PREMIÈRES RÉFLEXIONS ET ANALYSES SUR LES MÉTHODES DE TRAITEMENT DES DONNÉES DANS LES ENQUÊTES ANNUELLES D'ENTREPRISES

Dominique BONNANS - Emmanuel RAULIN

Introduction

Les travaux présentés ci-dessous s'inscrivent dans le cadre du projet de 4e génération des Enquêtes Annuelles d'Entreprises (EAE) ; un des axes de ce projet est, en effet, de doter toutes les EAE d'un même logiciel de traitement des données, incluant les phases de contrôle-apurement (avec recontact éventuel de l'entreprise), redressement et extrapolation. Il faut rappeler que les EAE c'est environ 230 000 entreprises interrogées tous les ans, plus de 100 données par questionnaire traitées par à peu près 200 gestionnaires mobilisés pour cette seule opération. Bien entendu, le travail de contrôle-apurement des données, le recontact d'entreprises pour cause de données manquantes ou suspectes représente la tâche principale des gestionnaires, au moins quant au temps qui y est consacré.

D'autre part, depuis plus de 10 ans, de nombreuses réflexions ou évaluations sur les méthodes de traitement des données ont été engagées par de nombreux instituts de statistique, aux États-Unis, au Canada, en Suède ou en Australie par exemple. Un des points communs à ces divers travaux est l'attention portée au travail confié aux gestionnaires d'enquête dans cette phase de contrôle-apurement des données par rapport à celui pris en charge automatiquement par l'ordinateur (correction automatique de données suspectes, imputation pour cause de non-réponse partielle...). Plus précisément, les efforts déployés dans divers instituts de statistiques portent sur un ciblage plus grand des recontacts d'entreprises demandés (ou suggérés) aux gestionnaires, en quelque sorte pour une intervention plus sélective des gestionnaires sur les données manquantes ou suspectes.

De ce point de vue, les chaînes de traitement mises en place dans les EAE Commerce et Services (et particulièrement Commerce) ont depuis longtemps intégré la non-nécessité de recontact systématique de l'entreprise en cas de non-réponse ou de donnée suspecte : les procédures de redressement automatique des données qui ont été développées permettent de résoudre correctement et automatiquement de nombreux cas. Il n'en reste pas moins, et c'est l'objet de la partie ci-dessous de montrer que des voies d'amélioration sensible apparaissent possibles, tant certaines modifications de données opérées par les gestionnaires, apparaissent, in fine, de peu d'utilité. Ce qui apparaît alors fortement en cause, c'est l'absence de prise en compte de l'impact de l'erreur supposée (donnée suspecte) sur le résultat agrégé (par rapport au domaine d'étude envisagé).

La partie 2 explore alors une méthode de contrôle des données qui serait organisée autour d'une sélection plus grande des recontacts d'entreprise. Sur la base des données transmises par l'entreprise (questionnaire), comme de celles disponibles dans la base de sondage ou dans des enquêtes précédentes, une sélection des entreprises est faite, à l'aide d'une règle de décision ; cette sélection oriente vers les gestionnaires les entreprises pour lesquelles un recontact peut s'avérer nécessaire, les autres entreprises ne faisant l'objet de corrections qu'automatiques. Afin de se prémunir contre des corrections automatiques aberrantes, un contrôle final sur données agrégées permet de sélectionner en fin de traitement un lot supplémentaire d'entreprises réclamant une analyse spécifique de la part du gestionnaire.

Seuls quelques-uns des scénarios testés sont présentés dans ce rapport. Pour de plus amples détails, se référer aux rapports disponibles à la division H3E. En l'état actuel, ces travaux apparaissent encourageants sans, toutefois, autoriser immédiatement la mise en place d'une nouvelle architecture du contrôle des données.

C'est l'objet de la partie 3 de lister les points d'approfondissement nécessaires si l'on veut faire évoluer nos méthodes de traitement vers des méthodes garantissant une qualité de résultats comparable à celle d'aujourd'hui, avec des coûts moindres, autorisant alors des délais de publications plus courts. Les points d'approfondissements encore nécessaires sont très importants et montrent clairement que le travail engagé n'en est qu'à une première phase.

* * *

Enfin, nous remercions Monsieur Hesse (Unité méthodologie de la Direction des Statistiques Économiques) pour les nombreuses et pertinentes remarques qu'il a faites sur l'ensemble de ce problème.

* * *

On utilise très souvent dans ce document deux expressions qu'il convient de bien préciser dès à présent.

***A priori** signifie en début de traitement, c'est-à-dire avant toute intervention du gestionnaire. On considère par exemple que les données disponibles a priori sont celles que l'on connaît par le fichier de lancement d'enquête ou par toute autre source extérieure, par le fichier de l'année précédente, par les résultats bruts transmis par l'entreprise, par les messages issus d'un éventuel contrôle automatique...*

***A posteriori** signifie en fin de traitement, c'est-à-dire après le passage de la procédure de contrôle-redressement-extrapolation. Les données a posteriori sont les données définitives.*

Tous les travaux qui suivent ont été réalisés à partir des données de l'Enquête Annuelle d'Entreprise dans le Commerce, essentiellement parce que c'est, aujourd'hui, l'EAE qui possède le plus d'informations sur le traitement subi par une donnée lors des phases de contrôle-redressement-extrapolation.

Les trois études réalisées sur les méthodes de traitement (rapports disponibles à la Division H3E), se sont efforcées d'évaluer des procédures de contrôle plus sélectives, fondées sur une articulation entre un contrôle individuel exercé a priori sur certaines entreprises répondantes et un contrôle agrégé appliqué en fin de traitement. L'évaluation s'est heurtée à l'utilisation de certaines procédures actuelles, en particulier celle de redressement, qu'il a fallu "forcer" pour l'adapter à nos simulations. La restriction à ce cadre particulier, qui préserve les procédures existantes, contribue à expliquer notre relative incapacité à valider à ce stade la méthode de contrôle proposée.

