Journées de méthodologie statistique de l'Insee 2022

LE REDRESSEMENT D'UNE ENQUÊTE ENTREPRISES AUPRÈS D'UNE POPULATION ATYPIQUE : LE CAS DE L'ENQUÊTE R&D AUPRÈS DES ENTREPRISES

Thomas BALCONE (*), Charles DEULIN (*)

(*) Sies, Département des études statistiques de la recherche

thomas.balcone@recherche.gouv.fr
charles.deulin@recherche.gouv.fr

Mots-clés : non-réponse, imputation, repondération, calage, groupe de réponse homogène (GRH)

Domaine concerné : Statistique d'entreprises, théorie des sondages aval

Résumé

La loi de programmation de la recherche (LPR) présentée le 19/03/2020 a pour objectif de porter l'investissement dans la recherche et le développement expérimental (R&D) à 3 % du PIB. Cet objectif s'inscrit dans la continuité de la stratégie Europe 2020 qui portait déjà cet objectif de 3 %. Or la R&D représente 2,2 % du PIB et 462 000 emplois en équivalent temps plein (ETP) en France en 2019[1]. Les enjeux autour de la recherche n'ont pas échappé à l'Etat et celui-ci a décidé de soutenir l'innovation (par exemple via le Crédit d'impôt recherche (CIR) ou plus récemment avec la LPR). Pour que cette politique soit efficace, il faut s'appuyer sur des statistiques fiables concernant entre autres les moyens consacrés à la R&D par les entreprises.

C'est la sous-direction des systèmes d'information et des études statistiques (Sies) qui est chargée de produire ces statistiques. Ces statistiques sont réalisées à partir de l'enquête annuelle sur les moyens consacrés à la R&D dans les entreprises implantées en France. Afin de produire les statistiques les plus fiables possibles à partir des données collectées dans le cadre de cette enquête, de nombreux traitements post-collecte ont été mis en place. Parmi ces traitements figurent ceux permettant de corriger la non-réponse des entreprises. Il est légitime, dans un souci de potentiellement mieux faire, de se poser la question suivante :

Quelles améliorations apporter aux traitements post-collecte actuellement mis en œuvre ?

Dans cet article, nous allons tout d'abord faire un état des lieux des traitements actuels visant à corriger la non réponse. Pour réaliser cet état des lieux, il est nécessaire de prendre du recul sur les données. Nous proposons donc dans un premier temps de s'intéresser aux entreprises étudiées. La population d'intérêt de cette enquête est constituée des sociétés ayant réalisé effectivement des travaux de R&D en interne au cours de l'année. Cependant, afin de minimiser le risque de défaut de couverture, la base de sondage est constituée des entreprises susceptibles de réaliser de la R&D en interne. Ce recul sur les données permet également de présenter le plan de sondage ainsi que l'échantillonnage.

[1]Données semi-définitives

Ceci fait, on peut décrire les traitements post-collecte actuels. Ils s'articulent autour de 5 grandes phases :

- une première phase durant laquelle les bases de production sont rassemblées en une seule base de travail,
- une phase de correction de la non-réponse totale qui s'effectue par repondération,
- une phase de changement de clef primaire,
- une phase de correction de la non-réponse partielle,
- une phase de mise en forme de la base pour la diffusion.

Plusieurs problèmes potentiels ressortent de cet état des lieux des traitements post-collecte actuels. Nous avons tenté de les résoudre en proposant nos propres traitements. Ainsi, les deux dernières parties de cet article décrivent les traitements que nous proposons. Ce sont :

- les traitements préalables à la correction de la non-réponse. Ils diffèrent des traitements actuels en présentant une autre approche de la construction de la base de travail,
- les traitements destinés à la correction de la non-réponse partielle. L'ensemble des imputations existantes a été repris. Nous avons adapté certains traitements et proposé de nouveaux types de correction comme l'imputation par la régression,
- les traitements destinés à corriger la non-réponse totale. Nous avons gardé la stratégie de repondération en l'affinant en mettant en place des groupes de réponse homogène (GRH),
- le traitement des valeurs influentes. Celles-ci étaient traitées arbitrairement. Nous avons utilisé la technique de winsorisation de type II avec un calcul des seuils à l'aide de la méthode présentée par Kokic et Bell,
- le calage sur marges. Ce traitement n'était pas présent dans les traitements post-collecte actuels.

Abstract

The Sies (the Higher education and research Ministerial Statistical Department) provides statistics on the resources allocated by companies to research and experimental development (R&D) in France. It is important for the state to be able to estimate these resources. They are estimated from the annual survey of the resources devoted to research and experimental development in companies. As with most surveys, there is the question of non-response. Non-response can be a source of bias. Various techniques can be used to correct it.

In current treatments, non-response is not neglected. The purpose of this work is to improve them. For this, we propose new treatments. These treatments are based on a different strategy. We propose a different construction of the working base. We also propose to treat the partial non-response before the total non-response. This allows to treat differently some large non-respondent companies.

We have adapted some existing treatments and proposed the use of new methods to correct partial non-response. Total non-response is corrected by re-weighting. We use the homogeneous response group method to improve it.

Those non-response correction treatments were completed by a treatment to deal with unusually large observations. This work also provided an opportunity to implement a margin calibration.

1. Présentation de l'enquête R&D auprès des entreprises

Cette première partie permet de comprendre le contexte dans lequel se déroule cette enquête et d'avoir un premier aperçu des traitements existants.

1.1. Présentation de l'enquête

L'enquête annuelle sur les moyens consacrés à la recherche et au développement expérimental (R&D) auprès des entreprises a pour but de mesurer les efforts de R&D réalisés par l'ensemble des entreprises implantées en France. La collecte est entièrement réalisée sur internet. Ce sont six agents qui travaillent sur cette enquête au quotidien afin de réaliser la conception, l'apurement et les traitements post-collecte.
11 500 sociétés sont interrogées en moyenne chaque année. Ce sont les moyens financiers et humains consacrés aux activités de R&D par les entreprises qui sont collectés. Cette enquête permet la production d'agrégats économiques suivis au niveau national et international (objectif Europe 2020 portant sur la R&D).

1.2. La non-réponse dans l'enquête

La non-réponse dans les enquêtes statistiques est un problème récurrent. Elle touche la plupart des enquêtes et sa non-prise en compte aboutit à une erreur dans l'estimation des agrégats recherchés. Cette erreur peut influer sur le biais (les entreprises répondantes ont un profil particulier, par exemple elles peuvent être très concernées par la R&D) et aussi sur la variance de ces estimateurs (la taille effective de l'échantillon diminue). On distingue deux types de non-réponse, la non-réponse partielle (seule une partie des informations fournies par l'entreprise est exploitable) et la non-réponse totale (toute ou la majeure partie des informations fournies par l'entreprise sont inexploitables). En ce qui concerne l'enquête R&D auprès des entreprises, les deux formes de non réponse sont présentes et non négligeables. En effet, la non réponse totale représente plus de 10 % des unités échantillonnées. La non réponse partielle touche 20 % des unités répondantes.

La correction de la non-réponse est donc une étape nécessaire pour produire des statistiques de la meilleure qualité possible. Elle implique des traitements adaptés aux deux formes de non réponse. La littérature sur ce sujet décrit différents types de méthodes à mettre en place suivant le type de non réponse ([1], [2] et [3]). Pour la non-réponse totale, des méthodes par repondération sont mise en place tandis que la non-réponse partielle est corrigée à l'aide de méthodes par imputation. Les traitements post-collecte de l'enquête annuelle sur les moyens consacrés à la recherche et au développement expérimental auprès des entreprises comprennent des traitement de ce type.

1.3. Les traitements existants

Afin de comprendre les traitements existants, il est nécessaire de comprendre les données sur lesquelles ils s'appliquent, donc de s'attarder sur le plan de sondage et l'échantillon. C'est sur le millésime 2016 qu'a été réalisé cet inventaire des traitements et les réflexions qui l'accompagnent.

1.3.1. Le plan de sondage

La population d'intérêt de cette enquête est constituée des sociétés réalisant des travaux de recherche et de développement expérimental en interne au cours d'une année donnée. Cette population est difficile à discerner. En effet, il n'y a pas de base de sondage permettant de lister l'ensemble de ces sociétés. Par exemple, aucune des variables du Système d'immatriculation au répertoire des unités statistiques (Sirus) ne permet de distinguer les Unités légales (UL) ayant réalisé effectivement de la R&D en interne au cours d'une année N des autres UL.

Pour pallier ce problème, une **base de sondag**e propre au Sies a été constituée. Afin de minimiser le risque de défaut de couverture, cette base est c**onstituée des sociétés susceptibles de réaliser de la R&D en interne**. Sa construction pour le millésime N s'appuie sur :

- la population post-collecte de l'enquête N-1,
- les bases du Crédit d'impôt recherche (CIR),
- la base regroupant les sociétés ayant bénéficié du statut de Jeune entreprise innovante (JEI),
- la liste des sociétés créées via les incubateurs publics,
- la liste des lauréats du concours i-Lab,
- la liste des sociétés ayant déclaré des dépenses intérieures de R&D (DIRD) non nulles dans la dernière enquête « Capacité à innover et stratégie » (CIS) disponible,
- Sirus afin de retirer les sociétés cessées durant l'année N.

L'échantillon de l'enquête est tiré dans cette base via un sondage stratifié. Cette base est constituée d'une strate exhaustive et d'une strate non exhaustive, chacune de ces strates étant composée de 3 sous-strates (cf. figure 1).

Figure 1 – Strates et sous-strate constituant la base de sondage de l'enquête R&D auprès des entreprises

La strate exhaustive est composée de trois sous-strates :

- « QG » : sociétés dont la dernière DIRD connue est supérieure à 2M€. Cette sous-strate est nommée « QG » car elle regroupe les sociétés qui répondent au « Questionnaire Général » de l'enquête,
- « QS exhaustif » : sociétés dont la dernière DIRD connue est comprise entre 400k€ et 2M€. Le sigle « QS » fait ici référence au « Questionnaire Simplifié » auquel répondent ces sociétés,
- « QS new » : sociétés nouvellement détectées comme susceptibles de réaliser de la R&D en interne pour l'année N. Ces sociétés ne sont pas présentes dans la population post-collecte de l'enquête N-1.

La strate non exhaustive est elle aussi composée de trois sous-strates qui ne regroupent que des sociétés dont la dernière DIRD connue est inférieure à 400k€ :

- « QS sondé - partie conservée (sondée en N-1 mais pas en N-2) » : sociétés interrogées pour le millésime N-1 mais pas pour N-2. Toutes ces sociétés sont réinterrogées pour le millésime N,
- « QS sondé - partie exclue (sondée en N-1 et en N-2) » : sociétés interrogées pour les millésimes N-1 et N-2. Aucune de ces sociétés n'est réinterrogée pour le millésime N,
- « QS sondé - partie renouvelée » : sociétés non interrogées pour le millésime N-1. Dans cette sous-strate, un échantillon est sélectionné via un tirage systématique pour le millésime N.

De ce sondage aléatoire stratifié découle les poids initiaux (cf. tableau 1). En attribuant la même pondération initiale à toutes les unités de la strate non exhaustive « QS sondé », le tirage dans cette strate est assimilé à un tirage aléatoire simple sans remise.

Tableau 1 - Taille des sous-strates et pondérations initiales pour l'enquête 2016

	QG	QS exhaustif	QS new	QS sondé - partie conservée, sondée en N-1	QS sondé - partie exclue, sondée en N-1	QS sondé - partie renouvelée
Dernière DIRD connue	≥ 2 M€	entre 400 k€ et 2 M€	inconnue	≤ 400 k€		
Taille de la population	2167	2890	3429	1887	1812	19805
Taille de l'échantillon 2016	2167	2890	3429	1887	0	2125
Pondérations initiales (2016)	1	1	1	5,87		

L'application de ce plan de sondage permet de tirer l'échantillon de l'enquête. Comme dit précédemment, la base de sondage est plus large que la population d'intérêt de l'enquête. Il est donc indispensable de pouvoir identifier les différentes populations qui composent cette base et notre échantillon.

