Esane : À la recherche d'une cohérence maximale des données multi-sources sur les entreprises par le biais de micro et macro contrôles

Auteur : Olivier Haag (Insee : Département «Répertoires, Infrastructure et statistiques structurelles »)

Email : olivier.haag@insee.fr

En 2006, un grand projet de Refonte de l'Élaboration des Statistiques ANuelles d'Entreprises (le programme Resane [1]) a été mis en œuvre à l'Insee. Ces statistiques se fondent sur la prise en compte de plusieurs sources d'informations ;

- des sources administratives (déclarations fiscales et sociales)
- une source statistique : l'Enquête Sectorielle Annuelle (ESA)[1] qui se décline par des questionnaires personnalisés selon les différents secteurs économiques.

Afin de conserver une certaine adaptabilité du système face à cette diversité de sources qui peuvent évoluer indépendamment de l'Insee, le système d'information est conçu de façon modulaire. Chaque source subit de façon indépendante, une chaine de contrôle. L'objectif de cette chaîne est d'optimiser le partage des tâches entre l'homme et la machine pour rechercher et corriger les erreurs qui se sont glissées dans l'information collectée. Ne sont confiés à l'action humaine que les dossiers pour lesquels l'automatisation est moins performante. L'automatisation trouve ses limites surtout en matière d'entreprises dont le poids économique relativement à l'ensemble de la population est important.

Une fois que les caractéristiques sont validées par les processus de contrôle de chaque sources, il convient d'assurer une cohérence globale de ces caractéristiques multi-sources. C'est le but du processus de réconciliation des données individuelles. Il s'assure que les caractéristiques communes à deux sources ont in fine la même valeur. Dans le cas où il existe une divergence initiale sur ces caractéristiques d' « accrochage », le processus affecte une valeur unique après un arbitrage automatique qui privilégie une source selon différentes règles de priorité. Cette réconciliation concerne le chiffre d'affaires de chaque entreprise et sa ventilation en type de ventes mais également l'emploi et les rémunérations. Ce processus automatique s'accompagne d'un contrôle manuel des plus gros écarts initiaux. Ces derniers sont identifiés par le biais de macro contrôles. Une fois la valeur unique déterminée, elle est réinjectée dans chaque source dans laquelle elle figure et d'éventuels redressements automatiques ont lieu afin de rétablir les cohérences intra-source.

Après avoir présenté les différentes sources de données cet article se focalisera sur la description des contrôles intra et inter sources. Enfin, en statistique d'entreprises, les restructurations économiques (fusion, scission etc.) ont une très forte influence sur les résultats diffusés. Le processus de production ESANE essaye d'appréhender ce phénomène dès le contrôle des données en analysant les réponses des entreprises par le biais d'une nouvelle unité statistique spécifique : l'enveloppe de restructuration. Sa prise en compte dès le contrôle des donnés est présentée dans la dernière partie de cet article.

1 Les différentes sources en présence

1.1 L'enquête sectorielle annuelle

L'Enquête Sectorielle Annuelle (ESA) est formellement nouvelle, bien qu'elle soit une redéfinition sensible des Enquêtes Annuelles d'Entreprises (EAE) sur les mêmes champs sectoriels, redéfinition menée dans le cadre du programme Resane.

[1] Pour les secteurs industriels cette enquête est remplacée par l'Enquête Annuelle de Production (EAP) qui permet de répondre à la fois à la demande des statistiques structurelles et au règlement européens sur les produits (Prodcom). Cette enquête est gérée de façon autonome est n'est pas décrite dans cet article. Les résultats de cette enquête interviennent toutefois au moment de la réconciliation des données.

Un des objectifs principaux de ce programme étant l'allègement de la charge de réponse des entreprises via une large mobilisation des sources administratives jusqu'alors collectées dans les EAE et qui sont de fait déjà disponibles dans diverses sources administratives, notamment fiscale, ne feront plus l'objet d'une enquête statistique. L'ESA correspond ainsi à une EAE très allégée, notamment de toutes les données comptables que l'on peut mobiliser sur les liasses fiscales[1], des principales données d'emploi et des données d'échanges extérieurs.
Les principaux objectifs de cette enquête sont les suivants :

- comme pour les EAE précédemment, de repérer les différentes activités exercées par les entreprises, via la ventilation de leur chiffre d'affaires en branche, et d'en déduire alors leur activité principale (APE). Cet objectif est premier tout d'abord en ce qu'il conditionne le bon classement sectoriel des entreprises et par conséquent la qualité des statistiques sectorielles. Ensuite, il conditionne un bon passage secteur/branches, sur lequel repose l'élaboration des comptes nationaux de la France.
- de repérer les restructurations juridiques qui affectent la vie des entreprises et dont le repérage est essentiel pour produire de bonnes statistiques en évolution.
- de compléter la liasse fiscale sur certains aspects liés à l'investissement, notamment dans sa composante immatérielle et par voie d'apport.
- de décrire au travers de variables spécifiques les principales caractéristiques de chaque secteur.

Cette enquête concerne les secteurs économiques du commerce, des services, des industries agro-alimentaires, des exploitations forestières et des scieries, des transports et de la construction. Toutes les tailles d'entreprise sont dans le champ de l'enquête.
Le nombre d'unités enquêtées (120 000 environ) est inférieur à celui des EAE. Il s'agit pour le moment d'une enquête postale. La possibilité de répondre par internet sera développée ultérieurement. Les questionnaires sont envoyés en 4 vagues successives. La première est envoyée en fin février et la dernière en fin juin. C'est la date de clôture d'exercice N-1 qui détermine l'appartenance d'une entreprise à une vague.

- La vague 1 regroupe les entreprises qui ont une date de clôture comprise entre septembre et décembre N ;
- La vague 2 regroupe les entreprises qui ont une date de clôture au 31/12 de l'année N ;
- La vague 3 regroupe les entreprises qui ont une date de clôture comprise entre janvier et mars N+1 ;
- La dernière vague regroupe les autres entreprises.

Le plan de sondage est « classique » : stratifié à un degré, en utilisant les critères « code APE » et tranche de taille (en effectifs salariés) comme critères de stratification. Au sein des strates ainsi définies, le même taux de sondage sera appliqué pour l'ensemble des régions françaises.

1.2 Les données fiscales

Les entreprises, sociétés ou entreprises individuelles ont l'obligation de tenir une comptabilité (article L123-12 du code du commerce).

Les entreprises doivent acquitter l'impôt chaque année à raison des bénéfices réalisés cette année là (article 12 du code des impôts)

- Impôt sur les sociétés (IS) pour les sociétés
- Impôt sur le revenu (IR) pour les entreprises individuelles

La DGFiP et l'INSEE ont une convention pour que la DGFiP fournisse ces données, brutes, à l'Insee, qui les utilise pour des besoins statistiques.
L'obligation comptable diffère selon le secteur, la forme juridique et la taille de l'entreprise

Il en résulte différents types de bénéfice :

- Les bénéfices agricoles (BA)
- Les bénéfices industriels et commerciaux (BIC)
- Les bénéfices non commerciaux (BNC)

Et différents régimes selon l'importance de l'entreprise :

[1] pour certaines catégories d'entreprises, non astreintes à dépôt de liasses fiscales (par exemple, certaines coopératives agricoles des IAA ou du commerce de gros), la collecte du compte de résultat et de la ligne des immobilisations est maintenue.

