Journal de la Société Française de Statistique
Vol. 155 No. 4 (2014)

Méthodes et pratiques des enquêtes entreprises à l'Insee

Title: Methodology and practice of business surveys at Insee

Elvire Demoly[1], Arnaud Fizzala[2] et Emmanuel Gros[1]

Résumé : La statistique institutionnelle distingue traditionnellement statistique auprès des entreprises et statistique auprès des ménages. Cette séparation apparaît clairement dans de nombreux organigrammes des instituts nationaux statistiques et se retrouve, plus ou moins explicitement, dans les colloques et les publications. Elle découle de l'existence de spécificités fortes, propres à chaque univers. Dans la sphère entreprise, ces spécificités – univers très hétérogène et de taille relativement restreinte, problème de charge statistique, collecte par voie postale ou par Internet, etc. – affectent l'ensemble du processus d'enquête et conditionnent les choix méthodologiques effectués lors des différents traitements.

Cet article a pour but de passer en revue les différentes phases du processus de production d'une enquête auprès des entreprises à l'Institut national de la statistique et des études économiques (Insee) – de la constitution de la base de sondage à la diffusion des résultats, en passant par la sélection de l'échantillon et les différentes étapes de redressement – et dresse un état des lieux des différentes techniques mises en œuvre à l'Insee pour les accomplir.

Abstract: Institutional statistics traditionally distinguish household statistics and business statistics. This distinction, clearly apparent in the organization charts of many national statistical institutes, and also present in conferences and publications, results from the existence of strong specificities, peculiar to each world. In the business statistics field, these specificities – very heterogeneous and relatively small population, problem of statistical burden, postal mail or internet data collection, etc. – impact the whole statistical process and determine the methodological choices.

This article aims to review the different steps of a business survey at the National Institute of Statistics and Economic Studies (Insee) – from the sampling to the results – and make a state of the art of the different techniques used at Insee.

Mots-clés : Statistiques d'entreprises, base de sondage, stratification, allocations, tirage d'échantillons, coordination d'échantillons, correction de la non-réponse, imputation, repondération, unités influentes, winsorisation, calage, secret statistique
Keywords: Business statistics, frame, stratification, optimum allocation, sampling, sampling coordination, non-response adjustment, imputation, weighting, influential units, winsorization, calibration, statistical disclosure control
Classification AMS 2000 : 62D05

[1] Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE
E-mail : elvire.demoly@insee.fr and E-mail : emmanuel.gros@insee.fr
[2] Drees, 14 avenue Duquesne, 75350 Paris 07 SP, FRANCE
E-mail : arnaud.fizzala@sante.gouv.fr

 ISSN: 2102-6238

Introduction

Environ 40 enquêtes auprès des entreprises sont réalisées chaque année par le service statistique public français [1], dont une quinzaine sont menées par l'Insee. Les processus de production de ces enquêtes – de la constitution de la base de sondage à la diffusion des résultats, en passant par la sélection de l'échantillon et les différentes étapes de redressement – possèdent de nombreuses similitudes et seront amenés à converger de plus en plus suite à la mise en place, en septembre 2012, d'une nouvelle direction consacrée à la méthodologie à l'Insee.

On dresse ici un état de l'art des différentes techniques mises en œuvre à l'Insee pour accomplir ces processus de production en les illustrant par ce qui est mis en pratique dans l'enquête sur les technologies de l'information et de la communication et le commerce électronique (enquête TIC).

L'enquête TIC

L'enquête communautaire sur les technologies de l'information et de la communication et le commerce électronique est une enquête annuelle auprès des entreprises d'au moins 10 personnes [a] des secteurs marchands hors entreprises agricoles, financières et d'assurance. Elle s'inscrit dans le dispositif d'enquêtes européennes [b], et vise à connaître le niveau d'informatisation et la diffusion des technologies de l'information et de la communication (TIC) dans les entreprises. C'est un ensemble d'enquêtes réalisées par les instituts nationaux de la statistique des pays membres de l'union européenne suivant un champ commun et un même modèle de questionnaire (à la traduction près). Elle cherche notamment à apprécier la place des outils nouveaux dans les relations externes de l'entreprise (internet, commerce électronique) et dans leur fonctionnement interne (réseaux, systèmes intégrés de gestion).
L'unité enquêtée est l'unité légale, à l'exception de quelques unités considérées au niveau groupe (unités profilées). La population cible compte environ 190 000 unités françaises, et 12 500 unités sont échantillonnées.
Une grande partie des questions posées sont qualitatives (équipements en TIC, usages, etc.), les questions quantitatives concernent essentiellement le commerce électronique (achats et ventes) et le chiffre d'affaires. Depuis l'édition 2012, la collecte est réalisée par internet. Elle a lieu chaque année de janvier à mi-mai.
Les résultats de cette enquête sont publiés par l'Insee pour le niveau France et par Eurostat pour tous les pays réalisant cette enquête (les 27 Etats membres et quelques pays supplémentaires).

[a] L'effectif est mesuré en nombre de personnes occupées, c'est-à-dire incluant les non salariés tels que le gérant, le dirigeant par exemple.
[b] En application du règlement européen n° 1006/2009 du 16 septembre 2009 amendant le règlement du 21 avril 2004.

L'article commence par une présentation du répertoire Sirus qui sert de socle à la constitution des bases de sondage. Une fois construite, la base de sondage est stratifiée et des taux de sondage sont appliqués dans chaque strate. La stratification et la détermination des allocations font l'objet des parties 2 et 3. Le tirage de l'échantillon, présenté dans la partie 4, est généralement coordonné avec d'autres échantillons afin de mieux répartir la charge de réponse entre les entreprises ou d'assurer au contraire une certaine continuité avec l'échantillon de l'édition précédente de l'enquête.

Après la collecte, les traitements visent à résoudre les problèmes de cohérence, de valeurs aberrantes ou influentes et la non-réponse. Parmi les unités pour lesquelles aucune information n'a pu être collectée, une distinction doit être réalisée entre les unités n'ayant pas répondu à l'enquête car elles n'étaient pas concernées et les autres ; c'est l'objet de la partie 5. Les traitements se poursuivent, pour les unités considérées dans le champ de l'enquête, par la correction de la non

[1] Le service statistique public français comprend l'Institut national de la statistique et des études économiques (Insee) et les Services statistiques ministériels (SSM).

 ISSN: 2102-6238

réponse-totale et partielle et la gestion des unités atypiques. Ces traitements sont présentés dans les parties 6 et 7. Les résultats corrigés de la non-réponse sont ensuite mis en cohérence avec des sources externes via des procédures de calage présentées en partie 8. Enfin, les résultats peuvent être diffusés en respectant les règles de confidentialité rappelées dans la partie 9.

1. Sirus : un nouvel outil permettant de construire des bases de sondage

Pour construire leurs bases de sondage, les statisticiens de l'Insee disposent depuis peu du répertoire statistique Sirus « Système d'Identification au Répertoire des Unités Statistiques ». Ce dernier recense les entreprises françaises et quelques-unes de leurs caractéristiques économiques ou statistiques. Parmi celles-ci, la localisation géographique, l'activité principale exercée, l'effectif salarié, le chiffre d'affaires annuel déclaré à l'administration, le statut en termes d'activité économique et la charge statistique pesant sur les unités sont les caractéristiques les plus utiles à la constitution des bases de sondage.

Ce répertoire statistique comporte quelques différences avec le répertoire administratif Sirene « Système Informatisé du Répertoire national des ENtreprises et des Établissements » qui a longtemps constitué le socle des bases de sondage des enquêtes auprès des entreprises réalisées à l'Insee. En particulier, le répertoire Sirus a pour vocation [2] d'identifier des unités statistiques, plus communément appelées entreprises [3], qui ont un sens économique, alors que Sirene identifie des unités légales [4], qui ont un sens juridique. Même si la grande majorité des unités légales peuvent être considérées comme des entreprises, il est souvent difficile d'en faire de même avec les unités légales appartenant à des groupes de sociétés car ces unités légales n'ont souvent pas d'autonomie de décision. Des travaux de « profilage », cherchant notamment à définir les liens entre entreprises et unités légales, sont en cours à l'Insee actuellement. La Société nationale des chemins de fer (SNCF), par exemple, correspond à plusieurs centaines d'unités légales qui ont été regroupées, dans le cadre de cette réflexion, en quelques entreprises.

Pour toutes ces unités, Sirus enregistre des caractéristiques économiques grâce à des mises à jour régulières provenant d'une multitude de sources, telles Sirene, le dispositif d'élaboration des statistiques sectorielles annuelles d'entreprise, les déclarations annuelles de données sociales, ou encore certains résultats d'enquêtes. En cas de divergence entre les sources, Sirus arbitre « la » valeur à retenir. De plus, Sirus calcule des caractéristiques statistiques comme la probabilité d'existence des unités en s'appuyant sur des sources externes. Par exemple, si une unité n'a effectué aucune déclaration à l'administration au cours des trois dernières années, le répertoire Sirus la considérera comme « statistiquement » inactive, évitant ainsi à l'Insee d'en tenir compte dans les procédures d'échantillonnage. Cela constitue une valeur ajoutée importante par rapport au répertoire Sirene qui, de par sa fonction administrative, attend la déclaration de cessation de l'unité légale pour enregistrer l'information.

Enfin, Sirus calcule la charge statistique des entreprises [5]. Cette information est mobilisée dans

[2] Fin 2013, le répertoire Sirus était essentiellement composé d'unités légales : seules une cinquantaine de grandes entreprises étaient profilées.

[3] Plus petite unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision.

[4] Entité juridique de droit public ou privé pouvant être une personne morale ou une personne physique.

[5] Il s'agit d'une fonction tenant compte notamment du nombre de fois où l'entreprise a été enquêtée, ainsi que de la durée moyenne de réponse à chaque enquête.

 ISSN: 2102-6238

le cadre de la nouvelle procédure de coordination des échantillons – cf. Guggemos and Sautory (2012).

2. Stratification des bases de sondage d'enquêtes auprès des entreprises : strates d'optimisation et strates de tirage

Les échantillons des enquêtes auprès des entreprises sont généralement tirés selon des plans de sondages aléatoires simples stratifiés. Le plus souvent, la population d'entreprises correspondant au champ de l'enquête est stratifiée en croisant deux [6] critères : un critère d'activité – utilisant des agrégats plus ou moins fin de la nomenclature d'activités française (NAF) – et un critère de taille – utilisant des tranches d'effectifs salariés et / ou des tranches de chiffres d'affaires. Se pose alors la question du nombre de strates à construire qui revient ici à choisir un niveau de détail pour nos deux critères.

En pratique, deux niveaux de stratification sont généralement retenus. Le premier niveau – indexé par la lettre h par la suite – correspond, à quelques aménagements près, aux croisements des domaines sur lesquels on souhaite diffuser des résultats. Comme on le verra dans la partie suivante, ce niveau sert au calcul de taux de sondage t_h assurant une certaine précision dans chaque domaine de diffusion envisagé. Le fait que ces strates d'optimisation soient relativement agrégées permet des estimations de dispersions (et donc des calculs de précisions anticipés) robustes, puisque basées sur un nombre important d'unités.

Le second niveau – indexé par la lettre t par la suite – qui est utilisé pour le tirage est éventuellement plus fin que le premier. Plus précisément, chaque strate de tirage t est incluse dans une strate d'optimisation h. On calcule le nombre d'unités à tirer n_t dans la strate de tirage t en y appliquant le taux de sondage t_h de la strate d'optimisation correspondante :

$$n_t = t_h * N_t$$

Cette procédure, basée sur les propriétés des allocations proportionnelles au nombre d'unités, permet d'améliorer la précision des estimations à venir si les critères de stratification sont liés aux paramètres que l'on souhaite mesurer [7].