Examen du processus actuel de contrôle-apurement dans l'EAE Commerce

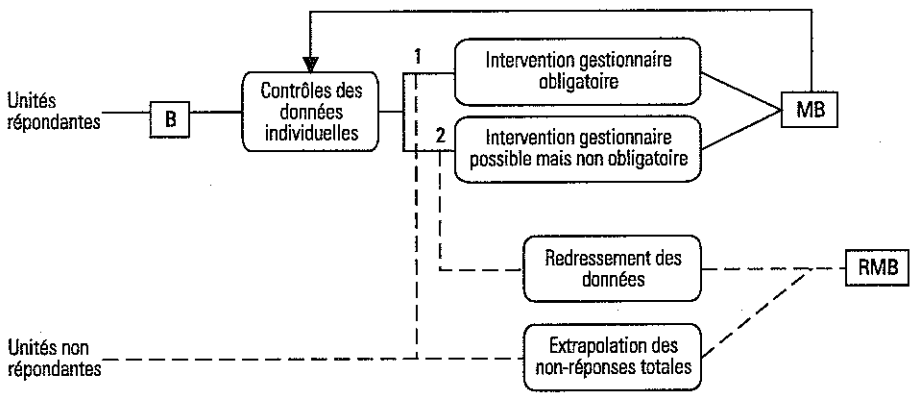
Le principe actuel du contrôle

Le contrôle actuel est un contrôle individuel. Une fois saisi, le questionnaire renvoyé par l'entreprise subit un premier contrôle, à l'issue duquel des messages précisent le type d'anomalies ou d'erreurs détectées. Une liste d'entreprises non valides est ainsi déterminée. Le gestionnaire peut alors intervenir directement sur les données et effectuer des mises à jour à l'écran.

Un code qualité (EQAA2) est attribué aux entreprises. Il définit le niveau de cohérence atteint par le questionnaire et il indique si une entreprise est ou non redressable, c'est à dire si les ultimes corrections de données jugées suspectes, ou renseignements de données manquantes peuvent être, sans risque trop grand, prises en charge par la procédure finale de redressement des données. Le code qualité tient compte de la taille de l'entreprise. Les unités redressables sont, en fin d'enquête, soumises à la phase d'imputation automatique. Celles qui n'atteignent pas un niveau d'apurement "suffisant" (au regard des seuils choisis par le responsable d'enquête) au moment de l'élaboration des résultats agrégés sont traitées comme des entreprises non répondantes.

Graphique 1

Cinématique du traitement actuel (contrôle des données, redressement des données, extrapolation des non-réponses totales)



_____ En cours d'année

_ _ _ _ En fin d'année

Avec : B données brutes
MB données obtenues après intervention du gestionnaire
RMB données définitives (où éventuellement MB = B = RMB pour les variables)

La répartition des unités entre **1** et **2** est principalement fonction du "verdict" du contrôle et de la taille de l'unité (contrôle de base pour toutes les entreprises, contrôle complémentaire pour les entreprises de plus de 50 salariés)

Il faut souligner ici que **l'ensemble du contrôle est largement paramétrable dans le processus existant**, c'est-à-dire que la plupart des tests peuvent être réglés, ainsi que la répartition des contrôles entre contrôle de base et complémentaire.

Si l'on s'interroge malgré tout sur les éventuelles failles du traitement actuel, c'est après avoir constaté *a posteriori* qu'un certain nombre d'interventions demandées aux gestionnaires se révélaient d'une faible efficacité au regard des résultats agrégés. C'est ce que l'on va développer au cours des paragraphes suivants.

On peut cependant dès à présent souligner deux traits de prudence du système actuel.

Tout d'abord, même si la taille de l'unité est prise en compte, dans le calcul du code qualité par exemple, l'effet d'impact potentiel sur le résultat agrégé n'est que partiellement intégré. En effet, le contrôle de base s'applique de la même manière à toutes les entreprises. Le système actuel ne fait pas de distinction entre les unités qui ont des messages "bloquants", c'est dire dont une ou plusieurs données sont en cause dans la non-vérification d'une règle du contrôle, cette règle faisant partie du groupe des règles incontournables.

Ensuite, le nombre de données suspectes n'intervient pas dans la sélection issue du contrôle des données brutes. Le code qualité n'indique pas parfaitement le degré d'incohérence du questionnaire.

On peut voir dans ces deux traits une faiblesse coûteuse à l'arrivée, dans la mesure où le gestionnaire est conduit à examiner de la même manière des unités pour lesquelles le caractère "suspect" est plus ou moins intense, ou plus ou moins susceptible d'influer sur le résultat agrégé.

Analyse des mises à jour effectuées par les gestionnaires

Une mise à jour est une intervention du gestionnaire, qui se traduit par une modification de la donnée brute issue du fichier de saisie.

Il faut souligner ici une limite importante à l'analyse effectuée, qui n'a pas examiné les interventions laissant inchangé un résultat qui aurait pu être modifié par les procédures de redressement, par la voie de la confirmation des anomalies[1]. Ce type d'intervention sera

(1) Une anomalie est le fait qu'un ratio sorte des limites autorisées. Le gestionnaire d'enquête peut, après examen confirmer la validité de ce ratio ou de modifier une des données en cause.

pris en compte par le contrôle alternatif présenté ultérieurement. Cette première phase de l'analyse s'est intéressée essentiellement à deux types de mise à jour : la modification d'une donnée brute est le renseignement d'une valeur manquante. Elle a en outre relevé un certain nombre d'opérations qui pourraient être évitées, comme par exemple celles de totalisation qui représentent encore en moyenne 3 % des interventions concernant l'effectif total ou l'investissement total, comme le prouve le *tableau 1*.