- Le régime normal
- Le régime simplifié
- Le régime micro (ou au forfait pour le BA)

Régime ou source	**Nb total d'entreprises**	**Nb d'entreprises du champ ESANE**	**% du CA total du champ ESANE**	**Nb variables disponibles**
Bénéfice réel normal (BIC-RN)	700 000	700 000	**90**	1000
Bénéfice réel simplifié (BIC-RSI)	1 200 000	1 200 000	5	400
Bénéfices non commerciaux (BNC)	500 000	500 000	2	100
Entreprises profilées	3	3	3	1000
Régimes micro (BIC-micro et BNC-micro)	400 000	400 000	ε	0
DGCP	500	500	ε	1000
Coop IAA non DGI	500	500	ε	200
Bénéfices agricoles (BA-RN/BA-RSI)	300 000	200	ε	400
TOTAL	3 100 000	**2 800 000**	**100**	**1200**

Ces données sont fournies en 4 vagues :
- les liasses dites « anticipées » livrées en juin N+1 pour les entreprises qui télé-déclarent,
- les liasses dites « intermédiaires » livrées en mi septembre N+1
- les liasses dites « normales » livrées en mi octobre N+1
- les liasses dites « complémentaires » livrées en mars N+2

et par quatre centres de saisie des impôts (Marseille, Clermont-Ferrand, Nantes et Reims). Le tout représente donc un nombre important de fichiers à traiter, d'autant qu'il y a autant de fichiers que de types de régime.

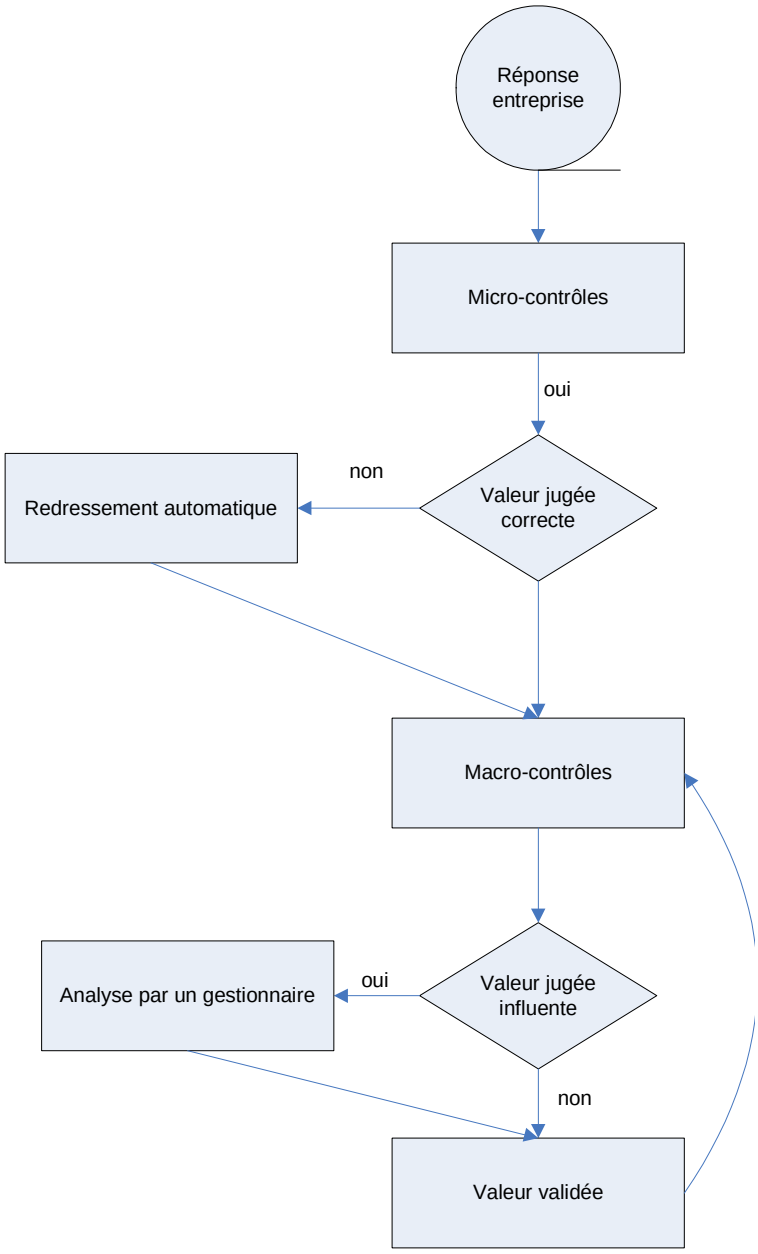
2 Des contrôles « intra sources » indépendants d'une source à l'autre

Cette chaîne de contrôle se déroule en deux étapes une fois que l'entreprise a bien été identifiée comme faisant partie du champ de l'enquête :

- la première a pour objectif de traiter la non-réponse partielle et de corriger des valeurs jugées aberrantes. Elle est entièrement automatique et met en œuvre des micro-contrôles de cohérence et de vraisemblance puis, le cas échéant, des redressements automatiques se basant sur la réponse de l'année précédente ou sur les réponses des autres entreprises appartenant à la même strate de contrôle (strate obtenue par croisement entre l'activité principale exercée (APE) et la tranche de taille de l'entreprise).
- la seconde vise à identifier les unités qui doivent être contrôlées par un gestionnaire. Il s'agit d'entreprises pour lesquelles la réponse (ou la non-réponse) est jugée atypique et suffisamment importante pour ne pouvoir être redressée automatiquement. Cette sélection s'opère à l'aide de macro-contrôles qui vont permettre d'identifier les unités les plus fortement contributrices aux agrégats diffusés.

Contrairement à ce qui était fait jusque là en statistiques d'entreprises, c'est donc bien les macro-contrôles qui sont le principal outil pour la mise en évidence des unités à traiter. C'est pourquoi, bien qu'ayant lieu chronologiquement après les micro-contrôles ils sont présentés en premier dans cet article.

Schéma de représentant la succession des contrôles des données dans RESANE.



2.1 Les Macro-contrôles

Les macro-contrôles ont donc pour objectif de repérer les entreprises dont les réponses ont une influence notable sur les agrégats qu'il est prévu de diffuser.
Dans le cadre de RESANE, les macro-contrôles servent à déterminer quelles entreprises sont traitées par les gestionnaires et, au sein d'un dossier d'entreprise, quelles variables sont à analyser.

Les macro-contrôles passent après la chaîne des micro-contrôles décrite ci-dessus, sur des données qui ont donc éventuellement pu être redressées automatiquement.

Le principe des macro-contrôles est le suivant. Il consiste à calculer pour chaque variable de chaque entreprise une contribution (cf. 2.1.3 pour les formules de calcul). Si la contribution de l'entreprise dépasse un seuil fixé, un message propre au contrôle est produit.
Ces messages permettront au gestionnaire de savoir quelles sont les variables à contrôler et donnent quelques éléments de diagnostic pour rappeler l'entreprise.

Exemple de message de macro-contrôle :
Le taux d'évolution de CA Total du secteur (Niveau sous-classe) est de -1.66%. Sans cette entreprise, ce taux d'évolution serait de -1.00%. L'entreprise fait évoluer son secteur de -0.664 point(s).

