Mise au point de critères d'arrêt pour la chaîne de traitement EPURE

Nathalie CARON(*) et Sylvie GRCIC(**)

(*)Insee, Unité Méthodes Statistiques, (**) Insee, Département de l'Emploi et des Revenus d'Activité

1. Introduction

Les employeurs affiliés au régime général de la sécurité sociale adressent des bordereaux récapitulatifs de cotisations aux URSSAF[1] chaque mois pour les établissements de plus de 10 salariés et chaque trimestre pour les établissements de moins de 10 salariés. Ces bordereaux, dont l'objet est de calculer les cotisations sociales dues par l'employeur, contiennent aussi des informations sur le nombre de salariés inscrits au dernier jour de la période, le nombre de salariés rémunérés pendant la période et les salaires versés sur la période considérée. Conformément à une convention entre l'ACOSS[2] et l'Insee finalisée en 1999, les 107 URSSAF transmettent trimestriellement à l'Insee des fichiers contenant, pour chaque établissement employeur, des renseignements d'identification (dont le SIRET), l'effectif inscrit, l'effectif rémunéré et la masse salariale.

Les données réceptionnées par l'Insee sont surtout utilisées pour le suivi conjoncturel des indices d'évolution d'emploi et de masse salariale. Il existe deux niveaux de publication : un niveau national et un niveau régional. Une chaîne de traitement appelée EPURE (Extension du Projet URSSAF sur les Revenus et l'Emploi) a été développée en conséquence et permet en particulier de réaliser des contrôles sur les données et de proposer des corrections si nécessaire. Celles-ci sont ensuite expertisées par des gestionnaires en directions régionales. Comme la vérification exhaustive des anomalies est impossible, les gestionnaires ne vérifient qu'une partie d'entre elles, la sélection se faisant actuellement sur un critère de taille et sur la gravité de l'anomalie.

Le comité de direction de l'Insee souhaite « *qu'une culture commune sur la qualité en statistique se développe et se diffuse à l'Insee et en particulier que des implications en terme d'exigence de qualité soient traduites en termes concrets à partir d'opérations de production existantes et/ou d'expérimentations à mener* ».

La chaîne de traitement Epure est une des applications qui a été retenue au titre des expérimentations sur des opérations statistiques.
L'analyse a été menée selon trois axes :

- Mieux connaître les attentes des utilisateurs, en mettant en regard les outils permettant d'assurer l'adéquation avec les résultats attendus.
- Identifier l'ensemble des processus mis en œuvre dans l'élaboration des indices.
- Améliorer l'efficacité du travail du gestionnaire en hiérarchisant les établissements à traiter et en recherchant un critère d'arrêt.

[1] Union de Recouvrement de cotisations de Sécurité Sociale et d'Allocations Familiales
[2] Agence Centrale des Organismes de Sécurité Sociale

L'objet de ce papier est de présenter la démarche suivie dans le cadre de ce troisième axe ainsi que les difficultés rencontrées. L'idée consiste à essayer de modéliser l'effectif final, utilisé dans les calculs d'indice, à partir de celui proposé par la chaîne de redressement automatique et des caractéristiques propres à l'établissement concerné, de manière à être capable, pour les trimestres ultérieurs, de prédire cet effectif. Les établissements seraient alors classés par ordre décroissant d'écart estimé par le modèle et il serait possible d'arrêter le travail des gestionnaires dès que cet écart est inférieur à une borne à déterminer.

Pour mener à bien l'expérimentation, il s'avère nécessaire de faire expertiser l'ensemble des anomalies des établissements. Ainsi, pour chaque trimestre on disposerait de trois effectifs : l'effectif inscrit déclaré par l'établissement, la valeur proposée par le redresseur (valeur 'redresseur') et celle corrigée par le gestionnaire, supposée être la cible (valeur 'vraie'). Ce test a été approuvé par le comité de direction du 20/01/2004. Compte tenu des contraintes de moyens, il a été restreint à quelques directions régionales. Les régions de Haute-Normandie, de Bretagne, de Picardie et de Midi-Pyrénées ont accepté de participer au test qualité sur les troisième et quatrième trimestre 2003. Le Centre, les Pays de la Loire, le Poitou-Charentes et le Limousin ont rejoint la démarche pour les données du quatrième trimestre 2003.

2. Le contexte : la source EPURE et la problématique

2.1 Présentation de la source Epure

La chaîne de traitement EPURE[3] mise en production à l'Insee en 1996 est fondée sur l'exploitation trimestrielle des bordereaux récapitulatifs de cotisations.

Les données transmises par les URSSAF sont, dans un premier temps, contrôlées par le « redresseur ». Il s'agit d'un ensemble de programmes informatiques qui affectent à chaque établissement une note établie principalement en fonction de la vraisemblance des évolutions constatées de l'emploi compte tenu de son passé. Les anomalies détectées peuvent être des anomalies d'identification[4] et / ou des anomalies statistiques (sur les effectifs, les masses salariales). La valeur de la note permet de savoir quel est le type de l'anomalie, la ou les variable(s) concernée(s) ainsi que la gravité de l'erreur (voir annexe 2). En cas de détection d'anomalie statistique, le redresseur propose une valeur redressée ou conserve la valeur brute (issue des URSSAF). Parallèlement, le redresseur exclut des établissements des calculs d'indice lorsqu'il les « juge » en anomalie grave.