1.3.2. L'échantillon et le champ de l'enquête

Pour identifier facilement ces populations, nous introduisons quelques notations :

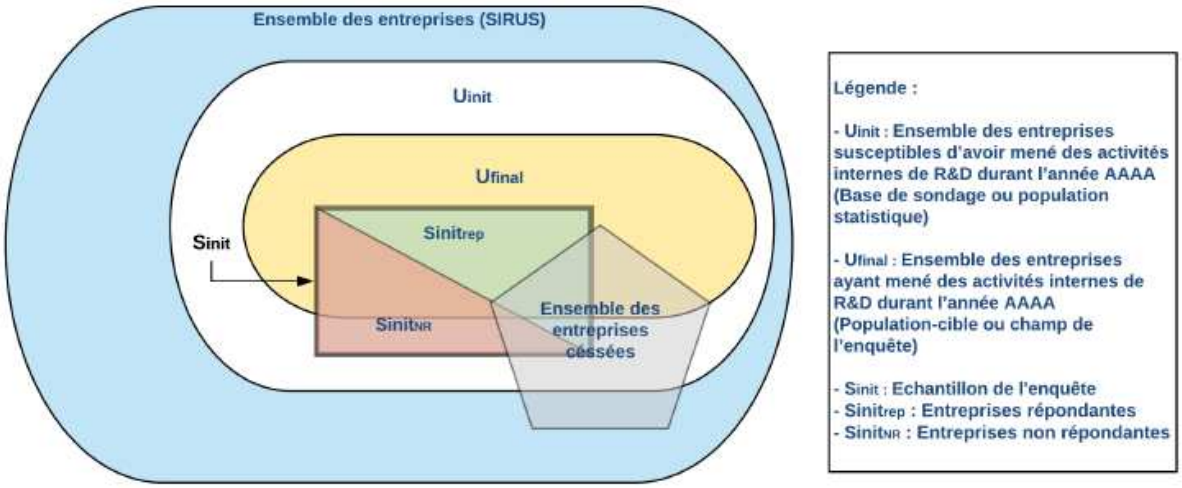
- **La population d'intérêt ou population cible (ou champ de l'enquête) U_{final}** visée par l'enquête portant sur l'année AAAA : elle regroupe « l'ensemble des entreprises ayant réalisé effectivement des activités de R&D en interne durant l'année AAAA »,
- **La population statistique ou base de sondage U_{init}** pour le millésime AAAA : elle regroupe « l'ensemble des entreprises susceptibles d'avoir réalisé de la R&D en interne durant l'année AAAA ».

La population statistique U_{init} ne coïncide pas avec la population d'intérêt U_{final} (cf. figure 2). Cette non coïncidence peut amener deux types de problèmes :

- Un problème de sous-couverture de la population U_{final} par la population U_{init} : $U_{final} \cap \overline{U_{init}} \neq \emptyset$. Ceci correspondrait au cas où des entreprises ayant réalisé une activité de R&D en interne durant l'année AAAA n'ont pas été repérées parmi les entreprises susceptibles de réaliser de la R&D en interne cette année-là. Cette potentiel défaut de couverture a été écartée lors d'une étude de validité de la base de sondage (U_{init}) confrontant l'enquête R&D auprès des entreprises et l'enquête CIS,

- Un problème de sur-couverture de la population U_{final} par la population U_{init} : $U_{init} \cap \overline{U_{final}} \neq \emptyset$. Ceci correspondrait au cas où des entreprises figurent dans la base de sondage mais n'appartiennent pas à la population cible U_{final}. Différentes sources traitent de ce sujet ([4] et [5]). Pour les sociétés échantillonnées et répondantes, il est facile de différencier celles appartenant à la population d'intérêt des autres par leurs réponses à l'enquête. Cet exercice est plus complexe pour les autres sociétés (i.e. celles non échantillonnées ou non répondantes). Un moyen de pallier cette difficulté est l'utilisation de sources externes.

Figure 2 - Population cible, population statistique et échantillon de l'enquête pour le millésime AAAA



Les sociétés cessées au cours de l'année AAAA représentent un cas particulier en ce qui concerne l'appartenance au champ. En effet, l'appartenance d'une telle société ayant réalisé des travaux de R&D en interne durant l'année AAAA à la population cible ne dépend que de sa date de cessation.

La non coïncidence entre la base de sondage U_{init} et la population d'intérêt U_{final} se retrouve logiquement dans **l'échantillon S_{init}** :
- certaines unités le composant sont hors champ : $S_{init} \cap \overline{U_{final}}$,
- les autres sont dans le champ : $S_{init} \cap U_{final}$.

1.3.3. Les traitements post-collecte actuellement appliqués

De nombreux traitements sont appliqués tout au long de la production de l'enquête. Nous nous focaliserons sur les traitements appliqués aux données après la collecte (de la collecte des données à leur diffusion, i.e. les traitements post-collecte). Ces traitements post-collecte s'articulent autour de **cinq grandes phases** :

- **Une première phase durant laquelle les bases de production sont rassemblées en une seule base de travail** en conservant les clefs primaires de production. Ces clefs primaires sont constituées des variables *ENT_CODE*, *BRA_NO* et *DPT_NO* où :
  - *ENT_CODE*[2] est la variable identifiante de l'unité interrogée. C'est une variable de production.
  - *BRA_NO* est la variable correspondant au numéro[3] de la branche de recherche [4].
  - *DPT_NO* est la variable correspondant au numéro du département où est localisée l'activité de R&D.

On enrichit cette base à partir de variables de production et on ajoute des données auxiliaires provenant de différentes bases comme Sirus pour détecter de possibles entreprises cessées ou encore imputer certaines données manquantes.

[2]Ce code entreprise est composé de 3 lettres et 5 chiffres. Les lettres décrivent la source qui a permis de détecter l'unité (CIR pour indiquer que la source est la base du crédit d'impôt recherche par exemple). Les 5 chiffres sont générés aléatoirement.

[3]Une entreprise peut consacrer des moyens dans plusieurs branches de recherche. Elle devra fournir des informations pour chacune d'entre-elles. Par exemple, une entreprise ayant une branche de recherche dédiée à l'extraction d'hydrocarbures (BRA_NO = 1) et une autre dédiée à l'industrie chimique (BRA_NO = 2) fournira des informations sur ces deux branches. C'est la variable BRA_NO qui permettra de distinguer ces informations pour une même entreprise.

[4]Une branches de recherche est une branches d'activité dans laquelle l'entreprise mène une activité de R&D

- **Une phase de correction de la non-réponse totale**. La correction de la non-réponse totale se compose de deux traitements distincts :

- Les unités non-répondantes des sous-strates « QG » et « QS exhaustif » sont imputées à partir des données de l'enquête précédente. La façon dont le traitement actuel est construit ne prend pas en compte une possible précédente reconduction des données. Ceci peut amener à reconduire d'année en année les mêmes données avec le risque d'imputer des valeurs moins plausibles.
- Les unités des strates « QS new » et « QS sondé » ne sont pas reconduites. Elles sont repondérées. Cette repondération s'effectue par l'application de la formule suivante :

$$poids_{final} = poids_{init} \times \frac{sech - sces}{srep}$$

où :

- $poids_{final}$ correspond à la pondération après correction de la non-réponse totale et prise en compte des UL cessées
- $poids_{init}$ correspond aux pondérations initiales
- sech est le nombre d'unités (*ENT_CODE*) échantillonnées $(\text{Card}[S_{init}])$
- sces est le nombre d'unités cessées
- srep est le nombre d'unités répondantes $(\text{Card}[S_{init_{rep}}])$

Cette formule peut être réécrite :

(1) $$poids_{final} = poids_{init} \times \frac{1}{Probabilité\,de\,répondre}$$

où $$Probabilité\,de\,répondre = \frac{srep}{sech - sces}$$

- **Une phase de changement de clef primaire**. Cette étape permet de passer de la clef primaire de collecte (*ENT_CODE*, *BRA_NO*, *DPT_NO*) à la clef primaire de diffusion (*SIREN*, *CODE_RECH2*, *DPT_NO*) où :

- *SIREN* est la variable identifiante de l'unité légale (UL).
- *CODE_RECH2* est le code issu de la NAF, révˆ2 correspondant à la branche de recherche.

Une fois cette phase effectuée, la table de travail contient autant de lignes que le nombre de branches de recherche de chaque UL dans chaque département (cf. encadré 1).

- **Une phase de correction de la non réponse partielle**. Ce sont des méthodes déterministes qui sont utilisées ici. Sont effectuées par ordre décroissant de priorité :

- des imputations déductives s'appuyant sur des relations connues entre variables,
- des imputations par les données antérieures. Selon les variables, les données antérieures peuvent être corrigées de l'évolution entre les deux années,
- des imputations par la moyenne observée pour les individus partageant des caractéristiques similaires (par exemple partageant la même branche de recherche).

- **Une phase de mise en forme de la base pour la diffusion**.

Encadré 1 - Une grandeur, trois variables

Les unités légales (UL) interrogées ont des profils variés et cela se reflète dans la base. Une UL peut consacrer des moyens dans différentes branches de recherche et dans différents départements. Le questionnaire tient compte de ces particularités. Cela débouche sur des variables "à plusieurs niveaux". Les variables préfixées par "DPT_" décrivent les informations concernant l'UL pour un département précis, celles préfixées par "DI_" ou "BRA_" décrivent les informations concernant l'UL pour une branche de recherche précise et celles non suffixée décrivent des informations concernant l'UL dans son ensemble.

Dans l'exemple ci-dessous, l'entreprise 1 a investi dans deux branches de recherche (2 pour « 0111Z » et 1 pour « 9900Z ») et est présente dans trois départements tandis que l'entreprise 2 a investi dans une unique branche de recherche et dans un seul département :

SIREN	CODE_RECH2	DPT_NO	DIRD (k€)	DI_DIRD (k€)	DPT_DI_DIRD (k€)
0000001	0111Z	01	3	2	1
0000001	0111Z	02	3	2	1
0000001	9900Z	03	3	1	1
0000002	0111Z	04	1	1	1

1.3.4. Bilan

Plusieurs points ressortent de cet état des lieux des traitements post-collecte actuels :

- Une grande partie des traitements (y compris le traitement de la non réponse totale) s'effectue avec les clefs primaires des bases de collecte (*ENT_CODE*, *BRA_NO*, *DPT_NO*) tandis que les autres traitements (ce sont principalement des traitements de correction de la non-réponse partielle) s'effectuent avec la clef de diffusion (*SIREN*, *CODE_RECH2*, *DPT_NO*). Il semble plus cohérent de basculer au plus tôt vers la clef de diffusion car les données auxiliaires (Sirus, base de diffusion de l'enquête de l'année passée) ont la variable *SIREN* comme identifiant au niveau entreprise. Nous avons donc choisi d'adopter une stratégie différente en établissant la clef de diffusion comme clef primaire dès le début des traitements.

- De nombreux changements de noms de variables viennent émailler les programmes rendant ceux-ci peu lisibles. Nous avons choisi de ne pas changer les noms des variables avant la fin des traitements afin d'avoir une meilleure traçabilité.

- En théorie, on ne reconduit des données que d'une année sur l'autre (pas de reconduction des données trois années de suite). En pratique, il peut être décidé d'utiliser les données concernant une unité légale (UL) sur une plus longue période pour éviter de ne pas prendre en compte l'entreprise (pas de repondération possible pour les grosses unités). Certaines données reconduites ne sont donc possiblement plus d'actualité et peuvent engendrer un biais. Nous proposons de ne pouvoir imputer l'année AAAA que par la reconduction des données de l'année AAAA-1 et pas par celles des années antérieures et de traiter la non réponse totale de la strate exhaustive par repondération si aucune donnée AAAA-1 n'est disponible.

- Le statut des unités dans le champ, ou hors champ, répondante ou non répondante est connu grâce à une variable de collecte qui représente l'état du questionnaire (la variable *SITU*). Cette variable est retravaillée au début des traitements afin de regrouper des modalités similaires et subit des modifications au cours des redressements rendant le suivi difficile. Elle devient ainsi une variable hybride entre variable de collecte et variable de traitement. Nous avons donc choisi de conserver

cette variable en l'état tout au long des traitements et de créer d'autres variables de traitement nécessaires aux redressements.