Tableau 1

Les entreprises répondantes de l'EAE Commerce selon leur situation après le contrôle-apurement effectué par les gestionnaires (avant les procédures de redressement automatique).

Situation	Rémunérations totales	Nb total d'heures travaillées	Effectif total	Chiffre d'affaires net	Achat de marchandises	Investissement
NM	81,3	73,0	78,1	83,3	76,4	49,9
NR	11,3	14,8	2,7	4,4	8,0	29,5
RG	6,2	9,8	12,3	8,6	10,0	13,0
MO	1,2	2,4	3,3	3,7	5,6	4,1
VE	-	-	0,6	-	-	0,1
CA	-	-	3,0	-	-	3,4
Total	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

NM : donnée non modifiée
NR : donnée manquante (non réponse partielle)
RG : donnée manquante puis renseignée
MO : donnée modifiée
VE : donnée totale non modifiée mais ventilation renseignée
CA : donnée calculée (totalisation par exemple)

La proportion faible de modifications effectuées par les gestionnaires (code M0), est à mettre en regard du très grand nombre d'entreprises dont un examen par le gestionnaire est demandé (voir *tableau 3* page 154). Ce phénomène illustre le grand nombre de cas où le gestionnaire confirme un ratio jugé anormal pour les règles présentes.

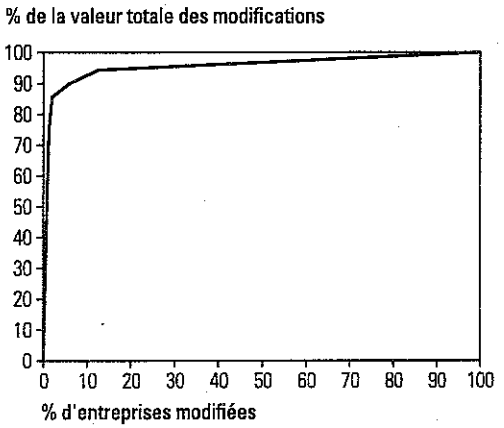
La valeur d'une modification est l'écart entre la donnée modifiée par le gestionnaire et la donnée brute. La valeur du renseignement d'une non-réponse partielle est l'écart entre la donnée renseignée par le gestionnaire et le résultat que fournirait un redressement des données brutes (imputation automatique). La valeur totale des mises à jour est la somme des valeurs absolues des mises à jour.

L'évaluation du "rendement" des mises à jour s'est posée en termes de concentration et de vitesse de convergence.

Seules les modifications ont permis d'établir des courbes de concentration, indiquant ce que x % de ce type de mise à jour représentait en pourcentage de la valeur totale des mises à jour. Les graphiques établis ont souligné le fort degré de concentration des modifications[1]

Graphique 2

Nombre total d'heures travaillées (6101)

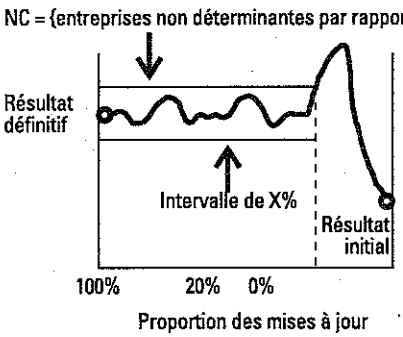


Lecture : 5,8 % des modifications représentent 90,8 % du volume total de ces modifications (concentration des modifications en valeur absolue).

Cette forte concentration s'explique en partie par les erreurs de saisie (multiplication par 100 ou 1000 de la donnée du questionnaire). Conjuguée avec l'analyse de la convergence des modifications (cf *graphique 3*), ce résultat montre que peu de modifications ont eu finalement un impact sur le résultat agrégé.

Graphique 3

Convergence vers le résultat définitif



Lecture : 20 % des interventions manuelles sont suffisantes pour approcher le chiffre définitif à x % près

(1) Cf. Projet EAE 4G "Réflexions sur les méthodes de traitement - I[ere] étape", novembre 1992.

L'approche en terme de convergence permet de suivre l'élaboration du résultat agrégé obtenu après le contrôle-apurement effectué par les gestionnaires, considéré comme étant le "vrai" résultat, valeur de référence, à partir de la valeur agrégée initiale.

Les entreprises sont classées par ordre croissant de la valeur absolue des mises à jour. L'écart à la valeur finale est calculé pas à pas, en cumulant les valeurs algébriques des mises à jour de chaque entreprise, en partant de la plus petite mise à jour. Il est ainsi tenu compte des éventuelles compensations entre les corrections apportées. La courbe de convergence est obtenue en intégrant successivement les mises à jour effectuées par les gestionnaires.

Cette courbe permet de déterminer le nombre de mises à jour "non contributives", si l'on autorise une plage de variation de x % autour de la valeur finale.

L'analyse a été menée pour chacun des deux types de mise à jour.

Vitesse de convergence des modifications

Dans ce cas, le résultat initial inclut les données brutes ainsi que les valeurs manquantes renseignées par les gestionnaires. Les entreprises qui participent alors à l'élaboration effective du résultat définitif (par rapport au résultat initial) sont donc uniquement celles qui ont été modifiées.

Sur les huit variables étudiées (rémunérations, heures travaillées, effectif salarié, effectif en fin de premier trimestre, effectif cadre, chiffre d'affaires net, achat de marchandises et investissement), en moyenne seules 15 % environ des modifications manuelles opérées sur les données brutes (par retour à l'entreprise), sont nécessaires pour approcher le résultat agrégé à 0,5 % près (cf. *tableau 2*). Ce qui signifie, par conséquent, que près de 85 % des modifications opérées par les gestionnaires ne sont que de faible intérêt quant à la précision du résultat.

