Esane, le dispositif rénové de production des statistiques structurelles d'entreprises

Philippe Brion, chef du département Répertoires, Infrastructures et Statistiques structurelles à la direction générale de l'Insee

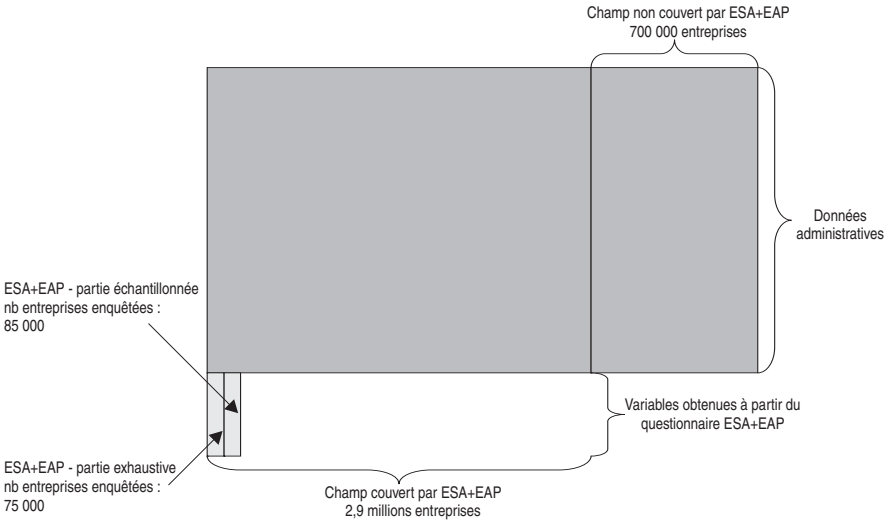
Ce papier est dédié à la mémoire d'Emmanuel Raulin qui a été l'initiateur de ce projet et a grandement contribué à sa mise en place.

L'Insee vient de publier les premiers résultats du dispositif rénové Esane (élaboration des statistiques annuelles d'entreprises) qui permet de produire les statistiques structurelles d'entreprises. La mise au point de ce dispositif a constitué un des chantiers majeurs récents de l'Institut et ce papier en décrit les principes généraux, ainsi que les innovations apportées par rapport à l'ancien système.

Les statistiques structurelles d'entreprises, photographie annuelle de la population des entreprises appartenant au système productif, constituent le matériau de base pour de nombreux utilisateurs de statistiques : statisticiens au premier rang desquels les comptables nationaux, Union européenne pour la réponse au règlement SBS (structural business statistics), cabinets ministériels, fédérations professionnelles, chercheurs, économistes, etc. En 2005, l'Insee a lancé un grand programme, le programme Resane (refonte des statistiques annuelles d'entreprises), dont l'objectif central était de redéfinir le processus de production de ces statistiques (Depoutot, 2010). La base de ce nouveau système d'information dénommé Esane (élaboration des statistiques annuelles d'entreprises) est l'utilisation renforcée des sources administratives.

L'utilisation des données administratives correspond à une tendance générale pour beaucoup d'offices statistiques nationaux, ceci principalement pour deux raisons : recherche d'économies pour l'administration au niveau de la collecte des données, allègement de la charge de réponse pour les unités interrogées. Si les données fiscales sont utilisées depuis longtemps par les statisticiens d'entreprises français, c'est le raccourcissement récent des délais de fourniture par l'administration fiscale qui a conduit à revoir le dispositif qui existait jusqu'à présent *(encadré 1)* en permettant de respecter certaines échéances importantes. Elles ont alors été intégrées dans un dispositif utilisant d'autres sources administratives ainsi que des données d'enquêtes statistiques. Cette utilisation conjointe de données administratives et de données d'enquête *(schéma)* n'est pas sans poser un certain nombre de difficultés méthodologiques.

Schéma - Esane, un dispositif multisources



ESA : enquête sectorielle annuelle
EAP : enquête annuelle de production
Note de lecture : ce schéma représente la manière dont les différentes sources sont utilisées pour produire les statistiques dans le dispositif Esane. L'axe horizontal est celui des entreprises, l'axe vertical celui des variables. Ainsi, les variables obtenues à partir du questionnaire ESA+EAP concernent les 2,9 millions d'entreprises du champ total des deux enquêtes et leur calcul se base sur les réponses de l'ensemble des 75 000 entreprises enquêtées chaque année (généralement ce sont celles de 20 salariés ou plus) et de la partie échantillonnée des entreprises (en général, des entreprises de moins de 20 salariés, soit environ 85 000 entreprises).

Les sources administratives utilisées dans Esane

La source principale est constituée par le fichier - en pratique les fichiers - des déclarations annuelles sur les bénéfices adressées chaque année, par les entreprises, à la direction générale des finances publiques (DGFiP) : BIC (bénéfices industriels et commerciaux), BNC (bénéfices non commerciaux), BA (bénéfices agricoles). L'utilisation de ces déclarations

est facilitée par le fait que les informations déclarées utilisent le plan comptable général (qui constitue aussi la référence pour les statisticiens), et par l'utilisation par la DGFiP du numéro Siren pour identifier les entreprises.