- La correction de la non réponse totale par repondération est traitée en utilisant la formule (1). C'est une formule classique. Cependant, l'estimateur de la probabilité de réponse semble perfectible étant donné que les unités cessées sont actuellement considérées comme répondantes.

- La correction de la non réponse totale est effectuée avant le traitement de la non-réponse partielle. Cet ordre pose problème en ce qui concerne les répondants non-exploitables qui ne sont pas traités comme non-réponse partielle. Avec les traitements existants, un répondant non-exploitable peut être sorti du champ en n'étant pas traité par les corrections de la non-réponse partielle. Il semble préférable de faire les traitements de la non-réponse partielle avec les unités exploitables avant ceux de la non-réponse totale afin de ne pas exclure les répondants inexploitables.

2. Proposition de nouveaux traitements de la non-réponse

Nous présentons ici les nouveaux traitements proposés pour corriger la non-réponse partielle et totale. Dans un premier temps, nous décrivons le passage des données brutes de l'enquête à la base de travail mise en forme, puis nous décrivons les traitements de la non-réponse partielle et enfin ceux de la non-réponse totale.

2.1. Les traitements préalables à la correction de la non-réponse

2.1.1. Passage des bases de données de collecte ACCESS à la base de travail SAS

L'apurement des données de l'enquête annuelle sur les moyens consacrés à la R&D dans les entreprises se fait avec une application Access. La base de données correspondante se compose de 9 tables ACCESS. Durant la conversion des tables ACCESS en tables SAS, nous veillons également à ce qu'à chaque *SIREN* ne corresponde qu'un seul *ENT_CODE*.
La grande majorité des variables d'intérêt de l'enquête sont quantitatives. C'est sur celles-ci que se concentrent les traitements de la non-réponse. Nous les extrayons de chacune des 9 tables de collecte ACCESS puis nous les regroupons en une seule table de travail dont la clef primaire reste la clef de collecte (*ENT_CODE*, *BRA_NO*, *DPT_NO*). Pour pallier une des faiblesses des traitements actuels (c.f. supra), nous proposons de basculer dès à présent de cette clef primaire de collecte vers la clef primaire de diffusion (*SIREN*, *CODE_RECH2*, *DPT_NO*).

Le basculement de la clé primaire de la base de collecte ACCESS à celle de la base de travail SAS suppose d'agréger certaines données. Par exemple, une UL ayant deux laboratoires de recherche dans le même département et qui répondent séparément à l'enquête (i.e. avec 2 *ENT_CODE* différents) correspond à deux lignes partageant le même triplet (*SIREN*, *CODE_RECH2*, *DPT_NO*) dans la base de collecte et doit correspondre ainsi à une seule ligne dans la base de travail. En effet, pour une telle entreprise, l'information que l'on cherche à connaître concerne les moyens consacrés à la R&D par cette entreprise dans ce département. Il convient donc d'agglomérer les résultats des deux laboratoires du même département.
On vérifie également la correspondance entre *ENT_CODE* et *SIREN*[5].

Après basculement, nous obtenons ainsi notre table de travail SAS contenant les variables numériques d'intérêt et ayant pour clé primaire (*SIREN*, *CODE_RECH2*, *DPT_NO*). Un contrôle est effectué sur les totaux des variables d'intérêt numériques afin de s'assurer que le basculement s'est bien déroulé.

2.1.2. Création d'indicatrices d'appartenance au champ et de résultat d'enquête

Plusieurs populations sont entremêlées au sein de notre base de sondage et de notre échantillon (cf. figure 2). Il convient de les distinguer à l'aide d'indicatrices. Leurs créations nécessitent certaines informations absentes de la base de travail. Cette dernière est donc enrichie de données auxiliaires provenant de Sirus[6]. Cet enrichissement est facilité par le passage à la clef de diffusion comme clef primaire. On vérifie à cette occasion que le numéro *SIREN* est bien codé sur 9 caractères.

Les indicatrices suivantes sont ensuite créées :

- *CESSE_AA* est l'indicatrice distinguant les unités cessées pour l'année AAAA. Elle est construite à partir des informations issues de Sirus et de la variable de production de l'état du

[5]Pour quelques ENT_CODE, le « contour réponse » est plus large que le seul SIREN.
[6]Sirus est le système d'immatriculation au répertoire des unités statistiques. « Il contient l'ensemble des unités productives marchandes et l'ensemble des unités employeuses pour constituer la référence de la statistique d'entreprises et de la statistique d'emploi. Pour toutes ces unités, il enregistre des caractéristiques comme le chiffre d'affaires, le classement sectoriel, l'effectif salarié, grâce à des mises à jour provenant d'une multitude de sources. » (Source : www.insee.fr)

questionnaire *SITU*. En effet, l'enquête peut permettre de repérer des unités légales (UL) cessées qui n'étaient pas déclarées comme telles dans Sirus,

- *POP_AA* est l'indicatrice d'appartenance à la population cible de l'enquête de l'année AAAA U_{final} . Elle est construite à partir de la variable de production de l'état du questionnaire *SITU* et de l'indicatrice de cessation *CESSE_AA*,
- *ECH_AA* est l'indicatrice d'appartenance à l'échantillon final pour l'année AAAA, i.e. $S_{init} \cap U_{final}$, noté S_{final} . Elle est ainsi définie à partir des variables *POP_AA* et *ECH*, l'indicatrice d'appartenance à l'échantillon S_{init} ,
- *REP_EXPL_AA* est l'indicatrice d'appartenance à la population des « unités répondantes exploitables ». Une telle indicatrice est utile pour traiter la non réponse partielle. Elle est construite à partir des variables concernant les moyens humains et financiers consacrés à la R&D à savoir la *DIRD* et les effectifs R&D. Ainsi, l'égalité « *REP_EXPL_AA* = 1 » signifie que l'on a accès à la Dird et à l'effectif de R&D de l'unité considérée soit directement soit indirectement.

2.2. Le traitement de la non réponse partielle – éléments généraux

2.2.1. Imputation de variables à partir de Sirus (« Colddeck »)

Les informations pertinentes concernant les variables d'activité (*APE2*), d'effectif total de l'UL en personnes physiques (*EFFECTIF*), de chiffre d'affaire (*CAHT* et *CA_EXP* (CA à l'export)) et de localisation géographique (*ENT_CP* : code postal) sont extraites de Sirus en considérant les données les plus récentes et intégrées à la base de travail. En effet, le répertoire Sirus est mis-à-jour avec les informations à sa disposition. Ainsi, certaines informations peuvent ne pas être encore disponibles dans la base Sirus de l'année AAAA. Les dates d'effets des différentes variables extraites de Sirus sont donc prises en compte afin de sélectionner les données cohérentes avec l'année de l'enquête.

A l'image de ce qui est fait dans les traitements actuels, on utilise la méthode « Colddeck » (c.f. [3]) pour imputer les variables *APE2*, *EFFECTIF*, *CAHT*, *CA_EXP* ou *ENT_CP*, en utilisant les données extraites de Sirus lorsque cela est possible : si une donnée est manquante pour une de ces variables, alors on utilise Sirus pour imputer la variable correspondante. Pour la variable *APE2* (l'APE de l'UL), on privilégie les données collectées via la variable de collecte *ENT_NAF2*.

2.2.2. Intégration des données de l'année précédente

Certaines sociétés sont réinterrogées d'une année sur l'autre. Les données de l'année précédente les concernant peuvent être utiles pour la correction de la non-réponse. C'est pourquoi, nous ajoutons ces données à la base de travail. Pour différencier les données antérieures des nouvelles, on ajoute le suffixe « _1 » aux noms de toutes les variables de la table de l'année AAAA-1. Cette table est ensuite scindée suivant les 3 niveaux (niveau « unité légale » : *SIREN*, niveau « branche de recherche » : (*SIREN* ; *CODE_RECH2*), niveau « département » : (*SIREN* ; *CODE_RECH2* ; *DPT_NO*)). Ces données antérieures sont ensuite réinjectées dans la table de travail en utilisant la clef primaire correspondante. Cette manière de faire est une nouveauté (cf. §1.3.3).

A la fin de ces traitements, on dispose ainsi d'une base de travail dont la clef primaire est la clef de diffusion (*SIREN* ; *CODE_RECH2* ; *DPT_NO*), de données antérieures pouvant être mobilisées pour la correction de la non-réponse et de différentes indicatrices permettant de « retracer » les données et d'identifier notamment la population cible et la population des « répondantes exploitables ».

2.2.3. Autres préparation préalables aux imputations

La variable de « rémunérations » *DI_REM_CS* (niveau « branche ») ne recouvre pas exactement le même concept selon le type de questionnaire (général « QG » ou simplifié « QS »). En effet, elle

correspond aux « rémunérations et charges sociales/fiscales du personnel pour ses activités de R&D » pour les « QG ». Au contraire, pour les « QS », la variable de « rémunérations » inclut également **les rémunérations immobilisées du personnel employé pour la R&D (variable *DI_IMMO_REM* (niveau « branche ») disponible uniquement dans les « QG »)**. Il apparaît indispensable que la variable de « rémunérations » ait exactement la même définition pour les « QG » et les « QS ». Nous définissons ainsi une variable « salaire » (*DI_D_SALAIRE* ; niveau « branche ») qui remplacera la variable « rémunérations » originelle dans les traitements :

DI_D_SALAIRE = DI_REM_CS + ***DI_IMMO_REM***

Il est alors nécessaire de définir une nouvelle variable « dépenses courantes » (*DI_D_COURANTE_SAL*, niveau « branche ») :

DI_D_COURANTE_SAL = DI_D_SALAIRE + DI_FRAIS_GEN

où *DI_FRAIS_GEN* correspond aux frais généraux (niveau « branche »).

Actuellement, dans les « QG », les « rémunérations immobilisées » apparaissent dans les « frais de recherche immobilisés » (variable *DI_IMMO_RD*[7] ; niveau « branche »), ces frais figurant dans les dépenses en capital (variable *DI_D_CAPITAL* ; niveau « branche »). Il est donc également nécessaire de définir une nouvelle variable « dépenses en capital hors rémunérations immobilisées » (notée *DI_D_CAPITAL_H_RIMMO ;* niveau « branche ») :

DI_D_CAPITAL_H_RIMMO = DI_D_CAPITAL - DI_IMMO_REM

De plus, pour chaque variable pouvant être redressée, on crée une indicatrice de redressement. Cette indicatrice possède le même nom que la variable considérée suivi du suffixe « _R ». On crée également, pour chaque variable, une variable qualitative indiquant le type de redressement. Elle a le même nom que la variable considérée suivi du suffixe « _RT ». Elle est initialisée avec la modalité « Pas de redressement ».

Nous créons également un compteur du nombre de département *NB_DPT* par branche de recherche afin de discriminer les branches monodépartementales des branches multidépartementales. Par définition, *NB_DPT* est le nombre de départements associés à un couple (*SIREN* ; *CODE_RECH2*).

On commence par traiter la non réponse partielle des variables définies au niveau « branche ». Nous isolons tout d'abord dans la table de travail les unités répondantes exploitables.

2.3. Le traitement de la non réponse partielle – Les imputations déductives pour les branches monodépartementales

2.3.1. Principe général

L'enquête R&D auprès des entreprises se concentre sur les moyens financiers et humains consacrés à la R&D par les sociétés implantées en France. Nous nous concentrons tout d'abord sur le redressement des 2 principales variables d'intérêt :

- la Dird
- les effectif R&D en équivalent temps plein (ETP)

[7] Si *DI_IMMO_REM* > *DI_IMMO_RD* alors on effectue le redressement suivant : *DI_IMMO_REM = DI_IMMO_RD*

Ce sont les composantes de ces deux variables que l'on impute dans cette partie. On distingue ainsi deux groupes de variables :

- la Dird et ses composantes
- les effectifs R&D en ETP et ses composantes

Dans cette partie, on se concentre uniquement sur les triplets (*SIREN* ; *CODE_RECH2* ; *DPT_NO*) tels que la réponse est jugée exploitable (*REP_EXPL_AA* = 1) et la branche de recherche n'est présente que dans un seul département (*NB_DPT* = 1).