Les strates de tirage correspondront ainsi au niveau de détail le plus fin possible [8] compte tenu de l'étendue du champ de l'enquête et de la taille d'échantillon envisagée. De cette façon, on espère obtenir des estimations au moins aussi précises que si l'on effectuait le tirage au niveau des strates d'optimisation. La figure 1 résume cette stratégie de tirage sur un exemple simplifié.

Lorsque le nombre de strates de tirage est important, des méthodes de contrôle des arrondis des nombres d'unités à tirer sont mises en œuvre afin de ne pas trop s'éloigner de la taille d'échantillon initialement visée [9].

[6] Un troisième critère de localisation géographique est parfois utilisé pour le tirage mais il n'est, en général, pas pris en compte lors de l'optimisation du plan de sondage.

[7] Même lorsque les critères de stratification ne s'avèrent pas liés aux paramètres que l'on souhaite mesurer, cette procédure ne dégrade pas (sauf cas très particuliers) les estimations à venir.

[8] En pratique, pour faciliter les traitements post-enquêtes, les responsables d'enquêtes souhaitent généralement imposer un nombre minimum d'unités tirées dans chaque strate de tirage.

[9] Par exemple, pour l'enquête sectorielle annuelle 2011, il y avait un écart de 1 500 entre la taille d'échantillon visée (125 000 unités) et la somme des nombres d'unités à tirer arrondis à l'entier le plus proche.

 ISSN: 2102-6238

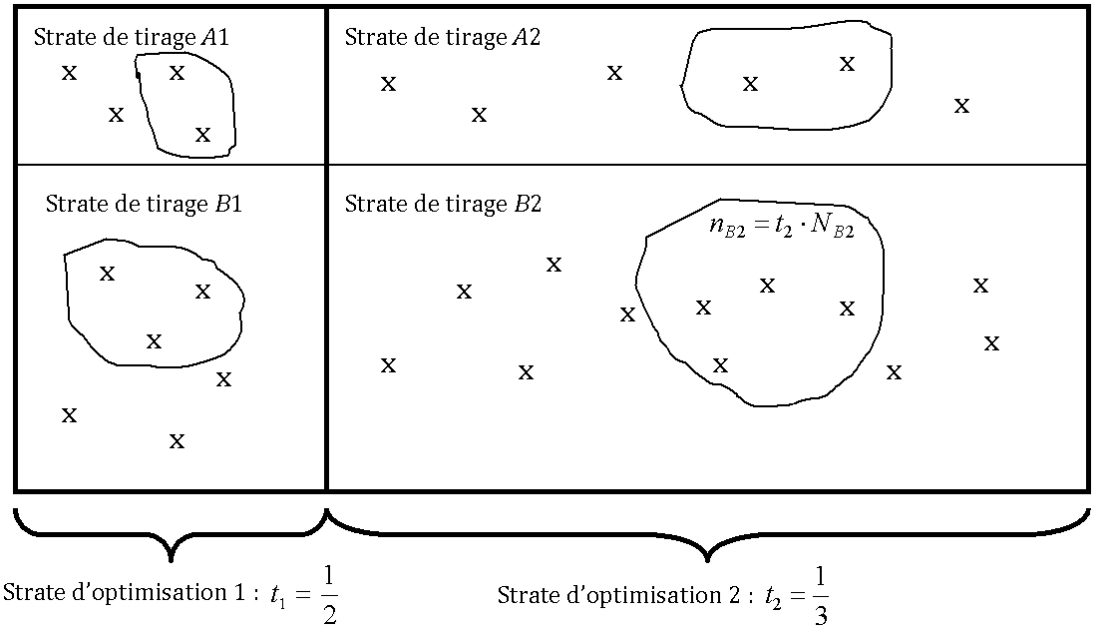

FIGURE 1. *Tirage d'un échantillon dans quatre strates de tirages incluses dans deux strates d'optimisation*

Lecture : les taux de sondage (t_1 et t_2) sont optimisés au niveau des deux strates d'optimisation (rectangles à la bordure épaisse) et l'échantillon (croix entourées) est tiré au niveau des quatre strates de tirage (rectangles à la bordure fine) en appliquant le taux de sondage de la strate d'optimisation correspondante.

La stratification de l'enquête TIC

L'échantillon de l'enquête TIC 2013 a été tiré suivant un plan de sondage stratifié selon le secteur d'activité, la tranche d'effectif de l'unité et le chiffre d'affaires, introduit depuis 2012 dans la stratification (cf. partie 3).
Les modalités des secteurs d'activité ont des niveaux d'agrégation très divers (de la classe au regroupement de sections de la NAF), correspondant en grande partie à des niveaux de restitution demandés par Eurostat.
Les tranches de taille sont au nombre de cinq : de 10 à 19 personnes occupées, de 20 à 49, de 50 à 249, de 250 à 499, 500 et plus.
À partir de 2012, un seuil de chiffre d'affaires dépendant de la tranche d'effectif de l'entreprise a été introduit comme critère d'exhaustivité (cf. partie 3).
Pour ce tirage d'échantillon, il n'y a qu'un niveau de stratification : les strates de tirage coincïdent avec les strates d'optimisation.

3. Détermination des taux de sondage : les allocations de Neyman sous contraintes de précision locales sont de plus en plus utilisées

Les enquêtes auprès d'entreprises réalisées à l'Insee ont généralement pour objectif de mesurer des quantités économiques (chiffre d'affaires consacré à un secteur d'activité, investissement consacré à un thème, etc.). Ainsi, les grandes entreprises, en raison de leur poids économique et de leur hétérogénéité, sont soumises à des taux de sondage beaucoup plus élevés que les petites. Dans la quasi-totalité des enquêtes auprès d'entreprises, une partie de la population, généralement

 ISSN: 2102-6238

composée de très grandes entreprises, est même intégrée d'office à l'échantillon, au sein de strates exhaustives. Le plus souvent, ces strates exhaustives, qui concernent parfois jusqu'à la moitié de l'échantillon, comprennent les entreprises dont le chiffre d'affaires et / ou l'effectif salarié dépasse un certain seuil. Généralement, ce seuil est fixé *a priori* et le plan de sondage est optimisé sur le reste du champ en tenant compte de ce seuil.

La définition de l'exhaustif dans l'enquête TIC

Pour l'enquête TIC, l'effectif et le chiffre d'affaires déterminent des seuils d'exhaustivité. Ainsi, les unités comptant 500 personnes ou plus sont échantillonnées exhaustivement. En outre, à partir de 2012, un seuil de chiffre d'affaires, dépendant de la tranche d'effectif de l'entreprise, a été introduit comme critère d'exhaustivité afin de limiter *a priori* l'apparition de valeurs influentes nécessitant une winsorisation (cf. partie 7 pour plus de détails) et ainsi améliorer et stabiliser les estimations des variables quantitatives.
Un cinquième de l'échantillon de l'édition 2013 a ainsi été selectionné de manière exhaustive.

L'allocation de Neyman selon une variable d'intérêt, très largement documentée dans la théorie des sondages et régulièrement utilisée [10] à l'Insee dans les plans de sondage des enquêtes auprès d'entreprises, optimise la précision de l'estimateur du total de cette variable d'intérêt au niveau de l'ensemble de la population.

Cette allocation, dans sa forme « traditionnelle », ne répond en général qu'en partie aux objectifs d'une enquête car, comme nous l'avons vu dans la partie 2, la publication des totaux de la variable est non seulement réalisée au niveau de l'ensemble de la population mais aussi à des niveaux intermédiaires appelés domaines de diffusion et correspondant à des sous-parties de la population (certaines activités seulement, certaines tailles d'entreprises seulement, etc.). Rien ne garantit que l'allocation de Neyman soit performante dans ces sous-parties. En particulier, les entreprises des secteurs correspondant à des montants peu importants (ou plus homogènes) relativement aux autres risquent d'être peu nombreuses dans l'échantillon et la précision des estimations limitées à ces entreprises peut s'avérer insuffisante.

Aussi, les taux de sondage correspondant aux enquêtes auprès d'entreprises réalisées par l'Insee sont de plus en plus basés sur une variante de l'allocation de Neyman introduisant des contraintes de précision locales. Cette variante, développée par Koubi and Mathern (2009), optimise la précision de l'estimateur du total de la variable d'intérêt au niveau de l'ensemble de la population en garantissant une précision minimale dans chaque domaine de diffusion.

Pour utiliser cette méthode, les statisticiens de l'Insee doivent estimer les dispersions de la variable sur laquelle sera basée l'allocation de Neyman sous contraintes. Si l'enquête porte sur un thème nouveau, la pratique la plus courante est d'optimiser l'allocation sur une variable connue dans la base de sondage (chiffre d'affaires ou effectif salarié) et censément liée aux variables d'intérêt de l'enquête. Lorsqu'il existe des éditions précédentes à l'enquête, les résultats de ces dernières sont généralement utilisés pour estimer les dispersions.

Le calcul des allocations de l'enquête TIC

Pour l'édition 2013, le choix a été fait d'une allocation mixte correspondant à une moyenne entre :

– une allocation proportionnelle au nombre d'unités en garantissant, pour chaque activité, une demi-longueur

[10] On trouve également souvent des allocations proportionnelles à des quantités économiques (par exemple l'effectif salarié ou le chiffre d'affaires), ce qui correspond à des cas particuliers de l'allocation de Neyman (lorsque le coefficient de variation empirique de la variable d'optimisation est le même dans chaque strate).

 ISSN: 2102-6238

de l'intervalle de confiance d'au plus 10 points pour l'estimation d'une proportion et en imposant un minimum de 10 unités tirées par strate (ou le nombre d'unités dans la strate si elle comporte moins de 10 unités dans la base de sondage) ;
– une allocation proportionnelle au nombre de personnes occupées (en imposant dans chaque strate le nombre minimum d'unités à tirer obtenu à l'étape précédente, qui garantit, pour chaque activité, une demi-longueur de l'intervalle de confiance d'au plus 10 points pour l'estimation d'une proportion).

La partie proportionnelle au nombre d'unités vise à répondre à un objectif de précision sur les variables de type proportion. Il s'agit d'un cas particulier d'allocation de Neyman sous contraintes locales. L'allocation de Neyman est calculée sur une variable indicatrice dont la dispersion (ou écart-type empirique) est estimée à 0,5 dans chaque strate [a], et les contraintes locales correspondent à une demi-longueur de l'intervalle de confiance de 10 points, par activité, pour l'estimation de la proportion correspondant à cette variable.
La partie proportionnelle au nombre de personnes occupées vise à répondre à un objectif de précision relatif aux variables de type montants (en favorisant les strates contenant les entreprises de grande taille). Le système d'allocation de Neyman sous contraintes locales n'a pas été mis en œuvre « directement » sur les variables d'intérêt quantitatives de l'enquête car elles ne concernent en général qu'une partie des unités d'une strate. La méthode et les outils utilisés à l'Insee n'ont pas pu être adaptés à temps pour l'édition 2013 de l'enquête.

[a] Il s'agit là d'une majoration de la dispersion d'une variable indicatrice.