L'administration fiscale procède, au cours de l'année *n+1*, à plusieurs livraisons de fichiers concernant les données de l'exercice *n* (Chami, 2010) : une première livraison à la mi-juin, limitée en nombre d'entreprises mais relative aux plus grandes (qui représentent environ les trois quarts du chiffre d'affaires total), suivie de deux livraisons en septembre et octobre (utilisées pour la production des résultats finaux), et une dernière en janvier *n+2*, qui, si elle n'est pas utilisée directement pour la production des résultats, est intégrée dans la base de données finale pour consolider la production de l'année suivante.

La source fiscale est complétée par les déclarations annuelles de données sociales (DADS) qui fournissent des données sur les effectifs employés et leurs rémunérations. Il a également été envisagé d'utiliser les déclarations douanières des entreprises, mais, au moment de la rédaction de cet article, des incertitudes pèsent sur la pertinence de l'opération en raison de l'importance des échanges intra-groupes au niveau international qui nécessite de mettre en place un dispositif d'une nature différente.

Une enquête statistique utilisée en complément

Les informations disponibles dans les sources administratives ne sont pas suffisantes à elles seules pour répondre à l'ensemble des besoins des utilisateurs de statistiques structurelles d'entreprise. Par exemple, l'information sur l'emploi non salarié n'est pas disponible dans les DADS ; d'autres variables non présentes dans les fichiers administratifs sont également demandées par les utilisateurs. Il est donc nécessaire de mener une enquête statistique, réalisée sur un échantillon d'entreprises (Haag, 2010) : c'est l'ESA, enquête sectorielle annuelle, qui a pris le relais des anciennes EAE (enquêtes annuelles d'entreprises). Des informations dites « sectorielles », relatives à ces données non disponibles dans les sources administratives, sont demandées *via* le questionnaire de l'enquête et sont, pour certaines, adaptées au secteur étudié : par exemple données sur l'équipement commercial dans le secteur du commerce, sur le type de clientèle pour le commerce et les services, ou encore sur la nature des ouvrages réalisés pour la construction.

Par ailleurs, une information est très importante dans le cadre d'un dispositif destiné à produire des statistiques sectorielles : celle qui permet de classer une entreprise dans un secteur d'activité donné. Rappelons la différence entre secteur et branche : une branche - branche d'activité - regroupe des unités

Encadré 1 - Le dispositif antérieur, et pourquoi fallait-il le modifier

On avait auparavant deux dispositifs fonctionnant en parallèle (Grandjean, 1996).

D'une part, les enquêtes annuelles d'entreprise (EAE : voir Raulin, 1992 ; Rivière, 1996) étaient menées sur un échantillon d'entreprises enquêtées annuellement via un questionnaire envoyé par courrier. Ces enquêtes étaient réalisées par différents services statistiques (Insee pour les secteurs du commerce et des services, et services statistiques en charge des domaines de l'agriculture pour les industries agro-alimentaires, de l'industrie pour l'industrie manufacturière, et de l'équipement pour les transports et la construction). Elles permettaient de récupérer de l'information sur un certain nombre de variables comptables, en particulier du compte de résultat, sur l'emploi, les investissements, les restructurations et sur un certain nombre de sujets sectoriels - le questionnaire étant donc décliné selon des versions différentes dépendant du secteur auquel on s'intéressait.

D'autre part, un dispositif s'appuyant sur les déclarations annuelles de bénéfices des entreprises existait : Suse (système unifié de statistiques d'entreprises). Il consistait en une mise en commun des fichiers des EAE et en une fusion/confrontation avec les données contenues dans les fichiers fiscaux. Les EAE étant réalisées sur un échantillon, la base de données Suse obtenue était non homogène quant à la richesse des informations concernant chaque entreprise : pour celles enquêtées par l'EAE, on tenait compte de l'actualisation de son code d'activité principale (APE) au travers de l'observation de la répartition de son chiffre d'affaires selon ses différentes activités, alors que pour les autres l'information se limitait à celle venant de la source fiscale ou de Sirene.

Un dernier dispositif, appelé SIE (système intermédiaire d'entreprise) servait à assurer le passage aux comptes nationaux, au travers d'un ensemble de retraitements.

Si la cohabitation des deux dispositifs EAE et Suse était consommatrice de moyens puisqu'on menait en parallèle un travail de contrôle sur deux sources, on se trouvait également face à l'existence de résultats non forcément identiques sur le même sujet : on avait ainsi des nombres d'entreprises par secteur différents selon les deux sources (et même la possibilité d'un troisième chiffrage, à partir de comptages réalisés sur le répertoire Sirene). Si la source Suse avait l'avantage de l'exhaustivité (aux données absentes des sources fiscales près, cependant), elle souffrait du fait qu'aucune inférence n'était réalisée sur la partie non enquêtée par l'EAE : outre le fait que le code APE des entreprises de cette partie n'était pas actualisé, d'autres données faisaient l'objet d'approximations - en particulier, les petites entreprises étaient considérées comme mono-actives (c'est-à-dire ayant une seule branche d'activité), qu'elles soient enquêtées ou pas par l'EAE, d'ailleurs.