La démarche suivie pour imputer a été la même pour les deux groupes de variables mentionnés ci-dessus :

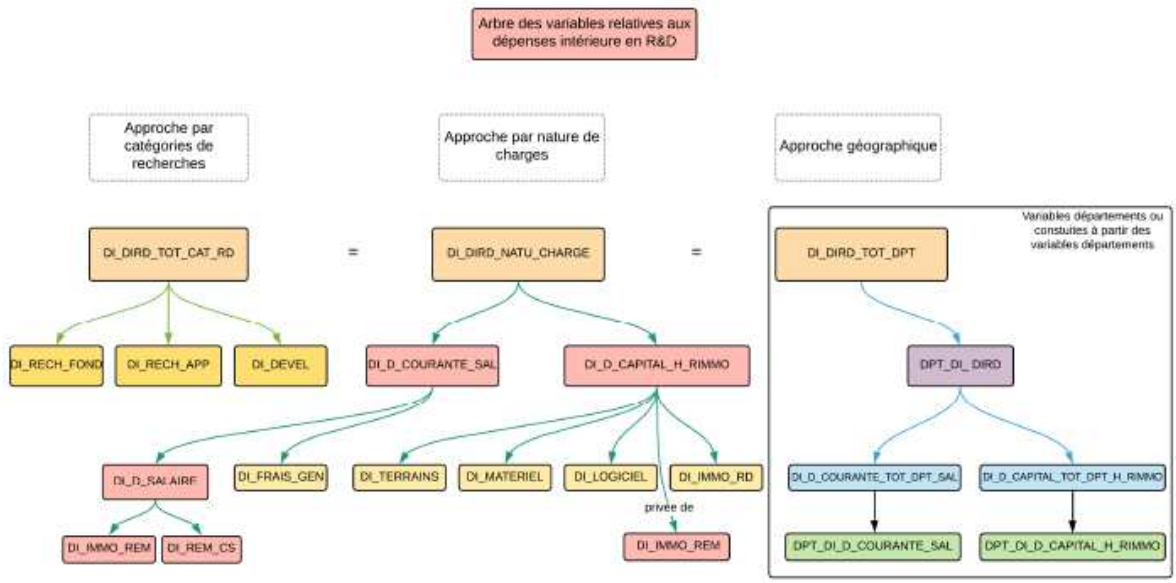
- référencer les liens entre la variable d'intérêt et ses composantes (constitution d'un arbre),
- imputer les variables en commençant par les plus fines,
- remonter l'arbre des variables

Dans cette partie nous n'effectuons que des imputations déterministes basées sur les relations existantes entre les différentes variables.

2.3.2. Imputations déductives des variables relatives à la Dird

Les imputations déductives des variables relatives à la Dird s'appuient sur les liens entre la Dird et ses composantes qui sont résumés dans la figure suivante :

Figure 3 – Liens entre la Dird et ses composantes



Trois décompositions de la Dird sont disponibles à partir de l'enquête :

- approche par nature de charges (*DI_DIRD_NATU_CHARGE* ; niveau « branche »),
- approche par catégorie de recherche (*DI_DIRD_TOT_CAT_RD* ; niveau « branche »),
- approche géographique (*DI_DIRD_TOT_DPT* ; niveau « branche »)

L'imputation des variables se fait dans l'ordre suivant :

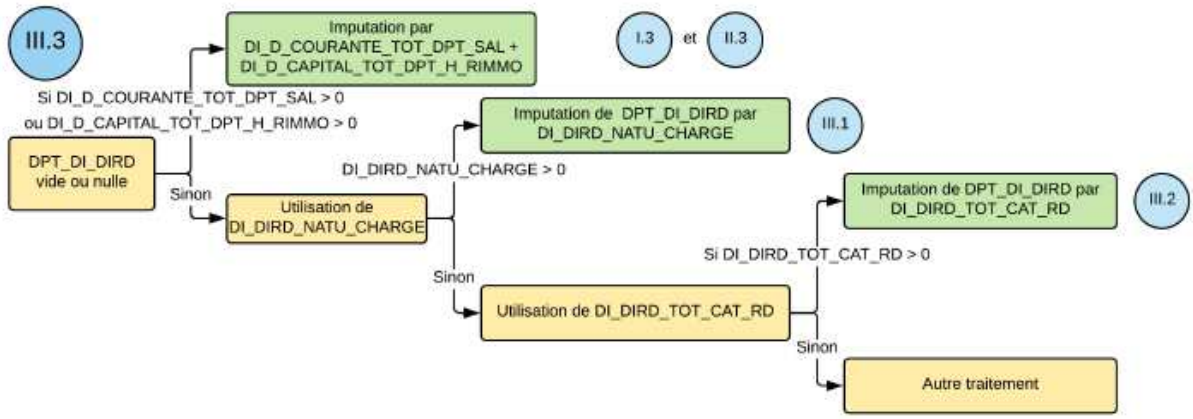
1. les variables « dépenses courantes avec salaires » :
   - *DI_D_COURANTE_SAL* (niveau « branche »),
   - *DPT_DI_D_COURANTE_SAL* (niveau « département »),
   - *DI_D_COURANTE_TOT_DPT_SAL* (niveau « branche »)

2. les variables « dépenses en capital sans rémunérations immobilisées » :
   - *DI_D_CAPITAL_H_RIMMO* (niveau « branche »),
   - *DPT_DI_D_CAPITAL_H_RIMMO* (niveau « département »),
   - *DI_D_CAPITAL_TOT_DPT_H_RIMMO* (niveau « branche »)

3. les variables « Dird » :
   - *DI_DIRD_NATU_CHARGE* (niveau « branche »),
   - *DI_DIRD_TOT_CAT_RD* (niveau « branche »),
   - *DPT_DI_DIRD* (niveau « département »),
   - *DI_DIRD_TOT_DPT* (niveau « branche »)

Toutes les imputations sont détaillées dans l'annexe 1. On ne propose ici que le processus d'imputation de la variable *DPT_DI_DIRD* (Dird au niveau « département ») :

Figure 4 – Processus d'imputation de la variable DPT_DI_DIRD (Dird au niveau « département ») pour les branches monodépartementales



Si la variable *DPT_DI_DIRD* est nulle ou vide :

1. si la décomposition directe par nature de charge est disponible, alors on utilise la relation :

   $$DPT\_DI\_DIRD = DI\_D\_COURANTE\_TOT\_DPT\_SAL + DI\_D\_CAPITAL\_TOT\_DPT\_H\_RIMMO$$

   pour l'imputer.

2. Sinon, si la variable *DI_DIRD_NATU_CHARGE* est disponible (i.e. >0), alors on impute la variable *DPT_DI_DIRD* de la manière suivante :

   $$DPT\_DI\_DIRD = DI\_DIRD\_NATU\_CHARGE$$

3. Sinon, si la variable *DI_DIRD_TOT_CAT_RD* est disponible (i.e. >0), alors on impute la variable *DPT_DI_DIRD* de la manière suivante :

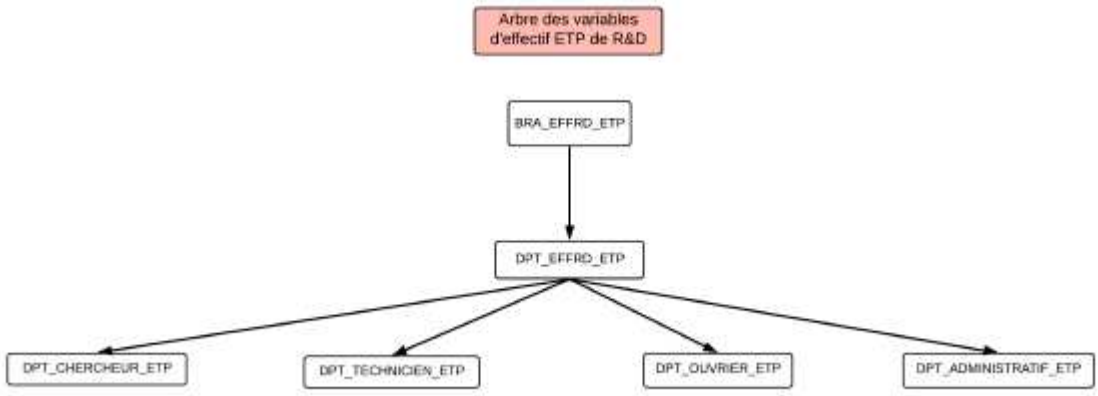
$$DPT\_DI\_DIRD = DI\_DIRD\_TOT\_CAT\_RD$$

4. Sinon, il n'est pas possible à ce stade d'imputer la variable DPT_DI_DIRD

2.3.3. Imputations déductives des variables relatives aux effectifs R&D en ETP

Les effectifs R&D en ETP sont définis au niveau « branche de recherche » (variables préfixées par *BRA_*) et au niveau « département » (*DPT_*). Ce sont les données niveau « branche » qui sont privilégiées.
Dans cette partie, nous considérons toujours uniquement les branches monodépartementales. Les imputations s'appuient sur les liens entre variables qui figurent dans l'arbre ci-dessous :

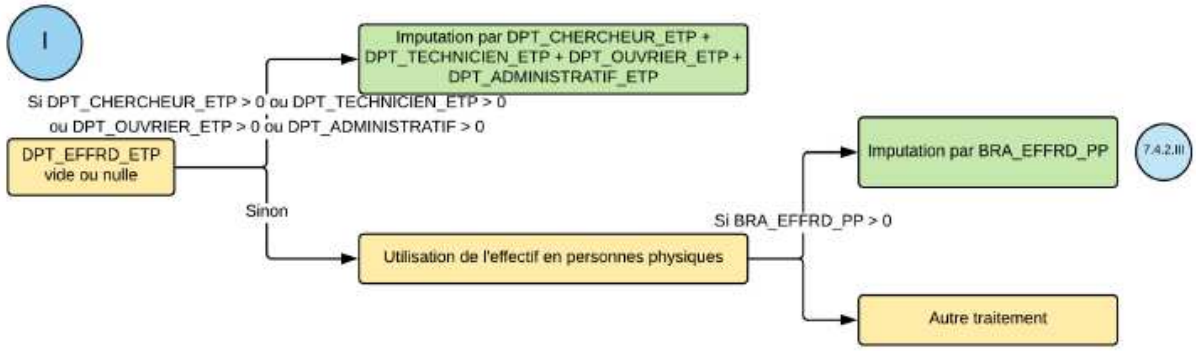
Figure 5 – Processus d'imputation de la variable BRA_EFFRD_ETP (Effectif R&D en ETP au niveau « branche ») pour les branches monodépartementales



On commence par imputer la variable *DPT_EFFRD_ETP* :

- en utilisant sa ventilation par catégorie de personnel si elle est disponible,
- sinon, on impute cette variable par l'effectif R&D en personnes physiques (*BRA_EFFRD_PP* ; niveau « branche »)

Figure 6 – Processus d'imputation de la variable DPT_EFFRD_ETP (Effectif R&D en ETP au niveau « département ») pour les branches monodépartementales



On impute ensuite la variable *BRA_EFFRD_ETP* par la variable *DPT_EFFRD_ETP*.

2.4. Le traitement de la non réponse partielle – Mise en cohérence des ventilations pour les branches monodépartementales

2.4.1. Principe général

Dans la partie précédente, on partait des variables les plus « fines » pour retrouver les agrégats correspondants. Visuellement, on « remontait » les arbres de variables afin d'utiliser toutes les réponses collectées. Dans cette partie, le cheminement est opposé, on part des agrégats pour imputer des valeurs plausibles aux variables plus fines. On utilise différentes méthodes d'imputation dans cette partie. Pour les mettre en place, nous nous sommes référés à la note [3]. La logique suivie dans cette étape pour tous les types de variables est la suivante :

1. privilégier les données issues de l'enquête. Par exemple si la ventilation d'une variable ne lui correspond pas, elle sera reventilée dans des proportions identiques,
2. sinon, on privilégie une méthode *Colddeck* en utilisant les données issues de l'enquête antérieure. Dans l'exemple des ventilations, si la ventilation de l'année AAAA est indisponible, on reconduira les proportions de l'année AAAA-1 et on les appliquera à l'agrégat de l'année AAAA,
3. sinon, on utilise les données d'autres unités répondantes pour imputer. Dans l'exemple de la ventilation manquante d'une variable connue, on peut appliquer les proportions observées pour les autres unités « similaires » ayant répondu.

