD, et s'établit à 2,5 % pour le nombre d'entrée en CDI.
Ces erreurs ne sont pas forcément le signe d'un biais de non-déclaration, et peuvent être le reflet de la variance des estimateurs. Aussi, les estimations, calculées avec les pondérations issus des modèles retenus (calées ou non), sont inférieures au nombre d'entrées relevées par les DPAE au premier semestre 2016.




*Graphiques 14, 15 et 16 – Comparaison des estimations effectuées à partir des déclarants, avec les grandeurs connues dans les bases de référence*

## 7. Mise en place d'un calage sur marge

Une fois le traitement de la non-réponse effectuée, il est nécessaire d'effectuer un calage pour assurer la cohérence entre les estimations obtenues à partir des établissements déclarants, et les totaux connus sur les bases de référence. Pour cela, un calage sur marge est choisi ; ce dernier permet en effet d'assurer, outre la cohérence des estimations, une réduction de la variance des estimateurs.

Dans le cadre de la correction des DSN, il a été choisi de modifier le jeu de pondération, afin de caler les estimations d'effectifs salariés (sur les déclarants), sur celles des bases de référence (appelées marges). Plus précisément, les marges sont les effectifs calculés à partir des bases de référence sur différentes strates obtenues par le croisement de trois variables : la taille de l'établissement, son secteur d'activité et sa localisation. Si mi-2015, les variables de stratification ont été considérées à un niveau agrégé de la nomenclature (52 marges), à partir de 2016, elles l'ont été à un niveau plus désagrégé (plus de 1000 marges à partir de janvier 2016), offrant par ailleurs un plus grand nombre d'axes d'analyse.

### a. Principe du calage mis en place

Pallier la non-déclaration de certains établissements nécessite la création d'un jeu de poids (noté $\{p_i\ ;\ i \in R\}$). Chaque poids est alors affecté à un établissement déclarant pour compenser la non-réponse des autres établissements (non-déclarants). Pour cela, il est nécessaire d'estimer la probabilité de déclaration des établissements à la DSN et de créer des groupes homogènes de répondants (cf. fiche 4). En outre, le premier jeu de pondération correspond à l'inverse des taux de réponses dans chacune des classes.

Pour chaque base de référence, différents totaux sont connus, notamment le total des effectifs sur l'ensemble des établissements (noté $E$). Généralement, l'estimation $\widehat{E_\pi}$ de l'effectif total, calculé à partir des déclarants, ne coïncide pas avec le « vrai » total $E$ ($\widehat{E_\pi} = \sum_{i \in R} p_i . E_i \neq E = \sum_{i \in bdr} E_i$). Il y a alors une incohérence entre les estimations et les totaux connus de la population.

Effectuer un calage consiste à ajuster le poids des établissements de manière à ce que les estimations faites sur la population d'établissements répondants coïncident avec les totaux connus pour l'ensemble de la population. Ce procédé peut également améliorer la précision des estimateurs[7].

De nombreux jeux de pondérations vérifient les contraintes de calage. Le principe est de minimiser la déformation du premier jeu de poids, sous la contrainte de respecter les marges.

[7] La réduction de variance est due à la corrélation entre les variables sur lesquelles on prend les marges et la variable d'intérêt : plus la corrélation est forte, plus la variance est diminuée. Plus précisément, l'estimateur par calage sur marge est équivalent à un estimateur par régression généralisée. Toutefois, si les variables de calage sont peu explicatives de la variable d'intérêt (voire une relation autre que linéaire), il est même possible d'observer une dégradation de la variance.

Différentes fonctions de distance peuvent être utilisées pour quantifier cette déformation. Sous des hypothèses assez générales, étant donnée la fonction de distance $G$, on cherche le jeu de poids $\{w_i\ ;\ i \in R\}$ résolvant le problème d'optimisation suivant :

$$\begin{cases} \text{Min}\left(\sum_{i \in R} p_i . G(^{w_i}/_{p_i})\right) \\ s.c. \sum_{i \in R} w_i . E_i - E = 0 \end{cases}$$

Le programme CALMAR de l'Insee est utilisé pour résoudre ce problème d'optimisation, en utilisant une fonction de distance de type *logit* (cf. Sautory 1993).

### b. Application

Étant donné que le nombre de mouvements est corrélé à l'effectif, les marges choisies ici correspondent aux effectifs salariés, calculés sur différentes strates, à partir des bases de référence, c'est-à-dire sur le champ des répondants et des non-répondants. Plus précisément, les strates sont construites à partir de trois variables : le secteur d'activité, la taille de l'établissement et la région de l'établissement.

Comme la convergence de l'algorithme[8] n'est pas toujours garantie, l'utilisation des variables de stratification à un niveau fin de nomenclatures peut s'avérer délicate. La recherche de la meilleure configuration des strates est faite avec des bornes inférieures et supérieures de la fonction d'ajustement (c'est-à-dire du rapport entre les poids avant et après calage) respectivement fixées à 1/4 et 4. Une fois les strates choisies, des bornes plus restrictives ont été testées (par exemple 1/3 et 3 ou encore 2/5 et 5/2). Si la convergence de l'algorithme est obtenue et que les traînes des distributions des rapports ancienne/nouvelle pondérations ne sont pas trop étalées, alors la configuration est retenue[9]. Dans le cas contraire, le processus est réitéré avec des nomenclatures plus agrégées.

Aussi, le nombre de marges retenues peut différer d'un mois à l'autre. Si le calage effectué implique 52 marges en avril 2015, ce nombre augmente selon les paliers de montée en charge de la DSN (cf. fiche 2), jusqu'à environ 1400 à la mi-2017. Les nomenclatures utilisées ne sont pas exactement celles qui sont définies aux tableaux 4 et 5 : quelques modifications ont pu être apportées afin d'assurer la convergence de l'algorithme. En particulier :

- pour certains secteurs considérés en A17, la distinction par taille d'établissements n'est pas effectuée (ex : les activités financières et d'assurance) ;
- les industries extractives et les télécommunications sont considérées sur l'échelon national ;

[8] L'algorithme de Newton-Raphson est utilisé pour résoudre le problème de minimisation présenté au paragraphe précédent.

[9] À noter qu'au deuxième trimestre 2015, et dans une moindre mesure au troisième trimestre 2015, les queues de la distribution premier/dernier poids apparaissent légèrement trop épaisses, notamment en raison du faible taux de déclaration. La qualité des poids finaux en est donc altérée.

- le secteur de la cokéfaction-raffinage est rattaché à celui de l'industrie chimique ;
- le secteur de la construction n'est pas considéré au niveau A38 (1 position) mais au niveau A88 (3 positions) ;
- à partir du quatrième trimestre 2015, les marges les plus petites (c'est-à-dire pour lesquelles le nombre de Siret servant au calcul est inférieur à un seuil) sont fusionnées selon leur région[10]. Par exemple, en octobre 2015, les établissements ayant entre 10 et 49 salariés du secteur industriel « fabrication d'équipements » (C3 au niveau A17) des régions Centre-Val de Loire, Bourgogne Franche-Comté, Normandie, Bretagne et PACA constituent une seule et même marge.

| périodes | secteurs | taille | région | seuil fusion | nombre moyen de marges |
|---|---|---|---|---|---|
| T2-2015 | A17 | 2 | 2 | | 52 |
| T3-2015 | A17 | 2 | 5 | | 150 |
| T4-2015 | A17 | 3 | 12 | 4000 | 464 |
| T1-T2-2016 | A38 | 3 | 12 | 1000 | 1068 |
| T3-2016 et après | A38 | 4 | 12 | 1000 | 1400 |

*Tableau 4 – Nombre de marges de calage et nomenclatures des variables de la stratification*

*Note de lecture : au troisième trimestre 2015, les strates de calages sont définies à partir (i) du niveau A17 de la nomenclature des activités (légèrement modifié cf. supra), (ii) des effectifs des établissements selon deux catégories, les moins de 50 salariés et les plus de 50 salariés, (iii) de cinq grandes régions, définies au tableau 5*

| taille | | | | |
|---|---|---|---|---|
| 2 | 1 à 49 | 50 et plus | | |
| 3 | 1 à 9 | 10 à 49 | 49 et plus | |
| 4 | 1 à 9 | 10 à 49 | 50 à 99 | 100 et plus |
| région | | | | |
| 2 | Nord/Sud | | | |
| 5 | Les nouvelles régions sont regroupées en cinq groupes [11] | | | |
| 12 | nouvelle région, la Corse étant rattachée à la région PACA | | | |

*Tableau 5 – nomenclatures des variables de la stratification*

*Note de lecture : lorsque la variable « taille de l'établissement » dans la stratification* ante *calage, est utilisée selon deux catégories, ces dernières séparent les établissements de plus de 50 salariés, de ceux qui ont moins de 50 salariés.*

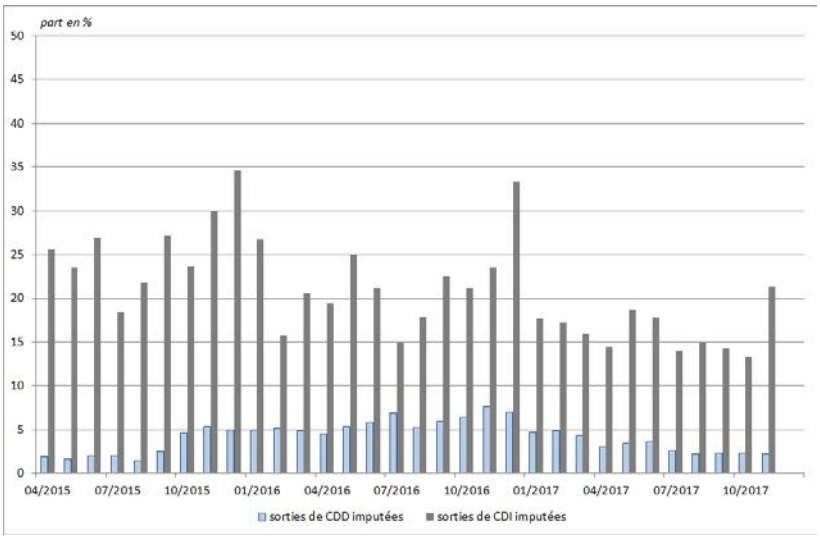
[10] Le maintien du secteur d'activité et de la taille de l'établissement dans la construction des strates est considéré comme prioritaire.

[11] Les nouvelles régions sont regroupées en cinq groupes :
- Bretagne, Normandie ;
- Occitanie, Nouvelle Aquitaine ;
- Grand-Est, Hauts-de-France ;
- Auvergne-Rhône-Alpes, Paca, Corse ;
- Ile-de-France, Centre-Val-de-Loire, Bourgogne-France-Comté.

# Fiche 3 – Imputation des valeurs manquantes

Pour reconstruire des MMO à partir des DSN, un travail important de mise en cohérence des données est nécessaire. L'enjeu est de pouvoir suivre de déclaration en déclaration (*i.e.* de mois en mois pour un établissement) la vie d'un contrat afin de détecter les embauches et les fins de contrat. Il s'agit d'un processus relativement complexe. Par exemple, si un établissement modifie la date de début de contrat d'un de ses salariés, il convient de prendre en compte cette modification en ne conservant qu'un seul contrat déclaré.

Les traitements nécessaires pour passer des données de « stock » de la DSN aux données de « flux » des MMO sont réalisés *via* une application développée par la Dares. Mais d'autres traitements statistiques sont également menés. En effet, dans certaines DSN, il arrive que des contrats actifs au cours d'un mois ne soient pas retrouvés le mois suivant, sans qu'aucune fin de contrat ne soit déclarée. Dans la plupart des cas, il s'agit bien de fins de contrat mais qui sont mal déclarées, de sorte que ni le motif ni la date de fin de contrat ne sont connues. En moyenne en 2017, ces anomalies représentent 16 % des fins de CDI, et un peu plus de 3 % des fins de CDD (cf. graphique 17). Le traitement de ces anomalies est détaillé ci-après.