4. Le tirage et la coordination de l'échantillon

Le système statistique public réalise chaque année un nombre important d'enquêtes auprès des entreprises [11]. Afin de réduire la charge statistique imposée aux petites et moyennes entreprises [12], des techniques de coordination négative d'échantillons sont régulièrement mises en œuvre. L'objectif de ces techniques est de favoriser, lors du tirage d'un échantillon, la sélection d'entreprises n'ayant pas déjà été sélectionnées lors d'enquêtes récentes, tout en conservant le caractère sans biais des échantillons.

La méthode utilisée jusqu'à fin 2013 à l'Insee [13], mise au point par Cotton and Hesse (1992), permet de fournir des échantillons d'entreprises coordonnés négativement deux à deux. Plus précisément, elle permet, lors du tirage de l'échantillon d'une enquête 2, de minimiser le recouvrement avec l'échantillon, déjà tiré, d'une enquête 1. Pour être mise en œuvre, il est nécessaire [14] que l'échantillon de l'enquête 1 ait été tiré selon la méthode du tri aléatoire, qui consiste à attribuer à chaque unité de la base de sondage un nombre aléatoire compris entre 0 et 1, puis à sélectionner les n_t unités possédant les plus petits nombres aléatoires de chaque strate de tirage t. Basée sur cet algorithme de tirage, cette méthode de coordination négative d'échantillons consiste à échanger les nombres aléatoires entre les unités d'une même strate dans la base de sondage de l'enquête 1, de façon à ce que les unités sélectionnées dans l'échantillon 1 récupèrent les plus grands nombres aléatoires de leur strate d'origine (c'est-à-dire la strate correspondant à la base de sondage de l'enquête 1), comme illustré à la figure 2.

[11] Une quarantaine d'enquêtes auprès des entreprises : une quinzaine est menée par l'Insee, une petite dizaine par la Direction de l'animation de la recherche, des études et des statistiques (DARES), une grosse dizaine par le Service de l'Observation et des Statistiques (SoeS) et quelques autres par des organismes hors du service statistique public comme par exemple le ministère de la recherche ou le ministère de l'écologie.
[12] Les grandes entreprises, à partir d'un certain seuil, étant systématiquement interrogées dans la plupart des enquêtes.
[13] Date à laquelle une méthode plus générale, permettant la prise en compte de la charge cumulée des entreprises et la coordination globale de l'ensemble des enquêtes, est entrée en production à l'Insee – cf. Guggemos and Sautory (2012).
[14] En particulier, la méthode n'est pas compatible avec l'algorithme de tirage systématique.

 ISSN: 2102-6238

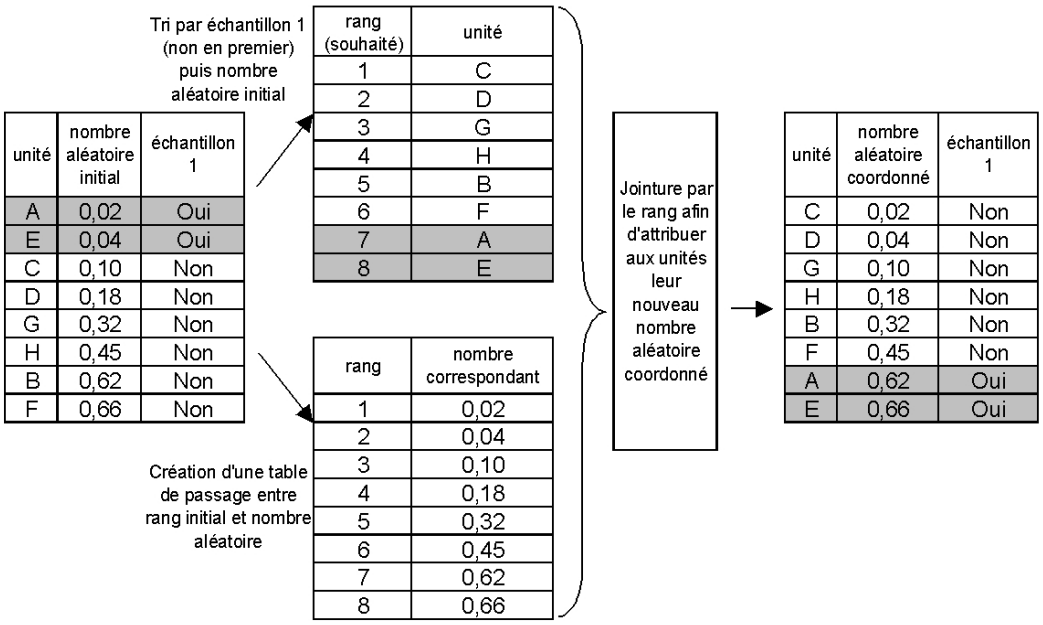

FIGURE 2. *Algorithme de permutation des nombres aléatoires dans une strate de l'enquête 1*

Cette permutation, appliquée dans chacune des strates de la base de sondage de l'enquête 1, attribue à chaque unité de cette base un nombre aléatoire "coordonné".

Pour procéder au tirage de l'échantillon de l'enquête 2, on affecte à toutes les unités de la base de sondage de l'enquête 2 :

- le nombre aléatoire coordonné, si l'unité était présente dans la base de sondage de l'enquête 1 ;
- un nombre aléatoire généré dans le cas contraire.

Il ne reste alors plus qu'à sélectionner les unités ayant les plus petits nombres aléatoires dans chaque nouvelle strate de la base de sondage de l'enquête 2.

Il est important de remarquer que l'échange de nombres aléatoires a lieu dans la base de sondage de l'enquête 1 avec la stratification correspondante : c'est ce qui va permettre d'assurer le caractère sans biais des estimations issues de l'échantillon de l'enquête 2 (car dans une strate donnée de la base de sondage de l'enquête 2, chaque entreprise aura la même probabilité de sélection marginale), au prix d'un possible [15] recouvrement entre les deux échantillons.

Lorsque les enquêtes sont répétées dans le temps et que l'on s'intéresse à des évolutions, l'échantillon n'est généralement pas renouvelé intégralement à chaque édition de l'enquête. On parle alors d'échantillon rotatif. La technique la plus couramment utilisée à l'Insee pour gérer ces rotations d'échantillon se base sur un numéro permanent – dit numéro hexal – attribué aléatoirement à chaque unité de la base de sondage correspondant à la première enquête. On

[15] Plus les stratifications seront différentes, plus le recouvrement risque d'être important.

présente dans la suite le cas le plus courant qui correspond aux renouvellements par moitié, mais la méthode se généralise sans difficultés aux renouvellements par tiers, par quart, par cinquième et par sixième.

La base de sondage de l'année initiale est divisée aléatoirement en deux parties équivalentes via l'attribution aléatoire, à chaque unité, d'un numéro "hexal" permanent compris entre 1 et 60 [16]. Les unités possédant un numéro hexal impair forment la première partie de la base de sondage, et les unités possédant un numéro hexal pair la seconde. L'échantillon de l'année initiale se trouve de fait lui aussi partitionné [17] en deux sous-échantillons de tailles équivalentes.

Par suite, pour la gestion courante de l'échantillon une année T donnée, le renouvellement par moitié des unités de l'échantillon consiste alors :

- d'une part à conserver l'une des deux parties de la base de sondage (dite partie conservée) de l'année T-1, privée des unités qui ne sont plus dans le champ de l'enquête en T, ainsi que la partie correspondante de l'échantillon (dite partie conservée de l'échantillon) ;
- et d'autre part à tirer un nouvel échantillon dans une base de tirage constituée des unités légales de l'autre partie de la base de sondage (dite partie renouvelée) de l'année T-1, également privée des unités qui ne sont plus dans le champ de l'enquête en T, auxquelles sont ajoutées les unités nouvelles, c'est-à-dire entrées dans le champ de l'enquête en T. On attribue à ces unités nouvelles un numéro hexal qui peut être pair ou impair et qui sera conservé lors des prochains tirages.

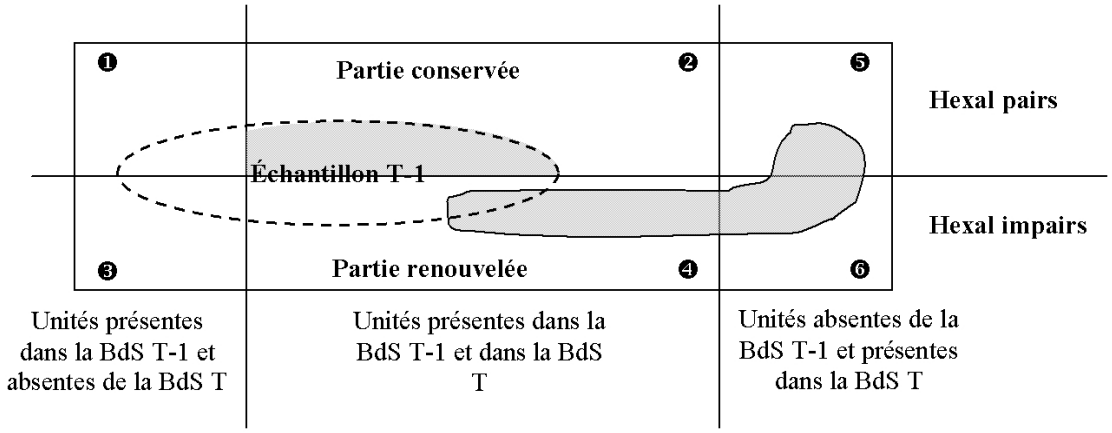

FIGURE 3. *Renouvellement par moitié concernant les numéros hexal impairs*

Dans la figure 3, l'échantillon de l'année T est constitué de l'ensemble des parties grisées. Le tirage effectivement réalisé l'année T n'a concerné que les unités de n° hexal impair présentes à la fois dans la base de sondage de l'année T-1 et de l'année T – partie ❹ –, ainsi que les unités nouvelles, c'est à dire absentes de la base de sondage en T-1, que leur numéro hexal soit

[16] Ce choix de base s'explique par le fait que 60 est divisible par tous les entiers compris entre deux et six : un numéro hexal compris entre 1 et 60 permet donc de partitionner la base de sondage en deux, trois, quatre, cinq ou six parties et donc de gérer des renouvellements par moitié, par tiers, par quart, par cinquième ou par sixième.

[17] En fonction de la parité des numéros hexal des unités le composant.

 ISSN: 2102-6238

pair ou impair – parties ❺ et ❻. Ces unités effectivement tirées en T correspondent donc à la « banane grise » et sont extrapolables à l'univers ❹ + ❺ + ❻. Les unités de la partie conservée de l'échantillon en T (la partie grisée de l'échantillon T-1 sur le schéma) sont extrapolables à l'univers ❷.

Toujours dans une démarche de réduction de la charge statistique imposée aux entreprises, cette méthode est généralement utilisée en coordonnant négativement la partie de l'échantillon renouvelée l'année T avec l'échantillon de l'année T-1. Dans le cas des renouvellements par moitié, on cherche ainsi à interroger les unités deux ans, mais pas plus.