2.4.2. Imputation des composantes de la Dird « catégorie de recherche » et de la Dird « approche géographique »

Trois décompositions de la Dird sont disponibles à partir de l'enquête (cf. §2.3.2). Il s'agit ici d'imputer les différentes composantes de la Dird. En cas de différences entre les 3 variables de Dird disponibles au niveau « branche » (*DI_DIRD_NATU_CHARGE*, *DI_DIRD_TOT_CAT_RD*, *DI_DIRD_TOT_DPT*), on privilégie l'approche par nature de charges (variable *DI_DIRD_NATU_CHARGE*) à condition que la variable correspondante soit non nulle car cette variable a la particularité de présenter une décomposition très fine au niveau « branche ». **La Dird correspondant à l'approche par nature de charges (variable *DI_DIRD_NATU_CHARGE*) sera ainsi notre Dird de référence dans les traitements suivants**.

- Imputation de la Dird « approche géographique » :

Si la Dird calculée via l'approche géographique ne correspond pas à celle calculée via l'approche par nature de charges et que cette dernière est strictement positive, alors on impute la Dird « approche géographique » par la Dird « nature de charges ». On impute ensuite ses composantes (les Dird départementales) en conséquence de manière déductive. Cette imputation est facilitée ici par le fait que l'on ne considère dans cette partie que les branches monodépartementales.

- Imputation de la Dird « catégorie de recherche » :

Si la DIRD calculée via l'approche par catégorie de recherche ne correspond pas à celle calculée via l'approche par nature de charges alors :

- si la ventilation de la Dird par catégorie de recherche (recherche fondamentale, recherche appliquée, développement expérimental) est disponible, alors on impute la Dird « catégorie de recherche » par la Dird « nature de charges » puis on redistribue les trois composantes « catégorie de recherche » dans les mêmes proportions,
- sinon, si la ventilation de la Dird par catégorie de recherche est disponible pour l'année antérieure et que ces données sont indiquées comme diffusables, alors on impute la Dird « catégorie de recherche » par la Dird « nature de charges » puis on redistribue les trois composantes « catégorie de recherche » dans les mêmes proportions que celles de l'année antérieure,
- Sinon (i.e. si on ne dispose ni de la ventilation de la Dird par catégorie de recherche pour l'année en cours ni pour l'année précédente), alors on impute la Dird « catégorie de recherche » par la Dird « nature de charges » puis on calcule la répartition moyenne de la Dird dans les trois catégories de recherche que l'on applique à la Dird « catégorie de recherche » (**imputation par la moyenne** - cf. encadré 2). Nous proposons également une variante : une **imputation par hot-deck aléatoire** en utilisant les mêmes classes d'imputation que dans le cas de l'imputation par la moyenne. Pour rappel, cette méthode consiste à sélectionner aléatoirement un répondant et d'utiliser sa réponse pour imputer la valeur manquante ([3]). Nous avons veillé ici à utiliser le même donneur pour imputer les trois composantes de la Dird « catégorie de recherche ».

Encadré 2 – Imputation par la moyenne et par hot-deck aléatoire

Ces méthodes d'imputation sont utilisées en découpant la population en classes d'imputation formant une partition de l'ensemble des unités. Les réponses des unités répondantes de chaque classe sont ensuite utilisées pour imputer les non-répondants des mêmes classes.
Les classes d'imputation sont ici construites à partir des variables de tirage et de branche de recherche en les croisant et en regroupant certaines modalités. Ces regroupements (cf. annexe 2) sont construits en conciliant deux critères :

- avoir au moins 30 « donneurs » dans chaque classe,
- les branches de recherche agglomérées doivent être proches dans leurs activités

Les « donneurs » sont les unités telles que :

- la Dird « catégorie de recherche » est positive
- la somme des trois composantes de la Dird « catégorie de recherche » est égale à la Dird « catégorie de recherche »

Il reste à imputer la Dird « nature de charges » et ses composantes.

2.4.3. Imputation des composantes de la Dird « nature de charges »

Pour réaliser l'imputation de la Dird « nature de charges » et de ses composantes, nous souhaitons utiliser l'effectif R&D en ETP et ses composantes par catégorie de personnel. Nous procédons ainsi tout d'abord à leur imputation.

- Imputations des composantes par catégorie de personnel de l'effectif R&D en ETP :

Les effectifs R&D en ETP sont disponibles au niveau « département » (*DPT_EFFRD_ETP*). Or dans ce paragraphe, nous considérons uniquement des branches monodépartementales. Ainsi, les effectifs R&D en ETP de la branche de recherche sont égaux à ceux du département. Pour rappel, ces effectifs R&D en ETP sont ventilés par catégorie de personnel (chercheur, technicien, ouvrier et administratif - cf. §2.3.3). Il s'agit dans cette partie d'imputer ces composantes lorsque celles-ci sont manquantes ou ne correspondent pas à l'effectif aggloméré (*DPT_EFFRD_ETP*). La logique de ces imputations est similaire à celles réalisées dans la partie précédente. A cette étape, pour toutes les unités de la table de travail (i.e. les UL répondantes exploitables), la variable correspondant aux effectifs R&D en ETP par département (*DPT_EFFRD_ETP*) est non nulle. Si tel n'est pas le cas, les unités concernées seront traitées comme une non réponse totale (cf. infra).

L'imputation des composantes par catégorie de personnel de l'effectif R&D en ETP se fait selon l'algorithme suivant :

- Si on dispose d'une décomposition antérieure considérée comme diffusable, alors on impute la ventilation dans les mêmes proportions que celle de l'année précédente,
- Sinon, on calcule la répartition moyenne dans les différentes catégories de personnel par classe d'imputation. On l'applique ensuite à l'effectif total R&D en ETP de la branche. Ces classes d'imputation sont similaires aux classes d'imputation de la partie précédente. Ici, les « donneurs » vérifient que la somme des composantes des effectifs R&D en ETP est égale à l'effectif R&D en ETP du département. Comme pour la partie précédente, une variante est proposée (imputation par hot-deck aléatoire en utilisant les mêmes classes d'imputation).

- Imputation des variables de salaire, de dépense courante, de capital et de frais généraux :

Dans cette partie on souhaite imputer les composantes de la Dird « nature de charges » (*DI_DIRD_NATU_CHARGE* ; niveau « branche ») et plus précisément :

- les dépenses de R&D en capital hors rémunérations immobilisées (*DI_D_CAPITAL_H_RIMMO* ; niveau « branche »)
- les dépenses courantes « avec salaires » (*DI_D_COURANTE_SAL* ; niveau « branche ») qui sont composées :
  - des salaires des personnels de la branche rémunérés directement par la société (*DI_D_SALAIRE* ; niveau « branche »)
  - des frais généraux (*DI_FRAIS_GEN* ; niveau « branche »)

La variable *DI_D_SALAIRE* permet de calculer la rémunération moyenne des personnels dans une branche toutes catégories de personnel confondues pour une année AAAA donnée (*REM_MOY_BRUT_AA* ; niveau « branche »).

L'imputation des variables de salaire, de dépense courante, de capital et de frais généraux se fait selon l'algorithme suivant :

- **1er cas : la Dird « nature de charges » (*DI_DIRD_NATU_CHARGE*) est connue contrairement aux salaires (*DI_D_SALAIRE*) :**
  On calcule une estimation de la rémunération moyenne à partir des effectifs R&D en ETP, de la Dird « nature de charges » et des dépenses de R&D en capital hors rémunérations immobilisées
  - Si cette estimation ne s'éloigne pas « excessivement » de la rémunération moyenne de l'année précédente pour cette même branche de recherche (on vérifie au préalable que les données de l'année précédente sont mobilisables), alors on fait des imputations déterministes pour les salaires et les dépenses courantes « avec salaires » en utilisant les égalités reliant ces variables.

  - Sinon, si la rémunération brute moyenne est connue pour l'année antérieure et si ces données ont été indiquées comme diffusables, alors les salaires sont imputés à l'aide des données antérieures. On impute également en conséquence les autres composantes de la Dird « nature de charges » de la manière suivante :
    - si la dépense en capital est positive et si la dépense courante est inférieure à la Dird « nature de charges », alors on impute la dépense en capital par la différence entre la Dird « nature de charges » et la dépense courante,
    - Sinon, si la dépense en capital est connue, on la compare à la Dird « nature de charges » :
      - si la dépense courante est inférieure ou égale à la Dird « nature de charge », on impute en utilisant les égalités entre variables,
      - sinon, on a fait le choix de fixer les frais généraux à 0 et d'imputer les autres composantes en conséquence.

  - Sinon (i.e. si la rémunération brute moyenne est inconnue pour l'année antérieure ou si les données antérieures ne sont pas mobilisables car non diffusables), alors on modélise le salaire en fonction des effectifs R&D en ETP des différentes catégories de personnel. **On impute alors par régression les salaires**. Pour les unités de la strate « QG », on utilise le modèle « SALAIRE QG » (cf. annexe 3) et pour les autres unités le modèle « SALAIRE QS » (cf. annexe 4).
    On impute également en conséquence les autres composantes de la Dird « nature de charges » de la même façon que dans le cas précédent.

- **2ème cas : la Dird « nature de charges » est inconnue :**

Suite aux imputations déductives des variables relatives à la Dird (cf. §2.3.2), si la Dird « nature de charges » est nulle ou vide, il en est de même de ses composantes. On lui substitue donc la Dird « approche géographique » ou la Dird « catégorie de recherche » :

- si la Dird « approche géographique » est connue :
  - si les données antérieures sont mobilisables, alors on impute les composantes en respectant les mêmes proportions que l'année précédente,
  - sinon les composantes sont imputées en respectant les moyennes des proportions données par les classes d'imputation utilisées précédemment (**imputation par la moyenne**),

- sinon, si la Dird « catégorie de recherche » est connue, on la substitue à la Dird « nature de charges ». Les composantes sont imputées de manière similaire au cas précédent,

- sinon, l'unité est considérée comme relevant de la non réponse totale.

- **3ème cas : la Dird « nature de charges » est différente de la somme de ses composantes :**

  - Si la dépense courante est positive, alors les composantes de la Dird « nature de charges » sont redressées dans les mêmes proportions,
  - Sinon, elles sont imputées par la moyenne observée dans les classes d'imputation utilisées précédemment (**imputation par la moyenne**).

- Imputation des composantes des salaires :

Les salaires (*DI_D_SALAIRE*) sont la somme des rémunérations du personnel de R&D (DI_REM_CS) et des rémunérations immobilisées pour la R&D (*DI_IMMO_REM*). Si cette ventilation ne correspond pas aux salaires, alors :

- si une ventilation a été donnée par l'UL, alors on redistribue dans les mêmes proportions les composantes,
- sinon si les données de l'année précédente sont mobilisables, alors on redistribue dans les mêmes proportions que l'année précédente,
- sinon on choisit de faire porter tout le poids des salaires par les rémunérations des personnels de R&D.

- Imputation des composantes de la dépense en capital :

Le détail des dépenses en capital n'est disponible que pour les UL répondants au questionnaire général (« QG »). C'est pourquoi, dans cette partie, seules les unités « QG » sont concernées.
Pour ces unités :

- si le détail des dépenses en capital est inconnue alors que celle-ci est positive, on impute les composantes pour qu'elles correspondent en proportion aux unités de la même classe d'imputation (**imputation par la moyenne**),
- sinon, si la somme des composantes ne correspond pas, on les redistribue suivant la même répartition que celle observée.

2.4.4. Redressement des variables de dépense extérieure de R&D et des variables de ressources

En ce qui concerne les variables liées à la dépense extérieure de R&D (Derd) ou aux ressources effectives directes consacrées à la R&D, on réapplique les traitements déjà existants qui consistent en divers contrôles de cohérence ainsi qu'en diverses redéfinitions de variables.