La raison pour laquelle ces deux dispositifs parallèles avaient été mis en place était que la date de disponibilité de la source fiscale était trop tardive pour répondre à un certain nombre de demandes ; c'était le cas en particulier pour celle concernant les comptes nationaux, ou encore celle concernant le règlement européen SBS - structural business statistics - qui demande des résultats préliminaires sur l'année n avant la fin octobre n+1. Cette date de disponibilité ayant été considérablement avancée[1], il était possible de rénover le système, et par là-même de l'unifier en un système cohérent[2]. Et tout en cherchant cette mise en cohérence des différentes sources (par rapport à l'ancien système où on juxtaposait les résultats des deux sources), on a également visé un raccourcissement des délais de production, en se fixant comme objectif de produire des résultats définitifs, pour une année n, à la fin de l'année n+1.

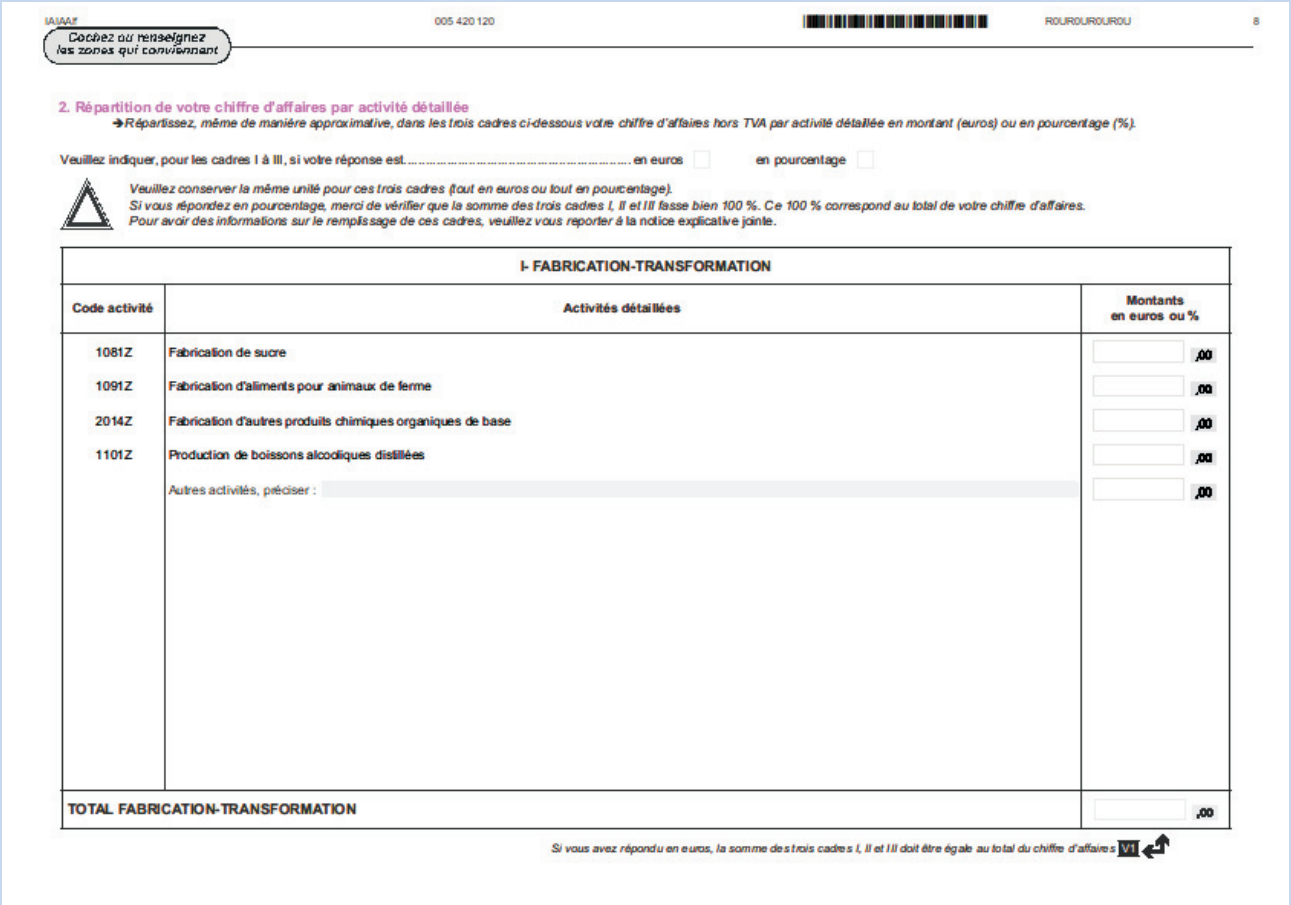
1. En parallèle, la qualité de la « sirénisation » des fichiers fiscaux a progressé, rendant l'utilisation de ces derniers par les statisticiens plus facile.
2. Le travail de contrôle redressement des données a été également centralisé, au sein des directions régionales de l'Insee des Pays de la Loire et de Basse Normandie, alors qu'il était réparti sur plusieurs ministères auparavant.