Pour conclure sur la gestion des échantillons rotatifs, notons que dans certains cas – en cas de modifications substantielles du plan de sondage d'une enquête ayant un impact important sur l'échantillon –, la méthode du numéro hexal décrite précédemment ne convient pas au renouvellement partiel de l'échantillon du fait de l'impossibilité d'en modifier la partie conservée. Par exemple, dans le cas extrême où la taille de l'échantillon devrait être divisée par deux une année T donnée, on devrait se limiter à une partie renouvelée de l'échantillon de quelques unités seulement (voire aucune s'il n'y a pas eu d'unités entrées ou sorties du champ de l'enquête), créant ainsi un déséquilibre les années suivantes entre les tailles des deux parties de l'échantillon. Lorsque ces cas problématiques se présentent, on utilise plutôt, pour renouveler partiellement l'échantillon, une technique plus « souple » dite de coordination positive. Cette technique ressemble dans sa mise en œuvre à la technique de coordination négative décrite précédemment, mais plutôt que de permuter les nombres aléatoires des unités entre l'année T-1 et l'année T, on les conserve à l'identique pour un certain nombre d'unités en fonction du recouvrement visé entre les deux échantillons. Ainsi, certaines unités possédant de petits nombres aléatoires l'année T-1 les conservent l'année T, et ont donc à nouveau une probabilité élevée d'être tirées. L'inconvénient de la technique de coordination positive réside principalement dans le fait qu'il est difficile de bien maîtriser le recouvrement entre les deux échantillons et la durée de présence des unités dans le panel. C'est pourquoi, lorsqu'il n'y a pas de modification de l'enquête ayant un impact important sur l'échantillon, on lui préfère généralement la méthode du numéro hexal.

Renouvellement et coordination de l'échantillon de l'enquête TIC

Pour l'enquête TIC, annuelle, on souhaite assurer une continuité des données et utiliser les données de l'année précédente pour effectuer certains traitements post-collecte. Ainsi, pour la partie de l'échantillon tirée aléatoirement, on conserve habituellement la moitié de l'échantillon de l'année précédente grâce à la technique du numéro hexal.

La partie renouvelée de l'échantillon est coordonnée négativement avec l'échantillon de l'année précédente selon le procédé exposé précedemment (permutations de nombres aléatoires). Pour l'édition 2013, sur les 6 000 unités tirées dans la partie renouvelée, 1 200 (dont 1 100 dans la partie exhaustive de l'échantillon 2013) appartiennent à l'échantillon de l'édition 2012. Le résultat de la coordination est satisfaisant puisque les recouvrements sont inévitables lorsqu'une strate est exhaustive. Sans coordination, le recouvrement dans la partie non exhaustive passerait de 100 à 600 unités.

5. La caractérisation des unités : une étape préalable à la correction de la non-réponse

Les enquêtes menées par le système statistique public français auprès des entreprises sont réalisées soit par courrier, soit via Internet. Les entreprises non répondantes sont relancées pendant la collecte, de manière systématique par courrier, et de manière plus ponctuelle par téléphone, fax

 ISSN: 2102-6238

ou courriel [18]. Mais il arrive un moment où il est nécessaire d'arrêter les relances et de clore le processus de collecte. À ce stade, la collecte a abouti à un retour d'information pour une partie des entreprises : réception d'un questionnaire (exploitable ou non), éléments permettant de justifier que l'entreprise n'a pas retourné de questionnaire (entreprise ayant cessé leur activité ou hors du champ de l'enquête, etc.). Les autres unités de l'échantillon, qui n'ont pas retourné de questionnaire et pour lesquelles aucune information n'a permis en cours de collecte de justifier cette absence de réponse (que l'on appellera par la suite les « non-retours »), sont de deux types : d'une part celles qui font partie du champ de l'enquête et pour lesquelles on aurait dû recevoir un questionnaire – c'est-à-dire les non-réponses – et d'autre part celles qui ne font pas partie du champ, en particulier les entreprises cessées, et pour lesquelles il est légitime de n'avoir aucun questionnaire en retour. Il est donc impératif, préalablement à toute correction de la non-réponse, de faire la part, parmi les non-retours, entre les unités appartenant au champ de l'enquête et relevant de la non-réponse et les unités hors champ.

Or, du fait de l'absence de contact direct avec les entreprises échantillonnées induite par le mode de collecte, aucune information quant à l'appartenance ou non au champ de l'enquête des non-retours n'est disponible à l'issue du processus de collecte. Dès lors, la caractérisation de ces unités ne peut se faire qu'au travers de la mobilisation de sources d'information externes à l'enquête, la figure 4 résumant la situation.

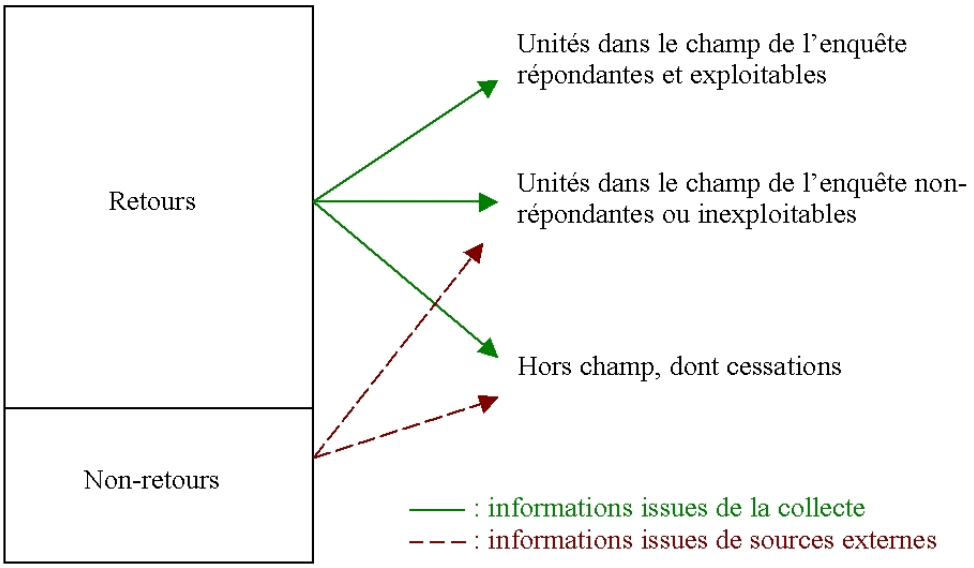

FIGURE 4. *Répartition des unités en trois catégories pour le traitement de la non-réponse*

Les sources d'information externes à l'enquête mobilisables sont de nature diverse :

[18] Notons qu'à l'Insee, les entreprises les plus importantes (en termes de chiffre d'affaires et / ou d'effectif) peuvent faire l'objet de procédures de relances spécifiques en face-à-face par le réseau des enquêteurs entreprises de l'Insee.

 ISSN: 2102-6238

- il peut s'agir de résultats d'autres enquêtes relatives à la même date d'intérêt et portant en partie sur des unités identiques. À noter que cette approche est rendue possible par l'existence de strates exhaustives dans la quasi-totalité des enquêtes entreprises ;
- les informations administratives relatives à la démographie des entreprises – date de création, date de cessation économique, date de cessation administrative, date de réactivation, etc. – contenues dans le répertoire Sirus peuvent également permettre de statuer sur le caractère inactif d'une partie des non-retours, à savoir les « faux actifs » [19] de l'échantillon. Cette approche n'est réellement efficace que lorsqu'un laps de temps suffisamment important sépare la phase de redressement de l'enquête de la constitution de la base de sondage ;
- de nombreuses sources administratives, plus ou moins exhaustives, fournissent également des informations qui permettent d'affiner la connaissance des entreprises, au moins quant à leur caractère actif ou inactif : liasses fiscales adressées par les entreprises à la Direction générale des finances publiques (DGFiP), déclarations infra-annuelles de taxe sur la valeur ajoutée (TVA), déclarations d'effectifs à l'Union de recouvrement des cotisations de sécurité sociale et d'allocations familiales (Urssaf), etc. Le recoupement de ces informations peut alors permettre d'élaborer, au prix de certaines hypothèses, une indicatrice de présomption d'activité / inactivé et donc de statuer sur le caractère « non-répondant dans le champ de l'enquête » versus « hors champ, dont cessations » des non-retours.
  Ainsi par exemple, dans l'Enquête sectorielle annuelle (ESA), sont présumées cessées, parmi les non-retours, les unités non affiliées aux régimes d'imposition micro, sans liasse fiscale pendant trois années successives et n'ayant effectué aucune déclaration de TVA l'année d'intérêt de l'enquête. Les autres non-retours de l'enquête [20] sont quant à eux présumés actifs dans le champ de l'enquête et donc considérés comme non-répondants.
  Ce recours à des sources administratives externes aux enquêtes pour caractériser le statut des unités en termes d'activité ou d'inactivité présumée tend à se généraliser, et même à se déporter plus en amont du processus d'enquête, dès la phase d'échantillonnage : ainsi, le calcul dans Sirus d'une probabilité d'existence des unités à partir de sources d'informations externes offrira à l'avenir la possibilité d'exclure de la base de sondage les unités présumées cessées et donc de ne tirer l'échantillon d'une enquête que parmi les unités actives ou présumées actives ;
- enfin, une solution pour affiner la connaissance relative aux non-retours peut être de mener une investigation complémentaire sur tout ou partie de ces derniers afin de collecter l'information relative à leur appartenance au champ de l'enquête. Les résultats de cette enquête complémentaire [21] – beaucoup plus légère que l'enquête principale puisque limitée à la seule question de l'appartenance au champ – permettent alors de caractériser les non-retours, soit directement si l'enquête complémentaire est exhaustive, soit par inférence si cette dernière a été menée sur un sous-échantillon de non-retours. Cette approche s'avère particulièrement

[19] Il s'agit d'unités économiquement cessées à la date d'intérêt de l'enquête – donc hors du champ de celle-ci – mais considérées comme actives au moment du tirage – et donc présentes à tort dans la base de sondage de l'enquête – du fait du décalage existant entre cessation économique et cessation juridique.

[20] À savoir l'ensemble des entreprises affiliées aux régimes d'imposition micro ainsi que les entreprises ayant effectué au moins une déclaration de TVA l'année d'intérêt de l'enquête ou ayant renvoyé au moins une liasse fiscale au cours des trois dernières années.

[21] Usuellement appelée « enquête auprès des non-répondants » par abus de langage.

 ISSN: 2102-6238

adaptée aux enquêtes portant sur un champ spécifique – entreprises appartenant à une filière donnée par exemple – et pour lesquelles nombre de non-retours sont constitués d'entreprises actives mais hors dudit champ.

À l'issue de cette étape de caractérisation des non-retours, l'ensemble des unités de l'échantillon sont réparties entre les trois catégories « répondantes », « non-répondantes » et « hors champ », et on peut alors procéder à la correction de la non-réponse à proprement parler.