Pour les contrôles de données dans RESANE, trois types de macro-contrôles sont mis en œuvre :
- contemporain (par exemple : rapport Chiffre d'affaires sur Effectif en N)
- en évolution (par exemple : Chiffre d'affaires entre N − 1 et N, ou rapport Chiffre d'affaires sur Effectif entre N − 1 et N), utilisant éventuellement les enveloppes (cf. chapitre 4)
- par différence, pour contrôler les redressements individuels (par exemple : mesure de l'écart entre le chiffre d'affaires initial et le chiffre d'affaires redressé automatiquement).

2.1.1 Macro-contrôles de type écart[2]

Il s'agit des macro-contrôles contemporains ou en évolution et il convient de distinguer :
- les entreprises qui participent au calcul des agrégats en N (voire N-1). On les appellera l'« assiette du macro-contrôle » dans la suite de l'article
- et les entreprises qui peuvent être sélectionnées par le macro-contrôle. On les appellera « entreprises éligibles » par la suite.

Un macro-contrôle est lié à un niveau d'agrégation. Dans le cadre de RESANE, deux niveaux ont été définis :
- le niveau APE (niveau à 3 ou 5 caractères)
- le niveau APE (niveau 3 caractères) × tranche de taille (en 16 tranches de tailles)

2.1.1.1 Calcul des agrégats et des contributions

Le détail du calcul des agrégats et des contributions est décrit ci-dessous.
Les formules présentées concernent le cas d'un échantillon (enquête ESA). Dans le cas de données exhaustives (cas du contrôle des liasses fiscales), il suffit de remplacer l'échantillon par la population totale et de considérer que le poids de chaque unité est à 1 dans les formules.

Le calcul des agrégats utilise les valeurs imputées des caractéristiques des entreprises en non-réponse totale auxquelles est appliqué le poids de sondage au lancement de l'enquête, appelé par la suite « poids avant calage ». En revanche comme ceci a déjà été évoqué plus haut, seules les entreprises répondantes se voient calculer des contributions.

2.1.1.2 Exemple de calcul des agrégats sectoriels en évolution (N / N-1)

Assiette

Pour le calcul des agrégats, les créations/cessations observées l'année N (y compris en cours d'année) ne sont pas prises en compte. Les entreprises ayant changé de secteur entre N et N − 1 seront considérées comme faisant partie du secteur d'arrivée (N) pour les deux années : leurs données N − 1 seront donc incluses dans le calcul de l'agrégat N − 1 du secteur d'arrivée N.
De même, les unités de la partie renouvelée de l'échantillon sont prises en compte, mais chacune pour le compte d'un seul agrégat (N − 1 ou N) . En revanche, les cessations au cours de l'année N de la partie conservée de l'échantillon ne sont pas prises en compte, ni pour l'agrégat N ni pour l'agrégat N − 1.

[2] Drop-out en anglais.

Par principe on fait l'hypothèse que le processus de production permet d'identifier l'ensemble des restructurations d'une année. Les enveloppes (cf. chapitre 4) ont donc un poids de 1. De ce fait, pour le calcul des valeurs des caractéristiques d'une enveloppe, le poids de chaque entreprise de l'enveloppe devient égal à un 1.

Formule de calcul

Les formules de calcul des agrégats en N et N − 1 de la grandeur Y sur le secteur S sont :

$$Y_S^N = \sum_{i \in \text{Échantillon N}} (\text{POIDS\_AVT\_CALAGE}_i^N \times Y_i^N \times I_{i \in S}^N \times I_{i \notin \text{nais}}^N \times I_{i \notin \text{env}}^N) + \sum_{e \in \text{Enveloppes}} (Y_e^N \times I_{e \in S}^N)$$

$$Y_S^{N-1} = \sum_{i \in \text{Échantillon N}-1} (\text{POIDS\_APRES\_CALAGE}_i^{N-1} \times Y_i^{N-1} \times (I_{i \in S}^N + I_{i \in S}^{N-1} \times I_{i \in \text{Échantillon N}}) \times I_{i \notin \text{mort}}^N \times I_{i \notin \text{env}}^N) + \sum_{e \in \text{Enveloppes}} (Y_e^{N-1} \times I_{e \in S}^N)$$

où $\text{POIDS\_AVANT\_CALAGE}_i$ et $\text{POIDS\_APRES\_CALAGE}_i$ sont les poids de sondage,

$$I_{i \in S}^N = \begin{cases} 1 \text{ si l'entreprise (} \textit{ou l'enveloppe} \text{) x } \textit{a}\text{ppartient au secteur S en N} \\ 0 \text{ sinon} \end{cases}$$

$$I_{i \in S}^{N\text{-}1} = \begin{cases} 1 \text{ si l'entreprise i appartient au secteur S en N - 1} \\ 0 \text{ sinon} \end{cases}$$

$$I_{i \notin \text{Échantillon N}} = \begin{cases} 1 \text{ si l'entreprise i n'appartient pas à l'échantillon en N} \\ 0 \text{ sinon} \end{cases}$$

$$I_{i \notin \text{nais}}^N = \begin{cases} 1 \text{ si l'entreprise i n'est pas née en N} \\ 0 \text{ sinon} \end{cases}$$

$$I_{i \notin \text{mort}}^N = \begin{cases} 1 \text{ si l'entreprise i n'est pas morte en N} \\ 0 \text{ sinon} \end{cases}$$

$$I_{i \notin \text{env}}^N = \begin{cases} 1 \text{ si l'entreprise i n'appartient pas à une enveloppe en N} \\ 0 \text{ sinon} \end{cases}$$

Contribution

La formule de calcul d'une contribution pour une unité i_0 de la variable Y est :

$$\text{CTR}(i_0) = \left| \frac{Y_S^N}{Y_S^{N-1}} - \frac{Y_S^N - \text{POIDS\_AVT\_CALAGE}_{i_0}^N \times Y_{i_0}^N}{Y_S^{N-1} - \text{POIDS\_AVT\_CALAGE}_{i_0}^N \times Y_{i_0}^{N-1}} \right|$$

2.1.2 Macro-contrôles par différence

Le principe est de vérifier, pour chaque unité et pour chaque caractéristique, que les redressements effectués n'ont pas été « trop forts », c'est à dire visibles au niveau de l'agrégat. Le redressement va en effet naturellement « normaliser » la réponse des entreprises, mais il ne faudrait pas arriver à l'extrême inverse à savoir supprimer les réponses atypiques mais correctes. L'objectif de ce contrôle est de détecter les entreprises et les variables pour lesquelles le redressement même s'il est fort est juste car il corrige bien une erreur de saisie ou d'unité de réponse mais pas un comportement atypique. Pour cela, pour toute unité i, une contribution est calculée à partir de l'écart entre la valeur redressée de la variable Y et sa valeur brute.

La formule de calcul est la suivante :

$$CTR(i_0) = POIDS\_AVT\_CALAGE_i \times \left| Y_i^{red} - Y_i^{brute} \right| / Y_S^N$$

où Y_S^N est l'agrégat en niveau du secteur S pour l'année N de la variable Y. Cet agrégat peut être peut être calculé au niveau groupe, sous-classe ou groupe croisé avec les tranches de taille.