*Graphique 17 – Part des sorties avec une date de fin et motif imputé, selon la nature de contrat*

## *La correction des dates de fin de contrat*

Deux cas sont à distinguer dans l'assignation de dates de fin dans les contrats :

- Les CDD, pour lesquels une date de fin prévisionnelle est souvent renseignée à la signature du contrat. C'est donc cette date qui est retenue pour les fins de CDD non déclarées.
- Les CDI, ou les CDD pour lesquels aucune date de fin prévisionnelle n'est renseignée. Dans ce cas, une date de fin de contrat est sélectionnée aléatoirement sur le mois de disparition.

### *L'imputation des motifs de fin de contrat*

Dans le cas d'une sortie non déclarée, un motif de fin est imputé à cette fin de contrat. Trois variables corrélées aux motifs de rupture de CDI[12] sont sélectionnées pour créer des classes d'imputation :

- La durée du contrat, en trois catégories (moins de 365 jours, entre 365 et 1 700 jours, plus de 1 700 jours)[13] ;
- L'âge du salarié, en trois catégories définies selon les terciles de la variable observées sur les déclarants (moins de 31 ans, entre 31 et 45 ans, plus de 45 ans) ;
- La catégorie socioprofessionnelle du salarié en quatre positions (cadres, professions intermédiaires, employés, ouvriers).

Choisir des variables corrélées au type de motif permet de conserver le plus possible le lien entre la variable imputée et les variables servant à la construction des classes. En effet, dans chacune des classes créées, la distribution des motifs imputés est semblable à la distribution des motifs observée dans la classe. Par exemple, si :

- Une classe a 45 % démissions, 40 % de fins de période d'essai, et 15 % de ruptures conventionnelles ;
- 45 % des motifs de sortie imputés dans cette même classe seront des démissions, 40 % des fins de période d'essai, et 15 % des ruptures conventionnelles.

[12] Pour les CDD, seule la variable de la durée est prise en compte pour les classes d'imputation : le motif principal de fin de contrat est l'arrivée à l'échéance du contrat (à 98 % sur les sorties de CDD observées).
[13] La durée médiane des CDI rompus après une année est d'environ 1700 jours.

# Fiche 4 – Traitement des ruptures de séries des mouvements de main-d'œuvre

La prolongation des séries de MMO issues des sources historiques par les séries reconstituées à partir des DSN n'est pas directe. Les taux d'entrée et de sortie présentent en effet des ruptures de séries. Ces dernières sont à la fois constatées sur les séries qui concernent les CDD, et celles qui concernent les CDI. Elles sont de plus grandes ampleurs pour les établissements de moins de 50 salariés que pour les établissements de plus de 50 salariés.

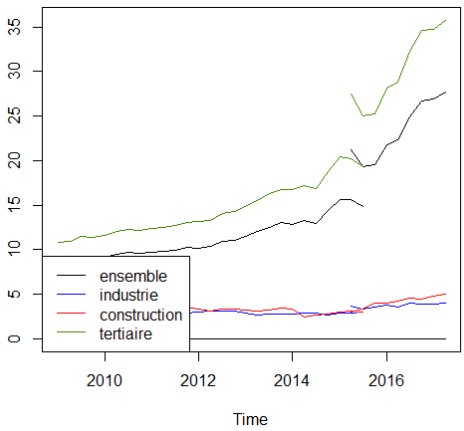
Cette fiche présente, dans un premier temps, les différents éléments qui expliquent les ruptures de série, c'est-à-dire la sous-déclaration des mouvements dans les sources historiques des MMO et la fusion de CDD très courts en DSN lors de la montée en charge. Dans un second temps, elle présente les méthodes de correction utilisées pour assurer la continuité des séries de mouvements de main-d'œuvre : elles consistent à créer différents jeux de poids, qui offrent une cohérence entre données détaillées au niveau établissement et données agrégées.

La correction de la sous-déclaration des mouvements dans les données est satisfaisante pour les établissements de plus de 10 salariés. Des séries longues sont donc diffusées (à partir de 1993). Pour les établissements de moins de 10 salariés, seules les données concernant les CDI semblent correctement corrigées. Pour analyser et corriger correctement les taux d'entrée en CDD et les taux de fin de CDD dans les très petits établissements, il parait nécessaire d'avoir davantage de recul.

## 1. Les ruptures de séries observées en 2015 sont expliquées par les imperfections des MMO historiques

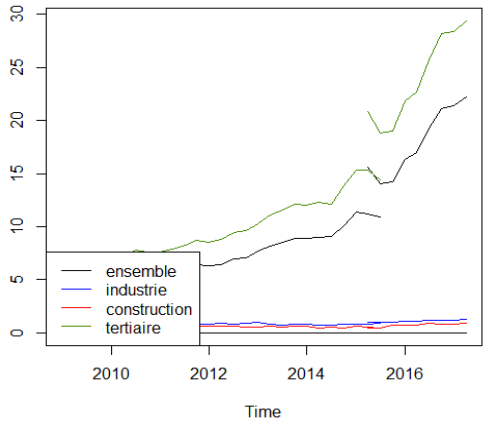
### a. Une déclaration plus systématique des CDD très courts en DSN

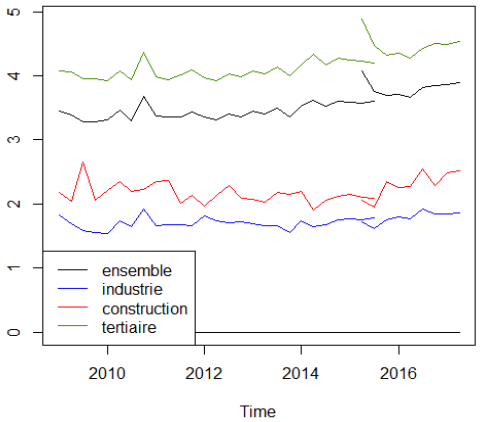
Le passage de la source historique des MMO à la DSN implique un rehaussement moyen de 29 % du nombre d'entrées en CDD sur le deuxième et troisième trimestre 2015. Cette réévaluation est particulièrement importante dans le secteur des services (cf. graphique 18).




*Graphique 18 - Taux d'entrée en CDD des séries non rétropolées*

Ce plus grand nombre d'embauches en CDD recensé en DSN s'explique notamment par une déclaration plus systématique des contrats très courts (de moins d'un mois). Les séries concernant les fins de CDD de plus d'un mois ne présentent pas de rupture de série, contrairement à celles des fins de CDD de moins d'un mois (cf. graphiques *19* et *20*).




*Graphiques 19 et 20 –Taux de sortie des CDD de moins d'un mois (à gauche) et plus d'un mois (à droite),*

### b. Un problème avéré de qualité de déclaration en DSN en 2015 biaise les évolutions conjoncturelles du taux d'entrée en CDD

En 2015, le changement de source s'accompagne, en plus d'une rupture en niveau, d'une plus forte hausse du nombre d'embauches en CDD. En effet, le taux d'entrée trimestriel moyen en CDD en 2014 a augmenté de 1,0 point par rapport à 2013 (après +1,6 et +0,7 point en 2013 et 2012 ; cf. tableau 6 et graphique 18). Cette hausse s'est ensuite accélérée en 2015, à +2,4 points. Avec les données issues de la DSN, la hausse demeure également soutenue (+3,0 points en 2016, puis +1,3 point en 2017). Il s'agit alors de savoir si ce changement de rythme est imputable au cycle conjoncturel, ou à la période de transition entre les deux supports de déclaration.

| | **MMO** | | | | **DSN** | | |
|---|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | 2015 | 2015 | 2016 | 2017 |
| taux d'entrée trimestriel moyen (en %) | 10,6 | 12,3 | 13,3 | 15,7 | 21,0 | 23,9 | 25,2 |
| Evolution moyenne en points | 0,7 | 1,6 | 1,0 | 2,4 | | 3,0 | 1,3 |

*Tableau 6 – dynamique des taux d'entrée en CDD*

Ces évolutions peuvent être comparées à celles des déclarations préalables à l'embauche (DPAE), autre source de données mesurant les embauches. Il s'avère que le changement provient probablement d'une sous-déclaration des CDD en DSN à la mi-2015 :

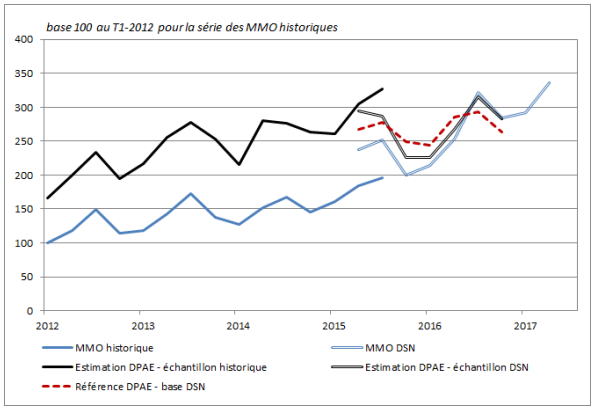
- Les estimations du nombre d'embauches en CDD dans les DPAE sur le champ MMO (restriction sur l'échantillon des déclarants en DSN) sont proches des valeurs référentes correspondantes : le redressement appliqué aux données DSN s'avère donc de bonne qualité (cf. graphique 21).

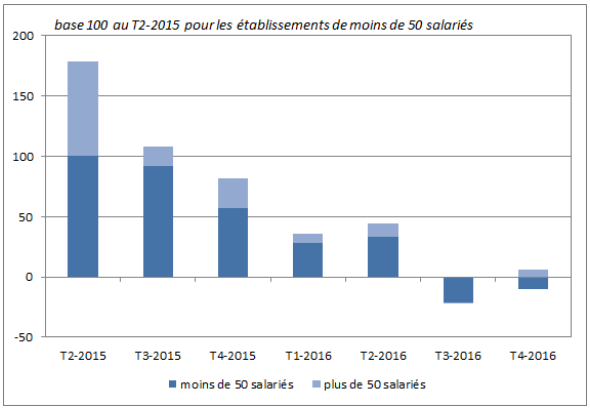
- Toutefois, un écart important subsiste entre les estimations du nombre d'embauches en CDD déclarées en DSN, et celles du nombre d'embauches en CDD dans les DPAE. Cet écart particulièrement élevé au deuxième trimestre 2015 s'atténue les trimestres suivants (cf. graphique 21). Ces écarts s'interprètent comme une sous-déclaration des CDD en 2015. Plus précisément, des contrats très courts ont pu être fusionnés. Par exemple, pour une personne qui a eu cinq contrats sur une semaine, un seul est comptabilisé sur cette même période. Des campagnes de communication, qui ont débuté fin 2015, ont incité les établissements à déclarer l'ensemble de leurs contrats, et par conséquent à éviter les fusions de contrats.
  En conséquence, la plus forte progression des embauches en CDD n'est pas imputable au cycle conjoncturel, mais serait expliquée par une amélioration de la qualité des déclarations. En particulier, la rupture en niveau en 2015, mise en évidence au paragraphe précédent, aurait dû être plus importante.

La sous-déclaration des CDD en DSN est surtout portée par les établissements de moins de 50 salariés. Plus précisément :

- Au deuxième trimestre 2015, les établissements de moins de 50 salariés ont quasiment autant sous-déclaré que les établissements de plus de 50 salariés. En effet, l'écart entre l'estimation du nombre d'embauches en CDD dans les DPAE et en DSN est porté à 56 % par les établissements de moins de 50 salariés (cf. graphique 22).