de production homogènes, c'est-à-dire qui fabriquent des produits (ou rendent des services) qui appartiennent au même item de la nomenclature d'activité économique considérée. Au contraire, un secteur regroupe les entreprises possédant la même activité principale. En France, la détermination de l'activité principale des unités légales est de la responsabilité des statisticiens : c'est l'APE, activité principale exercée, qui se réfère à la nomenclature d'activités françaises (NAF). On pourrait penser qu'utiliser le code d'activité principale disponible dans le répertoire Sirene, est suffisant pour produire les statistiques structurelles ; il n'en est rien car la valeur du code dans le répertoire peut être ancienne, et donc obsolète si l'entreprise a procédé à une réorientation de ses activités depuis sa création par exemple, sans que la valeur du code APE dans le répertoire ait été nécessairement revue. On demande donc à chaque entreprise interrogée dans l'enquête statistique de fournir la répartition de son chiffre d'affaires selon ses différentes activités élémentaires *(encadré 2)*. En effet, beaucoup d'entreprises ne sont pas mono-actives et peuvent avoir plusieurs activités distinctes produisant différents produits, ou encore combiner des activités de fabrication et des activités de commerce ou de services. C'est à partir de l'information fournie par l'entreprise que l'on réévalue son code APE selon un algorithme étudiant la part de chaque poste. Les statistiques sectorielles seront fondées sur ce classement de l'entreprise, découlant d'une observation réalisée au moment de l'enquête, et non sur son classement résultant du code APE disponible dans le répertoire.

Encadré 2 - La ventilation du chiffre d'affaires

Dans le questionnaire de l'enquête statistique ESA (enquête sectorielle annuelle), on demande aux entreprises de détailler le montant de leur chiffre d'affaires selon les différentes activités qu'elles exercent. Ceci se fait au travers de trois tableaux à remplir, l'un ayant trait aux activités de fabrication, un autre aux activités commerciales et le troisième aux activités de services.

L'exemple fourni ci-dessous est relatif à certaines entreprises de l'agro-alimentaire et donne le cadre dédié aux activités de fabrication - transformation. Un certain nombre de lignes sont pré-imprimées et correspondent aux activités exercées traditionnellement par les entreprises de cette catégorie (ceci fait que le questionnaire des enquêtes ESA et EAP est décliné selon une variété de modèles adaptés à chacune de ces catégories). Mais une entreprise peut exercer une activité non indiquée dans la liste ; dans ce cas elle indique le montant correspondant et le descriptif de l'activité en question dans la ligne « autres activités, préciser ». Lorsque l'entreprise remplit cette ligne, il devient nécessaire de « coder » l'activité correspondante, travail qui peut être relativement chronophage pour les statisticiens.



IAIAAI 005 420 120 ROUROUROUROU 8

Cochez ou renseignez les zones qui conviennent

2. Répartition de votre chiffre d'affaires par activité détaillée

➔Répartissez, même de manière approximative, dans les trois cadres ci-dessous votre chiffre d'affaires hors TVA par activité détaillée en montant (euros) ou en pourcentage (%).

Veuillez indiquer, pour les cadres I à III, si votre réponse est..... en euros ☐ en pourcentage ☐

Veuillez conserver la même unité pour ces trois cadres (tout en euros ou tout en pourcentage).
Si vous répondez en pourcentage, merci de vérifier que la somme des trois cadres I, II et III fasse bien 100 %. Ce 100 % correspond au total de votre chiffre d'affaires.
Pour avoir des informations sur le remplissage de ces cadres, veuillez vous reporter à la notice explicative jointe.

I- FABRICATION-TRANSFORMATION

Code activité	Activités détaillées	Montants en euros ou %
1081Z	Fabrication de sucre	,00
1091Z	Fabrication d'aliments pour animaux de ferme	,00
2014Z	Fabrication d'autres produits chimiques organiques de base	,00
1101Z	Production de boissons alcooliques distillées	,00
	Autres activités, préciser :	,00
TOTAL FABRICATION-TRANSFORMATION		,00

Si vous avez répondu en euros, la somme des trois cadres I, II et III doit être égale au total du chiffre d'affaires VI

La ventilation du chiffre d'affaires est également fondamentale pour l'élaboration des comptes nationaux, celle-ci passant par la réalisation de comptes de branches qui nécessitent d'avoir une information sur chacune des activités de l'entreprise. La partie du questionnaire de l'enquête statistique relative à cette ventilation constitue donc le « cœur » de l'enquête.

Une spécificité doit être mentionnée pour le secteur de l'industrie manufacturière (hors industries agro-alimentaires) : sur ce champ, l'enquête sert à la fois au dispositif Esane et d'enquête de production, ceci afin de répondre au règlement européen Prodcom sur les produits. L'enquête pose des questions à un niveau de détail plus fin que celle réalisée sur les autres secteurs. Elle est baptisée EAP (enquête annuelle de production). L'EAP utilise une collecte par internet, alors que l'ESA est réalisée pour l'instant par voie postale.