L'examen des principales modifications montre que, dans la majorité des cas, il n'est pas permis de conclure à l'existence d'un biais systématique. Si l'on observe fréquemment que les corrections les plus importantes se font toutes dans le même sens (en général à la baisse), c'est en réalité imputable aux erreurs d'unité ou de saisie.

Les erreurs d'unité sont par exemple fréquentes pour les variables "rémunérations totales", "nombre total d'heures travaillées" et "chiffre d'affaires net", quel que soit le secteur étudié.

Tableau 2

Pourcentage de modifications effectuées à l'intérieur d'un intervalle de 0,5%

Variables	6101	6243	6411	58	62	64	61 + 62
Rémunérations totales	94,8	90,0	100,0	**100,0**	100,0	95,3	99,2
Nb total d'heures travaillées	53,8	54,5	75,0	78,3	69,9	57,8	74,2
Effectif total	98,4	67,9	92,3	76,3	85,6	91,0	93,9
Chiffre d'affaires net	**100,0**	95,5	**100,0**	99,0	98,7	99,4	99,1
Achat de marchandises	97,8	**33,7**	99,4	88,7	70,2	99,4	99,2
Investissement total	70,0	87,5	67,4	89,6	89,2	79,9	83,2

Vitesse de convergence des interventions consistant à renseigner les valeurs manquantes

L'analyse n'a pu être menée que sur un extrait de données provisoires de l'EAE Commerce portant sur l'exercice 1991. D'après cette simulation, le renseignement manuel (après retour à l'entreprise) des valeurs manquantes pourrait être **dans plus de la moitié des cas**, remplacé par une imputation automatique, sans perte de précision.

Bilan de cette première partie

Les résultats de cette première étape, présentés dans le rapport sur l'analyse de l'efficacité du mode actuel de traitement des données, doivent être bien entendu nuancés, pour tenir compte d'un certain nombre de difficultés relevées au cours de l'étude et surtout parce qu'ils ne permettent d'établir qu'un constat *a posteriori* du "rendement" du mode de contrôle actuel.

Ils ont cependant paru suffisamment prometteurs pour encourager la recherche d'un mode de contrôle plus "efficace" susceptible de repérer *a priori* (cf remarque préliminaire) les unités justifiant un examen du gestionnaire.

L'examen de la procédure actuelle doit donc être interprété essentiellement comme un moyen d'éclairer les différents phénomènes intervenant dans les phases de traitement d'enquête et de jeter les bases d'une réflexion sur une méthode alternative de contrôle.

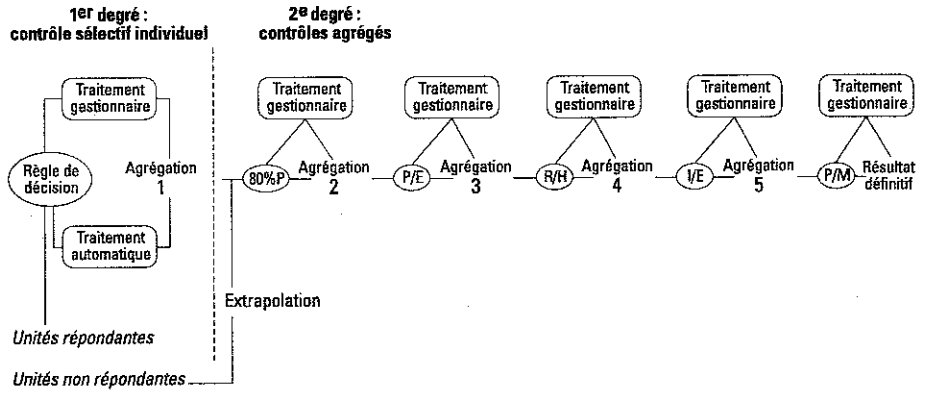
Proposition d'une méthode de contrôle alternative

Architecture du contrôle proposé

Le schéma de contrôle envisagé est à deux degrés : l'un concerne le contrôle individuel sélectif s'appliquant *a priori* aux unités répondantes et gouverné par une règle de décision, l'autre est relatif au contrôle agrégé fonctionnant *a posteriori* et pouvant comporter plusieurs étapes.

Graphique 4

Procédure d'un traitement à 2 degrés et 6 étapes



La règle de décision pilotant le contrôle sélectif individuel intègre toutes les informations disponibles *a priori*, c'est-à-dire avant toute intervention du gestionnaire, en particulier les données brutes, les éventuelles données de l'année précédente ainsi que les informations contenues dans le fichier de lancement d'enquête.

L'objectif de ce type de contrôle est de repérer les unités susceptibles de ne pas être bien traitées par une procédure automatique, afin de les confier à l'examen du gestionnaire et de soustraire à cet examen manuel les unités pouvant être convenablement traitées automatiquement. Il faut souligner toutefois que même pour ces entreprises exclues du contrôle gestionnaire, la cohérence interne des données est assurée.

On peut remarquer dès à présent le rôle déterminant joué par la règle de décision qui doit prendre en charge l'essentiel de la sélection. En effet l'orientation vers le contrôle gestionnaire doit être effectuée pendant toute la durée du traitement pour lisser

la charge de travail et non se concentrer en fin de période. Les contrôles agrégés ne doivent isoler qu'un petit nombre d'entreprises et permettre uniquement de se garantir contre des dérives, préjudiciables aux chiffres agrégés, de l'imputation automatique.

Le contrôle sélectif individuel

Il s'applique aux entreprises répondantes dont les données ont déjà subi, comme dans la procédure actuelle, un contrôle automatique permettant d'affecter certains codes (dont le code qualité) et de repérer certaines anomalies.