La gestion de la collecte et la caractérisation des unités dans l'enquête TIC

La collecte de l'enquête TIC est suivie par une équipe de gestionnaires, dont le rôle est de répondre aux demandes des entreprises, de relancer certaines unités et de contrôler la cohérence et la complétude des réponses. Des contrôles automatiques sur le contenu des questionnaires signalent les incohérences ou anomalies et guident ainsi le travail des gestionnaires, qui peuvent contacter des entreprises afin de vérifier certaines réponses ou pour les inciter à retourner un questionnaire.
Après la collecte, les traitements visent à traiter les problèmes de cohérence, de valeurs aberrantes ou influentes et la non réponse. Les deux derniers points seront traités dans les parties suivantes.
Première étape, la phase d'apurement ("editing" en anglais) consiste à supprimer ou corriger certaines incohérences ou valeurs aberrantes. Ainsi par exemple lorsqu'une entreprise a déclaré posséder au moins un ordinateur et zéro utilisateur d'ordinateurs, cette deuxième réponse est supprimée. La valeur manquante générée ainsi sera traitée ensuite, à l'étape de correction de la non-réponse. Pour les variables quantitatives, l'apurement consiste aussi à repérer et éventuellement supprimer les valeurs aberrantes, c'est-à-dire les valeurs supposées erronées. Un cas courant est l'erreur d'unité ou erreur de mesure. Par exemple, le chiffre d'affaires dans l'enquête TIC est demandé en millier d'euros, mais il arrive que le répondant indique un montant en euros. Ainsi, une entreprise dont le chiffre d'affaires est de 1 million d'euros devrait rapporter 1 000 dans le questionnaire, mais indique 1 million. Donc la valeur collectée est de 1 000 millions, ce qui est faux. Le repérage de telles erreurs se fait par comparaison avec des données disponibles dans la base de sondage (chiffre d'affaires par exemple), de la précèdente édition de l'enquête ou encore par contact avec l'entreprise.
Avant de traiter la non-réponse, la caractérisation des unités est nécessaire, c'est-à-dire leur répartition entre répondantes, non répondantes et hors champ. Une partie des unités hors champ est repérée lors de la collecte, notamment les unités déclarant un effectif inférieur à 10 personnes ou une activité qui aurait changé et ne serait plus dans le champ. Après la collecte, les fichiers Sirus et de déclaration de TVA permettent de repérer des unités cessées, les autres étant ensuite classées par défaut non répondantes du champ.

6. La correction de la non-réponse

6.1. La non-réponse dans les enquêtes statistiques

La non-réponse est un problème « universel » de la statistique d'enquête, au sens où elle concerne aussi bien les enquêtes auprès des entreprises que les enquêtes auprès des ménages, qu'il s'agisse d'enquêtes par sondage aléatoire, d'enquêtes par sondage empirique (comme les enquêtes par quotas) ou même de recensements.

La non-réponse dans les enquêtes peut se définir comme l'incapacité à obtenir des réponses utilisables, pour tout ou partie des questions de l'enquête. On distingue deux grands types de non-réponse : la non-réponse totale et la non-réponse partielle. Il y a non-réponse totale lorsqu'aucune information exploitable n'a pu être recueillie pour une unité échantillonnée. La non-réponse est partielle lorsque l'unité sélectionnée répond seulement à une partie de l'enquête mais pas à l'ensemble des questions [22].

[22] En réalité, la frontière entre non-réponse totale et non-réponse partielle n'est pas aussi facilement identifiable. Ainsi,

 ISSN: 2102-6238

Or la présence de données manquantes résultant de la non-réponse influe sur la qualité de l'inférence. En effet, l'utilisation des formes habituelles d'estimateurs en se restreignant à la population des seuls répondants pose problème du fait que ces derniers présentent en général des caractéristiques différentes de celles non-répondants [23]. Ces estimateurs sont donc biaisés, le biais étant d'autant plus important que les comportements ou les caractéristiques des non-répondants sont différents de ceux des répondants et que le taux de non-réponse est élevé. De plus, la non-réponse induit également une variance totale plus grande et donc une perte de précision des estimations du fait de la réduction de la taille de l'échantillon exploitable.

Afin de pallier ces problèmes, et tout particulièrement celui du biais de non-réponse [24], il est impératif de gérer ces unités non-répondantes via des procédures de correction de la non-réponse. Celles-ci s'appuient sur deux techniques différentes, la repondération et l'imputation :

- les méthodes de repondération – quasi-exclusivement utilisées pour compenser la non-réponse totale – consistent à augmenter, de façon judicieuse, les poids des unités répondantes au sein de l'échantillon pour prendre en compte les unités non-répondantes. Lors de l'exploitation des résultats, les unités non-répondantes sont alors ignorées ;
- les méthodes d'imputation – qui peuvent être employées pour traiter la non-réponse partielle comme pour la non-réponse totale – consistent à remplacer une donnée absente ou invalide dans un fichier d'enquête par une donnée « plausible », obtenue en général à partir des individus répondants. Les unités incomplètes restent alors présentes dans le fichier d'exploitation, avec des données imputées en lieu et place des données manquantes.

Pour appliquer ces méthodes, il est indispensable de disposer de données connues individuellement à la fois sur les unités répondantes et sur les unités non répondantes. En conséquence, les procédures de correction de la non-réponse s'appuient généralement sur des données issues de la base de sondage ou de sources administratives exhaustives, sur des paradonnées ou sur des données historiques.

6.2. La correction de la non-réponse partielle dans les enquêtes auprès des entreprises

Le traitement de la non-réponse partielle dans la sphère entreprises présente peu de spécificités : il s'effectue exclusivement par imputation, selon des procédures relativement classiques. Les principales méthodes d'imputation mises en œuvre sont les suivantes – cf. Haziza (2009) pour plus de détails :

- imputation déductive : la donnée manquante est déduite sans ambiguïté des réponses aux autres questions. Cette méthode est fréquemment utilisée lorsqu'il existe des relations liant

lorsque l'on n'obtient que trop peu de réponses à un questionnaire, il est souvent préférable de considérer cette non-réponse partielle comme une non-réponse totale et de traiter l'unité en conséquence.

[23] Un exemple classique dans la sphère ménages est celui des enquêtes relatives aux revenus ou aux patrimoines, pour lesquelles les individus disposant de revenus ou patrimoines élevés ont une propension à répondre plus faible que les autres. Par conséquent, l'estimateur classique de revenu – ou de patrimoine – moyen par individu obtenu à partir des seuls répondants sous-estime le vrai revenu – ou patrimoine – moyen.

[24] S'il ne s'agissait que du problème de perte de précision liée à la réduction de la taille de l'échantillon exploitable, une simple augmentation de la taille de l'échantillon initial en fonction des taux de non-réponse anticipés suffirait à le résoudre.

 ISSN: 2102-6238

différentes variables du questionnaire entre elles. Par exemple, lorsqu'une variable apparaît dans une équation comptable, la donnée manquante est calculée automatiquement à partir des valeurs des autres variables de l'équation ;

- cold-deck : la donnée manquante est remplacée par une valeur relative à la même unité issue d'une source d'information extérieure à l'enquête. On peut par exemple utiliser l'effectif salarié présent dans le répertoire Sirus pour traiter la non-réponse sur cette variable ;
- imputation par la moyenne par classe : pour une unité appartenant à une classe d'imputation donnée, la valeur manquante est remplacée par la moyenne observée sur les répondants de sa classe d'imputation. Les classes d'imputations sont constituées de manière à regrouper des unités présentant des caractéristiques similaires, l'objectif étant d'obtenir des groupes d'unités homogènes en termes de comportement de non-réponse et / ou de valeurs de la variable d'intérêt. En effet, dans ces conditions, l'impact de la procédure d'imputation, en termes de biais, sur la qualité d'estimations de totaux ou de moyennes sera approximativement nul. Notons qu'il est préférable de constituer des groupes homogènes en termes de valeurs de la variable d'intérêt puisque cette stratégie permet de réduire la variance liée à la non-réponse, ce que n'assure pas la constitution de groupes homogènes en termes de comportement de non-réponse ;
- imputation par le ratio par classe : pour une unité k appartenant à une classe d'imputation donnée, la valeur manquante est remplacée par la valeur d'une variable auxiliaire « actualisée » par un ratio moyen calculé sur les répondants de sa classe d'imputation :

$$y_k^* = \frac{\overline{y}_r}{\overline{x}_r} x_k$$

Cette méthode, qui mobilise de l'information relative à l'unité imputée via la variable auxiliaire X, fournit en général une imputation de meilleure qualité que l'imputation par la simple moyenne, dès lors que les variables Y et X sont suffisamment liées. Un cas particulier de cette méthode, fréquemment utilisé dans les enquêtes répétées dans le temps, est l'imputation par la valeur précédente actualisée :

$$y_k^* = \frac{\overline{y}_{r,t}}{\overline{y}_{r,t-1}} y_{k,t-1}$$

- imputation par la tendance unitaire : pour les enquêtes répétées dans le temps, la valeur manquante est remplacée par la valeur déclarée lors de l'enquête précédente et actualisée selon la tendance de l'unité observée sur une variable auxiliaire X positivement corrélée à Y :

$$y_k^* = \frac{x_{k,t}}{x_{k,t-1}} y_{k,t-1}$$

Cette méthode, qui ne fait intervenir que des données relatives à l'unité imputée, est celle qui donne *a priori* les meilleurs résultats ;

- hot-deck aléatoire par classe : pour une unité appartenant à une classe d'imputation donnée, la valeur manquante est remplacée par une valeur choisie au hasard parmi les répondants de sa classe d'imputation. Cette méthode permet de préserver la distribution des variables, au prix d'une plus grande variabilité des estimations.

 ISSN: 2102-6238

En pratique, on opte en priorité pour les méthodes d'imputations mobilisant un maximum d'informations relatives à l'unité imputée : imputation déductive, cold-deck, imputation par la tendance unitaire [25]. À défaut, le choix entre les autres méthodes d'imputation s'effectue en fonction du type d'analyses que l'on désire faire avec les données de l'enquête : lorsque les statistiques d'intérêt principales sont des totaux ou des moyennes, on privilégie les méthodes d'imputation déterministes – imputation par la moyenne, par le ratio – qui introduisent peu de variance dans les estimations. En revanche, elles présentent l'inconvénient de déformer les distributions des variables imputées, et sont donc inappropriées dès lors que l'on s'intéresse à des paramètres de distribution tels les quantiles. Dans ce cas, il est préférable de procéder à une imputation selon une approche stochastique, tel le hot-deck. Ces procédures d'imputation aléatoire présentent en outre l'avantage d'être utilisables aussi bien pour des variables quantitatives que qualitatives, alors que les méthodes déterministes ne sont valides que pour des données quantitatives.

La correction de la non-réponse partielle dans l'enquête TIC

Dans l'enquête TIC, plusieurs types de corrections de la non-réponse partielle sont appliqués selon le type de variable et les disponibilités d'informations externes.
L'imputation déductive, basée sur des règles déterministes, est utilisée lorsque la réponse à une question peut se déduire d'une autre. Par exemple, si une entreprise n'a pas répondu à la question suivante : « en 2012, votre entreprise a-t-elle passé des commandes de biens ou services via un site web ou un message de type EDI ? » et qu'elle a répondu par ailleurs n'avoir ni accès internet ni pratique d'échanges de type EDI, et qu'elle n'a indiqué que des valeurs nulles ou aucune valeur à la question sur les montants de ces achats, alors la réponse « NON » est imputée à la variable correspondante (l'entreprise ne fait pas d'achats par voie électronique) et la variable de montant des achats électroniques est mise à blanc.
Certaines variables quantitatives, comme le chiffre d'affaires, sont présentes dans d'autres sources (Sirus, liasses fiscales, fichier TVA, édition précèdente de l'enquête TIC). Un ordre de priorité est donné à chacune de ces sources selon sa fiabilité supposée et la recherche est faite dans cet ordre afin de trouver une valeur qui servira à l'imputation de la valeur manquante de l'unité (méthode du cold-deck). Si aucune de ces sources ne permet d'imputer une valeur, on affecte à l'unité concernée un chiffre d'affaires calculé au prorata de son effectif, en utilisant le chiffre d'affaires moyen par effectif dans sa strate de diffusion [a] (imputation par le ratio par classe).
Les variables qualitatives sont redressées selon la méthode du donneur (hot-deck) ce qui permet notamment de corriger plusieurs variables, qui seraient liées entre elles, à l'aide d'un donneur unique. Ainsi, pour chaque variable non renseignée, la réponse est récupérée chez une unité donneuse qui appartient à la même classe d'imputation (i.e. qui a les mêmes caractéristiques, ou qui vérifie les mêmes modalités pour les variables auxiliaires retenues). Les caractéristiques ou variables auxiliaires prises en compte peuvent être des caractéristiques de l'entreprises – connues avant l'enquête – ou des réponses à d'autres questions du questionnaire, ce qui permet de conserver un lien statistique entre certaines variables.