Le champ du dispositif

Le dispositif permet de produire des statistiques sur le champ des entreprises marchandes, à l'exception des entreprises du secteur financier (observé par l'Autorité de contrôle prudentiel) et des exploitations agricoles (couvertes par de nombreuses enquêtes gérées par le service statistique du ministère de l'Agriculture). Ce champ est essentiellement défini à partir de codes de la nomenclature d'activités NAF.

Cependant, il existe une partie du champ du dispositif sur lequel on se contente d'utiliser les données administratives sans réaliser une enquête en complément : il est constitué, pour l'essentiel, de secteurs de services, particulièrement dans les domaines de la santé, de l'éducation et de l'action sociale (pour leur partie considérée comme marchande), sur lesquels il est fait l'hypothèse que les données administratives sont suffisantes pour les statistiques qu'on cherche à produire. On considère en particulier implicitement que les entreprises de ces secteurs ne sont pas pluriactives[3]. En termes de valeur ajoutée, le poids de ces entreprises est de l'ordre de 10 %.

Sur l'ensemble des 3,6 millions d'entreprises concernées par le dispositif, environ 2,9 millions appartiennent au champ sur lequel on réalise l'enquête statistique.

Les premiers résultats publiés

Les premiers résultats produits ont concerné les données de l'année 2008 et ont été publiés sur le site web de l'Insee en 2010. Ils concernent les agrégats sectoriels pour une dizaine de variables[4]. Ces premiers résultats 2008 ont été complétés par des agrégats du même type calculés pour d'autres variables, de nature comptable, et publiés dans la base de données Alisse, disponible également sur le site web de l'Insee.

Le traitement des données de l'année 2008 n'a cependant pas permis de travailler aux niveaux de finesse et de complétude des variables prévus pour la diffusion courante. En effet, on a été conduit à ne travailler, pour la première année de mise en production, qu'au niveau « trois caractères » de la NAF pour le contrôle des données décrit plus loin dans cet article, donc à ne procéder à des vérifications des statistiques qu'à ce niveau de diffusion[5]. D'autre part, des arbitrages ont dû être rendus au niveau du développement et de la mise en production des différentes parties du système et ont conduit à reporter à l'année suivante la production de statistiques à partir des variables du questionnaire de l'enquête autres que la ventilation du chiffre d'affaires en branches (qui, elle, est essentielle pour le classement sectoriel des entreprises), par exemple certaines dépenses professionnelles particulières à un secteur donné, ou le type de clientèle.
Les résultats diffusés pour l'année 2009 et les suivantes seront donc plus complets. Il faut noter en particulier que de premiers résultats 2009 ont été transmis à Eurostat à la fin du mois d'octobre 2010 pour les résultats de l'année 2009.

Un projet complexe et innovant

Le programme Resane comprend de nombreux niveaux d'innovations et a dû prendre en compte de manière simultanée des questions d'ordre statistique, des aspects organisationnels et le développement d'un logiciel informatique associé[6]. Un choix fondamental a été opéré dès le départ : le principe de modularité, qui consiste à traiter les différentes sources selon des sous-processus indépendants, ceci afin de ne pas être trop dépendant des options prises dans les différentes sources et de pouvoir prendre en compte plus facilement un changement au niveau d'une de ces sources. Par exemple, un module pour traiter les données du questionnaire de l'enquête statistique ESA, un autre pour traiter quelques variables « principales » de la liasse fiscale, un autre encore - se déroulant plus tard - pour traiter le reste des variables de la liasse ont été définis et conçus. Ce principe de modularité est également lié aux deux questions suivantes : d'une part, les données disponibles dans les différentes sources n'arrivent pas au même moment, ce qui induit de facto un découpage temporel du travail des gestionnaires ; d'autre part, la livraison de résultats à différentes dates nécessite de prioriser les travaux relatifs au contrôle-redressement des données.

3. Plus précisément, on considère que la seule distinction de branches que l'on peut faire est celle qui est contenue dans la liasse fiscale, laquelle distingue les ventes de marchandises (activité commerciale), les ventes de biens et celles de services (production de biens ou production de services).
4. Les résultats sont disponibles à l'adresse http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=esane.
5. Ceci signifie qu'on n'étudie l'impact d'une donnée jugée possiblement en erreur que sur des statistiques publiées à ce niveau de diffusion.
6. Développé au centre informatique de l'Insee de Nantes, et s'appuyant sur le langage de spécification d'enquêtes (L.S.E.) mis au point à l'Insee (Bouchet et al., 2011).