2.2 Les micros Contrôles

2.2.1 Le principe

Les micro-contrôles attribuent des notes aux variables sur lesquelles ils sont lancés, permettant ainsi, en sommant ces notes, de savoir si ces variables sont qualifiées ou non (voir partie Qualification), et donc si elles subiront ou non un redressement automatique (voir partie Redressement). Suivant la note attribuée, un message de contrôle pourra être généré, cependant l'affichage de ce message au gestionnaire sera conditionné par le résultat des macro-contrôles correspondant à la variable. Ce sont en effet les macro-contrôles qui vont permettre de sélectionner les unités et les variables à contrôler manuellement. Les messages des micro-contrôles sont à considérer comme des aides au diagnostic du gestionnaire et non comme des éléments déclencheurs des contrôles manuels.

Le principe de calcul des micro-contrôles a été repris de l'EAE ([2], [3]).

Les micro-contrôles sont de quatre types :

- **contrôles internes de cohérence** (CIC) :
  - ils permettent de détecter des incohérences au sein du questionnaire (ou de la liasse fiscale) d'une entreprise, en comparant deux valeurs, dont l'une, en général, est la somme de l'autre.
  - ils ne nécessitent pas de paramètres ; une marge d'erreur de 5 ou 10 % est accordée
  - Les notes accordées à ce type de contrôle sont fortes (en général 10 si l'écart est acceptable et 0 sinon). Ainsi, par exemple, les données seront jugées correctes au niveau de la phase des micro-contrôles à partir du moment où les équilibres comptables les impactant sont respectés.

- **contrôles internes de vraisemblance** (CIV)
  - ils permettent de détecter des invraisemblances entre la valeur d'un ratio pour une entreprise et les valeurs de ce ratio dans la strate à laquelle se trouve cette entreprise. Les CIV nécessitent l'utilisation d'une table de paramètres contenant :
    - les bornes Q1 et Q3, correspondant aux premiers et troisièmes quartiles de la distribution
    - les bornes B1 et B3, correspondant aux premier et neuvième déciles de cette même distribution
  - la note attribuée à la variable dépendra de la valeur du ratio par rapport aux bornes :
    - entre Q1 et Q3, la note sera élevée

        - entre B1 et B3 mais hors de l'intervalle Q1 Q3, la note sera moyenne (en général : 4)
        - en dehors de l'intervalle B1 B3, la note sera nulle.
    - ces bornes seront calculées en début d'enquête grâce aux données N − 1 puis recalculées en cours d'enquête
    - on utilisera des strates du type : 3 premiers chiffres de l'APE x tranche de taille (en 4 tranches de tailles) pour le calcul des bornes[3]
    - les messages seront susceptibles d'être activés dès que le ratio de l'unité sera hors des bornes Q1 Q3
    - on pourra utiliser des estimations pour données incomplètes, par exemple en ramenant la durée d'exercice à 12 mois

— **contrôles temporels de vraisemblance** (CTV) :
    - ils permettent de détecter des invraisemblances entre la valeur d'une variable ou d'un ratio d'une entreprise en N et sa valeur en N − 1
    - Le principe de notation est le même que celui décrit ci-dessus pour les CIV
    - Les notes attribuées aux contrôles temporels sont plus fortes que celles des contrôles internes. Ainsi, la cohérence temporelle est jugée plus fiable que la cohérence interne. Autrement dit une entreprise ayant le même comportement atypique deux années de suite sera considérée comme « normale » la 2e année.

Remarque : les contrôles temporels ne sont faits que pour les entreprises présentes en N et N − 1 et n'appartenant pas à une enveloppe de restructuration en N (cf. chapitre 4 de cet article).

— **contrôles temporels de cohérence** (CTC) :
    - Ces contrôles ne sont effectués que pour les liasses fiscales. Ils permettent de repérer les incohérences d' « accrochages temporels » des liasses fiscales. Ainsi par exemple, le total des immobilisations de fin d'exercice de l'exercice N-1 doit être égal au total des immobilisations de début d'exercice N.
    - ils ne nécessitent pas de paramètres ; une marge d'erreur de 5 ou 10 % sera accordée

2.2.2 La Qualification

Cette phase de qualification n'a lieu que dans l'ESA.
La qualification d'une variable consiste à sommer les notes des micro-contrôles qui concernent la variable et à comparer ce total à un seuil.

- Si le total des notes est supérieur ou égal au seuil alors la variable est dite qualifiée et sa valeur est conservée pour la suite des traitements
- Si le total des notes est strictement inférieur au seuil alors la variable est redressée automatiquement (cf. paragraphe suivant). C'est le cas notamment des variables en non-réponse partielle pour lesquelles la somme des notes est nulle.

Dans le cadre d'ESANE, le seuil de qualification des variables est fixé à 10.
La note des contrôles est ensuite adaptée de telle sorte qu'une variable est en générale qualifiée si :

- elle obtient la valeur maximale à un CIC ;
- elle obtient deux notes moyennes à au moins 2 contrôles (CTV et / ou CIV).

Par rapport à l'EAE, la qualification des variables de l'ESA est moins stricte. L'idée est de redresser le moins possible les réponses initiales des entreprises. Les macro-contrôles permettront dans un deuxième temps de repérer des problèmes éventuels.
Cette phase de qualification vise essentiellement à corriger la non-réponse partielle d'une part les « grosses » erreurs du type erreur de saisie ou erreur dans l'unité de réponse (réponse en € au lieu de K€ par exemple).

[3] Si la strate contient moins de 3 unités alors le calcul des bornes se fera sur la strate élargie obtenue en croisant les deux premiers chiffres de l'APE avec la tranche de taille. S'il y a encore moins de 3 unités dans cette strate, des bornes par défaut sont utilisées.

2.2.3 Les Redressements

Dans la chaîne de contrôle de l'ESA, à la suite de la qualification des variables a lieu la phase de redressements des variables non qualifiées et donc notamment des variables en non-réponse partielle.

Plusieurs redressements, par ordre de priorité, sont successivement testés pour une même variable. Dès que l'ensemble des conditions nécessaires à l'application d'un redressement est réuni, ce dernier est validé sinon, un autre type redressement est tenté, et ainsi de suite jusqu'à temps que les conditions (de moins en moins exigeantes) nécessaires à un des redressements de la liste soient réunies. Quoi qu'il arrive, en fin de course un redressement aura toujours lieu (si un redressement est requis, évidemment).

Il existe différents types de redressements, qui sont tentés dans cet ordre :

— des **redressements déterministes** :
  - ils consistent à remplacer la valeur de la variable par une autre valeur ou une combinaison de valeurs propres à l'entreprise
  - par exemple : nombre total de magasins remplacé par la somme des nombres de magasins déclarés par taille de surface

Vu que le questionnaire ESA évite au maximum toute redondance d'informations, ce genre de redressement est assez rare.

— des **redressements par estimation** :
  - **par tendance auxiliaire** :
    - ils consistent à faire évoluer la valeur N-1 de la variable à redresser selon le taux d'évolution observée sur une autre variable du questionnaire (appelée variable auxiliaire) corrélée à la variable à redresser
    - $X_N^{red} = X_{N-1} * \frac{Y_N}{Y_{N-1}}$ *où Y est la* var*iable auxiliaire corrélée à X*
    - ces redressements supposent que la valeur en N − 1 de la variable à redresser n'avait pas subi de redressement ou avait été redressée par tendance auxiliaire
    - ils supposent aussi que les valeurs en N et N-1 de la variable auxiliaire sont de bonne qualité (variable auxiliaire qualifiée en N et N-1)

  - ou **par moyenne de strate** :
    - La valeur redressée est obtenue en multipliant la valeur N d'une caractéristique auxiliaire par le ratio obtenu en divisant le total de la caractéristique à redresser par le total de la caractéristique auxiliaire calculée au niveau de la strate d'appartenance de l'unité (groupe de l'APE croisé avec la tranche de taille) [4]
    - $X_N^{red} = Y_N * \frac{\sum_{strate} X_N^{qualifié}}{\sum_{strate} Y_N^{qualifié}}$
    - Ce type de redressement suppose que la valeur en N de la variable auxiliaire est de bonne qualité (variable auxiliaire qualifiée en N)[5].