[a] Sur les unités n'ayant pas de valeur manquante, on calcule, par strate, le ratio de la somme des chiffres d'affaires sur la somme des effectifs par strate. Puis on utilise ce ratio pour affecter à chaque unité devant être imputée un chiffre d'affaires au prorata de ses effectifs salariés et selon sa strate d'appartenance.

[25] La statistique d'entreprise fournit en particulier un cadre d'application privilégié à cette dernière méthode : en effet, de nombreuses enquêtes auprès des entreprises sont des enquêtes répétées dans le temps, dont les échantillons sont en général sélectionnés via un schéma rotatif permettant d'assurer un recouvrement contrôlé entre les échantillons relatifs à deux millésimes successifs d'une même enquête.

 ISSN: 2102-6238

6.3. La correction de la non-réponse totale dans les enquêtes auprès des entreprises

La correction de la non-réponse totale dans les enquêtes auprès des entreprises s'effectue en général par repondération – cf. Caron (2005) pour plus de détails. L'idée de l'approche par repondération consiste à assimiler la non-réponse à une seconde phase de tirage – tirage du sous-échantillon de répondants R au sein de l'échantillon initial S – malheureusement non contrôlée, effectuée selon des probabilités de tirage p_k inconnues et correspondant aux probabilités de réponse des unités. La figure 5 résume la situation.

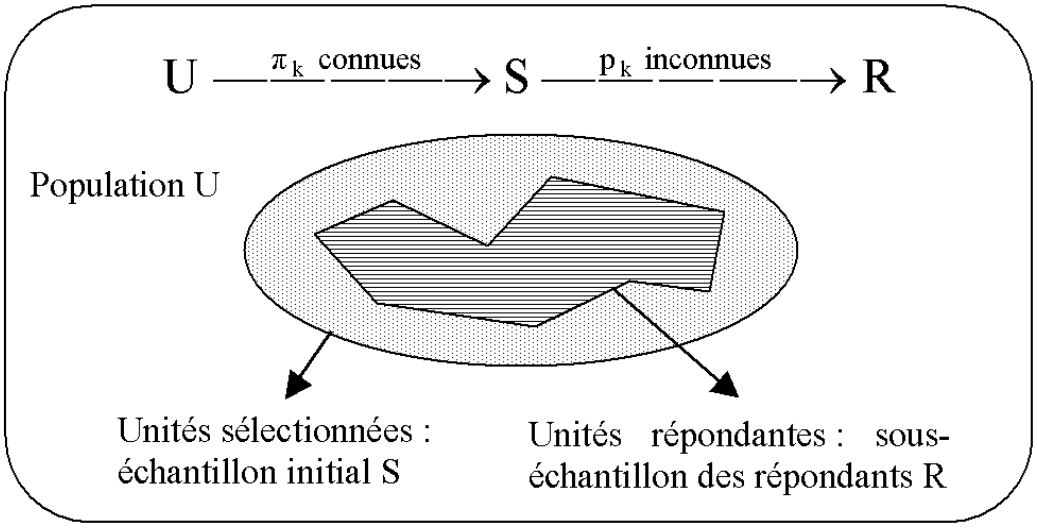

FIGURE 5. *La non-réponse totale vue comme une seconde phase de tirage*

Si les probabilités p_k étaient connues, on pourrait alors estimer sans biais le total de n'importe quelle variable Y à l'aide de l'estimateur en expansion [26] suivant [27] :

$$\hat{Y}_R = \sum_{k \in R} \frac{1}{\pi_k \, p_k} \, y_k \tag{1}$$

Malheureusement, les probabilités de réponses p_k ne sont pas connues, et toute la difficulté consiste à les estimer. Pour ce faire, on procède à une modélisation du mécanisme de non-réponse, en analysant le critère « répond / ne répond pas » en fonction de variables connues pour les répondants et les non-répondants. Cette analyse se fait le plus souvent au moyen d'une régression logistique appliquée sur l'échantillon pour déterminer les variables auxiliaires les plus explicatives de la non-réponse. Une fois cette modélisation effectuée, plusieurs stratégies sont envisageables pour estimer les probabilités de réponses p_k :

[26] Estimateur classique dans le cadre des sondages en plusieurs phases.

[27] Notons au passage que cet estimateur conduit bien à une augmentation du poids des unités répondantes, comme annoncé en 6.1

 ISSN: 2102-6238

- il est tout d'abord possible d'utiliser directement les probabilités de réponses issues de la régression logistique. Cependant, cette méthode est rarement utilisée en pratique. En effet, elle présente l'inconvénient majeur de ne garantir aucun contrôle sur les valeurs des $\hat{p}_k$ estimés, qui peuvent ainsi être très dispersés. En particulier, si la probabilité de réponse estimée d'un répondant est très faible, la pondération de ce répondant devient alors par construction très importante, ce qui conduit à un estimateur relativement instable. En conséquence, on lui préfère quasi-systématiquement la méthode suivante ;
- l'alternative consiste à partitionner, à partir des résultats de la régression logistique, l'échantillon [28] en sous-populations supposées homogènes en termes de comportement de réponse, appelées Groupes de Réponse Homogènes (GRH). Pour ce faire, plusieurs approches sont possibles :
  - la « méthode des scores » consiste à classer les individus selon leur probabilité de réponse estimée via la régression logistique et à constituer les groupes de réponse homogènes en divisant l'échantillon en groupes de tailles suffisantes et approximativement égales. Par construction, la première sous-population rassemble les unités pour lesquelles les probabilités de réponse estimées sont les plus faibles et ainsi de suite ;
  - la « méthode par croisement » consiste à définir comme sous-populations celles qui correspondent aux croisements de toutes les modalités des variables explicatives qui interviennent dans la régression logistique [29]. Afin d'assurer une certaine stabilité des estimations de probabilités de réponse, on peut être amené à effectuer des regroupements de modalités en vue d'obtenir des GRH finaux ayant une taille suffisamment importante ;
  - enfin, la « méthode de segmentation par arbre » s'appuie sur des procédures de segmentation de données, tel l'algorithme CHAID. Partant de l'ensemble de l'échantillon, la méthode détermine à chaque étape la variable et les modalités qui séparent le mieux l'ensemble des unités, relativement au fait de répondre ou non. La procédure est itérative, et permet de construire un « arbre » dont les « feuilles » (i.e. les nœuds terminaux) sont les différents GRH.

  Une fois ces GRH constitués, le mécanisme de réponse est supposé indépendant d'un GRH à l'autre et uniforme au sein de chaque GRH. La probabilité de réponse au sein d'un GRH g donné est alors estimée par le taux de réponse pondéré observé dans ce GRH :

$$\hat{p}_g = \frac{\sum_{k \in R} d_k \ \mathrm{1I}_{k \in GRH_g}}{\sum_{k \in R \oplus NR} d_k \ \mathrm{1I}_{k \in GRH_g}}$$

  avec R l'échantillon des répondants à l'enquête, NR celui des non-répondants, d_k les poids de sondage initiaux et $\mathrm{1I}_{k \in GRH_g}$ l'indicatrice d'appartenance de l'unité k au GRH g.
  Cette méthode d'estimation des probabilités de réponses au sein de groupes homogènes permet de mieux contrôler la dispersion des $\hat{p}_k$ estimés, et apporte une certaine robustesse aux estimations même si le modèle est mal spécifié.

[28] Privé des unités hors champ, qui n'interviennent pas dans l'étape de correction de la non-réponse.
[29] Si certaines variables de la régression logistique sont continues, alors on les catégorise avant de constituer les GRH.

 ISSN: 2102-6238

À l'issue de cette étape, on dispose donc de valeurs estimées $\hat{p}_k$ des probabilités de réponses, valeurs qui seront utilisées en lieu et place des vraies valeurs inconnues dans l'estimateur en expansion (1). La validité de l'inférence, c'est-à-dire l'extrapolation des résultats obtenus sur l'échantillon des seuls répondants à l'ensemble de la population à l'aide des poids corrigés de la non-réponse, dépend alors de la validité des hypothèses retenues dans la modélisation : si le modèle est bien spécifié, et en particulier s'il prend bien en compte l'ensemble des variables explicatives de la non-réponse, alors les $\hat{p}_k$ estimés seront très proches des vraies probabilités de réponses p_k, et l'estimateur retenu sera approximativement sans biais.

Enfin, rappelons que les enquêtes entreprises se caractérisent, entre autre, par l'existence quasi-systématique de strates exhaustives. Dans le cas d'enquêtes répétées dans le temps, cette particularité peut influer sur le choix de la méthode de correction de la non-réponse totale et amener à mettre en place deux procédures de correction de la non-réponse totale distinctes pour l'exhaustif et la partie échantillonnée de l'enquête :

- pour l'exhaustif, l'hétérogénéité des unités et l'existence de données historiques peut conduire à privilégier l'utilisation des techniques d'imputation s'appuyant sur les réponses des unités à l'enquête précédente, au premier rang desquelles l'imputation par la tendance unitaire exposée au paragraphe précédent. L'expérience montre en effet que ces redressements temporels sont de bien meilleure qualité que les autres, car ils s'appuient sur des informations provenant de l'unité elle-même ;
- pour la partie échantillonnée en revanche, on procède le plus souvent à une correction de la non-réponse totale par repondération, selon les techniques exposées ci-dessus.

La correction de la non-réponse totale dans l'enquête TIC

Dans l'enquête TIC, certaines unités particulières sont repérées dans l'échantillon. Ces unités, appelées non substituables, sont considérées comme atypiques *a priori* et font l'objet d'une procédure de correction de la non-réponse spécifique afin de ne pas « propager » leur comportement au reste de la population.
La définition d'unité légale non substituable s'appuie sur deux critères : l'importance de l'entreprise dans sa strate, et son importance au regard du commerce électronique. Ainsi, une unité légale est non-substituable si les conditions suivantes sont réunies :

- elle fait partie de la strate exhaustive (poids de sondage égal à 1) ;
- au moins une des conditions suivantes est vérifiée :

  a) chiffre d'affaires dans la base de sondage supérieur à 4 milliards d'euros ou représentant plus de 10 % du total non pondéré des chiffres d'affaires de sa strate de diffusion ;

  b) effectif dans la base de sondage supérieur à 10 000 ou représentant plus de 10 % du total non pondéré des effectifs de sa strate de diffusion ;

  c) achats totaux dans la base de sondage supérieurs à 5 milliards d'euros ou représentant plus de 15 % du total non pondéré des achats totaux de sa strate de diffusion ;

  d) ventes via un site web estimées avant la collecte supérieures à 150 millions d'euros ;

  e) ventes électroniques (c'est-à-dire via un site web ou d'autres réseaux comme EDI, etc.) estimées avant la collecte supérieures à 600 millions d'euros ;

  f) achats électroniques estimés avant la collecte supérieurs à 600 millions d'euros.