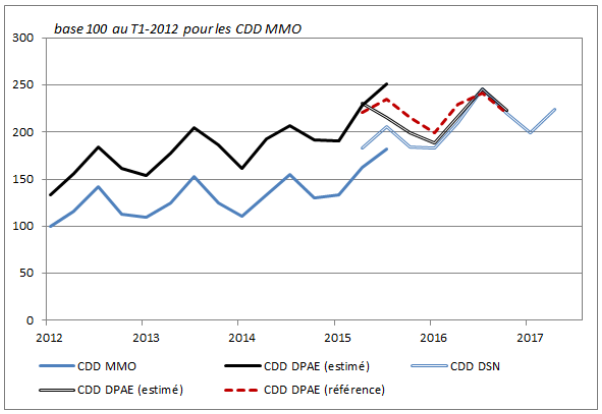
- Par la suite, les établissements de plus de 50 salariés contribuent pour un peu moins du quart de l'écart entre estimation du nombre d'entrées en CDD en DPAE et du nombre d'entrées en CDD en DSN. Ainsi, le problème de sous-déclaration est surtout observé pour les petits établissements (cf. graphiques 23 et 24).

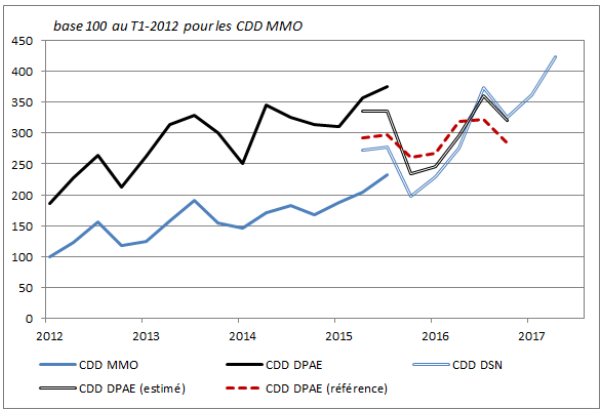



*Graphiques 21 et 22 - Nombre d'entrées en CDD estimé et observé pour l'ensemble des établissements (à gauche), et estimation de la sous-déclaration des CDD en DSN (à droite)*

*Note : la sous-déclaration des entrées en CDD est estimée en faisant la différence entre l'estimation du nombre d'embauches en CDD à partir des DPAE, avec celle obtenue à partir des DSN, sur le champ MMO*



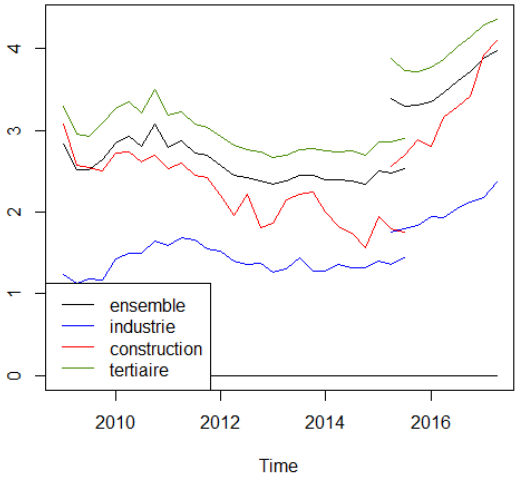



*Graphiques 23 et 24 - Nombre d'embauches en CDD, pour les établissements de plus de 50 salariés (à gauche), et les établissements de moins de 50 salariés (à droite)*

### c. Une sous-déclaration des embauches en CDI, particulièrement marquée dans l'EMMO

#### *Évaluation de la sous-déclaration*

Avec le changement de source, un plus grand nombre d'entrées en CDI est recensé, et ce quel que soit le secteur d'activité considéré (cf. graphique 25). Sur l'ensemble des secteurs, cette différence correspond à un rehaussement de 27 % du nombre d'entrées en CDI, par rapport aux données historiques.




*Graphique 25 - taux d'entrée en CDI,*

Quatre raisons peuvent être envisagées pour expliquer cet écart :

(i) Les déclarants omettaient, en MMO, de déclarer certains mouvements en CDI : il s'agirait d'un problème dit de déclaration individuelle ;
(ii) Le redressement mis en place en MMO n'était pas efficace ;
(iii) Le redressement de non-réponse mis en place en DSN augmente artificiellement le nombre de CDI et biaise les données ;

(iv) Eventuellement, un problème de sur-déclaration en DSN, la référence pour ces analyses étant les données de la DPAE (déclaration préalable à l'embauche).

Des analyses ont été menées pour évaluer laquelle de ces 4 raisons est la plus pertinente. Les résultats sont présentés dans le tableau ci-après :

| | **DSN – MMO** |
|---|---|
| **Ecart DSN – MMO moyen sur les T2-T3 2015** | **100** |
| *Contributions* | |
| **(i) Problème de déclaration individuelle en MMO[14]** | **91** |
| (ii) Problème de redressement des données en MMO[15] | -14 |
| (iii) Problème de redressement des données en DSN[16] | -5 |
| (iv) Sur-déclaration en DSN (par rapport aux DPAE) | 27 |

*Tableau 7 - Décomposition de la réévaluation du nombre de CDI avec le changement de source (en point)*

Sous l'hypothèse que les DPAE donnent la bonne estimation du nombre d'embauches en CDI, il apparait alors que le plus faible nombre d'entrées en CDI en MMO qu'en DSN s'explique essentiellement par une sous-déclaration de ce type de contrat dans le processus historique des MMO, et plus particulièrement d'une sous-déclaration observée dans l'EMMO (91 % de l'écart ; cf. tableau 7).

D'autres facteurs contribuent également à la différence de niveau entre nouvelle et ancienne séries, mais leurs effets sont plus marginaux :

- La méthode historique de redressement des MMO a deux effets opposés. D'un côté, elle limite la rupture de série, en surestimant le « vrai » nombre de CDI en DPAE (minimisant ainsi la rupture de série de 27 points). De l'autre, l'utilisation du traitement particulier des franchissements de seuil[17] pour la construction des séries accentue la rupture de série (contribution de +14 points ; cf. graphique 26).
- La correction apportée aux données DSN semble satisfaisante car elle ne contribue que faiblement à la rupture de série (-5 points ; effet évalué par comparaison du nombre estimé d'embauches en CDI en DPAE sur le champ MMO, au vrai nombre d'embauches sur le champ MMO). En outre, la légère surestimation du nombre d'entrées en CDI en

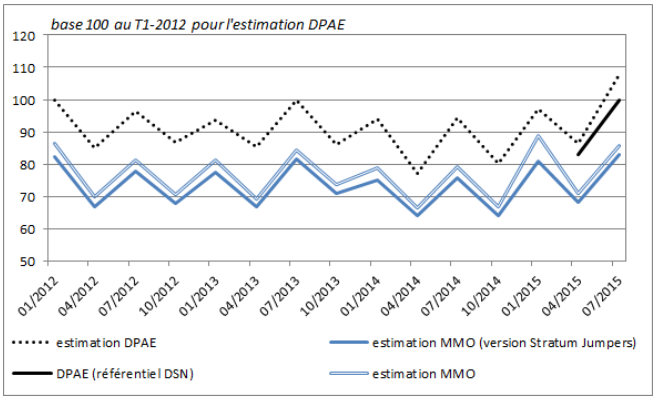
[14] Pour obtenir cet indicateur, les séries MMO historiques sont comparées avec des séries recalculées suivant la même méthodologie mais en substituant les données individuelles MMO par les données de la DPAE.

[15] Cet indicateur est calculé en comparant les séries DPAE calculées sur le champ MMO à des estimations de séries DPAE calculées sur le champ MMO en utilisant la méthode de redressement des données MMO.

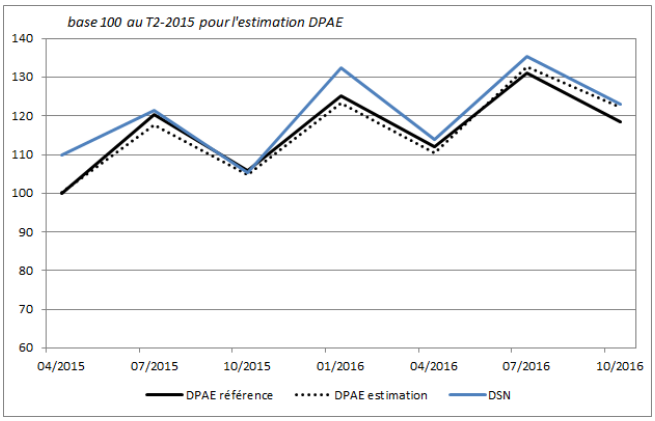
[16] Cet indicateur est calculé en comparant les séries DPAE calculées sur le champ MMO à des estimations de séries DPAE calculées sur le champ MMO en utilisant la méthode de redressement des données DSN.

[17] Un établissement retenu dans l'EMMO avec un effectif en fin d'année précédente inférieur à 10 aura un poids associé conséquent (d'environ 50), car le taux de sondage des TPE est très faible dans l'enquête (environ 2 %). Si sur l'année de l'enquête, l'établissement passe dans la strate exhaustive des DMMO, sa contribution en terme de nombre d'entrées dans sa strate sera d'au moins 50 x 50 = 2500 entrées, ce qui est peu crédible. Cet établissement est alors considéré comme un *stratum jumper*, et nécessite un traitement particulier afin de diminuer sa contribution dans le total des embauches. Ainsi, le poids et les mouvements de cet établissement sont répartis entre des établissements « fictifs » (leur nombre dépend du nombre de strates franchies par le *stratum jumper*).

DSN relativement aux DPAE (contribution de +27 points à la rupture de série) ne semble être que ponctuelle et imputable au deuxième trimestre de 2015 (cf. graphique 27)[18].




*Graphique 26 – Embauches en CDI recensées en MMO et DPAE (sur champ MMO)*




*Graphique 27 - Embauches en CDI recensées en DSN et DPAE (sur champ MMO)*

Ainsi, la comparaison des données MMO et DPAE met en exergue que la réévaluation de nombre d'embauches en CDI provient essentiellement d'une sous-estimation du nombre d'embauches en CDI dans le processus historique des MMO.

### *La sous-déclaration dans le processus historique se limite aux établissements de moins de 50 salariés*

Pour mieux comprendre l'origine de cette sous-déclaration des embauches en CDI dans le processus historique, une dissociation des données selon les sources (DMMO ou EMMO) est effectuée. Sur la période 2012-2014, il apparaît alors que :

- Sur le champ des DMMO, l'estimation du nombre d'embauches en CDI est légèrement inférieure mais relativement proche de celle des DPAE ;

[18] Un écart est également visible sur le premier trimestre 2016. Ce dernier s'explique par un plus large problème de qualité des données en décembre 2015, qui rend difficile l'appariement des DSN de décembre 2015 et celles de janvier 2016 provoquant dès lors une surestimation de mouvements entrants en janvier 2016 (la reconstitution des MMO inclut alors de « faux » mouvements).

- Sur le champ de l'EMMO, l'estimation du nombre de CDI est très inférieure à celle des DPAE. À noter que cet écart n'est pas imputable au jeu de poids utilisé pour l'enquête, puisque l'écart apparaît à la fois sur les données non pondérées et les données pondérées (cf. tableau 8).