Après cette phase de redressements automatiques de variables, un lissage automatique permettra de s'assurer que toutes les ventilations d'une variable du questionnaire sont égales au total de cette variable si elle figure dans le questionnaire.
Il s'agit généralement de ventilations du chiffre d'affaires (par région, ou type d'ouvrages etc.)

[4] En début d'enquête cette moyenne est calculée à partir des résultats de l'enquête N-1. Elle est ensuite recalculée une fois que le nombre de réponses « qualifiées » en N dans la strate est jugé suffisant

[5] Si aucune variable auxiliaire n'est trouvée, la caractéristique est redressée en affectant sa moyenne dans la strate. Ce qui revient à prendre la variable « 1 » comme variable auxiliaire !

2.2.4 Messages de contrôle des données associés aux micro-contrôles

Chaque micro-contrôle en erreur (c'est à dire un micro-contrôle pour lequel la note obtenue n'est pas maximale) donne lieu à un message d'erreur à l'attention des gestionnaires.

Ces messages sont considérés comme des « aides au diagnostic » pour les gestionnaires.
Toutefois pour une liasse ou un questionnaire à contrôler donné, ne sont présentés aux gestionnaires que les messages des micro-contrôles qui sont associés à un macro-contrôle en erreur, c'est à dire à une variable qui doit être contrôlée manuellement.

2.3 Indicateur de priorité de traitement

Comme vu précédemment, ce sont les macro-contrôles qui structurent le plus profondément le travail des gestionnaires.
Pour gérer la production, il est nécessaire de calculer une note globale pour chaque entreprise qui permettra de mettre en place des priorités de traitement ; l'objectif étant de faire traiter en priorité par les gestionnaires les dossiers jugés les plus douteux. Ainsi, si pour une raison ou pour une autre, les gestionnaires n'ont pas le temps de traiter l'ensemble des problèmes, les plus important l'auront été. Si ce vœux pieux est on ne peut plus légitime, il est loin d'être évident à réaliser vu le nombre de variables à contrôler dans le cadre d'ESANE.
Est-il préférable de contrôler en priorité les dossiers ayant une grosse contribution pour une variable donnée ou au contraire les dossiers présentant plusieurs problèmes. La solution proposée va plutôt dans le sens de la deuxième hypothèse mais force est de constater qu'elle a malheureusement ses limites (cf. chapitre 5).
Lorsqu'une entreprise est transmise à un gestionnaire pour examen, les différents messages de contrôle des données (des macro et micro-contrôles associés à ces variables) lui indiquent quelles sont les caractéristiques qu'il doit contrôler.
A l'inverse, si le dossier est jugé non prioritaire, il ne sera jamais à traiter par un gestionnaire bien qu'au moins une variable de sa réponse ait pu dépasser le seuil de contribution d'un macro-contrôle la concernant.

Le principe de calcul de l'indicateur de priorité de traitement retenu est le suivant :

- Pour chaque ratio contrôlé par un macro-contrôle de type écart et pour chaque macro-contrôle par différence, deux seuils ont été mis en place : un seuil élevé SE et un seuil plus modéré SM [4].

- Pour les entreprises ou les enveloppes dont la contribution dépasse le seuil élevé pour la variable d'intérêt *Xi*, le dossier est caractérisé pour cette variable par un « état », noté E(*Xi*) qui prend alors la modalité « I ». Si la contribution est inférieure à SE mais supérieure à SM, alors E(*Xi*) prend la modalité « S ».

- Pour une variable donnée, on retient le niveau de gravité maximal des macro-contrôles qui concernent la variable.

- On agrège ensuite les états de toutes les variables avec la formule suivante pour définir un état global quantitatif QEG :

$$QEG = \frac{\sum_i 1\{E(X^i) = I\} \bullet A \bullet Ki + \sum_i 1\{E(X^i) = S\} \bullet Ki}{(1+A)\sum_i Ki}$$

  avec A>1 et Ki fixé pour chaque variable.

  A représente l'importance qu'on donne à un seuil élevé par rapport à un seuil modéré et Ki représente l'importance de chaque variable.

- Un état global EG est ensuite calculé :

  EG=P(rioritaire) si $QEG > \alpha$,
  EG=I(mportant) si $\alpha > QEG > \beta$,
  EG=S(econdaire) si $\beta >= QEG > \gamma$
  EG=A(automatisable) si $\gamma >= QEG$

  avec α, β,.γ compris entre 0 et 1.

  Seules les entreprises dont la priorité globale est suffisante (EG parmi P, I ou S) sont examinées par les gestionnaires. En revanche, lorsqu'une entreprise est à contrôler, toutes les variables pour lesquelles un macro-contrôle dépasse le seuil bas sont examinées.

  α, β,.γ sont des paramètres globaux de réglage que peut utiliser l'encadrement centre chargé du contrôle des données.

Un tel indicateur est calculé pour chaque source. Un autre sera également calculé au moment de la réconciliation des données (cf. paragraphe 3).

2.4 Le travail du gestionnaire d'enquête

A l'issue de la chaine de contrôle, la population des entreprises est donc divisée en deux.

- **Les entreprises validées automatiquement** pour lesquelles la réponse est jugée de qualité suffisante et pour laquelle aucune intervention du gestionnaire n'est demandée.
- **Les entreprises à contrôler manuellement**. Il s 'agit des dossiers pour lesquels une intervention du gestionnaire est requise.

2.4.1 Les entreprises validées automatiquement

Ces entreprises peuvent être classées en trois catégories :

- Les réponses parfaites. C'est à dire les réponses pour lesquelles d'une part, aucun redressement automatique n'a eu lieu et d'autre part aucun macro-contrôle ne dépasse le seuil minimum décrit dans le paragraphe 2.3.
- Les entreprises pour lesquelles des redressements automatiques ont pu avoir lieu mais pour lesquelles aucun macro-contrôle n'a dépassé le seuil minimum. Dans ce cas, les redressements automatiques sont considérés comme validés et les messages des micro-contrôles en erreur sont inactivés et ne sont donc pas consultables par les gestionnaires.
- Les entreprises pour lesquelles il y a eu ou non des redressements automatiques, qui ont eu au moins un macro-contrôle actif (id est dont la contribution a dépassé le seuil minimum) mais dont l'indicateur de priorité de traitement prend la valeur « automatisables » (cf. paragraphe 2.3 ci dessus). Dans ce cas, les messages des micro-contrôles et des macro-contrôles en erreur sont inactivés et ne sont pas consultables par les gestionnaires.

2.4.2 Les entreprises à contrôler manuellement

Il s'agit des entreprises ayant au moins un macro-contrôle actif et dont la priorité de traitement est au moins égale à « secondaire ».
Ces unités sont donc à traiter par le gestionnaire et pour lesquelles il dispose :

- De l'ensemble des messages des macro-contrôles actifs
- De l'ensemble des messages des micro-contrôles en erreur qui concernent une variable pour laquelle il existe au moins un macro-contrôle actif. A contrario, s'il existe en plus des micro-contrôles en erreur concernant des variables non associées à un macro-contrôle actif, ces derniers sont « inactivés » et les messages d'erreurs associées ne sont pas accessibles aux gestionnaires. L'objectif est bien de focaliser les travaux du gestionnaire sur les variables « repérées » par les macro-contrôles.