Il est gouverné par une règle de décision. La recherche de cette règle s'est inspirée des caractéristiques d'entreprises identifiées comme "déterminantes" *a posteriori* : unités pour lesquelles l'imputation automatique fournit une valeur "éloignée" du résultat indiqué par le gestionnaire et qui ne permettent pas de rester dans une fourchette de 0,5 % du résultat agrégé définitif.

Cependant, comme l'objectif n'est pas de repérer exactement ces unités "déterminantes" *a posteriori* mais un ensemble d'entreprises "critiques", c'est-à-dire **présentant a priori un risque potentiel d'erreur importante**, les recherches reposant sur des analyses discriminantes se sont révélées peu fructueuses. Elles n'ont en particulier pas permis d'identifier une fonction de score **linéaire** apte à scinder correctement les unités répondantes.

Le choix des critères de sélection a été guidé par les trois hypothèses faites sur les entreprises que l'on ne sait pas "bien" redresser au regard du résultat définitif, c'est-à-dire pour lesquelles l'écart constaté *a posteriori* entre la donnée redressée à partir des valeurs brutes et la donnée définitive est significatif :

1 – ce sont les entreprises qui présentent le plus d'incohérences ;
2 – ce sont les entreprises qui ont une forte contribution au résultat agrégé ;
3 – ce sont les entreprises dont l'évolution contribue le plus à faire évoluer le résultat global.

Plusieurs règles de décision ont été testées[1], soit en organisant les critères de sélection en filtres successifs, soit en les combinant dans une fonction de score. Deux règles ont été retenues. Elles accordent une grande importance aux contributions.

Règle F3 : traitement gestionnaire pour les entreprises fortement incohérentes et pour les très grandes entreprises incohérentes (code EQAA2 égal à 1 ou 2) ou

(1) Cf. Projet EAE 4G "Réflexions sur les méthodes de traitement - 2e étape", mars 1993.

fortement contributives (données brutes de l'année N ou données définitives de l'année N-1 représentant plus de 0,05 % de l'effectif total, des rémunérations, du chiffre d'affaires ou de l'investissement du résultat agrégé définitif de l'année N-1).

Règle F10 : traitement gestionnaire pour les entreprises fortement contributives.

Le *tableau 3* présente les écarts au résultat définitif relevés après l'application de la règle F10 à différents secteurs du niveau intermédiaire (3 chiffres) de la nomenclature.

On peut dès lors observer que le volume d'unités confiées au gestionnaire est souvent réduit de manière importante. Il faut en outre rappeler que "contrôle gestionnaire" ne signifie pas "intervention du gestionnaire" : la règle de décision, fondée sur une stratégie de prudence, soumet à l'examen manuel toutes les entreprises importantes, même si leurs données ne présentent pas d'incohérences. Cette orientation délibérée pour la prudence peut constituer un frein à "l'efficacité" de la méthode proposée. On verra plus tard s'il est possible d'envisager d'autres scénarios plus hardis.

Cette diminution du nombre d'entreprises confiées au traitement gestionnaire s'effectue parfois au détriment de la précision.

Tableau 3

Application de la règle F10 à différents secteurs

Secteur	Nr	Nca	H	P	R	S	T	M	E	I	Nm
641	2 550	2 196	1,1	0,5	- 2,9	- 0,4	0,9	- 0,3	0,6	- 2,5	845
642	2 067	1 888	1,0	0,4	1,1	- 2,3	0,8	0	0,3	- 2,0	785
643	609	486	0	- 0,1	- 0,3	0,7	0	- 0,2	0	0,1	558
644	3 237	3 064	1,3	1,7	1,3	- 0,5	1,4	1,6	1,0	- 2,4	1 087
621	1 030	902	- 0,3	0,9	0,5	2,7	- 0,3	0,1	0	0,5	694
624	2 914	2 744	2,2	1,1	3,1	2,7	1,1	0	1,9	0,5	1 247
610	3 797	3 614	1,2	- 1,7	- 1,2	- 1,3	- 1,5	22,4	- 2,4	73,9	343
580	2 765	2 482	1,6	- 0,6	- 0,6	- 2,2	- 0,3	- 0,2	- 0,7	- 6,0	695
581	975	884	0,1	- 8,8	0,3	0,3	0,4	- 11,9	0,5	- 2,8	698

Avec : Nr = nombre d'entreprises répondantes, Nca = nombre d'entreprises actuellement soumises au contrôle gestionnaire, Nm = nombre d'entreprises soumises au contrôle gestionnaire dans la procédure de traitement proposée

Les contrôles agrégés

Ils constituent le deuxième degré de la procédure de traitement envisagée. Les résultats individuels élaborés après passage de la règle de décision, sont alors agrégés et permettent l'extrapolation des entreprises non répondantes. On obtient ainsi un premier

résultat agrégé. Ce résultat est soumis à un macro-contrôle qui permet de réorienter vers l'examen manuel certaines unités déterminantes. Un second résultat agrégé peut ensuite être calculé. Cette procédure est ainsi répétée en utilisant à chaque étape un contrôle différent.

Les macro-contrôles envisagés reposent sur un taux de couverture ou sur la distribution d'une variable importante. Leur succession permet d'enrichir à chaque passage l'information à l'aide des corrections induites. Les scénarios testés ont montré que l'utilisation successive des filtres est toujours préférable à leur application simultanée.