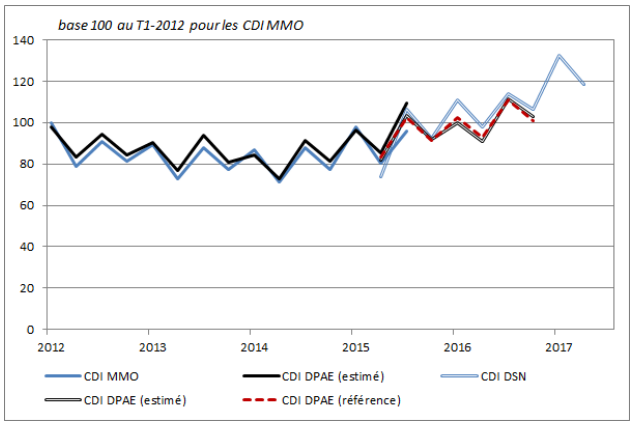
| | CDI | | CDD | |
|---|---|---|---|---|
| | pondéré | non pondéré | pondéré | non pondéré |
| **DMMO** | -3,1 | -2,8 | -39,9 | -38,5 |
| **EMMO** | -29,2 | -28,1 | -86,5 | -89,0 |

*Tableau 8– écarts moyens (en %) entre le nombre d'entrées en CDI estimés en MMO et en DPAE (période 2012-2014)*

*Note : la comparaison des agrégats peut se faire avec l'utilisation (ou non) du jeu de poids issu du processus de redressement (pondéré/ non pondéré)*

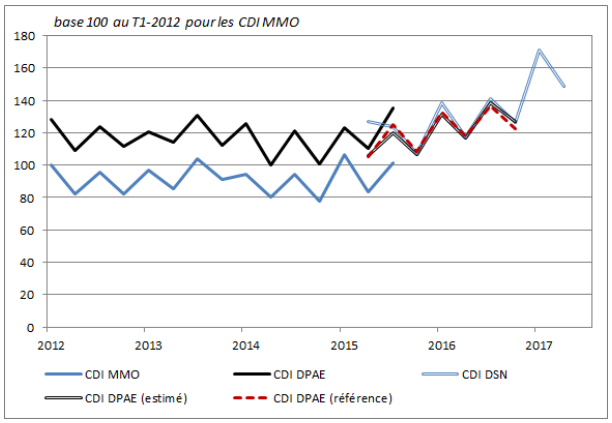
Ce constat est vérifié lorsque les entrées sont différenciées selon l'effectif de référence[19]. En effet, pour les établissements de plus de 50 salariés, la continuité de la série des embauches en CDI est visuellement assurée, tout comme la cohérence avec les données DPAE (cf. graphique 28).

À l'inverse, la rupture de série pour les établissements de moins de 50 salariés est avérée (cf. graphique 29). Au vue de l'écart entre l'estimation des entrées en CDI en MMO/DPAE, les embauches en CDI seraient sous-déclarées en EMMO.




*Graphique 28 - Nombre d'entrées en CDI dans les établissements de plus de 50 salariés*

[19] Un établissement qui déposait des DMMO n'avait pas obligatoirement un effectif de référence supérieur à 50 (environ 7 % des cas). À l'inverse, tous les établissements relevant de l'enquête EMMO avaient un effectif inférieur à 50 salariés. La différenciation ne se fait donc pas sur la source, mais plutôt sur l'effectif de référence (plus ou moins de 50 salariés), afin de garantir une homogénéité du champ dans le temps (processus historique / DSN).




*Graphique 29 - Nombre d'entrées en CDI dans les établissements de moins de 50 salariés*

En résumé, une discontinuité du nombre d'entrées en CDD (écart de 29 % en moyenne T2/T3 2015) est avérée et s'explique par une déclaration plus systématique des contrats courts en DSN. À cela, s'ajoute une sous-déclaration des CDD en DSN à la mi-2015, en particulier dans les établissements de moins de 50 salariés. De même, une rupture de série sur le nombre d'entrée en CDI (écart de 27 %) est observée. Elle s'explique également par une sous-déclaration des contrats dans le processus historique des MMO, et plus particulièrement du côté de l'EMMO.

Deux conséquences découlent de ces constats : la rétropolation des données historiques des MMO et la correction de la sous-déclaration des CDD en DSN entre mi-2015 et mi-2016.

## 2. Rétropolation des données historiques des MMO

Conformément aux constats précédents, il convient de corriger dans les données historiques :

- Les entrées en CDI. Comme la correction des entrées de CDI ne permet pas de corriger de manière satisfaisante les sorties de CDI, qui étaient sujettes à la sous-déclaration, une rétropolation spécifique y est appliquée.
- Les mouvements associés aux CDD de moins d'un mois arrivés à terme. Dans les sources historiques des MMO, les entrées et les sorties ne sont pas reliées et la distinction des CDD selon leur durée n'est possible que sur les sorties. La rétropolation des séries de fin de CDD très courts implique également une rétropolation des entrées en CDD, considérées dans leur ensemble.

Toutes les séries longues considérées ici sont construites à partir des sources historiques des MMO jusqu'au deuxième trimestre 2015. À partir du troisième trimestre 2015, les mouvements de main-d'œuvre sont reconstitués à partir des DSN.

Ces rétropolations des données historiques des MMO se font au niveau des établissements. Cette approche a l'avantage de maintenir une cohérence entre les données individuelles et toutes les données agrégées publiées, quel que soient les échelons considérés (sectoriels et ou géographiques). Ce travail est jalonné par deux grandes étapes : tout d'abord l'estimation des ruptures de séries sur un ensemble de strates préalablement fixé et commun à toutes les séries ;

puis la création de nouveaux jeux de poids. L'ensemble des étapes de rétropolation est résumé en encadré 4.

Encadré 4 – Résumé des différentes étapes de la rétropolation des données de mouvements de main-d'œuvre

Étape 1 – Les sorties sur la période 2001-2015, issues des sources historiques des MMO, sont ventilées selon la nature des contrats (CDD/CDI). Un peu moins de 2 % des natures de contrats sont indisponibles. Dans cette étape préliminaire, un traitement est mis en place pour leur imputer une nature de contrat.

Étape 2 – L'estimation de l'ampleur de la rupture

- Des strates de rétropolation sont tout d'abord créées, par le croisement de deux variables, le secteur d'activité et la taille de l'établissement.
- Les séries recensant les fins de CDD très courts, les embauches en CDI et les ruptures de CDI, sont ensuite constituées sur chaque strate de rétropolation. De début 2009 au deuxième trimestre 2015, les séries sont issues des sources historiques des MMO. À partir du troisième trimestre 2015, elles proviennent des DSN.
- L'ampleur des ruptures en niveau est estimée économétriquement.

Étape 3 – Une mise en cohérence des données détaillées avec les agrégats est ensuite menée. À cette fin, des nouveaux jeux de poids sont créés. L'objectif est d'attribuer des poids aux mouvements au sein des établissements qui appartiennent à des strates affectées par la sous-déclaration sur le passé.

Des étapes supplémentaires sont nécessaires pour rétropoler les données antérieures à 2009.

Étape 4 – Avant 2009, une rétropolation des secteurs d'activité des établissements est effectuée pour prendre en compte le changement de nomenclature intervenu en 2008.

Étape 5 – La rétropolation des données sur la période 1998-2000

- Les durées des contrats, sur la période 1998-2000, sont imputées : environ 15 % des valeurs sont manquantes. Cette variable de durée est notamment utilisée pour distinguer les fins de CDD de moins d'un mois de celles de plus d'un mois.
- Sur cette période, la nature des contrats rompus n'est pas connue. Il s'agit alors d'adapter la correction appliquée sur les fins de CDI sur la période 2001-2015, en utilisant des hypothèses *ad hoc*. On ne rétropole pas les fins de CDI sur cette période, mais les ruptures anticipées de contrat (CDI et CDD confondus).

Étape 6 – La rétropolation des données sur la période 1993-1997. Sur cette période, seuls les établissements de plus de 50 salariés sont considérés.

- Les données brutes de la DMMO ont été récupérées. Elles ont tout d'abord été apurées. Puis elles ont été corrigées : d'une part de la non-réponse totale par affection d'un poids de redressement, d'autre part de la non-réponse partielle par imputation. Enfin, les équations comptables reliant flux de main-d'œuvre et effectifs ont été reconstituées.
- La rétropolation des données sur cette période implique deux étapes supplémentaires. Premièrement, la nature des contrats rompus n'est pas connue : le traitement appliqué sur la période 1998-2000 est répliqué. Deuxièmement, la durée des contrats n'est pas connue : en l'absence de la distinction des CDD selon leur durée, les CDD sont considérés dans leur ensemble pour la rétropolation ; l'ampleur des corrections à apporter est alors adaptée à cet élargissement de champ sur la période.

### a. Étape 1 : imputation des données manquantes sur les données historiques (2001-2015)

Les séries finales distinguent les taux de sortie de CDI des taux de sortie de CDD. Pour cela, la nature des contrats associée aux sorties doivent être connues. Or, dans le processus historique de production des mouvements de main-d'œuvre, un peu plus de 2 % des natures de contrat associées aux sorties sont inconnues (cf. tableau 9). Il est alors nécessaire de leur affecter une nature de contrat.

| type d'entrée | part des sorties (en %) |
|---|---|
| CDD | 73,2 |
| CDI | 24,7 |
| Inconnu | 2,1 |

*Tableau 9 – Part moyenne de fins de contrat entre le premier trimestre 2001 et le premier trimestre 2015 (avant rétropolation ; en %)*

Une méthode d'imputation aléatoire est appliquée en tenant compte à la fois du motif de sortie et de la durée du contrat selon trois classes :

- moins de 13 jours ;
- entre 13 et 147 jours ;
- plus de 147 jours.

La méthode du hot-deck aléatoire consiste à sélectionner, pour chaque sortie « receveuse » dont la nature de contrat est inconnue, une sortie « donneuse » dont la nature du contrat est connue, qui lui sera associée et dont la valeur sera la valeur imputée. Plus précisément, l'imputation s'appuie sur la répartition des types de contrat sur chaque domaine. Un domaine est défini par une date, un motif de fin de contrat et une durée de contrat. Par exemple, pour un domaine donné, si $x$ % des contrats, parmi ceux qui ont leur nature renseignée, sont des CDI, et (1-$x$ %) sont des CDD, alors, après imputation, la répartition $x$ % de CDI et (1-$x$ %) de CDD est respectée sur l'ensemble des sorties du domaine (cf. exemple au tableau 10).

| | nature de contrat à imputer | nature de contrat renseignée |
|---|---|---|
| nature de contrat | part des sorties (en %) | part des sorties (en %) |
| CDD | 24,8 | 24,9 |
| CDI | 75,2 | 75,1 |

*Tableau 10 – stratégie d'imputation des natures de contrat pour le domaine suivant : démission du premier trimestre 2014 pour des contrats ayant duré entre 100 et 300 jours.*

### b. Étape 2 : estimation de l'ampleur de la rupture

Différentes strates, communes à chaque type de série, sont préalablement construites, et validées *a posteriori* : le niveau de rétropolation retenu doit être le plus fin possible, mais doit également permettre une détection et une estimation satisfaisante des ruptures de série, sans apporter d'instabilité dans les données (cf. *infra*).

Les strates retenues sont construites à partir de deux variables, qui sont le secteur d'activité et la taille de l'établissement :

- La taille de l'établissement est divisée en trois catégories (1-9 salariés, 10-49 salariés, plus de 50 salariés) ;
- Les secteurs d'activité des établissements de moins de 10 salariés sont considérés au niveau A17 de la NAF rév.2, les secteurs de l'industrie sont regroupés (cf. tableau 11); pour les autres établissements, les secteurs d'activité sont considérés au niveau A38 de la nomenclature[20]. Certains secteurs sont toutefois fusionnés (cf. tableau 12).

| |
|---|
| - Industries extractives, énergie, eau, gestion des déchets et dépollution (DE) ;<br>- Fabrication de denrées alimentaires, de boissons et de produits à base de tabac (C1) ;<br>- Cokéfaction et raffinage (C2) ;<br>- Fabrication d'équipements électriques, électroniques, informatiques ; fabrication de machines (C3) ;<br>- Fabrication de matériels de transport (C4) ;<br>- Fabrication d'autres produits industriels (C5). |

*Tableau 11 – Secteurs d'activité du niveau A17 de la nomenclature qui ont été regroupés pour la rétropolation des données historiques des MMO*

| |
|---|
| Première fusion :<br>- Cockéfaction/raffinage (CD) ;<br>- Industrie chimique (CE) ;<br>- Industrie pharmaceutique (CF). |
| Deuxième fusion :<br>- Fabrication de textiles, industries de l'habillement, industrie du cuir et de la chaussure (CB) ;<br>- Travail du bois, industries du papier et imprimerie (CC) |

[20] Le secteur de l'hébergement et de la restauration est divisé en deux secteurs : d'un côté la restauration, de l'autre l'hébergement. Cette distinction correspond au niveau A88 de la nomenclature NAF rév.2.

| |
|---|
| Troisième fusion :<br>- Fabrication de produits en caoutchouc et en plastique ainsi que d'autres produits minéraux non métalliques (CG) ;<br>- Métallurgie et fabrication de produits métalliques à l'exception des machines et des équipements (CH). |
| Quatrième fusion :<br>- Industries extractives (BZ) ;<br>- Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné (DZ) ;<br>- Production et distribution d'eau ; assainissement, gestion des déchets et dépollution (EZ). |

*Tableau 12 – Fusion de quelques secteurs d'activité du niveau A38 de la nomenclature pour la rétropolation des données historiques des MMO*

Par la suite, la rétropolation se fait sur chacune des strates. Les séries nationales, ou les séries recouvrant l'ensemble des secteurs, sont obtenues par agrégation en fin de processus.