2.4.5. Dernières imputations pour les branches monodépartementales

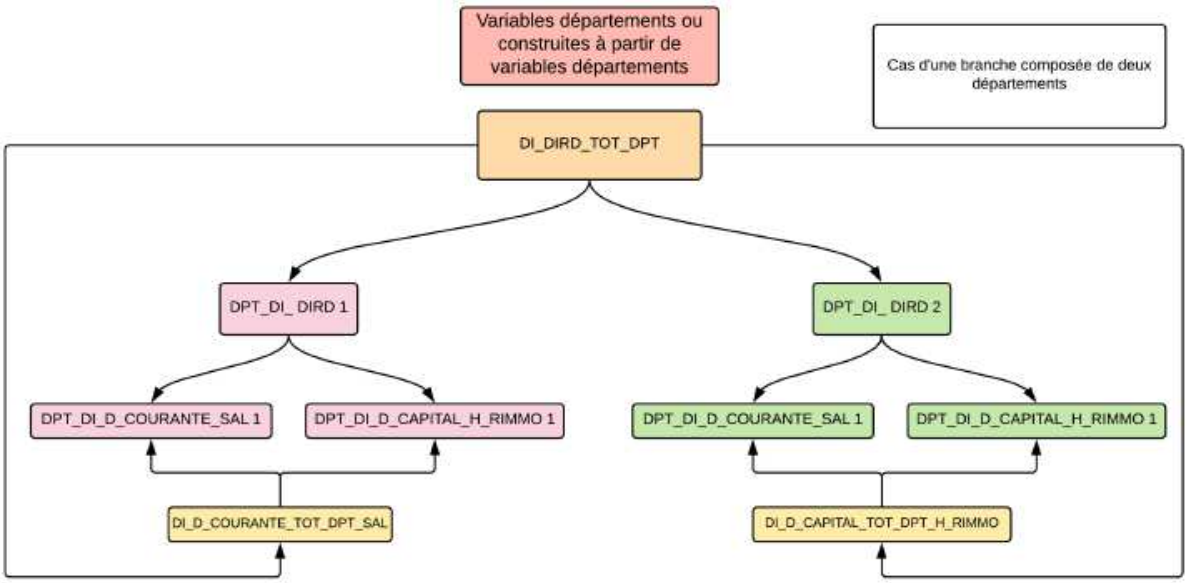
A cette étape, il peut rester une part de non réponse partielle. Par exemple, une unité pour laquelle seule la Dird « catégorie de recherche » est connue est considérée comme répondante exploitable. Cependant, du fait de l'ordre des traitements, les composantes de la Dird « nature de charges » n'ont pas été imputées. On réitère donc ici les traitements de la non-réponse partielle décrits aux §2.3 et §2.4.

A l'issue de ces derniers traitements, la non réponse partielle concernant les branches monodépartementales est traitée. Il reste à traiter la non réponse partielle des branches multidépartementales.

2.5. Le traitement de la non réponse partielle – Les imputations déductives pour les branches multidépartementales

Les relations liant les variables relatives à la Dird pour les branches monodépartementales (cf. figure 3) sont identiques pour les branches multidépartementales hormis pour les variables en lien avec l'approche géographique. On peut illustrer les liens entre ces variables pour une branche présente dans deux départements ainsi :

Figure 7 – Liens entre la Dird « approche géographique » et ses composantes dans le cas d'une branche bidépartementale



Il en découle des imputations déductives similaires à celles des branches monodépartementales (cf. §2.3).

2.6. Le traitement de la non réponse totale

2.6.1. Reconduction des données des unités de la strate exhaustive

Dans la lignée de l'existant, on reconduit les données concernant les unités de la strate exhaustive (i.e. les unités tirées dans les sous-strates « QG » et « QS exhaustif ») lorsqu'elles sont non répondantes exploitables. Contrairement aux anciens traitements, on ne permet pas que des données déjà reconduites le soient une fois de plus. Les unités dont les données n'ont pas été reconduites seront intégrées dans les traitements du §2.6.2 tandis que les unités « reconduites » ne seront réintégrées dans la base de travail qu'à la toute fin du traitement de la non réponse totale.

2.6.2. La correction de la non réponse totale par repondération

La correction de la non réponse totale par repondération consiste à faire porter le poids des non répondants par les répondants. Le but de cette enquête est d'estimer des totaux (celui de la Dird par exemple) de la forme :

$$t_y = \sum_{k \in U_{init}} y_k \times \mathbb{1}_{k \in U_{final}}$$

On souhaite estimer ce total à partir des données collectées auprès des répondants. On modélise le mécanisme aléatoire menant à l'échantillon des répondants comme un échantillonnage en deux phases :

1. sélection de l'échantillon S_{init} : ce mécanisme est connu, il découle du plan de sondage présenté au §1.3.1. On dispose des probabilité d'inclusion π_k pour toutes les UL,
2. sélection du sous-échantillon des répondants $S_{init_{REP}}$ parmi les UL échantillonnées. On note p_k la probabilité d'appartenir au sous-échantillon des répondants conditionnellement à S_{init}. Pour corriger la non réponse, on souhaite utiliser l'estimateur par expansion :

$$\hat{t_y} = \sum_{k \in S_{init_{REP}}} \frac{y_k}{\pi_k \times p_k} \times \mathbb{1}_{k \in U_{final}}$$

Cependant, les probabilités de réponse p_k ne sont pas connues. On fait l'hypothèse qu'il existe un modèle de réponse de la forme :

$$p_k = f(z_k ; \beta_0)$$

Où :

- z_k : un vecteur de variables auxiliaires connus sur S_{init}
- β_0 : un paramètre inconnu

Il a donc fallu estimer ces probabilités de réponse :

- <u>Travaux préliminaires :</u>

La variable *REP_EXPL_AA* que nous avons utilisée pour traiter la non réponse partielle permet de différencier les réponses exploitables des autres. Elle ne permet cependant pas de discriminer les répondants et les non répondants. Il est ainsi nécessaire de mettre en place une indicatrice de réponse prenant en compte les UL répondantes non exploitables : c'est l'indicatrice *REP_AA*.

- <u>Création des groupes de réponse homogène (GRH) :</u>

Estimer la probabilité de réponse pour chaque individu peut sembler ambitieux. En pratique, on choisit de mettre en place des groupes de réponse homogène. On fait ainsi l'hypothèse que la probabilité de réponse p_k est constante dans chaque partie d'une partition de l'ensemble S_{init}. Ces parties sont les groupes de réponse homogène (GRH).

C'est la méthode par croisement qui est utilisée ici pour construire ces GRH. Elle consiste à sélectionner les variables auxiliaires les plus significatives de la régression logistique expliquant la probabilité de réponse en fonction des différentes variables auxiliaires à notre disposition (les variables *TIRAGE* et *SECT_PUB2* correspondant à la strate de tirage et à la branche de recherche principale de l'UL (cf. annexe 5)). Le croisement de ces variables permet de constituer les GRH. Certains croisements présentent peu d'individus. Dans la note méthodologique [2], il est recommandé que chaque GRH contienne plus d'une centaine d'individus en évitant des groupes de moins de cinquante. Ces recommandations ont été suivies lors du regroupement des modalités pour aboutir aux GRH finaux.

- Calcul des nouveaux poids :

Pour chaque GRH i, noté GRH_i, la probabilité de réponse p_{k,GRH_i} de l'UL k présente dans le groupe GRH_i est estimée par la relation suivante :

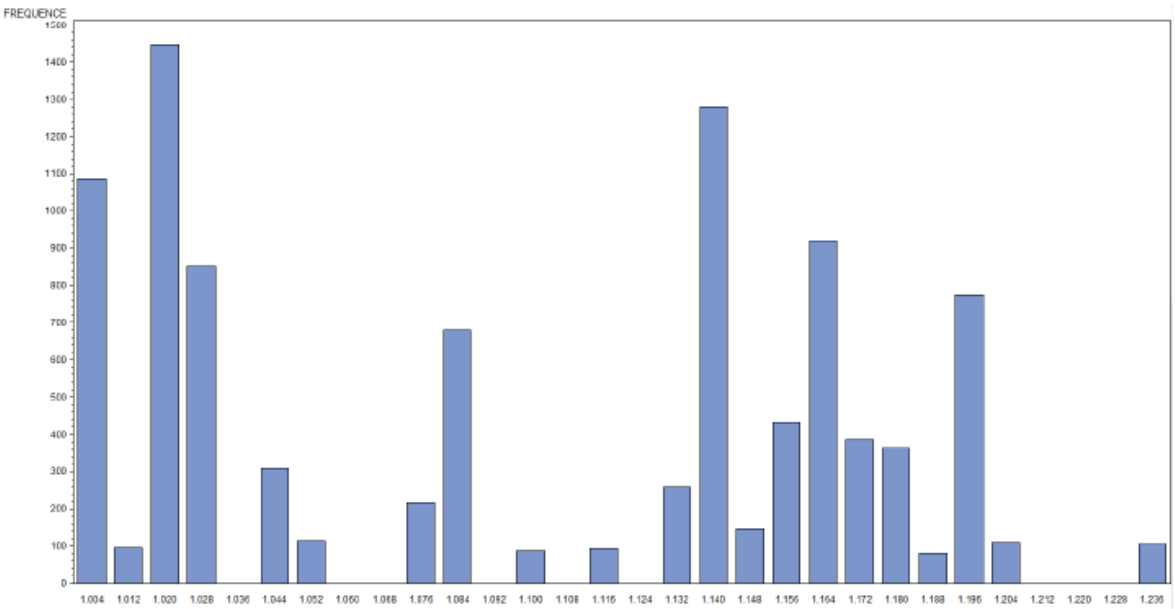
$$\forall k \in GRH_i, \hat{p_{k,GRH_i}} = \frac{nombre\ d'UL\ répondantes\ dans\ GRH_i}{nombre\ d'UL\ présentes\ dans\ GRH_i}$$

Les poids redressés de la non réponse totale, noté w_{k,GRH_i} sont alors donnés par la relation suivante :

$$\forall k \in GRH_i, w_{k,GRH_i} = \frac{1}{\pi_k \times \hat{p_{k,GRH_i}}}$$

On illustre ce traitement par le rapport des poids redressés de la non réponse totale sur les poids initiaux (cf. figure 8). Ce rapport est compris entre 1 et 1,24.

Figure 8 – Rapport des poids redressés de la non réponse totale sur les poids initiaux



3. Autres traitements post-collecte

3.1. Le traitement des valeurs influentes

Certaines unités ont une contribution très forte à la construction des agrégats comme la Dird. Il arrive cependant qu'elles ne représentent qu'elles-mêmes. Aussi, est-il important de réévaluer la pondération qui leur est associée. C'est ce que nous nous proposons de faire dans cette partie.

3.1.1. La méthode de Winsorisation

On fait référence à deux types de winsorisation :

- la winsorisation de type I que l'on peut réduire à un recalcul des poids :

$$\widetilde{d_i} = d_i \frac{min(y_i, \frac{K}{d_i})}{y_i}$$

- la winsorisation de type II que l'on peut aussi réduire à un recalcul des poids :

$$\widetilde{d_i} = 1 + (d_i - 1)\frac{min(y_i, \frac{K}{d_i})}{y_i}$$

avec :

- y_i: la valeur de la variable d'intérêt y (la Dird ici) pour l'unité i,
- d_i: le poids de l'unité i,
- K : une constante positive appelée seuil de winsorisation.

Dans ces mécaniques, le seuil de winsorisation est central. Cyril Favre-Martinoz et Thomas Deroyon [6] proposent trois méthodes pour le déterminer :

- à dire d'expert,
- en minimisant l'erreur quadratique moyenne estimée de l'estimateur robuste. Par exemple, la méthode de Kokic et Bell,
- en choisissant le seuil minimisant le maximum des influences calculées sur l'estimateur robuste (Beaumont et al., 2013).