Un certain nombre de filtres ont été testés[1]. Cinq ont été retenus pour simuler les procédures de traitement pour divers secteurs :

80 % P : couverture de 80 % du chiffre d'affaires provisoire
P / E : chiffre d'affaires / effectif total
R / H : rémunérations / heures travaillées
I / E : investissement / effectif total
P / M : chiffre d'affaires / achat de marchandises

Le premier filtre oriente vers le contrôle gestionnaire les entreprises participant à la couverture de 80 % du chiffre d'affaires provisoire. Les quatre filtres reposant sur des ratios sélectionnent les entreprises situées dans les "queues de distribution" (définies par les centiles 5 % et 95 %).

La procédure globale de traitement

On a retenu l'application du contrôle à un niveau intermédiaire (3 chiffres) de la nomenclature. C'est donc à ce niveau d'observation que l'on détermine les unités confiées au contrôle gestionnaire et celles qui sont soumises à une procédure automatique. On examine ensuite les résultats obtenus au niveau agrégé (2 chiffres) et au niveau détaillé (4 chiffres) de la nomenclature.

Les résultats sont présentés au *tableau 4*.

Le gain en volume n'est pas négligeable. Il ne suffit cependant pas à évaluer le gain exprimé en charge de travail ou en coûts.

En revanche, la perte en précision atteinte pour certaines variables et certains secteurs n'est pas satisfaisante. Un examen plus précis des écarts les plus importants

(1) Cf. Projet EAE 4G "Réflexions sur les méthodes de traitement - 3^{e} étape", octobre 1993.

Tableau 4

Impact du contrôle appliqué au niveau intermédiaire (3 chiffres) de la nomenclature

Secteur	Nr	Nca	H	P	R	S	T	M	E	I	Nm	Nm/Nr %
64	8 463	7 636	- 0,2	- 0,7	- 0,1	- 1,3	- 0,3	- 1,0	- 0,4	- 1,6	5 054	59,7
641	2 550	2 196	0	- 0,6	0	- 0,8	- 0,1	- 1,0	- 0,2	- 2,0	1 403	55,0
6411	1 742	1 490	0	- 0,6	- 0,1	- 0,4	- 0,2	- 1,0	- 0,3	- 1,6	1 011	58,0
6412	297	256	- 0,2	- 0,2	0	- 2,8	0,1	- 0,7	0	- 1,8	158	53,2
642	2 067	1 888	0	- 0,7	0,1	- 2,7	- 0,2	- 1,0	- 0,5	- 1,6	1 146	55,4
6422	486	448	- 0,8	- 0,8	- 0,6	- 2,2	- 0,7	- 1,0	- 0,9	- 3,0	289	59,5
6424	453	411	- 0,3	- 1,0	- 0,2	- 5,0	- 0,1	- 1,2	- 0,4	- 1,8	275	60,7
643	609	486	- 0,1	- 0,1	- 0,1	0,2	- 0,2	- 0,1	- 0,2	0	577	94,7
644	3 237	3 064	- 0,5	- 1,3	- 0,4	- 1,3	- 0,5	- 1,6	- 0,6	- 1,9	1 928	59,6
6443	604	578	0	- 1,1	- 0,1	1,4	- 0,3	- 1,3	0,3	- 1,0	356	58,9
6445	237	185	- 0,2	- 0,9	- 0,2	- 1,5	- 0,4	- 0,7	- 0,4	- 2,0	151	63,7
6449	465	454	- 1,6	- 1,8	- 1,4	- 1,3	- 2,5	- 2,0	- 2,1	- 0,6	238	51,2
62	3 995	3 696	0	- 0,9	0,3	0,1	- 0,2	- 1,1	- 0,3	- 1,0	2 893	72,4
621	1 030	902	- 0,1	- 0,3	0,1	- 1,0	0	- 0,5	- 0,1	- 0,7	792	76,9
6211	474	419	0	- 0,3	- 0,1	- 1,0	- 0,1	- 0,7	0	0	381	80,2
622	21	20	0	0	0	0	0	0	0	0	20	95,2
623	30	30	0	0	0	0	0	0	0	0	30	100,0
624	2 914	2 744	0,1	- 1,1	0,3	0,4	- 0,2	- 1,4	- 0,3	- 1,1	2 051	70,4
6243	1 704	1 643	- 0,1	- 1,0	0,3	1,6	- 0,2	- 1,1	- 0,2	- 0,9	1 208	70,9
61	3 797	3 614	- 1,0	- 1,8	- 0,8	- 0,9	- 0,9	- 1,3	- 1,1	- 4,8	1 821	48,0
6101	3 315	3 143	- 1,7	- 2,9	- 1,4	- 2,5	- 1,5	- 2,2	- 2,0	- 13,8	1 384	41,7
6103	409	403	- 0,5	- 1,1	- 0,4	- 0,2	- 0,5	- 0,8	- 0,5	- 1,2	378	92,4
58	3 740	3 366	- 0,3	- 2,8	- 0,3	- 1,1	- 0,2	- 3,2	- 0,5	- 3,2	2 194	58,7
580	2 765	2 482	- 0,4	- 0,9	- 0,4	- 1,3	- 0,2	- 0,6	- 0,6	- 3,6	1 437	52,0
5804	724	647	- 0,7	- 0,8	- 0,6	- 0,7	- 0,5	- 0,6	- 0,6	- 3,9	421	58,1
5806	465	428	0,9	- 1,7	0,8	- 1,1	1,1	- 1,0	- 0,2	- 4,3	278	59,8
581	975	884	- 0,1	- 9,4	- 0,1	- 0,7	- 0,1	- 12,5	- 0,2	- 2,0	757	77,6

propose comme explication prépondérante l'incapacité du redressement à renseigner des non-réponses partielles, alors que l'intervention du gestionnaire apporte une information.

Bilan de cette deuxième partie

L'analyse que l'on peut faire de la méthode de contrôle proposée se trouve limitée par le cadre très particulier dans lequel elle s'inscrit et qui a cherché à préserver le mode de fonctionnement actuel et à utiliser les procédures existantes de redressement et d'imputation.