En se plaçant sur une strate donnée, une série historique des MMO (CDD de moins d'un mois arrivés à terme, entrées en CDI, ou sorties de CDI), prolongée par la série reconstituée à partir des données de la DSN s'écrit sous la forme suivante :

$$Y_t^{(1)} = \begin{cases} Y_t^{(MMO)}, \text{ t}\leq 2015\ t2 \\ Y_t^{(DSN)}, \text{ t}>2015\ t2 \end{cases}$$

Pour juger de l'existence d'une discontinuité, et le cas échéant son ampleur, la dynamique temporelle de chaque série (en niveau ou différenciée à l'ordre 1) est modélisée selon l'équation (*) : il s'agit d'une modélisation de type ARIMA (cf. Gourieroux et Monfort, 1995), préalablement augmentée :

- D'une indicatrice valant 0 avant le deuxième trimestre 2015, puis 1 après ;
- D'une tendance conjoncturelle (*TC*) captée, soit par l'emploi salarié, soit par les DPAE du grand secteur d'activité associé à la strate (industrie, construction ou tertiaire). Le choix entre les deux séries est basé sur un critère d'information (AICc ; cf. schéma 1).

$$\Delta_i Y_t^{(1)} = \sum_{p\geq 1} \theta_p . \Delta_i Y_{t-p}^{(1)} + \varepsilon_t + \sum_{q\geq 1} \sigma_q . \varepsilon_{t-q} + C + K . \Delta_i . \mathbf{1}_{t>2015\ t2} + L . TC \ ; \qquad (*)$$

$$\text{avec } i = \{0,1\} \ ; \ \forall\, (p,q), \{\theta_p, \sigma_q\} \in \mathbb{R}^2 \ ; \ \{C, K, L\} \in \mathbb{R}^3 \ ; \ \{\varepsilon_t\}_t \text{ un bruit blanc}$$

Avant de tester l'existence d'une rupture de série (*i.e.* significativité du paramètre *K*), la légitimité de la tendance dans les modèle est vérifiée : si le paramètre associé *L* est significatif, le modèle est retenu. Dans le cas contraire, le même modèle sans la tendance conjoncturelle est conservé. La procédure complète est résumée dans le schéma.

Ainsi, si dans le modèle final retenu, le test unilatéral d'absence de rupture est rejeté à l'ordre 10 %, la rupture de série est considérée comme significative, et d'ampleur $\widehat{K}$ (estimation du paramètre K).

Si la rupture de série est statistiquement avérée, son estimation est suivie d'une traduction en proportion de contrats manquants sur l'année 2014, qui permet d'aboutir à un jeu de quatre coefficients de passage (un par trimestre, pour respecter la saisonnalité des séries) :

$$\alpha_\tau = 1 + \frac{\widehat{K}}{Y_\tau^{(MMO)}}, \tau \text{ variant du } 1^{\text{er}} \text{ au } 4^{\text{e}} \text{ trimestre 2014}$$

Enfin, l'utilisation de ces coefficients de passage permet d'aboutir à la série rétropolée :

$$Y_t^{(2)} = \begin{cases} \alpha_\tau . Y_t^{(MMO)}, & \text{t}\leq 2015\ \text{t2}\ ;\ \tau \text{ trimestre correspondant à t} \\ Y_t^{(DSN)}, & \text{t}>2015\ \text{t2} \end{cases}$$

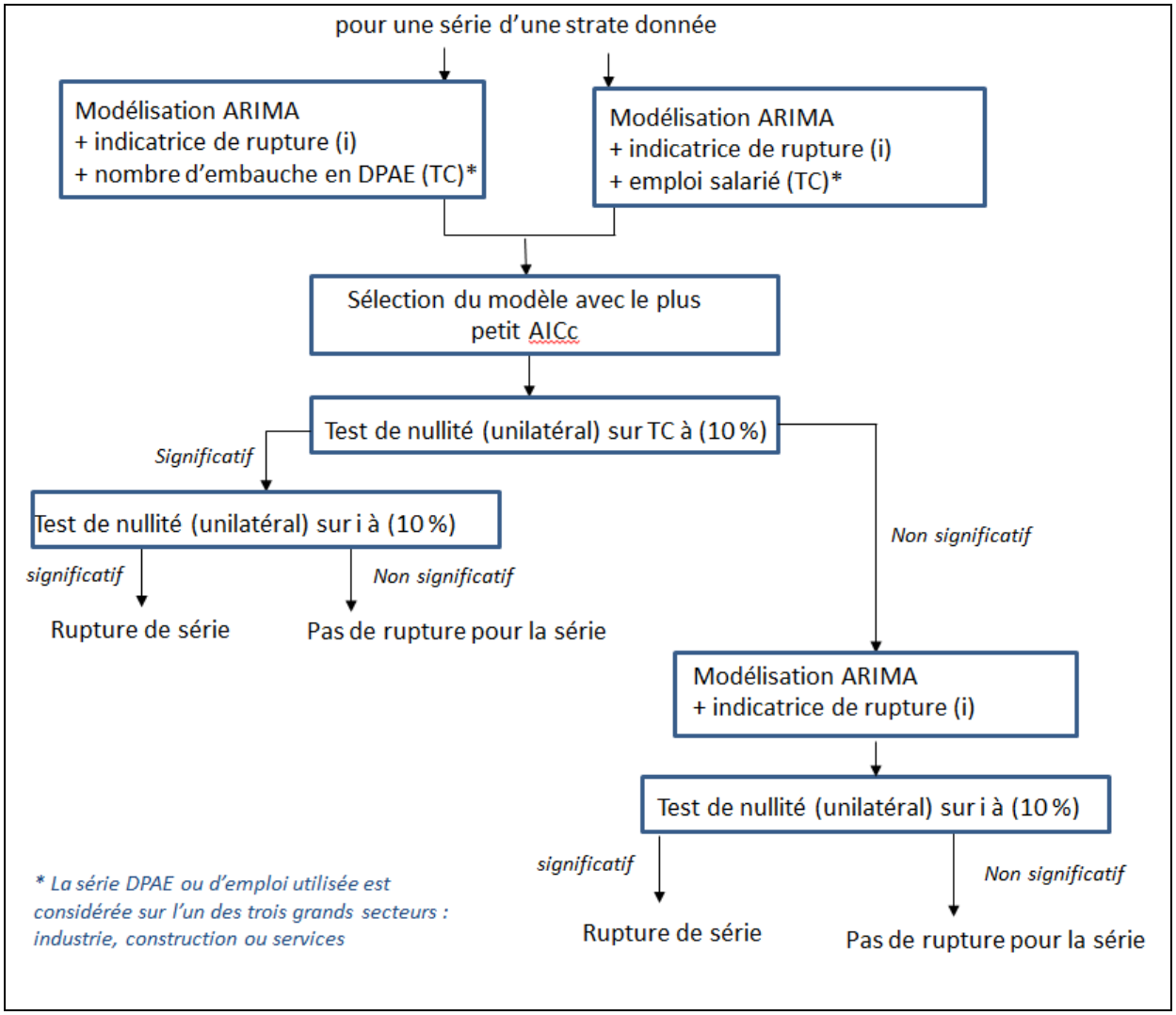



*Schéma – Algorithme pour la détection des ruptures de série (nombre d'embauches et de fins de contrats) par strate de rétropolation.*

### c. Étape 3 : mise en cohérence des données détaillées avec les séries rétropolées agrégées

À la fin de la deuxième étape de rétropolation présentée ci-dessus, il existe une incohérence entre les séries agrégées qui viennent d'être calculées ci-dessus et les données détaillées :

$$\forall\, t \leq 2015\ t2,\ \ Y_t^{(MMO)} = \sum_{i \in s^{(t)}} p_{i,t}.y_{i,t} \neq\ Y_t^{(2)} = \alpha_\tau.Y_t^{(MMO)}$$

avec :

- $s^{(t)}$ l'échantillon historique des établissements, au trimestre *t* ;
- $y_{i,t}$ le nombre de mouvements pour l'établissement *i*, sur le trimestre *t* ;
- $p_{i,t}$ le poids associé à l'établissement *i*, pour le trimestre *t*.

Afin de rendre cohérents les données détaillées et les nombres de mouvements, il convient de trouver un second jeu de poids $\{w_{i,t}\ ;\ i \in s^{(t)}\}$ tel que :

$$\forall\, t \leq 2015\ t2,\ \ Y_t^{(2)}\ =\ \sum_{i \in s^{(t)}} w_{i,t}.y_{i,t}$$

Pour établir ce nouveau jeu de poids $(w_{i,t})$, les propriétés de l'estimateur post stratifié sont utilisées, notamment pour en faciliter l'implémentation : les poids d'origine $p_{i,t}$ sont multipliés par $\alpha_\tau$. Pour un trimestre *t* antérieur au deuxième trimestre 2015, et *u* une strate de rétropolation (*U* étant l'ensemble des strates), la série rétropolée s'écrit :

$$\widehat{Y_t^{(2)}} = \sum_{i \epsilon s^{(t)}} w_{i,t}.y_{i,t} = \sum_{u=1}^{U} \sum_{i \epsilon s_u^{(t)}} w_{i,t}.y_{i,t} = \sum_{u=1}^{U} \sum_{i \epsilon s_u^{(t)}} \alpha_{u,\tau}.p_{i,t}.y_{i,t} = \sum_{u=1}^{U} \alpha_{u,\tau}.\widehat{Y_{u,t}^{(1)}} = \sum_{u=1}^{U} \widehat{Y_{u,t}^{(2)}}$$

Avec $\widehat{Y_{u,t}^{(1)}} = \sum_{i \epsilon s_u^{(t)}} p_{i,t}.y_{i,t}$ l'estimation du nombre de mouvements sur la strate *u* au trimestre *t*, et $s_u^{(t)}$ l'intersection de la strate *u* et de l'échantillon *s* au trimestre *t*).

La construction du jeu de poids consiste donc à augmenter (respectivement diminuer) l'importance des établissements qui appartiennent à des strates avec beaucoup (resp. peu) de sous déclaration de mouvements. En l'absence de sous-déclaration sur une strate, aucune dilation de poids n'est appliquée.

Pour valider le choix des strates de rétropolation, la déformation des pondérations est analysée. Il s'agit de vérifier que les bonnes propriétés des estimateurs sont conservées[21] et que les estimations ne sont pas instables (*i.e.* : dispersion « raisonnable » des nouvelles pondérations).

A la fin de cette seconde étape, il est aisé de déduire des taux d'entrée et de sortie (ventilés par motif, ou par caractéristiques des salariés). En notant $eff_{i,t}$ l'effectif moyen de l'établissement *i* au trimestre *t*, un taux d'entrée/sortie s'écrit :

[21] Ces « bonnes » propriétés concernent notamment l'absence de biais de sélection, propriété découlant de l'échantillonnage, mais également l'absence de biais de non-réponse, du fait du redressement appliqué.

$$tx_t^{(2)} = \begin{cases} \frac{\sum_{i \in s^{(t)}} w_{i,t}.\, y_{i,t}}{\sum_{i \in s^{(t)}} p_{i,t}.\, eff_{i,t}}, & \text{t}\leq\text{2015 t2} \\ tx_t^{(DSN)}, & \text{t>2015 t2} \end{cases}$$

Les séries des effectifs moyens sont conservées. Dès lors, le nouveau jeu de poids n'est utilisé que pour le calcul des entrées et des sorties.