3.1.2. La méthode de Kokic et Bell

La méthode de Kokic et Bell repose sur l'hypothèse qu'à l'intérieur d'une même strate, les valeurs y_i sont des réalisations indépendantes d'une même loi. Kokic et Bell ont montré ([7]) que le biais de l'estimateur winsorisé est le point où s'annule la fonction F définie par la relation suivante :

$$F(B) = -B[1 + \sum_h n_h E_h(\widetilde{J_h})] - \sum_h n_h E_h(\widetilde{Y_h}\widetilde{J_h})$$

avec :

- E_h : l'espérance selon la loi de Y dans la strate h,
- $\widetilde{Y_h} = (\frac{N_h}{n_h} - 1)(Y_h - \mu_h)$
- μ_h : l'espérance de Y dans la strate h

- $\widetilde{J}_h$ l'indicatrice de la condition $Y_h > K_h$
- K_h : seuil de winsorisation de la strate h. K_h est équivalent asymptotiquement, dans chaque strate h, à $-\frac{B}{\frac{N_h}{n_h}-1}+\mu_h$

Il s'agit donc dans cette méthode de trouver le nombre B qui permet d'annuler la fonction F et d'en déduire les seuils de winsorisation pour chaque strate h.

3.1.3. Application de la méthode

La méthode de Kokic et Bell est adaptée à notre cas ([6]). En effet :

- nous cherchons à estimer le total d'une variable d'intérêt positive (la Dird)
- l'échantillon est sélectionné par un sondage aléatoire simple stratifié à un degré
- nous disposons des données d'enquêtes précédentes

En appliquant la méthode de Kokic et Bell sur les données antérieures, on obtient les seuils suivants :

Strate	Seuil
QS sondé - partie conservée, sondée en N-1	4 087
QS sondé - partie renouvelée	4 105

3.1.4. Mise en place du mécanisme de winsorisation

Le mécanisme de winsorisation fait intervenir les poids des unités. Ceci amène une réflexion sur la place que doit occuper ce traitement vis-à-vis des autres traitements. Cette réflexion a été menée par Arnaud Fizzala ([8]) :

- Si on applique ce traitement sur les poids non corrigés, les hypothèses théoriques sont respectées. Cependant si une UL voit son poids augmenter du fait de la correction par repondération, sa nouvelle influence pourrait ne pas être détecter.
- A l'inverse, l'application de ce mécanisme à la suite de modifications de poids permet de se prémunir de la non-détection de ces unités. Le cadre théorique n'est en revanche plus respecté.

L'application de ces traitements sur les données 2016, quel que soit sa position par rapport aux autres traitements, amène à détecter et à corriger les poids des mêmes unités (au nombre de 10). Il a donc été décidé de placer ces traitements avant les traitements de correction de la non réponse et ainsi de se placer dans le cadre théorique.
A l'issue de ces traitements, on dispose dans notre table de travail des poids winsorisés $\widetilde{d}_i$ pour chaque unité i, obtenus en utilisant la winsorisation de type II.

3.2. Le calage

Le calage consiste en l'ajustement des poids pour obtenir des totaux connus grâce à des informations auxiliaires. Cette méthode permet de trouver de nouveaux poids W_k (poids de calage) qui sont aussi proches que possible, au sens d'une certaine « fonction de distance » D, des pondérations avant calage w_k. Ces nouveaux poids permettent également d'obtenir des totaux (X) connus grâce à des informations auxiliaires. Ainsi, comme le résume Pascal Ardilly ([9]) il s'agit de résoudre le problème de minimisation sous contraintes suivant :

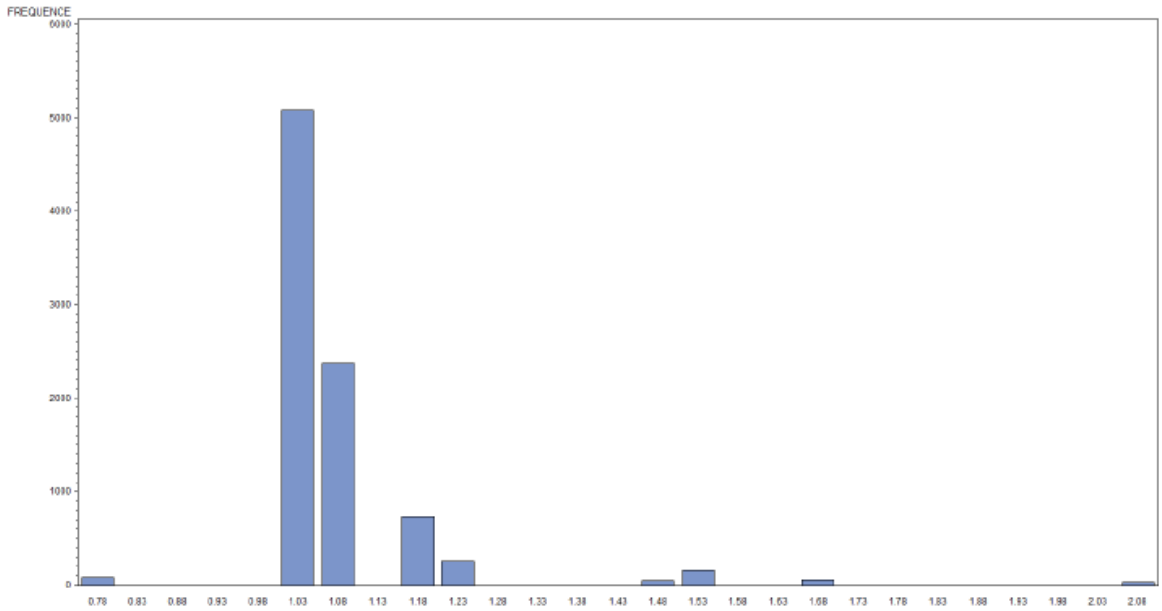
$$\min_{W_k} \sum_{k \in S} D(W_k, \omega_k)$$

avec :

$$\sum_{k \in S} W_k \times x_k = X$$

La macro %CALMAR2 est utilisée pour réaliser le calage sur marges. Les marges utilisées sont les effectifs totaux des différentes catégories de la variable *NIV 1*. Cette variable correspond au premier niveau (niveau « section ») de la nomenclature d'activité (NAF rév2, 2008). L'algorithme ne converge pas à un niveau plus fin. Plusieurs regroupements de modalités peuvent améliorer cette convergence. Ces regroupements étant assez nombreux, nous avons préféré opter pour le niveau « section ». On utilise la méthode « logit » afin de garder un contrôle sur les poids obtenus après calage. Le rapport des poids calés sur celui des poids avant calage est illustré par la figure suivante :

Figure 9 - Rapport des poids calés sur celui des poids avant calage



4. La suite

Le travail mené dans le cadre de cette étude a permis d'aboutir sur de nouveaux traitements post-collecte. En effet, il est apparu, après un inventaire des traitements actuels, que divers points pouvaient être améliorés. Nous proposons donc que les traitements exposés dans cette étude soient utilisés pour améliorer ou remplacer les traitements utilisés actuellement.
Cependant, il convient, avant d'utiliser ces nouveaux traitements en production de réaliser des calculs afin d'estimer le biais ainsi que l'impact sur la variance engendré par l'application de ces nouveaux traitements.

Bibliographie

[1] Béatrice NEITER et Benoît BUISSON. Comment redresser une enquête thématique ? *Série des documents de travail de la Direction des Statistiques d'Entreprises*, pages 8–10, Janvier 2010.

[2] Thomas DEROYON. La correction de la non-réponse par repondération. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[3] Thomas DEROYON et Cyril FAVRE-MARTINOZ. La correction de la non-réponse par imputation. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[4] Olivier SAUTORY. Les enjeux méthodologiques liés À l'usage de bases de sondage imparfaites. *Journées de méthodologie statistique*, pages 4–5, mars 2015.

[5] Nathalie CARON et Pascale PIETRI-BESSY Philippe BRION. Redresser la non-réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter. *Journées de méthodologie statistique*, pages 3–7, mars 2005.

[6] Cyril FAVRE-MARTINOZ et Thomas DEROYON. Traitement des valeurs influentes dans les enquêtes. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[7] Thomas DEROYON. Traitement des valeurs atypiques d'une enquête par winsorization – application aux enquêtes sectorielles annuelles. *Journées de méthodologie statistique*, pages 3–5, 2015.

[8] Arnaud FIZZALA. Adaptation of winsorization caused by weight share method. *Article ESANE*, octobre 2018.

[9] Pascal Ardilly. Compléments sur les redressements. *Cours de master*, 2019/2020.

[10] David HAZIZA, Jean-François BEAUMONT et Cyril FAVRE-MARTINOZ. Une méthode de détermination du seuil pour la winsorisation avec application à l'estimation pour les domaines. *Techniques d'enquête Vol. 41, Statistique Canada, n°12-001-X*, pages 59–79, juin 2015.

[11] Olivier Sautory. La macro calmar redressement d'un échantillon par calage sur marge. *Série des documents de travail de la direction des statistiques démographiques et sociales*, novembre 1993.

Annexes

Annexe 1 : Imputations déductives des variables relatives à la Dird pour les branches monodépartementales

Figure A1 – Processus d'imputation des variables « dépenses courantes avec salaires »

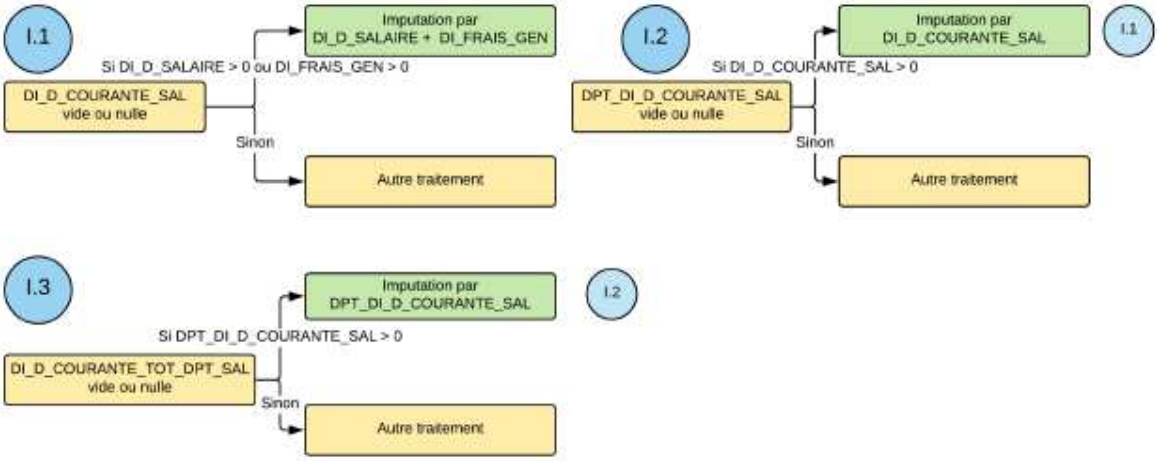


Figure A2 – Processus d'imputation des variables « dépenses en capital sans rémunérations immobilisées »