A partir de ces éléments, le gestionnaire corrige dans un premier temps les erreurs évidentes (erreur de saisie par exemple), puis s'il reste des problème il doit rappeler l'entreprise pour obtenir une explication.
Deux cas sont alors envisageables :

- l'entreprise confirme sa réponse initiale. Dans ce cas, le gestionnaire valide la réponse initiale de l'entreprise et essaye d'obtenir des éléments pouvant expliquer ce comportement jugé atypique, éléments qu'il consigne ensuite dans un commentaire.
- l'entreprise corrige sa réponse initiale, le gestionnaire modifie alors la réponse de l'entreprise.

Dans les deux cas, la valeur corrigée ou confirmée est considérée comme validée, elle ne sera donc en aucun cas redressée automatiquement. Elle pourra en revanche servir au redressement automatique d'autres variables et les macro-contrôles associés à cette variable seront « inactivés ».

3 Une réconciliation des sources validées ou les contrôles inter sources

3.1 Des caractéristiques d'accrochage

Une fois les caractéristiques des différentes sources validées, il convient d'assurer une cohérence globale de ces caractéristiques multi-sources. C'est le but du processus de réconciliation des données individuelles qui intervient donc en fin de traitement des données par les gestionnaires.

L'utilisation conjointe de données de l'enquête ESA d'une part et de données administratives d'autre part est un des points forts du dispositif ESANE, mais constitue en même temps un élément de complexité. A partir du moment où, pour une même entreprise, on peut opérer « l'accrochage » des sources, le dispositif acquiert de la robustesse : la détection d'une incohérence entre les sources conduit souvent à détecter un problème au niveau des données fournies par l'entreprise, pouvant être lié par exemple à un événement comme une restructuration. La réconciliation de l'ESA et de la liasse fiscale est faite à partir du chiffre d'affaires et de sa ventilation par ventes (biens, marchandises et services).

S'il y a une divergence entre les 2 sources, la valeur de la caractéristique réconciliée retenue est celle de la source jugée prioritaire. Ainsi, pour une entreprise répondante à l'ESA et ayant renvoyé une liasse fiscale, le chiffre d'affaires réconcilié est celui de la liasse fiscale alors que la structure de sa ventilation est celle observée dans l'enquête. En effet, si le chiffre d'affaires est une donnée sensible de la liasse fiscale, sa ventilation est peu utilisée par les impôts et donc pas toujours bien renseignée par les entreprises dans leur liasse fiscale. Au contraire, la ventilation par activité du chiffre d'affaires est l'une des priorités de l'ESA et elle est donc bien contrôlée dans cette source. Ceci explique donc les règles de priorités retenues. Il existe toutefois quelques exceptions mais qu'il n'a pas semblé intéressant de développer dans cet article.

En outre, comme pour le contrôle des données intra sources, il a été décidé d'appliquer des procédures de macro-contrôles conduisant à demander une expertise manuelle des gestionnaires uniquement pour les cas jugés les plus impactants[6].

[6] Par exemple les entreprises pour lesquelles l'écart de chiffre d'affaires entre les deux sources est supérieur à x% du chiffre d'affaires du secteur.

Pour une variable X donnée, on calcule la contribution d'une unité de la façon suivante :

$$score = \left| \frac{(X_{s1} - X_{s2}) * poids\_esa}{T(X_p)} \right|$$

où $\begin{cases} X_{S1} = valeur\ de\ la\ caracteristique\ X\ dans\ la\ source\ 1 \\ X_{S2} = valeur\ de\ la\ caracteristique\ X\ dans\ la\ source\ 2 \\ T(X_p) = total\ de\ la\ caracteristique\ X\ dans\ la\ source\ prioritaire\ au\ niveau \\ \qquad d'agrégation\ correspondant\ au\ niveau\ de\ controle \\ Poids\_esa = Poids\ de\ l'enquête \end{cases}$

Une réponse d'entreprise est à contrôler manuellement quand la contribution de l'entreprise pour un macro-contrôle dépasse un seuil lâche de 1%. Un deuxième seuil dit strict et égal à 2% permet également de repérer les entreprises fortement contributrices (cf. chapitre précédent sur l'indicateur de priorité de traitement). Enfin, pour éviter de traiter manuellement des petites entreprises de l'échantillon (dont la forte contribution serait essentiellement due à leur fort poids de sondage), il faut en plus que l'écart entre les chiffres d'affaires de l'ESA et celui de la liasse fiscale soit supérieur à 10% du chiffre d'affaires réconcilié.

3.2 Des caractéristiques liées

Sont également concernées par cette réconciliation « manuelle », des caractéristiques de la liasse ou de l'enquête qui sont fortement corrélées aux caractéristiques d'accrochage. Ceci, parce qu'un redressement automatique est jugé trop périlleux dans le cas d'une forte modification de la caractéristique d'accrochage. L'exemple type concerne les types d'achats (marchandises, biens et services) qui sont fortement corrélés aux ventes correspondantes.

Ainsi, lorsqu'un gestionnaire privilégie la valeur de l'ESA pour sa ventilation des ventes et que cette dernière est assez différente de celle observée sur la liasse fiscale, il lui est demandé de corriger également les achats correspondants, disponibles dans la liasse fiscale, en conséquence.

Dans le cas où la réconciliation est automatique, les caractéristiques liées sont redressées automatiquement. Ces redressements automatiques sont toutefois plus complexes que ceux décrits ci-dessus dans la partie intra source. Ils visent notamment à conserver au maximum la marge commerciale initiale.

3.3 Une mise en cohérence finale des données de chaque source

Pour les variables d'accrochage ainsi que pour les caractéristiques liées, une fois la valeur unique déterminée, elle est réinjectée dans chaque source dans laquelle elle figure. La chaîne de contrôle des données intra source est alors relancée dans chacune des sources afin de contrôler et éventuellement de redresser automatiquement les autres données de la source en prenant en compte le résultat de la réconciliation.

Les caractéristiques issues de la réconciliation (y compris les caractéristiques liées) sont toutefois considérées comme qualifiées. Elles ne pourront donc plus être redressées par la chaîne de contrôles intra sources. En revanche, elles pourront servir de base pour le redressement des autres caractéristiques afin que ces dernières soient cohérentes avec les valeurs définitives des caractéristiques réconciliées.

Ainsi, dans le cadre de l'ESA par exemple, le chiffre d'affaires peut avoir été modifié durant la phase de réconciliation des données et au moment de sa « réintroduction » dans le questionnaire ESA, toutes les ventilations du chiffre d'affaires[7] seront lissées sur cette nouvelle valeur.

[7] Ventilation par type clients, par régions, par type d'ouvrages par exemple

4 Le cas particulier des enveloppes de restructuration : où comment améliorer les contrôles temporels

4.1 Qu'est ce qu'une enveloppe de restructuration et pourquoi la prendre en compte dans les statistiques

Une restructuration économique se caractérise par un ensemble d'opérations qui a un impact sur l'activité productive courante (opérations de production et de formation brute de capital fixe) des entreprises par le biais de transferts d'activités.
Les restructurations les plus connues sont les fusions, les absorptions et les scissions partielles ou totales. Elles ont un impact non négligeable sur l'évolution entre deux années consécutives des agrégats sectoriels.