À ce stade, trois jeux de poids sont créés pour corriger les sous-déclarations : le premier pour les fins de CDD de moins d'un mois, le deuxième pour les entrées en CDI, et le troisième pour les sorties de CDI[22].

### *Mises en cohérence des entrées et sorties*

Le traitement décrit précédemment permet de corriger la discontinuité des séries de CDD de moins d'un mois arrivés à terme. Étant donné que les séries associées aux CDD de plus d'un mois ne sont pas corrigées, la démarche mise en place crée mécaniquement un déséquilibre entre entrées et sorties de CDD[23]. Pour pallier ce déséquilibre, un quatrième jeu de poids pour les entrées en CDD est construit par déduction comptable.

Pour cela, sur une strate de rétropolation donnée, on note pour un trimestre *t* antérieur au 2e trimestre 2015 :

- $S_t$ (resp. $E_t$) le nombre de sorties (resp. d'entrées) de CDD au trimestre *t* ; $\widetilde{S_t}$ (resp. $\widetilde{E_t}$ ) le nombre de sorties (resp. d'entrées) de CDD après repondération au trimestre *t* ;
- $S_t^- / E_t^-$ (resp. $S_t^+ / E_t^+$) les sorties/entrées de CDD de moins d'un mois (resp. de plus d'un mois) au trimestre *t* ; $\widetilde{S_t^-} / \widetilde{E_t^-}$ le nombre corrigé de fins/débuts de CDD de moins d'un mois ;
- $\gamma_t$ la part de CDD de moins d'un mois dans les sorties de CDD au trimestre *t* ; $\gamma_t = \frac{S_t^-}{S_t^+ + S_t^-}$ ;

Par ailleurs, la relation de rétropolation est donnée par : $\widetilde{S_t^-} = \alpha_\tau . S_t^-$ (cf. *supra*). Par prolongement, la relation de rétropolation sur les entrées de CDD de moins d'un mois est approchée par $\widetilde{E_t^-} = \alpha_\tau . E_t^-$

On peut approcher le nombre d'entrées en CDD de moins d'un mois par $E_t^- = \gamma_t . E_t$. Symétriquement, le nombre de CDD de plus d'un mois est approximé par $E_t^+ = (1 - \gamma_t) . E_t$.

---

[22] Le jeu de poids créé pour corriger la sous-déclaration des entrées en CDI aurait *a priori* pu corriger la sous-déclaration des sorties de CDI. Cependant, l'utilisation du jeu de poids associé aux entrées en CDI ne corrige que partiellement les ruptures de séries sur les fins de CDI. C'est pourquoi le traitement est appliqué à la fois sur les embauches et sur les ruptures de CDI.

[23] Dans les sources historiques des MMO, les entrées et les sorties ne sont pas reliées. Par conséquent, les entrées en CDD ne peuvent pas être ventilées selon la durée effective du contrat.

Sachant que $E_t = E_t^- + E_t^+$, on a aussi : $\widetilde{E_t} = \widetilde{E_t^-} + E_t^+$. Puis, en utilisant la relation de rétropolation, il découle :

$$\widetilde{E_t} = \widetilde{E_t^-} + E_t^+ = \alpha_\tau . E_t^- + E_t^+ = \alpha_\tau . \gamma_t . E_t + (1 - \gamma_t) . E_t$$

D'où :

$$\boxed{\widetilde{E_t} = (\alpha_\tau . \gamma_t + 1 - \gamma_t) . E_t}$$

Le quatrième jeu de poids, nécessaire à la correction du nombre d'entrées en CDD, est donc directement issu du jeu de pondération créé pour corriger les fins de CDD de moins d'un mois.

### *Conséquences potentielles de la méthode retenue*

Auparavant, des biais de non déclaration poussaient à sous-estimer fortement le nombre de mouvements. La création de nouveaux jeux de poids a alors pour rôle de corriger ce biais. Étant donné que les établissements sous-déclarants peuvent avoir des profils différents des autres établissements, leurs mouvements de main-d'œuvre associés pouvaient également avoir des caractéristiques particulières. Ainsi, l'utilisation de ces nouveaux jeux de pondération peut conduire à modifier les répartitions des mouvements selon les caractéristiques des salariés (âge, sexe, CSP, ...) mais aussi ceux des mouvements (durée du contrat, motif de sortie, etc …). Ces modifications s'apparentent également à une correction du biais de sous-déclaration.

## d. Étape 4 : gestion du changement de nomenclature de 2008

Jusqu'en 2008, les secteurs d'activité des établissements étaient exprimés en nomenclature d'activités française (NAF) rév. 1. À partir de 2009, ils sont exprimés en NAF rév. 2. Jusqu'à la dernière publication des MMO en 2015, cette discontinuité de nomenclature des secteurs d'activité étaient gérée à l'aide de tables faisant le passage entre l'ancienne et la nouvelle nomenclature pour les agrégats diffusés. Mais les données détaillées exprimées en ancienne nomenclature n'avaient pas été recodées dans la nouvelle.

Or la correction de la sous-déclaration des mouvements sur longue période nécessite de connaitre les secteurs d'activité des établissements en NAF rév. 2., puisque les strates de rétropolation sont construites sur les secteurs d'activité exprimés dans la nouvelle nomenclature.

C'est la raison pour laquelle un secteur d'activité A38 de la NAF rév. 2 est affecté à chaque établissement avant 2008, par application de la matrice de passage. Quels que soient l'année et trimestre, un seul et même secteur est affecté à un établissement.

Ainsi, cette rétropolation sectorielle s'applique, dans un premier temps, sur les données s'étalant de 1998 à 2008, puis dans un second temps, sur celle de 1993-1997 (cf. *infra*). Elle facilite l'utilisation des fichiers détaillés et garantit la cohérence avec les agrégats publiés.

### e. Étape 5 : rétropolation des données de 1998 à 2000

Comme évoqué précédemment, la rétropolation des données historiques suppose la création de nouveaux jeux de poids pour :

- Pallier la sous-déclaration des CDD très courts arrivés à terme (de moins d'un mois), et la mise en cohérence des entrées et sorties de CDD.
- Les entrées en CDI.
- Les sorties en CDI.

Toutefois, la nature de contrat sur les sorties n'est pas connue avant 2001. La rétropolation des données avant cette date nécessite donc des étapes supplémentaires pour corriger les fins de CDI. Par ailleurs, une consolidation des durées des contrats est également apportée.

#### *Une étape supplémentaire pour la rétropolation des données*

Sur la période 1998-2000, les données ne permettent pas de différencier les sorties selon la nature des contrats : à titre d'exemple, les démissions de CDI ne sont pas isolées des démissions de CDD. En conséquence, les coefficients de correction de la sous-déclaration des sorties de CDI ne peuvent plus être utilisés[24]. En les appliquant sur les séries de ruptures anticipées de contrat, une discontinuité serait introduite : par exemple, le nombre de démissions de CDI serait rehaussé (de la manière attendue), mais aussi les démissions de CDD, alors que cela n'est pas effectué sur la période *post* 2001.

Il s'agit alors de calculer la part de CDD dans les ruptures anticipées de contrats (*i.e.* en excluant les sorties avec motif « CDD arrivé à terme »), par strate de rétropolation. Cela permet de minimiser la dilatation des poids, et d'éviter l'introduction de ruptures de série : le principe est identique à celui adopté pour la mise en cohérence des entrées et sorties de CDD décrit au paragraphe 2.

Pour appréhender au mieux les saisonnalités des séries, la part de CDD dans les ruptures anticipées de contrats est différente selon le trimestre et calculée sur la période 2001-2002.

#### *Consolidation des durées de contrat*

Avant 2001, la variable portant sur la durée des contrats est uniquement disponible en mois à l'unité près, et non en jours. Par ailleurs, dans 15 % des cas, elle présente des valeurs manquantes ou négatives. Il convient donc de consolider cette variable en imputant des durées aux sorties n'ayant pas de durée.

Pour cela, des classes d'imputation sont construites à partir de quatre variables : le motif de sortie (démission, fin de CDD, fin de période d'essai, …), la catégorie socioprofessionnelle (en six modalités), l'âge du salarié (moins de 25 ans, 25-35 ans, plus de 35 ans), le secteur d'activité de

[24] Au contraire, la correction des entrées (CDI et CDD) sont directement applicables, tout comme la correction des CDD très courts arrivés à terme (car « CDD arrivé à terme » est un motif à part entière).

l'établissement (niveau A4 de la NAF rév.2). La création de ces classes permet de conserver le lien entre la durée des contrats et les variables utilisées pour la construction des strates. Ensuite, une méthode d'imputation par *hot-deck* aléatoire est utilisée. Elle compte les quatre étapes suivantes :

1. création des strates par croisement des quatre variables ;
2. pour une strate *C*, calcul du nombre *Nc* de durées à imputer ;
3. tirage d'un échantillon de taille *Nc* dans la strate *C'* (*i.e.* : le sous-ensemble de *C* constitué des sorties avec une durée renseignée), par un tirage aléatoire avec remise ;
4. affectation des durées échantillonnées dans *C'* aux mouvements sans durée de la strate *C*.

Pour les départs en retraite et les motifs « autres », l'âge n'est pas utilisé pour la stratification.

### f. Étape 6 : redressement et rétropolation des données de 1993 à 1997

Auparavant disponibles à partir de 1999, les statistiques des mouvements de main-d'œuvre sont désormais mises à disposition depuis 1998, pour les établissements de 10 salariés ou plus. Par ailleurs, les données brutes de la DMMO (établissements de 50 salariés ou plus) ont pu être récupérées, puis apurées, redressées et retropolées, permettant des analyses des mouvements de main-d'œuvre sur une fenêtre temporelle plus large, depuis 1993.

En outre, tous les établissements réputés concernés par la DMMO entre 1993 et 1997 n'ont pas transmis de DMMO. Un traitement de type « redressement de non-réponse totale » a donc été mis en place (*i.e.* affectation d'un poids à chaque établissement répondant afin de se prémunir de biais de non-déclaration). Par ailleurs, en dehors de l'absence des durées de contrat, les données disponibles sur 1993-1997 sont similaires à celles de 1998-2000. Certaines données sont mal renseignées[25] : un peu moins de 2 % des natures de contrat des entrées demeurent indéterminées. Il en va de même pour les motifs de sortie. Ces variables ont été consolidées par imputation : un type de contrat pour chaque entrée manquante, et un motif de fin de contrat pour chaque sortie non renseignée.

#### *Différentes étapes du redressement*

1. Des référentiels annuels, comportant la liste des établissements réputés concernés par la DMMO sur cette période, sont utilisés. Ils contiennent de l'information auxiliaire nécessaire aux différentes étapes de redressement des données (secteur d'activité, zone géographique, effectif de référence, …). Ces informations auxiliaires se doivent donc d'être connues pour tous les établissements (déclarants et non-déclarants), et identiques pour chaque trimestre d'une même année.

[25] Le sexe du salarié est correctement renseigné pour chaque mouvement, tout comme l'âge des salariés (un peu moins de 1 % de non-renseignés). Toutefois, entre 5 et 7 % des catégories socio-professionnelles ne sont pas renseignées ; la non-réponse de la famille professionnelle s'élève quant à elle aux alentours de 10 %. Aucun redressement n'a été apporté à ces variables pour le moment.