Des méthodes de contrôle rénovées

Le contrôle des données est un processus coûteux, beaucoup d'informations disponibles au travers de l'enquête statistique ou des sources administratives déclenchant des « signaux » d'erreur potentielle au travers des batteries de tests mises en place. Pour le processus Esane, qui traite un volume d'informations très important - de l'ordre de plusieurs milliards -, il a été décidé de mettre en place des contrôles situés en amont et opérant un tri sur ces informations suspectes conduisant à un classement de celles-ci dans deux groupes : données pour lesquelles l'erreur potentielle est jugée très impactante sur les statistiques calculées (et qui nécessite donc l'expertise manuelle d'un gestionnaire), données pour lesquelles on considère qu'un traitement plus fruste (sous forme de corrections « automatiques ») est suffisant pour la qualité des statistiques qu'on produit. Ces contrôles sont dits « macro-contrôles » au sens où ils ne procèdent pas à l'examen de la cohérence interne des données d'une seule entreprise, mais plutôt à la prise en compte d'un ensemble de données vues comme participant à l'élaboration globale de statistiques. Ils nécessitent de définir a priori le niveau de finesse qu'on se fixe pour la diffusion des résultats (Gros, 2009)[7]. Des priorités sont ainsi affectées aux unités détectées comme à expertiser manuellement, ce qui permet de piloter la charge de travail des équipes de gestionnaires chargées de ce traitement. Concrètement, pour une variable à contrôler, on calcule une statistique (par exemple, le taux d'évolution de l'agrégat relatif à la variable par rapport à celui de l'année précédente) avec l'ensemble des questionnaires (ou des données administratives) disponibles, et on calcule la même statistique avec tous les questionnaires sauf un : on pointe ainsi les questionnaires avec des données jugées suspectes par le fait que l'écart entre les deux estimations est grand (supérieur à un seuil spécifié par les méthodologues). Pour ces données jugées suspectes, le gestionnaire procède à une expertise manuelle, en contactant s'il le faut l'entreprise, ce qui va permettre soit de corriger une erreur, soit d'avérer une donnée jugée «étrange».

Le travail des deux équipes de gestionnaires, l'une située à la direction régionale de Basse Normandie et travaillant sur les données de l'EAP, l'autre située à la direction régionale des Pays de la Loire et ayant en charge le reste du processus, est donc ciblé sur les données « pointées » par ces macro-contrôles. Il peut conduire à plusieurs types de rappel de l'entreprise : pour compléter la réponse fournie à l'enquête statistique dans le cas d'une donnée manquante, en cas de donnée jugée suspecte, ou encore dans le cas d'un code APE de l'entreprise réévalué par l'enquête et ayant changé par rapport à la valeur du code disponible dans le répertoire. Ce travail s'échelonne tout au long de l'année, et une même entreprise peut être appelée plusieurs fois par les gestionnaires pour des demandes de précisions sur leurs données, par exemple au moment de l'examen des données de l'enquête statistique (les retours de questionnaires pouvant débuter en février), puis quand les fichiers des données fiscales sont traités (le fichier contenant la majorité des liasses fiscales est disponible en octobre).

Pour les équipes de gestionnaires, le passage au dispositif Esane a donc constitué une charge importante, en particulier du fait de la complexité d'un système multi-sources, et a nécessité la tenue de nombreuses sessions de formation centrées sur chacun des processus dans le dispositif. Un poste de travail spécifique dédié à l'application informatique servant à traiter les données a été mis au point et l'organisation générale du travail de l'équipe de Nantes, complexe du fait du traitement « multi-sources », a fait l'objet d'une approche Maiol (maîtrise d'œuvre en organisation locale) rassemblant différents représentants des équipes du dispositif précédent et de l'équipe de développement du projet, ainsi que des spécialistes de l'organisation et de la mise au point de projets.

Par ailleurs, les restructurations d'entreprises font l'objet d'un traitement spécifique et complexe prenant en compte, au niveau du contrôle des données, « l'enveloppe » de restructuration (à savoir l'ensemble constitué par les entreprises concernées), ceci afin de tenir compte de phénomènes liés à des flux entre unités pouvant apparaître, ou disparaître, d'une année sur l'autre. La bonne réalisation de ce traitement est primordiale afin de produire des statistiques en évolution qui soient pertinentes.

La mise en cohérence des données administratives et des données d'enquête

Un module dit de « réconciliation » de données individuelles permet de confronter les valeurs de variables connues à la fois par l'enquête statistique et par les sources administratives. En effet, l'utilisation conjointe de données d'enquête et de données administratives est un des points forts du dispositif, 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. C'est sur la valeur du chiffre d'affaires, ainsi que sur sa répartition en activités, que le travail dit de « réconciliation des données individuelles » provenant de ces différentes sources est opéré. Il y a également une réconciliation des données relatives à l'emploi, comparant la source sociale (venant des DADS) et la source fiscale.

7. Ceci constitue une innovation forte du système : le travail de contrôle des données est défini en fonction d'un niveau d'exigence fixé pour une statistique donnée : selon que ce niveau d'exigence est plus ou moins élevé (par exemple volonté de produire des résultats fiables au niveau trois caractères de la NAF ou au niveau cinq caractères), la quantité de travail nécessaire s'en trouvera modifiée.

Sur ce point comme pour le contrôle des données source par source, 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, par exemple si l'écart de chiffre d'affaires entre les deux sources est supérieur à x % du chiffre d'affaires du secteur.