Ces conditions amènent à considérer entre 100 et 200 unités comme non-substituables. Pour ces unités, la

 ISSN: 2102-6238

correction de la non-réponse totale s'effectue dans la mesure du possible par imputation à partir des données historiques : les informations de l'enquête TIC de l'édition précédente sont reprises à l'identique, sans même appliquer de coefficient d'évolution entre les données disponibles sur les deux années [a]. En l'absence de données historiques (non-réponse cumulée à l'absence de données l'année précédente), ces unités spécifiques sont soumises aux mêmes traitements que les unités substituables.
Pour les unités substituables, la correction de la non-réponse totale consiste en une repondération au sein de groupes de réponse homogènes (GRH). Ces groupes sont constitués à l'aide des variables repérées comme explicatives de la non réponse : la tranche de taille, la tranche de chiffre d'affaires, le secteur d'activité, la catégorie juridique, la localisation géographique, etc. Au sein de chaque groupe, les poids des unités sont divisés par le taux de réponse, pondéré par les poids de sondage, obervé dans le groupe. Jusqu'à l'édition 2012, les GRH étaient constitués par croisements. En 2013, la méthode de segmentation par arbre a été privilégiée.

[a] En effet, rien ne permet de préjuger que l'évolution des non-substituables (unités atypiques par définition) suit la moyenne des autres unités.

7. La gestion des unités atypiques

Comme nous l'avons vu précédemment, les échantillons des enquêtes entreprises sont quasi-systématiquement constitués selon un plan de sondage stratifié, les strates étant en général définies au minimum par le croisement des modalités de deux variables issues du répertoire Sirus : une variable de classement sectoriel et une variable de taille (il s'agit le plus souvent de l'effectif des unités).

Dès lors, un mauvais classement sectoriel au lancement de l'enquête ou l'évolution à la hausse en termes d'effectifs d'une unité sur l'année en cours (c'est-à-dire entre le lancement et la fin de l'enquête) sont autant de facteurs générant, dans les strates non exhaustives, des points qui s'avèrent à la fois atypiques et non aberrants : atypiques au sens où ils sont situés dans la queue de distribution de la strate à laquelle ils appartiennent, non aberrants car leur valeur est certifiée et ne résulte pas d'une erreur de mesure. Ces points atypiques non aberrants sont très problématiques dans la mesure où ils engendrent une forte variance dans les estimations et contribuent donc à augmenter fortement l'instabilité des résultats.

Il est donc nécessaire de gérer ces unités, en modifiant certaines de leurs caractéristiques. Cependant, cette opération doit être réalisée avec parcimonie et circonspection : en effet, les caractéristiques d'un point atypique non aberrant n'étant pas entachées d'erreurs de mesure, non seulement leur réévaluation n'est pas naturelle, mais elle introduit en outre un biais dans les estimations. Il s'agit donc de déterminer une procédure permettant de réaliser un bon compromis entre deux considérations contradictoires : d'une part détecter et gérer les points atypiques non aberrants de façon à réduire au maximum la variance engendrée par ces unités, et d'autre part minimiser le biais induit par le traitement de ces unités.

Une façon simple de traiter ces unités atypiques non aberrantes consiste à ramener leur poids de sondage à 1 afin de limiter leur influence dans les agrégats. Cette approche, relativement intuitive [30] et utilisée par le passé avec parcimonie en statistique d'entreprises, est cependant purement empirique, avec en particulier un choix arbitraire des unités considérées comme atypiques.

C'est pourquoi l'Insee privilégie depuis quelques années l'utilisation de techniques dites de

[30] Le fait de réduire le poids de sondage de telles unités à 1 peut en effet se justifier de la façon suivante : ces unités étant atypiques, elles ne sont pas « représentatives » des autres unités de leur strate de tirage ; au contraire, leurs caractéristiques ne valent que pour elles-mêmes, et il est donc légitime de leur attribuer un poids unitaire.

 ISSN: 2102-6238

winsorisation pour gérer ces unités atypiques non aberrantes – cf. Guggemos and Brion (2011). Le principe de la winsorisation est relativement simple : au sein d'une strate h donnée, sont considérées comme atypiques, au regard d'une variable Y donnée, les unités dont la valeur y_k dépasse un seuil K_h prédéfini. Pour une unité de poids w_k donné, winsoriser consiste alors à remplacer la variable initiale y_k par une nouvelle valeur y_k^* définie de la manière suivante :

$$y_k^* = \begin{cases} \dfrac{1}{w_k} y_k + \left(1 - \dfrac{1}{w_k}\right) K_h & \text{si } y_k \geq K_h, \text{ pour une unité } k \text{ appartenant à la strate } h \\ y_k & \text{sinon} \end{cases}$$

Il s'agit bel et bien d'une procédure de traitement des points atypiques, puisque la variable n'est modifiée que pour les entreprises pour lesquelles sa valeur est supérieure au seuil K_h. Pour ces unités, la variable winsorisée y_k^* constitue une révision à la baisse de la variable initiale y_k, réduisant ainsi la variance des données dans la strate h et, par suite, celle des statistiques produites.

En agissant de la sorte, on s'assure en outre du respect des objectifs fixés. En effet, l'impact de la procédure de winsorisation est d'autant plus important que la variable est atypique, i.e. que sa valeur initiale y_k s'éloigne du seuil K_h. Par ailleurs, plus le poids de sondage de l'unité est faible, plus l'aléa d'échantillonnage – donc l'instabilité des estimations – est faible, et moins le traitement des points atypiques non aberrants se justifie. La forme mathématique retenue pour y_k^* abonde en ce sens, puisque la correction apportée par la winsorisation est d'autant moins importante que le poids de sondage de l'unité se rapproche de 1.

Ainsi au final, l'ampleur de la correction apportée par la winsorisation est d'autant plus importante qu'elle porte sur des unités pour lesquelles elle se justifie, soit qu'il s'agisse d'unités particulièrement atypiques, soit qu'il s'agisse d'unités appartenant à des strates présentant des taux de sondage faibles et donc un aléa d'échantillonnage fort.

La recherche du compromis optimal entre l'objectif de réduction de la variance d'estimation et celui de limitation du biais repose alors sur le choix des valeurs à utiliser pour les seuils K_h au-delà desquels les entreprises sont considérées comme atypiques. La détermination de ces seuils est effectuée en suivant une procédure proposée par Kokic and Bell (1994) et qui consiste à choisir les valeurs des seuils qui minimisent l'erreur quadratique moyenne – somme de la variance et du carré du biais – des estimations relatives à la variable Y.

La démarche décrite jusqu'à présent permet de détecter et traiter les entreprises atypiques pour n'importe quelle variable quantitative d'une enquête. En théorie, on peut donc envisager d'appliquer cette procédure de winsorisation pour toutes les variables quantitatives d'une enquête. Ceci nécessite cependant de déterminer autant de jeux de seuils K_h que de variables à traiter, et présente en outre l'inconvénient de détruire la cohérence entre les traitements des différentes variables, puisqu'une entreprise peut alors être considérée comme atypique pour certaines variables, mais pas pour d'autres. Aussi distingue-t-on en pratique deux cas de figure, selon la nature de l'enquête considérée :

- soit cette enquête comporte essentiellement des variables qualitatives et peu de variables quantitatives, auquel cas on peut effectivement procéder à une gestion des unités atypiques variable par variable – c'est notamment le cas dans l'enquête TIC, cf. encadré ci-dessous ;

 ISSN: 2102-6238

- soit cette enquête comporte de nombreuses variables quantitatives. Dans ce cas, on préfère en général ne juger du caractère atypique des unités qu'au regard de la principale variable d'intérêt de l'enquête – noté Y dans la suite –, en lui appliquant la procédure de winsorisation décrite ci-dessus. L'impact de la winsorisation est ensuite « transféré » sur les poids de sondage, selon la formule suivante :

$$w_k^* = w_k \frac{y_k^*}{y_k} = \begin{cases} 1 + \dfrac{K_h(w_k - 1)}{y_k} & \text{si } y_k \geq K_h, \text{ pour une unité } k \text{ appartenant à la strate } h \\ w_k & \text{sinon} \end{cases}$$

L'utilisation de ces poids winsorisés dans les estimations permet alors de réduire « globalement » l'influence des unités atypiques au regard de la principale variable d'intérêt de l'enquête. La pertinence de cette approche repose sur l'hypothèse – plus ou moins bien vérifiée en pratique selon l'enquête considérée – d'une forte corrélation entre la principale variable d'intérêt et les autres variables quantitatives de l'enquête.

La gestion des unités atypiques dans l'enquête TIC

Une fois les opérations de correction de la non-réponse et de calage terminées, le fichier de données est expertisé, et les cas de valeurs aberrantes ou influentes sont repérés par l'observation de listes d'unités les plus contributrices aux agrégats correspondants.
Les valeurs aberrantes résiduelles, c'est-à-dire les erreurs, sont traitées « à la main » par modification des valeurs en utilisant des sources externes ou à dire d'expert.
Les valeurs influentes, qui contribuent fortement à l'estimateur et sont considérées comme parties intégrante de la population (i.e. ni erreur d'unité, ni déclaration erronée, ni erreur de saisie, etc.) font l'objet d'une winsorisation.
L'enquête TIC comprenant des variables de natures diverses (qualitatives sur les équipements TIC, quantitatives sur le commerce électronique par exemple) et principalement qualitatives, le choix a été fait de modifier les valeurs et non les poids des unités. Les variables concernées par la winsorisation sont les variables quantitatives sur le commerce électronique (ventes web ou via EDI, achats électroniques), ainsi que le chiffre d'affaires et les achats totaux de l'entreprise.
En pratique, les unités ayant un poids après calage supérieur à 1 et qui, pondérées, contribuent à plus de 1 % du total d'un estimateur sont examinées afin de déterminer la vraisemblance de la valeur déclarée. Pour chaque variable traitée, un seuil K correspondant à la valeur maximale jugée plausible est déterminé à dire d'expert. Si la valeur observée y_k pour une unité k dépasse ce seuil, alors elle est remplacée par la valeur winsorisée suivante :

$$y_k^* = \frac{1}{w_k} y_k + \left(1 - \frac{1}{w_k}\right) K$$

où w_k est le poids final de l'unité k, après repondérations (traitement de la non réponse et calage). La valeur winsorisée pondérée devient alors :

$$w_k y_k^* = y_k + (w_k - 1)K$$

Ainsi la valeur déclarée est conservée pour un poids de 1, i.e. représentant le seul individu l'ayant déclarée. La valeur seuil K est comptée pour $(w_k - 1)$ individus.

8. Le calage

Lorsqu'on réalise une enquête auprès des ménages ou des entreprises, il est rare de ne disposer d'aucune information extérieure à l'enquête sur les unités appartenant au champ de celle-ci. Cette information, appelée information auxiliaire, peut soit être connue à un niveau global, sous forme

 ISSN: 2102-6238

d'un total par exemple, soit être disponible au niveau individuel pour tous les individus de la base de sondage. Prendre en compte cette information auxiliaire permet, outre d'assurer la cohérence entre différentes sources, d'améliorer la précision des statistiques produites, dès lors que les variables auxiliaires considérées sont liées aux sujets d'intérêt de l'enquête. Pour ce faire, on utilise des procédures de calage [31] – cf. Deville and Särndal (1992) –, qui consistent à modifier les poids des unités de l'échantillon de façon à obtenir des poids après calage permettant d'estimer parfaitement les totaux des variables auxiliaires tout en s'éloignant le moins possible des poids initiaux.