2. Pour le référentiel de 1993, les secteurs d'activité, exprimés dans la nomenclature de 1973, sont tout d'abord convertis dans la nomenclature économique de synthèse (NES) en vigueur à partir de 1994[26], au niveau de l'établissement (code d'activité principale exercée, APE). Pour cela :
    a. le référentiel de 1993 est apparié au référentiel de 1994, afin de récupérer les secteurs d'activité des établissements concernés par la DMMO ;
    b. 83 % des établissements de 1993 sont retrouvés dans le référentiel de 1994 ;
    c. Pour les 17 % restants, un code APE de la nomenclature de 1994 est affecté aux établissements en respectant les répartitions « ancienne » / « nouvelle » nomenclatures observées sur la base des établissements de 1993 retrouvés en 1994 ;
    d. Enfin, la nomenclature NAF rév.1 est passée en NAF rév.2, en adoptant la méthode décrite au 3.d, en prévision de la correction de la sous-déclaration seulement.

3. Un poids est affecté à chaque déclarant en DMMO, pour pallier la non-déclaration de certains établissements. Les propriétés de l'estimateur post-stratifié sont de nouveau exploitées, notamment pour leur implémentation aisée et le calage sur marges (sur les effectifs de référence) qui en découle (cf. encadré 5) ;

4. Pour les motifs d'entrées et de sorties indéterminées, une imputation est effectuée par *hot deck* aléatoire dans des strates construites à partir du croisement de trois variables : âge (moins de 22 ans, 22-26 ans, 26-36 ans, plus de 36 ans), sexe, secteur d'activité de l'établissement (sur 16 positions, c'est-à-dire la NES16) ;

5. Les équations comptables entre les flux d'entrées, les flux de sorties et les effectifs sont reconstituées pour chaque établissement et chaque mois de la période. Les équations sont équilibrées, c'est-à-dire qu'en additionnant les effectifs de début de mois avec le nombre d'entrées sur le mois, et en retranchant les sorties, on retrouve les effectifs de fin de mois. Par ailleurs, les équations d'une même année sont chainées d'un mois à l'autre (cohérence entre les effectifs conjoncturels et les mouvements), mais la cohérence n'est pas garantie entre deux années. Ce chainage diffère des règles historiques : les données de 1998-2015 ne garantissaient aucune cohérence entre les équations de deux trimestres d'une même année pour un même établissement.

6. Enfin, la procédure de rétropolation décrite précédemment est appliquée. Néanmoins quelques traitements supplémentaires sont nécessaires :
    a. D'une part pour la correction de la sous-déclaration des CDI. En effet, la distinction des sorties par nature des contrats n'est pas applicable. Le traitement supplémentaire est donc le même que celui appliqué sur la période 1998-2000. Le principe est décrit au paragraphe 2.e ;

[26] La nomenclature économique de synthèse (NES) est utilisée à partir de 1994. Elle est notamment associée à la NAF rév.1 jusqu'en 2007.

b. D'autre part pour les fins de CDD. En effet, distinguer les CDD selon les durées n'est pas possible sur cette période. La rétropolation des CDD se fait alors en considérant les CDD dans leur ensemble : les coefficients de dilatation sont définis sur 1998, comme le rapport entre la série rétropolée de fin de CDD, et la série initiale correspondante (sur les mêmes strates de rétropolation avec un coefficient par trimestre pour respecter la saisonnalité).

Encadré 5 – Traitement de la non-réponse totale dans les DMMO (1993-1997)

Principe de la méthode retenue

Initialement, chaque établissement a un poids implicite de 1, puisque la DMMO est exhaustive. Étant donné que des établissements n'ont pas transmis de déclaration, un poids est affecté à chaque établissement déclarant. Un tel redressement permet d'aboutir à des données représentatives, car corrigées d'un éventuel biais de non-réponse.

À cette fin, des strates de non-réponse sont créées. Pour une correction satisfaisante, les variables servant à leur constitution doivent être corrélées aux probabilités de déclaration et/ou aux variables d'intérêt (ici la gestion de main-d'œuvre ; cf. fiche 2). Par la suite, tous les établissements répondants d'une même strate disposent d'un même poids.

En reprenant les notations de présentation du calage sur marge de la fiche 2, et en se plaçant pour un mois donné et sur une classe *u* :

- $\widehat{E_u}$ l'estimation de l'effectif de référence sur la strate, calculé à partir des répondants ;
- $E_u$ la « vraie valeur » de l'effectif de la strate : elle est connue pour tous les établissements *via* l'effectif de référence.
- Le poids affecté à chaque établissement de la classe vaut $p_u = \frac{E_u}{\widehat{E_{u,\pi}}}$.

Ainsi, plus une strate comporte de non-répondants, plus le poids $p_u$ est important et s'éloigne de 1, car l'estimation de l'effectif de référence sur la strate est sous-estimée. Implicitement, une plus grande importance est donnée aux établissements répondants ayant des caractéristiques similaires aux non-répondants. Au contraire, moins une strate comporte de non-répondants, plus $p_u$ se rapproche de 1.

Cette approche a l'avantage d'aboutir à des estimateurs calés. En effet, en notant *U* l'ensemble des classes de redressement, *R* l'ensemble des établissements répondants, *Ru* les répondants restreints à une classe *u*, l'estimation finale de l'effectif de référence à partir des déclarants coïncide parfaitement avec la valeur connue par ailleurs :

$$\widetilde{E_\pi} = \sum_R p_i . E_i = \sum_U p_u . \sum_{R_u} E_i = \sum_U p_u . \widehat{E_{u,\pi}} = \sum_U E_u = E$$

Les strates retenues

Les strates du redressement sont différentes des strates de rétropolation. Elles sont en effet le résultat du croisement de trois variables :

- L'effectif de référence, conformément aux quartiles de la variable :
  - 50-64 salariés ;
  - 65-94 salariés ;
  - 95-164 salariés
  - 165 salariés et plus.
- Les secteurs d'activité (en 16 postes de la NES) ;
- La zone géographique, divisée en 6 grandes régions :
  - Ile-de-France ;
  - Bretagne, Normandie ;
  - Occitanie, Nouvelle Aquitaine ;
  - Grand-Est, Hauts-de-France ;
  - Auvergne-Rhône-Alpes, Provence-Alpes-Côte d'Azur, Corse ;
  - Centre-Val-de-Loire, Bourgogne-France-Comté.

Ces strates de redressement doivent être construites au niveau le plus fin possible, mais doivent également respecter une « bonne proportion » de non-déclarants/déclarants afin d'éviter l'introduction d'instabilité dans les données finales. C'est pourquoi, quelques strates sont regroupées :

- Le secteur de l'énergie est considéré à l'échelon national ;
- Les secteurs industriels sont fusionnés pour la région Ile-de-France ;
- L'industrie des biens d'équipements est regroupée avec l'industrie automobile.

### g. Étape 7 : correction de la sous-déclaration des CDD en DSN, lors de la montée en charge du nouveau dispositif.

À la fin de l'étape précédente, les sous-déclarations liées au processus historique sont corrigées jusqu'en 1993 (pour les établissements de plus de 50 salariés), ou jusqu'en 1998 (pour les 10-49 salariés). Pour aboutir aux séries finales, il reste à traiter la sous-déclaration des CDD très courts sur les quatre trimestres de montée en charge des DSN (du troisième trimestre 2015 au deuxième trimestre 2016).

Sur les strates de rétropolation préalablement construites pour la correction des données historiques (cf. *supra*), les traitements suivants sont appliqués :

- La série[27] des fins de CDD de moins d'un mois (arrivés à terme) est obtenue en faisant la concaténation de la série historique rétropolée (jusqu'au deuxième trimestre 2015 ; notée $Y_t^{(retro)}$) et de la série reconstituée à partir des DSN (à partir du troisième trimestre 2015 ; notée $Y_t^{(dsn)}$) ;

[27] Notée $Y_t^{(2)}$ au 2.b.

- Les séries de chaque strate de rétropolation sont ensuite lissées à l'aide d'une moyenne mobile (notée $\widetilde{Y_t^{(dsn)}}$) ;
- Les écarts entre séries rétropolées et séries lissées sont déduits ($e_t = \widetilde{Y_t^{(dsn)}} - Y_t^{(dsn)}$) ;
- Les séries présentant de fortes saccades à la baisse sur la fenêtre 2015T3/2016T2 sont corrigées. Ainsi, si pour une date donnée sur la période de montée en charge du dispositif, la valeur de l'écart entre la série rétropolée et la série lissée est inférieure au double de l'écart-type de la série considérée sur toute la période, alors la valeur de la série finale pour cette date sera celle de la série lissée. Plus formellement, la série finalement retenue peut s'écrire :

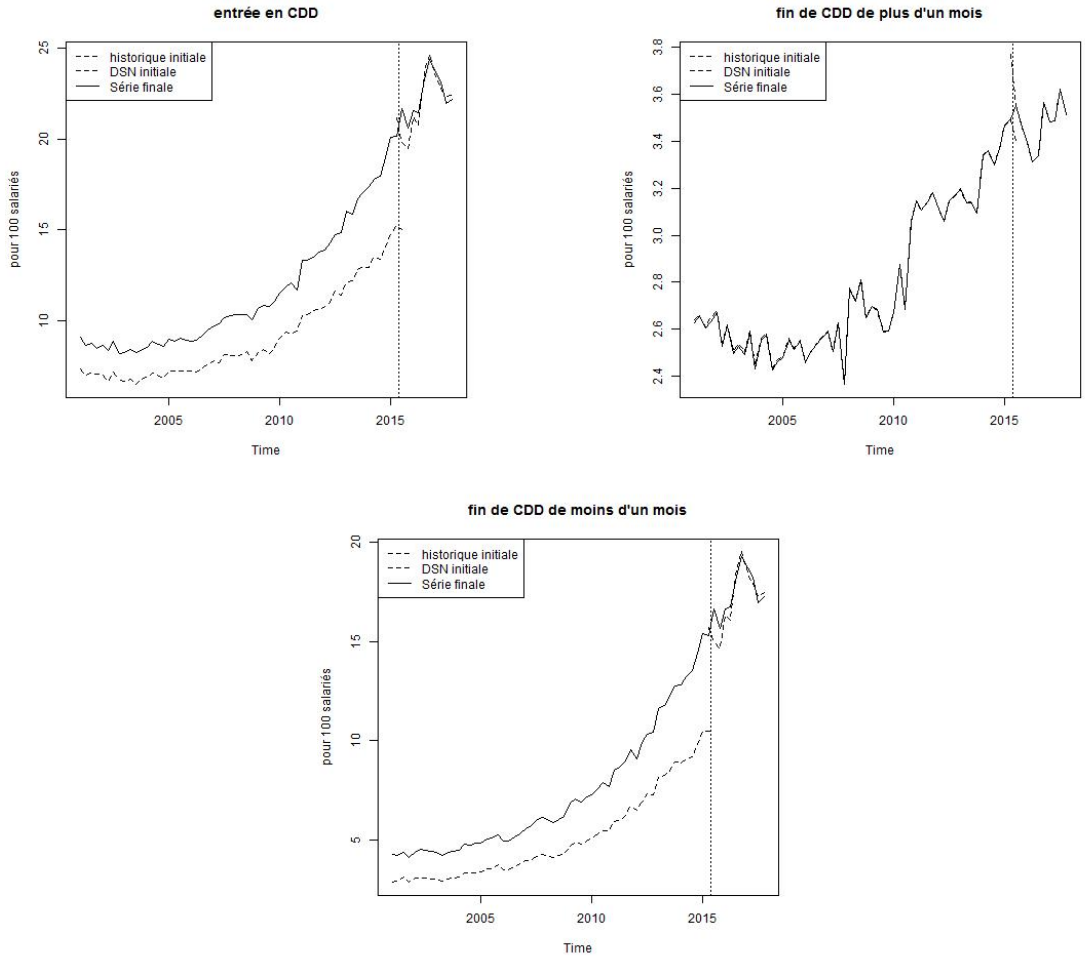
$$Y_t^{(fin)} = \begin{cases} Y_t^{(retro)} \text{ si t} \leq \text{t2 2015} \\ Y_t^{(dsn)} \text{ si t} > \text{t2 2015 et } e_t \geq -2.\sqrt{\sum_{t\in[2009T1\,;\,2016T3]} (e_t - \bar{e})^2} \\ \widetilde{Y_t^{(dsn)}} \text{ si t} > \text{t2 2015 et } e_t < -2.\sqrt{\sum_{t\in[2009T1\,;\,2016T3]} (e_t - \bar{e})^2} \end{cases}$$

- Les données détaillées sont repondérées afin d'être mises en cohérence avec les agrégats estimés précédemment. Deux nouveaux jeux de poids sont créés sur la période 2015T2-2016T2 :
  - Le premier correspond à la correction de la sous-déclaration des CDD de moins d'un mois arrivés à terme (avec le même procédé qu'au paragraphe 2.c.) ;
  - Le second pour la mise en cohérence entre embauches et fins de CDD (cf. paragraphe 2.c.).