Prenons l'exemple d'un secteur dans lequel toutes les unités voient leur chiffre d'affaires croître de 10% entre deux années consécutives et qu'en parallèle, la plus grosse entreprise du secteur qui pesait 50% du chiffre d'affaires en N-1 disparait car elle est absorbée par une autre unité d'un secteur différent.
Une simple comparaison des résultats en niveau des deux années consécutives peut conduire à un diagnostic erroné. En effet, entre les deux années, le chiffre d'affaires du secteur a baissé de 45 %. On pourrait donc conclure au fait que le secteur se porte mal alors qu'en réalité les entreprises du secteur sont toutes en croissance. Deux messages contradictoires sont donc possibles.

A partir de ce constat et afin de contrebalancer les discours qui se basent sur les évolutions apparentes[8], il a été décidé de prendre en compte, ou plutôt d'essayer de neutraliser les restructurations afin de rétablir des évolutions économiques sectorielles cohérentes entre 2 années successives. Pour se faire, l'objectif est de recalculer la valeur des caractéristiques en N-1 en se plaçant dans la même structure en termes d'entreprise qu'en N. il s'agit en quelque sorte de faire comme si la restructuration avait déjà eu lieu en N-1. Un nouveau concept d'unité statistique a donc été créé. Il s'agit de l'enveloppe de restructuration qui est géré à l'Insee dans un outil baptisé CITRUS [5].
Cette unité statistique se compose de deux listes d'entreprises ou Siren (entreprises de l'année N-1 avant la restructuration ; entreprises de l'année N après la restructuration).
Il est ensuite possible de calculer une APE pour l'enveloppe à partir des APE des unités légales qui la composent l'année N. Afin de neutraliser des changements de secteurs l'APE de l'enveloppe en N-1 est la même qu'en N.

Une liasse fiscale (respectivement un questionnaire ESA) est calculée en cumulant les liasses (respectivement les questionnaires) des unités légales qui composent l'enveloppe en N et en N-1.

Ces données « fictives » vont ensuite être utilisées dans les contrôles temporels en lieu et place des entreprises qui les composent.

4.2 La prise en compte des enveloppes dès la phase de contrôle

La prise en compte des enveloppes dès la phase de contrôle des données est chose nouvelle à l'Insee. Elle vise principalement trois choses :

- permettre d'avoir des agrégats en évolution entre N et N-1 plus proches des agrégats diffusés et ainsi d'améliorer la pertinence des contrôles temporels
- estimer des flux intra-enveloppes à gommer (ou ajouter) afin d'obtenir la meilleure estimation de l'évolution possible. Prenons le cas où A absorbe B en N et qu'en N-1, B vendait X millions d'euros de biens à A. Ces X millions disparaissent du marché en N. Pour raisonner à structure

[8] Par évolution apparente il faut comprendre l'évolution entre les agrégats en niveau calculé en N-1 et N.

constante il faut donc retirer ces X millions en N-1. Dans le jargon ESANE, X est qualifié de flux intra-enveloppe. Dans le cas des restructurations jugées importantes par les macro-contrôles, il est demandé aux gestionnaires d'essayer d'obtenir le montant de ces flux en contactant les entreprises prenant part à la restructuration.

— gommer des évolutions « atypiques » au niveau des unités légales en prenant en compte les enveloppes de restructuration et ainsi améliorer le processus d'identification des entreprises à contrôler. Ne pas utiliser l'enveloppe conduirait à contrôler manuellement à tort des entreprises ayant des évolutions surprenantes uniquement dues à la restructuration. Ainsi, une entreprise absorbante peut voir son chiffre d'affaires fortement augmenter d'une année sur l'autre alors que si on compare ce chiffre d'affaires à la somme en N-1 des chiffres d'affaires des entreprises absorbantes et absorbées, l'évolution paraitra beaucoup moins atypique.

5 Quelques difficultés rencontrés

La mise en œuvre de cette méthodologie de contrôle s'est heurtée à plusieurs difficultés dont les principales vont être détaillées ci-dessous.

5.1 Des agrégats initiaux non-stables

Le principe du macro-contrôle « drop out » tel qu'il a été décrit ci-dessus est de repérer les entreprises ayant un impact fort sur l'évolution d'un agrégat diffusé.
En d'autres termes, outre les entreprises ayant un fort poids dans l'agrégat, le macro-contrôle permet aussi d'identifier les unités qui ont une évolution très différente de celle de leur secteur et qui impactent cette dernière. Ceci suppose que l'évolution du secteur est stable afin de repérer les évolutions atypiques.
Or, dans certains cas, les premiers macro-contrôles sont passés sur des agrégats non suffisamment stabilisés, ce qui a pu générer la mise en évidence de dossiers à contrôler à tort.

Afin de pallier ce problème, lors de la 2^{e} campagne ESANE une première phase de « repérage » des évolutions atypiques a été mise en œuvre afin de ne commencer la phase des macro-contrôles qu'une fois les agrégats jugés stabilisés.

Enfin, ce problème est encore accentué dans le cadre du contrôle des résultats de l'enquête.
En effet, contrairement aux liasses fiscales pour lesquelles le premier chargement des données permet d'avoir des agrégats suffisamment stables en termes de recouvrement (on a plus de 80% de la valeur ajoutée[9]), pour l'enquête, les questionnaires rentrent au fil de l'eau et les plus grosses entreprises ne répondent par forcément les premières. Ainsi, il est nécessaire d'attendre d'avoir un minimum de réponses validées pour pouvoir avoir des agrégats suffisamment représentatifs et robustes. Ce seuil a été fixé à 30% de questionnaires validés mais ceci peut cacher des déséquilibres sectoriels importants. Et même si un traitement des non-réponses, basé sur les réponses N-1, permet d'obtenir un agrégat total du secteur, ce dernier reste fragile tant que les plus grosses entreprises n'ont pas répondu.

5.2 Des variables non adaptées à ce type de contrôle

Pour que les macros contrôles soient efficaces, il faut qu'il y ait :

— Soit au moins une caractéristique corrélée parmi les autres caractéristiques collectées afin de pouvoir mettre en place un macro-contrôle contemporain efficace ;
— Soit une cohérence temporelle dans la réponse à la variable pour pouvoir mettre en place un macro-contrôle temporel efficace.

Pour certaines caractéristiques telles que les investissements par exemple, aucune de ces deux conditions n'est vraiment respectée ce qui conduit à des contrôles inefficaces.

[9] Ceci vient du fait que les résultats des grosses entreprises figurent dans le fichier dit « des anticipées » car elles sont dans l'obligation de télé-déclarer, et leur traitement par la DGFiP est donc plus rapide.

Pour de telles variables un contrôle simple sur les plus gros contributeurs à l'agrégat en niveau ou la recherche de valeurs aberrantes (investissement nets fortement négatifs par exemple) semble plus pertinent.

Dans le même ordre d'idée, il est très difficile de contrôler par macro-contrôles des caractéristiques pour lesquelles l'agrégat est proche de 0 ou peut changer de signe. Des soldes, tels que l'Excédent Brut d'Exploitation par exemple, en sont le parfait exemple. Les contributions sont trop volatiles et leur interprétation devient impossible.
Pour de telles variables, il est préférable de se concentrer sur le contrôle des variables qui participent au calcul du solde et non au solde lui même.