Plus précisément, si l'on dispose de J variables auxiliaires [32] $X_1, \ldots, X_J$ disponibles au niveau individuel sur l'échantillon et dont on connaît les totaux sur la population T_{X_j}, le calage consiste à déterminer des poids après calage w_k qui :

- minimisent la distance avec les poids avant calage d_k ;
- vérifient les équations de calage suivantes : $\forall j = 1 \ldots J, \ \sum_{k \in S} w_k x_{jk} = T_{X_j}$.

Le choix de la fonction de distance servant à mesurer la proximité entre les poids avant calage d_k et les poids après calage w_k constitue un des paramètres principaux pour gérer la qualité de la procédure de calage. En particulier, certaines fonctions de distance – les plus utilisées en pratique à l'Insee – permettent de contrôler les déformations de poids maximales induites par le calage, et donc d'assurer notamment que la procédure de calage ne conduit pas à des poids finaux trop élevés préjudiciables à la robustesse des estimations.

Contrairement à la sphère ménage, où les informations auxiliaires utilisées pour les calages ne sont connues qu'au niveau global, au travers de totaux estimés à partir d'enquêtes importantes telles l'enquête emploi ou le recensement rénové de la population, les calages réalisés sur les enquêtes entreprises s'appuient sur des informations auxiliaires de niveau individuel : variables de la base de sondage ou données administratives disponibles pour toutes les unités de celle-ci.

Cette spécificité n'est pas sans conséquence sur la mise en œuvre du calage : elle implique en effet de faire participer au calage certaines unités « hors champ », à savoir les unités ayant été détectées comme hors champ par le seul biais de l'enquête. En effet, pour pouvoir exclure ces unités hors champ de l'échantillon de calage, il faudrait être capable d'exclure de la base de sondage les unités hors champ qu'elles représentent – de façon à assurer la cohérence entre le champ couvert par l'échantillon et le champ de la base de sondage –, ce qui est impossible puisque celles-ci ne sont pas identifiées. En conséquence, ces unités détectées comme hors champ suite à l'enquête sont prises en compte dans le calage – en étant conservées à la fois dans l'échantillon sur lequel le calage s'opère et dans le champ de calage servant au calcul des marges – et voient donc leur poids modifié par cette opération. En revanche, elles sont par la suite exclues comme il se doit lors de l'exploitation des résultats de l'enquête.

En pratique à l'Insee, la plupart des échantillons des enquêtes entreprises sont calés sur le nombre d'unités de la base de sondage [33], *a minima* au niveau global, et de plus en plus fréquemment ventilé par secteur d'activité et / ou tranche de taille des unités. Pour certaines

[31] Via la macro SAS CALMAR – pour CALage sur MARges – développée par l'Insee pour mettre en œuvre ces procédures.
[32] Il s'agit de variables quantitatives ou d'indicatrices associées aux modalités de variables catégorielles.
[33] Éventuellement privée des unités détectées comme hors champ par une source externe à l'enquête.

 ISSN: 2102-6238

enquêtes, un calage peut également être effectué sur des variables administratives, telles que l'effectif salarié des unités, le chiffre d'affaires ou encore la catégorie juridique. Ainsi par exemple, dans l'ESA (Enquête Sectorielle Annuelle), le calage est réalisé sur les variables « chiffre d'affaires par secteur » et « nombre d'entreprises par secteur [34] ». Plus précisément, les poids w_k sont ajustés de telle façon que, pour chaque groupe tel que défini par le répertoire, le chiffre d'affaires et le nombre d'entreprises de ce groupe soient retrouvés par l'échantillon d'enquête extrapolé.

Le calage dans l'enquête TIC

Dans l'enquête TIC, on procède à un calage sur le nombre d'unités légales par strate de tirage, éventuellement après regroupements. Les marges du calage correspondent ainsi à des croisements [secteur d'activité $\otimes$ taille], ou au fait d'avoir un chiffre d'affaires qui dépasse le seuil d'exhaustivité (dépendant de la taille de l'entreprise). Pour TIC 2012, ce calage a pu être réalisé en contenant la dispersion des rapports de poids (poids après calage / poids avant calage) entre les bornes 0,7 et 1,6.

9. La diffusion et la gestion de la confidentialité

La diffusion des résultats constitue la dernière étape du processus de production d'une enquête. Si la tabulation et la mise en forme des résultats ne pose en général aucun problème particulier, une fois les étapes précédentes de contrôles-redressements effectuées, se pose alors le problème de la gestion de la confidentialité dans les données diffusées, et tout particulièrement dans les tableaux diffusant des statistiques sur les entreprises – cf. Nicolas (2012) pour plus de détails sur le sujet.

En effet, le système statistique public a l'obligation morale et légale de garantir la confidentialité des informations qui lui ont été confiées par les répondants aux enquêtes. Cet impératif de confidentialité, vitale pour obtenir une bonne coopération des répondants et ainsi collecter un maximum de données de la meilleure qualité possible, se traduit à l'étape de diffusion par une obligation légale [35] de contrôler la divulgation statistique dans les informations mises à disposition, en minimisant le risque que des informations sensibles sur des individus ou des entreprises puissent être divulguées à partir des données diffusées.

La jurisprudence française comprend deux règles en vigueur, applicables à la diffusion des données de statistiques d'entreprises :

- règle des trois unités : une cellule d'un tableau ne doit pas être construite avec moins de trois unités. En effet, dans le cas contraire, des informations individuelles peuvent être déduites des agrégats diffusés : par tous les utilisateurs et de manière immédiate si une seule unité constitue l'agrégat ; par les unités impliquées dans un agrégat construit à partir de seulement deux unités, par déduction en fonction de l'agrégat diffusé et de leurs propres données ;
- règle des 85 % : aucune des unités contributrices à une cellule ne doit contribuer à plus de 85 % du total de celle-ci. Dans le cas contraire, les statistiques diffusées donnent une estimation trop précise des informations relatives à l'unité dominante.

Afin de respecter ces deux règles, il est donc nécessaire de masquer certaines cellules des

[34] Le niveau sectoriel retenu pour ce calage est le niveau « groupe », correspondant aux trois premiers caractères du code d'activité principale des entreprises en NAF Rév.2.

[35] Par application de la loi n° 51-711 du 7 juin 1951 modifiée relative à l'obligation, la coordination et le secret en matière de statistiques.

 ISSN: 2102-6238

tableaux diffusés. Pour ce faire, l'Insee, à l'instar des autres instituts nationaux de statistiques européens, utilise une méthode dite de « suppressions des cellules ».

Cette méthode consiste dans un premier temps à rechercher les cellules qui ne respectent pas les règles de confidentialité, puis à les supprimer : ce sont les suppressions dites primaires, correspondant au traitement du secret dit, lui aussi, primaire. Cependant, ces suppressions primaires s'avèrent nécessaires mais non suffisantes à la gestion du secret. En effet, les tableaux de données diffusées comportent en général des totaux en marges, et peuvent également être liés entre eux par certaines relations – cas des tableaux bidimensionnels liés par une des deux variables de ventilations qui ventilent une même quantité par exemple. Afin d'empêcher la reconstitution des suppressions primaires au moyen des marges ou des liens entre tableaux, il est donc nécessaire de supprimer des cellules supplémentaires : ce sont les suppressions dites secondaires, correspondant à la gestion du secret également dit secondaire. Au final, les suppressions primaires et secondaires sont masquées dans les tableaux diffusées en étant remplacées par un même symbole, ce qui renforce le secret posé.

Si la gestion du secret primaire ne pose pas de difficulté particulière au vu des deux règles en vigueur – il ne s'agit en fait que de calculer des fréquences et des contributions –, le traitement du secret secondaire est autrement plus ardu : en effet, la multiplicité des tableaux diffusés, le plus souvent liés entre eux, rend très complexe la détermination d'une structure de suppressions secondaires valide et qui ne masque pas trop d'information. Aussi la gestion du secret est-elle réalisée à l'aide du logiciel spécialisé τ-argus qui, couplé à un optimiseur, permet d'assurer un respect strict des règles de confidentialité tout en minimisant la perte d'information induite dans les tableaux diffusés.

Conclusion

Les méthodes décrites dans cet article sont celles mises en pratique dans les enquêtes entreprises à l'Insee fin 2013. Elles seront amenées à évoluer dans les prochaines années avec, entre autres, la montée en puissance de la procédure de coordination généralisée d'échantillons – permettant de répartir au mieux la charge statistique entre les unités et de mieux maitriser le recouvrement de l'ensemble des échantillons d'entreprises –, le développement de la collecte par internet – facilitant les possibilités d'interaction avec le répondant, via des contrôles à la saisie par exemple, qui permettent d'espérer une amélioration de la qualité des données collectées –, ou encore l'étude de méthodes de gestion de la confidentialité à un niveau fin, en vue de la mise à disposition de fichiers de données détaillées.

D'une façon plus générale, l'internationalisation de l'économie et la demande croissante de données chiffrées nécessitent des évolutions importantes de la statistique d'entreprise, au premier rang desquelles figure l'évolution du concept d'entreprise en tant qu'unité statistique. En effet, afin d'être au plus près de la réalité économique, l'unité statistique « entreprise », assimilée à l'unité légale jusque récemment, évolue pour se rapprocher de la définition économique de l'entreprise [36] et prendre en compte la dimension de groupe d'entreprises [37]. Le « profilage » des groupes au

[36] Plus petite unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision.

[37] Pour plus de détails sur l'impact de ce changement de concept sur les statistiques d'entreprises, on se réfèrera à Beguin et al. (2012).

 ISSN: 2102-6238

niveau national a commencé à l'Insee en 2010 par le profilage individuel et manuel des plus grands groupes. Il doit se poursuivre par l'automatisation de la procédure pour les groupes moins importants. Parallèlement, des travaux sont engagés au niveau européen afin de définir et prendre en compte les entreprises et groupes multinationaux.

Ces travaux sont nécessaires mais complexes. Ils doivent de plus être compatibles avec la demande croissante de données au niveau européen comme au niveau local et s'inscrivent dans un contexte difficile de réduction des budgets alloués à la Statistique publique française.

Références

Beguin, J., Hecquet, V., and Lemasson, J. (2012). Un tissu productif plus concentré qu'il ne semblait. *Insee Première*, 1399.

Caron, N. (2005). La correction de la non-réponse par repondération et par imputation. *Document de travail n°M0502 de la série "Méthodologie Statistique", Insee*.

Cotton, F. and Hesse, C. (1992). Tirages coordonnés d'échantillons. *Document de travail INSEE de la Direction des Statistiques Economiques*, 9206.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418) :376–382.

Guggemos, F. and Brion, P. (2011). Winsorisation sur les enquêtes annuelles auprès des entreprises françaises. *Pratiques et méthodes de sondage, sous la direction de M.E. Tremblay, P. Lavallée et M. El Haj Tirarni*, pages 195–200.

Guggemos, F. and Sautory, O. (2012). Sampling coordination of business surveys conducted by insee. *Proceedings of the Fourth International Conference of Establishment Surveys*.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics. Sample Surveys : Design, Methods and Applications*, 29 :215–246.

Kokic, P. and Bell, P. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of official statistics - Stockholm* -, 10 :419–435.

Koubi, M. and Mathern, S. (2009). Résolution d'une des limites de l'allocation de neyman. *Actes des X^{èmes} Journées de Méthodologie Statistique*, 2009 :1.

Nicolas, J. (2012). Traitement de la confidentialité statistique dans les tableaux : expérience de la direction des statistiques d'entreprises. *Actes des XI^{èmes} Journées de Méthodologie Statistique*.

 ISSN: 2102-6238