Cependant, ce travail de réconciliation ne peut être mené sur l'ensemble des entreprises : il est, pour sa partie relative au chiffre d'affaires et à sa ventilation en activités, nécessairement limité aux entreprises faisant partie de l'échantillon des enquêtes ESA et EAP. Or, si toutes les entreprises d'une certaine taille (en général, celles ayant au moins 20 salariés) appartiennent à l'échantillon, ce n'est pas le cas pour la population des petites entreprises, pour laquelle les taux de sondage sont faibles (pouvant être inférieurs au 1/50e pour certaines catégories). Il faut donc procéder à de l'inférence sur l'ensemble de la population des entreprises à partir de ce que l'on a observé sur l'échantillon.

Des procédures d'estimation statistique complexes

La production d'estimations statistiques à partir d'un dispositif multi-sources n'est pas immédiate. D'une part, on dispose de données plus ou moins riches selon que l'entreprise appartient à l'échantillon de l'ESA ou de l'EAP ou pas (auquel cas on ne dispose que de ses données administratives). D'autre part, comme il a été indiqué précédemment, pour certaines variables, on dispose de valeurs dans les deux sources : le cas du chiffre d'affaires et de sa ventilation vient d'être présenté, mais la situation est du même type pour le classement sectoriel. En effet, pour celui-ci on dispose du code APE du répertoire et de la valeur réévaluée à l'occasion de l'enquête ; c'est cette dernière qui doit être utilisée pour produire des statistiques sectorielles, et il est donc nécessaire de tenir compte des changements sectoriels observés sur l'échantillon ESA+EAP pour les « extrapoler » à l'ensemble des entreprises. Ceci se fait au travers de méthodes d'estimation statistiques assez complexes, proposant des estimateurs combinés ainsi qu'un calage de l'échantillon destiné à « ajuster » celui-ci sur la population totale des entreprises afin de rendre cohérentes les estimations utilisant des données recueillies uniquement sur échantillon et celles produites à partir des données administratives *(encadré 3)*.

Un allègement de la charge statistique et du coût global de l'enquête

Le programme Resane a été lancé en particulier pour répondre à l'objectif d'alléger la charge statistique pesant sur les entreprises (Béguin, 2009), demande récurrente de la part de celles-ci : à partir du moment où une donnée était disponible dans une source administrative, elle ne devait plus faire l'objet d'une demande via un questionnaire statistique.

La réduction de la charge profite aussi à l'administration puisque les coûts d'enquête sont réduits ; c'était d'ailleurs un autre objectif important en soi du programme de refonte. On estime que, pour les entreprises, le nombre de variables demandées a été réduit environ de

Encadré 3 - Les estimateurs statistiques

Produire des estimations statistiques à partir d'un matériau composite (données d'enquête recueillies sur échantillon et données administratives en principe exhaustives[8]) n'est pas facile ; de plus, il faut intégrer dans les estimations l'apport de l'enrichissement mutuel des deux sources dû à leur confrontation sur quelques variables. Des études méthodologiques ont donc été menées pour proposer des méthodes d'estimation adaptées (Brion, 2007).

Pour simplifier, on peut s'intéresser à la production de statistiques relatives aux chiffres d'affaires des différents secteurs. On pourrait se contenter de faire la somme des chiffres d'affaires (fournis par les données fiscales) des entreprises classées, selon leur code APE disponible dans Sirene, dans un secteur *X* donné :

$$\sum_{i\,/\,APESIRENE=X} CA(i)$$

Mais l'enquête révèle un certain nombre de mauvais classements : des entreprises classées dans Sirene dans le secteur *X* ont, pour l'année considérée, une activité différente et, au contraire, certaines entreprises classées dans Sirene hors du secteur *X* se révèlent appartenir à ce secteur. Il va donc falloir évaluer les montants de chiffres d'affaires correspondant à ces « départs » et à ces « arrivées ». Ceci va se faire grâce à l'échantillon, dans lequel à chaque unité enquêtée est affecté un poids de sondage .

L'estimation du chiffre d'affaires du secteur sera donc :

$$\sum_{i\,/\,APESIRENE=X} CA(i) + \sum_{arrivantdansX} w_i CA(i) - \sum_{sortantdeX} w_i CA(i)$$

Le même type d'estimation peut être proposé pour d'autres variables (Brion, 2009).

On doit également appliquer la méthode aux divergences constatées lors de la phase de réconciliation des données individuelles (sur le chiffre d'affaires ou sa ventilation, par exemple) : cette phase consiste à confronter la valeur disponible dans la source fiscale, par exemple, avec celle disponible dans le questionnaire d'enquête, et en cas de divergence, à proposer une valeur arbitrée. Les divergences observées sur l'échantillon doivent être « répercutées » sur l'ensemble des entreprises, et ceci se fait au travers des poids de sondage.

La réalité est plus complexe, car il faut également tenir compte des informations manquantes, celles-ci correspondant à des unités non répondantes à l'enquête, mais également à des « trous » dans les données administratives

Enfin, une procédure de calage, consistant à travailler sur les poids d'échantillonnage, permet d'assurer la cohérence entre des estimations produites à partir des données du seul échantillon et les estimations venant des estimateurs composites présentés ci-dessus.