5.3 Trop peu de réponses pour certaines variables au niveau des contrôles

Par définition, les agrégats sont d'autant plus robustes qu'ils sont calculés à partir des réponses d'un nombre important d'entreprises.
Ainsi, il devient difficile pour des enquêtes par sondage de prévoir des macro-contrôles sur certaines caractéristiques pour lesquelles le taux de réponses strictement positives est faible au niveau des strates de contrôles. L'exemple le plus marquant est la variable « Personnel prêté à l'entreprise » pour laquelle la médiane du nombre de réponse par groupe d'APE est 23.
Dans ces cas en effet, toutes les unités sont à contrôler ou aucune. Le « bon » seuil de contrôle devient quasiment impossible à trouver.
Là encore, pour ce genre de variables, des micro-contrôles cherchant à repérer une réponse aberrante semble être un processus plus efficace.

5.4 Difficultés concernant l'identification et l'ordre des entreprises à contrôler

Le pilotage du contrôle manuel des dossiers par les macro-contrôles a pour but d'identifier les entreprises à contrôler afin de s'assurer d'un niveau de qualité minimum pour chaque variable.
L'hypothèse de départ dans RESANE était donc de définir un niveau de qualité minimum (mesuré en termes de variance) qui se traduisait par la définition d'un seuil pour chaque macro contrôle [4]. Par la suite, la priorisation des dossiers à contrôler se faisait à partir de deux leviers :

- une pondération des variables (cf. Ki décrit au paragraphe 2.3) devait permettre de favoriser le contrôles de variables jugées plus stratégiques ;
- des macro-contrôles lancés à des niveaux de moins en moins agrégés. Une première vague de macro-contrôle était lancée au niveau groupe de la NAF pour permettre une diffusion anticipée de résultats à ce niveau d'agrégation. Une deuxième vague était ensuite calculée au niveau sous-classe et groupe de la NAF croisé avec des tranches de taille pour la diffusion définitive des résultats.

Ce pilotage « idéal » de la production par la qualité n'a toutefois pas pu être mis en œuvre pour les raisons suivantes.

5.4.1 Trop de dossiers à contrôler a priori

A cause du retard pris lors du lancement de la première campagne mais également à cause de la sous-estimation du nombre de réponses incomplètes à l'ESA (essentiellement au niveau de la ventilation en branches), la ressource disponible pour les contrôles ne permettait de vérifier tous les dossiers identifiés a priori. C'est donc plus la ressource disponible qui a piloté les contrôles que le niveau de qualité souhaité au départ. Les seuils des macro-contrôles ont donc été corrigés afin de répondre à la contrainte liée aux ressources disponibles.
De plus, il y avait beaucoup trop de variables à contrôler au niveau de la liasse fiscale. Le nombre de contrôles était trop important et il a donc été nécessaire de limiter le nombre de variables contrôlées (certains Ki ont été mis à 0) d'une part et de supprimer certains macro-contrôles jugés peu efficaces d'autre part.

5.4.2 Difficulté pour classer les dossiers par ordre de priorité de traitement

L'algorithme retenu pour la priorisation des dossiers à contrôler favorisait le contrôle des dossiers présentant plusieurs problèmes plutôt que les dossiers ayant une seule erreur quel qu'en soit son importance. En effet, une fois qu'une contribution dépasse le seuil le plus élevé, elle « compte » pour autant quelle que soit la valeur de sa contribution.
Cette logique se comprend parfaitement sachant que l'objectif des contrôles n'était pas uniquement de repérer des entreprises à corriger mais aussi d'obtenir des informations qualitatives sur les raisons d'un éventuel comportement atypique des unités. Il devenait alors logique de traiter en priorité des entreprises ayant un comportement « globalement » atypique.
Toutefois, à partir du moment où la ressource ne permet plus le contrôle de l'ensemble des unités qui auraient dû l'être, ce principe de priorisation pose problème car il peut conduire à ne pas corriger une unité ayant une seule valeur aberrante qui n'aurait pas été redressée automatiquement. Or la priorité, avant d'obtenir d'éventuelles raisons expliquant l'évolution atypique de tel ou tel agrégat est bien d'obtenir un agrégat correct.
Ce problème s'est encore accentué pour la validation des données définitives. En effet, pour la liasse fiscale, le nombre de variables à contrôler est tel (plus d'une centaine) que la probabilité de passer à côté d'une valeur aberrante devient non négligeable. C'est d'ailleurs ce qui a pu être observé au cours de la première campagne où certaines liasses d'entreprises ont dû être corrigées manuellement, hors chaîne de traitement, car elle comprenait une variable atypique et une seule.
Afin de pallier ce problème, une nouvelle formule de l'indicateur de priorité de traitement va être testée cette année. Elle prend en compte « la taille de l'erreur » et sa formule devient (cf. paragraphe 2.3 pour la signification des différentes variables) :

$$QEG = \frac{\sum_{\text{var i}} K_i . \text{score\_max}(X_i) . 1\!\!1_{\{E(X_i)=S \text{ ou } I\}}}{\sum_{\text{var i}} K_i}$$

où score_max(X_i) est le score maximal obtenu par la variable dans les différents macro-contrôles qui la concerne. Les scores des différents macro-contrôles seront, bien entendu, normalisés afin d'être comparables entre eux.

5.4.3 Les problèmes liées à l'estimateur composite [6]

Le nouvel estimateur retenu pour le calcul des agrégats donne la part belle à certaines entreprises plus qu'à d'autres.
Ainsi par exemple, les entreprises qui changent d'APE ou qui ont des corrections importantes au moment de la réconciliation des données vont peser fortement sur le calcul des agrégats. Il conviendrait donc de plus vérifier ces entreprises.
Or si le poids de sondage est bien pris en compte pour le calcul des contributions des caractéristiques de l'ESA, il n'en va pas de même pour l'examen des caractéristiques fiscales. Ceci pose des problèmes car dans certains cas, la contribution de la réponse non-pondérée d'une entreprise est dérisoire, alors que son impact dans le calcul de l'agrégat définitif est très important. L'exemple typique est celui d'une petite entreprise quittant un secteur dans lequel le taux de sondage est faible.
Là encore, des expérimentations vont être menées dans les campagnes à venir afin de favoriser, au moment du contrôle des liasses fiscales, les unités qui sont dans la partie échantillonnée de l'enquête ESA.

Références

[1] ***Esane, le dispositif rénové de production des statistiques structurelles d'entreprises,*** P. Brion, Courrier des statistiques n°130, 2011

[2] ***Pour une nouvelle génération d'enquêtes annuelles d'entreprise***, E. Raulin, Courrier des statistiques n°64, 1992

[3] ***Enquêtes annuelles d'entreprises à la recherche du 4e type***, P. Rivière, Courrier des statistiques n°78, 1996

[4] ***Setting cut off scores for selective editing in structural business statistics: an automatic procedure using simulation study***, E. Gros, UN/ECE work session on statistical data editing, 2009

[5] ***CITRUS - Système d'information sur les restructurations d'entreprises*** , M. Beauvois, Courrier des statistiques n°95-96, 2000

[6] ***L'utilisation combinée de données d'enquête et de données administratives pour la production des statistiques structurelles d'entreprises,*** P. Brion, papier présenté aux JMS de l'Insee, 2009