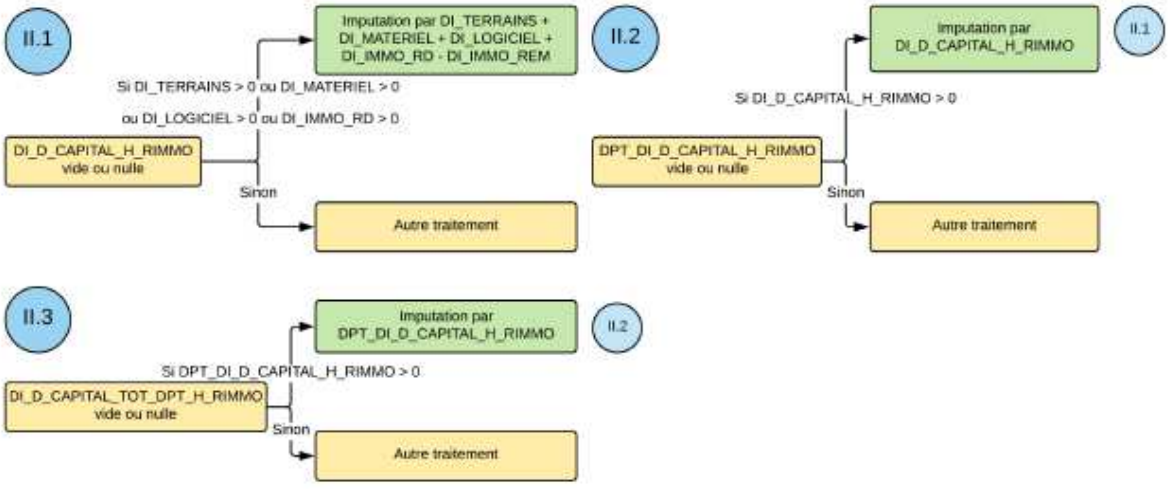
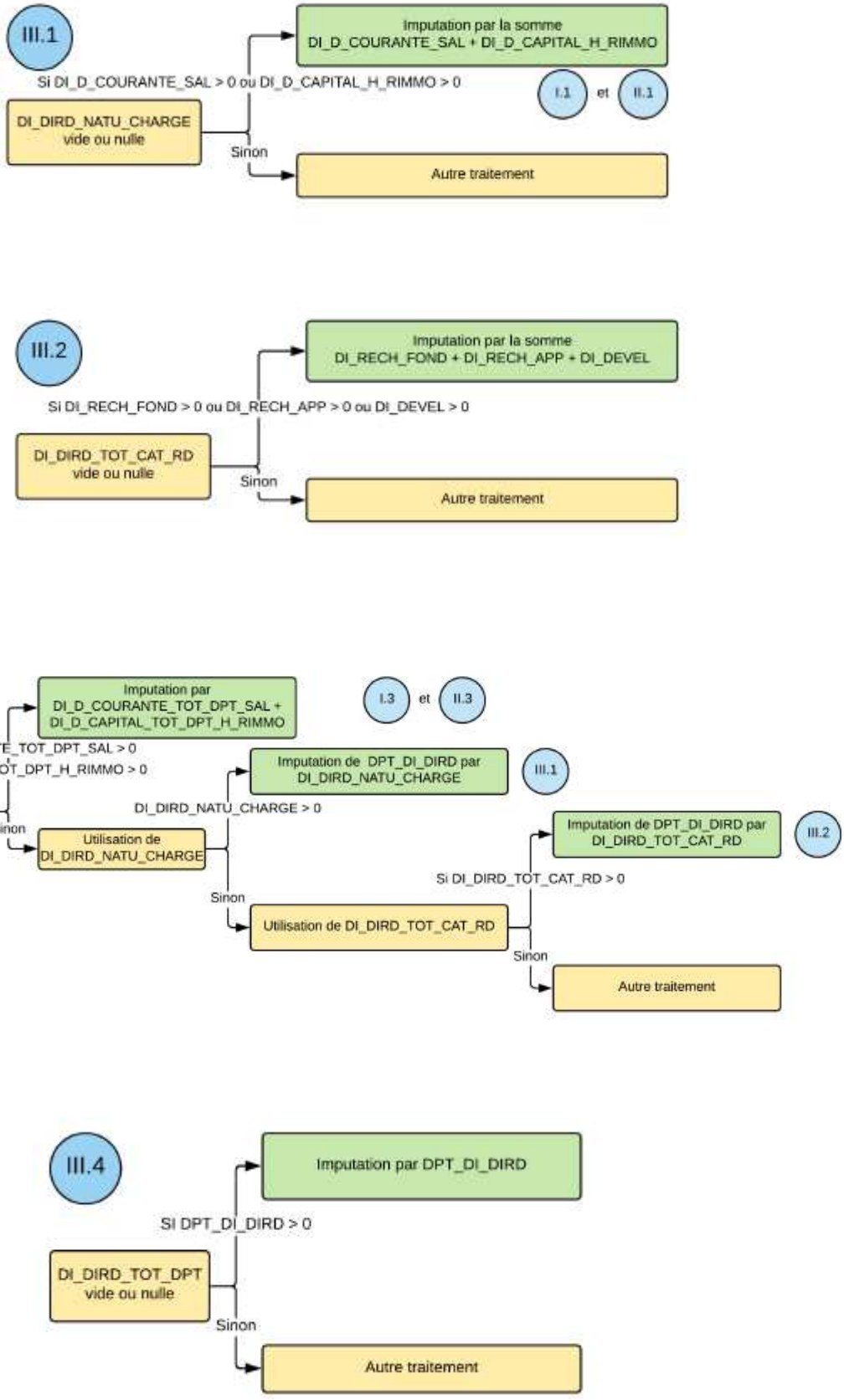


Figure A3 – Processus d'imputation des variables « Dird »



Annexe 2 : Construction des classes d'imputations

Les classes d'imputation sont construites en croisant :

- la variable de tirage (*TIRAGE_MOD*)
- la variable « branche de recherche regroupée » (*NF_MOD*)

Figure A4 – Modalités de la variable de tirage (TIRAGE_MOD)

Modalité $TIRAGE\_MOD$	Modalité $TIRAGE$
"QG"	"QG"
"QS exhaustif"	"QS exhaustif"
"QS sondé - partie conservée, sondée en N-1"	"QS sondé - partie conservée, sondée en N-1"
"QS sondé - partie exclue, sondée en N-1"	"QS sondé - partie exclue, sondée en N-1"
"QS new ou QS sondé - Partie renouvelée"	"QS new"
	"QS sondé - partie renouvelée"

Figure A5 – Modalités de la variable « branche de recherche regroupée » (NF_MOD)

Modalité NF_MOD	Modalité *tableau_NF*	libellé *tableau_NF*
NF_02	NF_02	Industrie automobile
NF_03	NF_03	Construction aéronautique et spatiale
NF_04	NF_04	Industrie pharmaceutique
NF_05	NF_05	Industrie chimique
NF_06	NF_06	Fabrication d'instruments et d'appareils de mesure, d'essai et de navigation ; horlogerie
NF_07	NF_07	Composants, cartes électroniques, ordinateurs, équipements périphériques
NF_08	NF_08	Fabrication de machines et équipements n.c.a.
NF_09	NF_09	Fabrication d'équipements électriques
$NF10\_11$	NF_10	Fabrication d'équipements de communication
	NF_11	Autres branches des industries manufacturières
NF_12	NF_12	Primaire, énergie, construction
NF_14	NF_14	Activités informatiques et services d'information
NF_15	NF_15	Activités spécialisées, scientifiques et techniques
NF_16	NF_16	Édition, audiovisuel et diffusion
$NF17\_18$	NF_17	Télécommunications
	NF_18	Autres branches de services

Annexe 3 : Modèle de régression « SALAIRE QG »

Le modèle de régression « SALAIRE QG » permet d'expliquer les salaires (*DI_D_SALAIRE*) à partir des variables d'effectif R&D en ETP suivantes :

- *DPT_CHERCHEUR_ETP*
- *DPT_TECH_SUPPORT_ETP_QG*
  (où *DPT_TECH_SUPPORT_ETP_QG*
  = DPT_TECHNICIEN_ETP + DPT_OUV RIER_ETP + DPT_ADMINISTRATIF_ETP).

Ce modèle se base sur les branches monodépartementales :

- ayant des salaires non manquants et non nuls (i.e. *DI_D_SALAIRE* ∉ (0 ; .))
- et dont le détail des effectifs R&D en ETP est connu et entièrement rémunérés par la société (i.e. *DPT_CHERCHEUR_ETP + DPT_TECHNICIEN_ETP +DPT_SUPPORT_ETP* ∉ (0 ; .) et *BRA_EFFRD_REM* = 100).

A ces unités ont été retranchées les unités aberrantes.

Le coefficient de détermination de ce modèle est proche de 1 (R^2 = 0,8686). Les coefficients sont :

Variable	Valeur estimée des paramètres	P - value
Intercept	−944.99	< 0.0001
DPT_CHERCHEUR_ETP	101.17	< 0.0001
DPT_TECH_SUPPORT_ETP	101.17	< 0.0001

Annexe 4 : Modèle de régression « SALAIRE QS »

Le modèle de régression « SALAIRE QS » permet d'expliquer les salaires (*DI_D_SALAIRE*) à partir des variables d'effectif R&D en ETP suivantes :

- *DPT_CHERCHEUR_ETP*
- *DPT_TECH_SUPPORT_ETP_QS*
  (où *DPT_TECH_SUPPORT_ETP_QS = DPT_EFFRD_ETP – DPT_CHERCHEUR_ETP*)

Ce modèle se base sur les branches monodépartementales :

- ayant une dépense courante non nulle (i.e. *DI_D_COURANTE_SAL* ∉ (0 ; .))
- et dont le détail des effectifs R&D en ETP est connu et entièrement rémunérés par la société (i.e. *DPT_CHERCHEUR_ETP + DPT_TECH_SUPPORT_ETP* ∉ (0 ; .) et *BRA_EFFRD_REM* = 100).

A ces unités ont été retranchées les unités aberrantes.

Le coefficient de détermination de ce modèle est proche de 1 (R^2 = 0,8628). Les coefficients sont :

Variable	Valeur estimée des paramètres	P - value
Intercept	−3.23	0.4124
DPT_CHERCHEUR_ETP	67.44	< 0.0001
DPT_TECH_SUPPORT_ETP	61.85	< 0.0001

Annexe 5 : Modélisation de la non réponse

On modélise la probabilité de répondre à l'aide d'une régression logistique. Deux variables sont significatives au niveau 5 %. Ce sont les variables :

- *TIRAGE* correspondant à la strate de tirage
- *SECT_PUB2* qui correspond à la branche principale de recherche de l'UL.

On utilise le croisement de ces deux variables pour créer les groupes de réponse homogènes (GRH).

Figure A6 – Le modèle de régression logistique de la non réponse

Test de l'hypothèse nulle globale : BETA=0

Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	556.8104	36	<.0001
Score	497.0684	36	<.0001
Wald	393.5006	36	<.0001

NOTE: No effects for the model in Step 2 are removed.

NOTE: All effects have been entered into the model.

Récapitulatif sur la sélection Stepwise

Etape	Effet Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2	Libellé de la variable
1	TIRAGE		4	1	401.1064		<.0001	TIRAGE
2	SECT_PUB2		32	2	87.7947		<.0001	

Analyse des effets Type 3

Effet	DDL	Khi-2 de Wald	Pr > khi-2
SECT_PUB2	32	82.2656	<.0001
TIRAGE	4	282.6029	<.0001

Le croisement des variables *TIRAGE* et *SECT_PUB2* nous permet de construire les GRH. On veille à ce que ceux-ci soient constitués d'au moins 50 unités en regroupant des modalités si besoin. C'est ainsi que nous construisons la variable *BLOC_EXPLICATIF* dont les modalités sont indiquées ci-dessous :

BLOC_EXPLICATIF	Fréquence
NF023345_QS sondé - partie renouvelée	105
NF02345_QS new	96
NF02345_QS sondé - partie conservée, sondée en N-1	129
NF023_QG	103
NF023_QS exhaustif	74
NF04_QG	97
NF04_QS exhaustif	82
NF05_QG	75
NF05_QS exhaustif	135
NF0678_QS sondé - partie conservée, sondée en N-1	192
NF0678_QS sondé - partie renouvelée	166
NF067_QS new	55
NF06_QG	57
NF06_QS exhaustif	100
NF07_QG	58
NF07_QS exhaustif	79
NF08910_QS new	131
NF08_QG	85
NF08_QS exhaustif	182
NF091011_QS sondé - partie conservée, sondée en N-1	435
NF091011_QS sondé - partie renouvelée	482
NF0910_QS exhaustif	111
NF09_QG	64
NF1011_QG	243
NF11_QS exhaustif	504
NF11_QS new	516
NF12_QG	57
NF12_QS exhaustif	98
NF12_QS new	221
NF12_QS sondé - partie conservée, sondée en N-1	84
NF12_QS sondé - partie renouvelée	101
NF14_QG	124
NF14_QS exhaustif	386
NF14_QS new	499
NF14_QS sondé - partie conservée, sondée en N-1	297
NF14_QS sondé - partie renouvelée	322
NF15_QG	146
NF15_QS exhaustif	540
NF15_QS new	761
NF15_QS sondé - partie conservée, sondée en N-1	417
NF15_QS sondé - partie renouvelée	445
NF161718_QG	115
NF16_QS exhaustif	237
NF16_QS new	189
NF16_QS sondé - partie conservée, sondée en N-1	130
NF16_QS sondé - partie renouvelée	131
NF1718_QS exhaustif	109
NF1718_QS new	507
NF1718_QS sondé - partie conservée, sondée en N-1	95
NF1718_QS sondé - partie renouvelée	134