8. Cette exhaustivité n'est que théorique. Le champ du dispositif est défini *a priori*, à partir du répertoire Sirene, mais des déclarations sont manquantes quand on reçoit les fichiers de la DGFIP, pour différentes raisons. Il faut donc, pour chacune d'entre elles, décider si on la considère manquante car n'ayant plus d'activité économique, ou « réellement » manquante (par exemple parce qu'elle est encore en cours de traitement par la DGFiP au moment où celle-ci nous transmet le fichier).

moitié par rapport à l'ancien système : cette réduction résulte à la fois d'un allègement du questionnaire de l'enquête statistique, mais également d'un travail d'optimisation de l'échantillon (Bauer, Brilhault, Gros, 2009). Cette optimisation, rendue possible par l'utilisation conjointe de données d'enquête et de données administratives, a conduit à diviser par deux la taille de l'échantillon pour la partie échantillonnée du champ de l'enquête - les seuils de taille au-dessus desquels l'enquête est exhaustive ont été maintenus, la partie échantillonnée concernant grosso modo la moitié de l'échantillon, la réduction globale de l'échantillon est donc de l'ordre du quart. La quantité de travail de contrôle de ces données a, de facto, également été réduite.

Un chantier à venir : la prise en compte de la dimension « groupes »

Au sein du programme Resane, il est prévu une deuxième phase de rénovation : celle qui consiste à prendre en compte les structures complexes existant entre les sociétés pour donner plus de pertinence aux statistiques publiées. Celles-ci sont essentiellement situées dans les groupes d'entreprises. Elles posent différents types de problèmes qui ne seront qu'effleurés dans cet article : flux internes entre sociétés qui peuvent conduire à des doubles comptes ; classement de certaines sociétés (comme des bureaux d'étude) à l'heure actuelle dans le secteur correspondant à leur code APE, alors qu'elles travaillent exclusivement pour les sociétés du groupe auquel elles appartiennent. Ces questions doivent être traitées grâce à des méthodes dites de « profilage » (Hecquet, 2010), qui consuisent à définir des « entreprises » différentes des unités utilisées à l'heure actuelle dans le système (qui sont les unités légales enregistrées dans le répertoire Sirene). Plus précisément, ces entreprises redéfinies sont des regroupements d'unités légales jouissant d'une certaine autonomie de décision, notamment pour l'affectation de leurs ressources courantes.

L'Insee met en place à l'heure actuelle une équipe dédiée à ce traitement qui devrait conduire dans les années qui viennent à une lecture rénovée de l'économie, telle que proposée par les statistiques structurelles d'entreprises. ■

Bibliographie

✓ **Bauer P., Brilhault G., Gros E.**, 2009 : *Le plan de sondage de l'ESA*, papier présenté aux Journées de méthodologie statistique (JMS) de l'Insee.

✓ **Béguin**, 2009 : *Refonte des statistiques structurelles d'entreprises en France : comment réduire la charge de réponse des entreprises (entre autres objectifs) ?*, papier présenté à la 57e réunion plénière de la conférence des statisticiens européens, Conseil économique et social des nations Unies, 8-10 juin 2009.

✓ **Bouichet A., Chami S., Haag O., 2011** : *The LSE : a standardized specification language for statisticians*, papier présenté à la conférence NTTS de février 2011, Bruxelles.

✓ **Brion Ph.**, 2007 : *Redesigning the French structural business statistics, using more administrative data*, papier présenté à la conférence ICESIII, Montréal.

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

✓ **Chami S.**, 2010 : *Reengineering French structural business statistics : an extended use of administrative data,* papier présenté à la conférence Q2010, Helsinki.

✓ **Depoutot R.**, 2010 : *Reengineering French structural business statistics : an overview*, papier présenté à la conférence Q2010, Helsinki.

✓ **Grandjean J.-P.**, 1996 : « Le système statistique d'entreprises », in *Courrier des statistiques* n°78.

✓ **Gros E.**, 2009 : *Setting cut off scores for selective editing in structural business statistics : an automatic procedure using simulation study*, papier présenté à la conférence sur le data editing organisée par la Commission statistique des nations Unies à Neuchâtel en octobre 2009.

✓ **Haag O.**, 2010 : *Rengineering French structural business statistics : redesign of the annual survey*, papier présenté à la conférence Q2010, Helsinki.

✓ **Hecquet V.**, 2010 : *Le profilage et son impact sur la représentation de l'appareil productif*, papier présenté au colloque de l'Association de comptabilité nationale.

✓ *Nomenclatures d'activités et de produits françaises, Nafrév2 - CPF rév2,* édition 2008, Insee.

✓ **Raulin E.**, 1992 : « Pour une nouvelle génération d'enquêtes annuelles d'entreprise », in *Courrier des statistiques* n° 64.

✓ **Rivière P.**, 1996 : « Enquêtes annuelles d'entreprises : à la recherche du 4e type », in *Courrier des statistiques* n° 78.