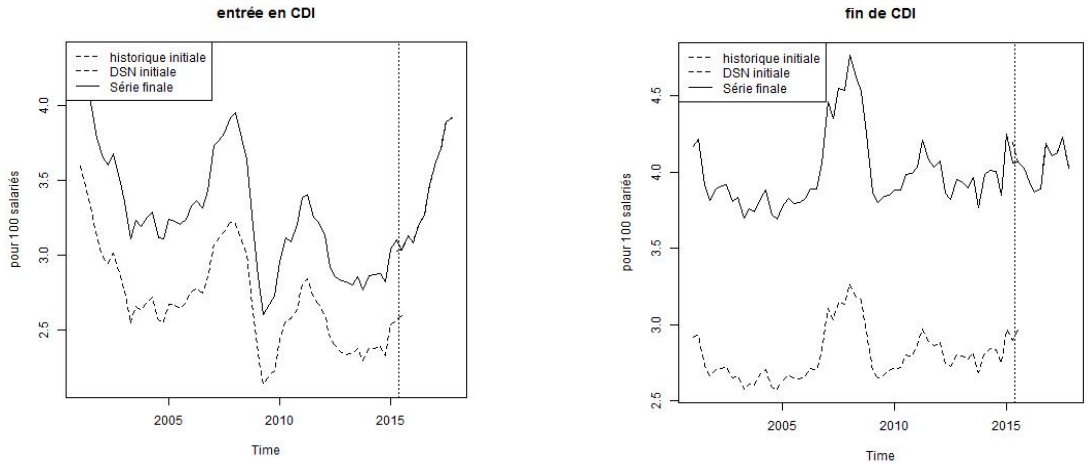
### h. Bilan de la rétropolation et de la correction des sous-déclarations

Les traitements décrits ci-dessus permettent de corriger de manière satisfaisante les données sur les établissements de plus de 10 salariés. Les indicateurs de mouvement de main-d'œuvre apparaissent en effet continus dans le temps (cf. graphiques 30 à 32). Le taux d'entrée est rehaussé d'un peu moins de 2 points par trimestre en 2001, et d'un peu plus de 4 points en 2014. Le taux d'entrée en CDI est revu d'un peu moins d'un point sur l'ensemble de la période (cf. graphiques 33 et 34). Enfin, conformément aux précédentes analyses, le taux de fin de CDD de plus d'un mois n'est pas modifié.

Ainsi, des séries longues démarrant en 1993 (respectivement 1998) sont mises à disposition pour les établissements de plus de 50 salariés (resp. de plus de 10 salariés).



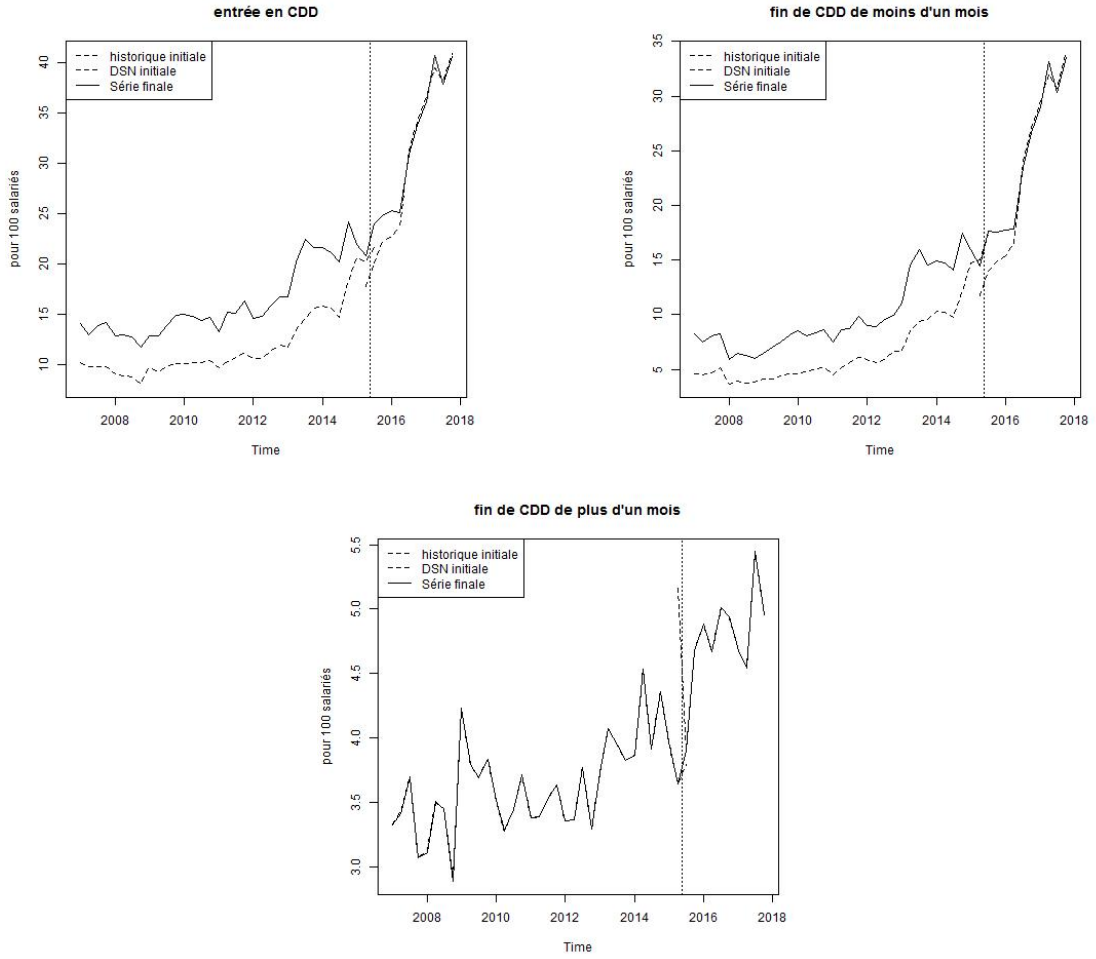

*Graphiques 30 à 32 – taux d'entrée en CDD et taux de fin de CDD, pour les établissements de 10 salariés ou plus*



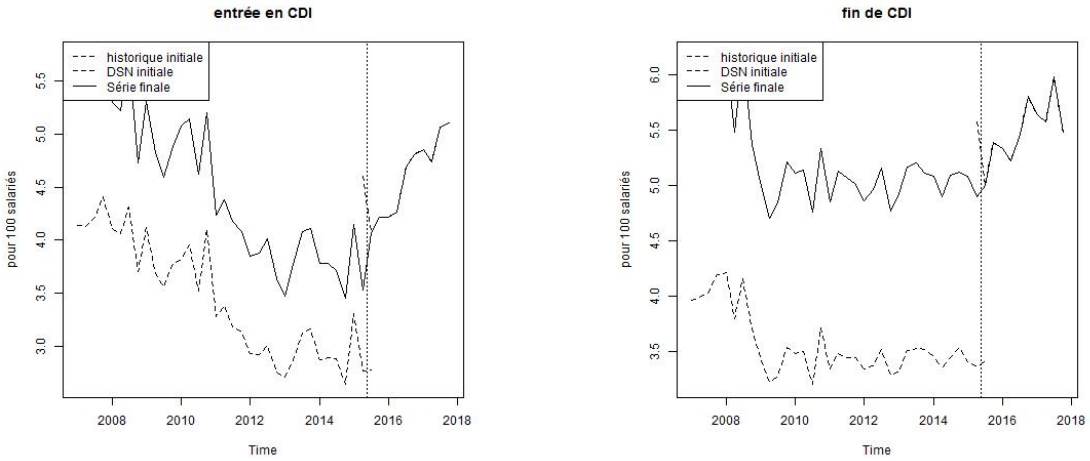

*Graphiques 33 et 34 – taux d'entrée en CDI et taux de sortie de CDI, pour les établissements de 10 salariés ou plus*

En ce qui concerne les établissements de moins de 10 salariés, les ruptures de série associée aux CDD ne sont visuellement pas entièrement corrigées (cf. graphiques 35 à 37). Les trajectoires des taux d'entrée sont déconnectées de celles publiées par l'Acoss à partir des DPAE. Par conséquent, il est nécessaire d'avoir une plus grande profondeur temporelle pour mieux corriger la sous-déclaration des CDD dans la source MMO. Les séries longues correspondantes, qui auraient débutées en 2007 soit l'année d'introduction de ces établissements dans l'EMMO, ne sont donc pas mises en avant dans les publications de la Dares. L'analyse de la gestion de la main-d'œuvre en CDD des établissements de moins de 10 salariés est toutefois toujours possible *via* des indicateurs structurels, en coupe sur la fin de période.

Enfin, les taux d'entrée en CDI et les taux de fin de CDI dans les établissements de moins de 10 salariés semblent plus satisfaisants. Ils sont relevés de plus d'un point par trimestre (cf. graphiques 38 et 39).




*Graphiques 35 à 37 – taux d'entrée en CDD et taux de fin de CDD, pour les établissements de moins de 10 salariés*




*Graphiques 38 et 39 – taux d'entrée en CDI et taux de sortie de CDI, pour les établissements de moins de 10 salariés*

### i. Les séries mises à disposition

Les séries des mouvements de main-d'œuvre, en fréquence trimestrielle et annuelle, sont désormais disponibles sur le site internet de la Dares. Elles débutent en 1993 pour les établissements de plus de 50 salariés, et en 1998 pour les établissements ayant entre 10 et 49 salariés. Elles recouvrent les trois grands secteurs d'activité : industrie, construction et tertiaire. Les agrégats d'ensemble sont également mis à disposition.

Plus précisément, les séries de mouvements de main-d'œuvre comprennent :

- le taux d'entrée en CDD et/ou en CDI ;
- le taux de sortie de CDD et de CDI confondu :
  - le taux de sortie de CDI est ventilé par motif de fin de contrat,
  - le taux de sortie de CDD arrivés à terme distingue notamment les CDD de plus d'un mois des CDD de ceux de moins d'un mois ;
- le taux de rotation, qui synthétise l'information portée par le taux d'entrée et le taux de sortie ;
- la part des embauches en CDD dans les entrées ;
- la part de CDD de moins d'un mois dans les fins de CDD.

Enfin, toutes les données peuvent être visualisées et téléchargées aisément à partir d'une application dédiée : dataviz.dares.travail-emploi.gouv.fr/MMO/

# Bibliographie


- Ardilly P. (2006), « Les techniques de sondage », *édition Technip*, juin
- Da Silva D.N., Opsomer J.D. (2009), « Nonparametric Propensity Weighting for Survey Nonresponse Through Local Polynomial Regression », *Survey Methodology*
- Gouriéroux C., Monfort A. (1995), « Séries temporelles et modèles dynamiques », *édition Economica*
- Haziza, D., Beaumont, J.-F. (2007), « On the construction of imputation classes in surveys », *International Statistical Review*, mars
- Sautory O. (1993), « La macro CALMAR : redressement d'un échantillon par calage sur marges », *document de travail,* Insee, n°9310, novembre