La procédure de contrôle à deux étapes ici présentée ne peut par conséquent recevoir de validation à ce stade de l'analyse. En effet, il n'est pas ici possible de séparer l'effet "conservation du cadre existant" de l'effet "performances de la méthode proposée".

Toutefois, même si la procédure présentée n'assure pas un degré de précision suffisant pour certaines variables, elle propose sans doute la trame du scénario alternatif recherché. En effet, les divers développements statistiques réalisés ont permis d'exhiber un certain nombre de mécanismes contribuant à conforter cette hypothèse. Il convient donc d'explorer d'autres axes de réflexion, qui ont parfois déjà été évoqués sans être traités en profondeur et qui seraient susceptibles d'enrichir la trame mise en évidence afin d'élaborer un scénario performant.

Réflexions non développées

Information insuffisante sur les données brutes

Certaines entreprises déterminantes échappent au contrôle gestionnaire issu de la règle de décision parce que les variables utilisées comme filtres ne sont pas renseignées. Cette remarque est en particulier essentielle pour les entreprises nouvellement interrogées qui ne peuvent être identifiées que par les variables de l'année courante ; si celles-ci sont manquantes, ni les informations relatives à l'année précédente ni les variables d'évolution ne peuvent les repérer.

Les données redressées à partir des données brutes pourraient fournir une information supplémentaire, susceptible de pallier les défaillances des données brutes. Les résultats obtenus en utilisant cette information dans les filtres de décision n'ont révélé aucune amélioration sensible par rapport aux scénarios issus des données brutes. Une des explications réside sans doute dans les limites du programme de redressement "forcé" qui ne parvient pas à estimer certaines variables comme l'investissement.

Cette information supplémentaire apportée par le redressement des données brutes pourrait cependant être mobilisée d'une autre manière, dans une approche alternative de l'appréciation du risque d'erreur *(cf. § page suivante)*.

Il faudrait toutefois réfléchir au mode d'organisation que suppose ce type de démarche, intégrant un redressement avant toute intervention du gestionnaire. Il implique notamment de ne pouvoir démarrer les opérations de contrôle gestionnaire qu'une fois un certain nombre de questionnaires rentrés, pour que ce redressement ait un sens (calcul des moyennes de strates par exemple). Il faudrait comparer les "coûts" de cette organisation au gain en précision. Une autre solution peut consister à utiliser en début de traitement les moyennes de strates de l'année précédente et de n'y substituer les moyennes courantes qu'une fois un nombre suffisant de questionnaires reçus (voir Greenberg et Petkunas – Bureau du Census).

Retour systématique à l'entreprise

Les études effectuées ont montré que le redressement "forcé" appliqué aux données brutes échoue fréquemment à renseigner une non-réponse partielle concernant certaines variables comme l'investissement.

On peut envisager un scénario dans lequel l'existence d'une non-réponse partielle sur quelques variables décisives (chiffre d'affaires, investissement...), repérée par le pre-

mier examen automatique du questionnaire, donnerait lieu à l'édition automatique d'une demande d'information complémentaire adressée à l'entreprise. Ce scénario ne peut malheureusement pas être évalué actuellement. Seule une enquête témoin menée sur quelques secteurs permettrait de chiffrer le coût supplémentaire induit et de le confronter au volume de renseignements ainsi collectés, en tenant compte des délais de réponse.

Ce retour à l'entreprise pourrait se révéler très intéressant dans la mesure où il permettrait de suffisamment compléter les données brutes pour assurer un "bon" fonctionnement de la procédure de redressement.

Approche alternative du risque d'erreur

On n'a jusqu'à présent abordé l'estimation du risque d'erreur qu'à l'aide des notions de contribution et d'incohérence interne du questionnaire. Cette approche est encore très fruste, notamment elle ne prend pas en compte les prévisions d'erreur de la donnée brute que l'on peut faire.

Le risque d'erreur induit par la méthode de contrôle proposée dépend en fait de la capacité du redressement automatique à approcher la "vraie" valeur obtenue après intervention du gestionnaire. Comme le redressement fonctionne essentiellement à partir d'estimations par la moyenne, le risque encouru résulte en partie de la faculté à discerner *a priori* les unités atypiques des unités "moyennes".

Deux pistes peuvent être explorées, qui véhiculent la même idée mais utilisent des outils différents.

La première reposerait sur une comparaison entre la valeur brute et la valeur obtenue après un premier redressement de la valeur brute, permettant d'approcher pour une variable son "écart à la moyenne". Cet écart devrait être rapporté au résultat agrégé afin de tenir compte de l'impact final. Un seuil de tolérance étant fixé, cet écart relatif pourrait intervenir comme un des critères de la règle de décision. On peut alors calculer un score pour chaque unité, combinaison, linéaire ou non, de ces différents écarts, pondérés eux-mêmes selon l'intérêt de la variable. Seules les entreprises dont le score dépasserait un seuil donné, seraient alors orientées vers un examen par le gestionnaire.

La deuxième s'intéresserait à la distribution des variables. C'est la notion de "profil moyen". On pourrait utiliser des ratios identiques à ceux retenus dans les contrôles agrégés et confier à l'examen du gestionnaire les unités se situant dans les "queues de distribution". La stratégie devrait distinguer les entreprises interrogées l'année précédente des entreprises nouvelles. Pour les premières, les ratios établis en N-1 serviraient

de critères, pour les autres, les ratios seraient calculés à partir des données brutes et comparés aux distributions de N-1.

Ces deux pistes devront être développées, même si l'on peut, dès à présent, présager quelques difficultés. En effet, la première réflexion risque d'être limitée par la procédure existante de redressement des données brutes et la seconde est subordonnée, pour les unités nouvellement interrogées, au degré de renseignement des questionnaires. Dans les deux cas, on risque de ne pas pouvoir repérer quelques unités influentes.

Évaluation de la méthode

Robustesse, flexibilité

On a souvent évoqué dans les dossiers d'analyse ces deux exigences sans jamais les explorer vraiment puisque l'on n'a pas réussi à proposer une méthode de contrôle totalement satisfaisante. Il faut donc rappeler que toute procédure proposée devra, pour être validée, satisfaire aux contraintes suivantes :

- sensibilité aux paramètres : il conviendra de s'assurer que le déplacement marginal des frontières de sélection (sous l'effet de la modification des seuils par exemple) n'engendre pas de variations de trop forte amplitude sur la précision des résultats. Les rapports d'étape ont montré comment les phénomènes de compensation des erreurs et d'impact indirect de l'intervention du gestionnaire pouvaient accroître la sensibilité des résultats aux paramètres ;

- universalité : il s'agit de vérifier si la méthode peut être adaptée à différents secteurs ou différents domaines. Les spécificités en termes de taille, de concentration, de taux de non-réponse, d'exhaustivité devront être prises en compte.

Efficacité

Dans les différents rapports d'étape, l'évaluation du contrôle proposé est envisagée en termes d'arbitrage entre perte en précision et gain en volume d'unités traitées. C'est une approche grossière du débat qualité / prix, ou plus généralement coût / EQM :

- prix : la seule notion de "coûts" retenue est celle qui repose sur le volume d'unités traitées. On n'a pas tenu compte des disparités entre les unités ; réduire de 50 % le volume d'entreprises traitées ne diminue pas forcément d'autant la charge de travail. Il faudrait pouvoir affecter à chaque questionnaire un indicateur d'estimation du temps de traitement manuel. La taille du questionnaire, la complexité de l'entreprise, le volume d'anomalies ou d'erreurs détectées pourraient participer à l'élaboration

de cet indicateur. Il faudrait en outre intégrer les coûts de fonctionnement induits par l'adoption d'une méthode alternative ;

- qualité : la qualité a été évaluée à la lecture de la précision des résultats agrégés. On ne peut cependant pas mesurer les gains en précision devant résulter d'une attention accrue apportée aux entreprises les plus importantes.

Il faudrait toutefois développer la réflexion sur le niveau de précision souhaitable, sachant que les erreurs de mesure résultant de l'assouplissement du contrôle sont à rapprocher des imprécisions dues à l'échantillonnage et à l'extrapolation des non réponses.

Pour cela il semble indispensable de préciser le cadre théorique plus général dans lequel se situe notre problématique. Les analyses réalisées ont fourni des informations ponctuelles sur les écarts à une valeur de référence (valeur définitive). Il s'agira de traduire en termes d'impact sur la variance totale (ou plutôt sur l'EQM) les effets d'une méthode alternative de traitement. En s'inspirant des travaux statistiques effectués, qui sont en mesure d'orienter le choix des critères de sélection, leur hiérarchie et leur combinaison au sein d'une règle, on pourra dresser un certain nombre de caractéristiques susceptibles de décrire une sous-population d'unités "contribuant peu" à la variance totale. Il faudra alors, à partir de la confrontation entre volume d'unités traitées manuellement et impact sur la variance, essayer de définir la "meilleure" règle de décision permettant d'orienter en début de traitement les entreprises répondantes vers un contrôle gestionnaire ou un contrôle automatique.

Ce débat sur la précision est en outre indissociable de la réflexion sur le niveau d'agrégation retenu pour la publication des résultats. Il faudra pouvoir définir quelle précision souhaitée pour quel niveau d'observation.

BIBLIOGRAPHIE

ANDERSON, K. : "Average Weekly Earnings – 4th quarter 1988, comparison of edited data from the current input editing system with an experimental (extreme outlier) input editing system", output edit study, *Statistics Sweden*, septembre 1989.

BOUCHER, L. : "Micro-editing for the Annual Survey of Manufactures: what is the value-added ? ", *Statistics Canada.*

BOUCHER, L. : "ASM selective editing project", memorandum, *Statistics Canada*, février 1992.

BOUCHER, L. : "Selective editing for the Annual Survey of Manufactures", project description, *Statistics Canada*, avril 1992.

COTTON, C. : "SGVI – description des fonctions du système généralisé de vérification et d'imputation", *Statistics Canada*, juillet 1991.

GRANQUIST, L. : "On the need for generalized numeric and imputation systems", *Statistics Sweden*, Séminaire de méthodologie statistique, Commission Économique pour l'Europe, Nations Unies, Genève, décembre 1987.

GRANQUIST, L. : "Data editing activities at Statistics Sweden", report, *Statistics Sweden*, avril 1989.

GREENBERG, B. et PETKUNAS, T. : "An evaluation of edit and imputation procedures used in the 1982 Economic Censuses in Business Division", *Bureau of the Census.*

HIDIROGLOU, M. A. et BERTHELOT, J.M. : "Contrôle statistique et imputation dans les enquêtes-entreprises périodiques", techniques d'enquêtes, vol. 12, n°1, pp. 79-89, *Statistics Canada*, juin 1986.

LATOUCHE, M. – BERTHELOT, J.M. : "Stratégie de suivi pour les enquêtes économiques", Recueil du Symposium 90 de Statistique Canada "Mesure et amélioration de la qualité des données", octobre 1990.

LATOUCHE, M. – BERTHELOT, J.M. : "Use of a score function to prioritize and limit recontacts in editing business surveys", *Journal of official statistics*, Vol 8, n° 3, 1992, pp 389-400.

STATISTICAL JOURNAL of the United Nations Economic Commission for Europe, *Special issue on data editing*, vol. 8, n° 2, 1991.