Les travaux manuels d'expertise et de redressement visent à réinclure dans le calcul de l'indice des établissements exclus par le redresseur ainsi qu'à expertiser et affiner les corrections automatiques. Ces travaux sont réalisés par les directions régionales[5] qui traitent les fichiers de leur ressort géographique. Près de 75 gestionnaires en équivalent temps complet sont affectés à cette opération. Ceux-ci ont la possibilité de valider la correction proposée par le redresseur, de proposer une correction ou encore d'exclure l'établissement concerné du calcul final des évolutions d'emploi.

Une vérification exhaustive des corrections apportées par le redresseur à l'ensemble des établissements repérés en anomalie est impossible ; celle-ci est donc sélective. Pour les anomalies non vues par les gestionnaires, les propositions du redresseur sont retenues. La règle de traitement des anomalies repose actuellement sur un critère de taille. En effet, l'objectif national de publication impose aux gestionnaires de traiter dans les dix semaines les établissements de plus de 50 salariés. Lorsque cet objectif est atteint, les directions régionales ont alors la possibilité de fixer leurs propres objectifs de traitement des anomalies en fonction des utilisations régionales attendues des données. Les traitements manuels des bordereaux récapitulatifs de cotisations

[3] Extension du projet URSSAF sur les revenus et l'emploi
[4] Anomalies sur le SIRET, sur la commune et / ou sur l'activité économique.
[5] Hormis les directions régionales d'Ile de France et la Dirag pour lesquelles les traitements sont centralisés au sein du pôle « Epure » à la direction régionale des Pays de la Loire.

concernant le trimestre T commencent la dixième semaine après la fin du trimestre considéré. L'objectif national implique de traiter les établissements de plus de 200 salariés entre la dixième et la treizième semaine après la fin du trimestre T et ceux de plus de 50 salariés (et de moins de 200) entre la treizième et la vingtième semaine après la fin du trimestre T. Le traitement des établissements de moins de 50 salariés, pour les besoins régionaux, prend place aussi dans cette période de 10 semaines. On constate des comportements très différents dans le pourcentage d'anomalies traitées par DR liés aux objectifs propres à chaque direction régionale.

Outre le redresseur qui vérifie la cohérence interne et temporelle des données de chaque établissement, il existe aussi[6] des contrôles au niveau des données agrégées, appelés plus communément macro-contrôles, dans le processus de production actuel. Ceux-ci reposent sur deux éléments : les résultats d'une analyse de la variance et la recherche des secteurs connaissant une évolution « anormale ». L'analyse de la variance cherche à expliquer les évolutions obtenues entre deux trimestres consécutifs dans les croisements département/tranche d'effectif/activité en fonction de la région, l'activité et la taille exprimée en tranche. Les secteurs retenus en anomalie correspondent :

- aux secteurs dont la différence en valeur absolue entre l'évolution obtenue et celle prédite par le modèle d'analyse de la variance est supérieure à 10 ;
- et à ceux vérifiant les 2 conditions suivantes : l'évolution entre 2 trimestres consécutifs est supérieure en valeur absolue à 5 % et l'écart entre l'évolution relative à 2 trimestres consécutifs et celle relative aux mêmes trimestres de l'année précédente.

Il est important de rappeler que les macros contrôles ne servent pas actuellement à classer et par conséquent à traiter les établissements en priorité. Ils sont utilisés plusieurs fois tout au long de la période de travail manuel, mais a posteriori, donc en tant qu'outil de validation des résultats agrégés. Les macro-contrôles permettent ainsi de détecter des évolutions trop importantes dans certains secteurs et les gestionnaires sont alors invités à corriger les établissements qui impactent l'évolution. Deux fichiers Excel sont communiqués aux gestionnaires en région : l'un contient pour chaque département d'une région donnée la liste des secteurs en anomalie et l'autre pour chaque secteur en anomalie la liste des établissements dont l'évolution en valeur absolue est supérieure ou égale à 5 quelle que soit la taille.

Juste avant la diffusion des chiffres nationaux, un dernier type de contrôle qualité a été mis en place. Ces contrôles sont « lancés » plusieurs fois par période de gestion et permettent d'obtenir une liste d'établissements impactant fortement l'indice national. Cette liste est envoyée aux gestionnaires pour expertise et correction éventuelle.

2.2. La problématique

Au 1er trimestre 2004, 351 931 établissements ont été détectés en anomalie par la chaîne, ce qui correspond à 18,5 % de l'ensemble des établissements traités (voir tableau 1 ci-dessous). En régime courant, les équipes EPURE vérifient environ 30 % des établissements en anomalie pendant les 10 semaines de travail manuel (France entière). Comme signalé précédemment, les objectifs nationaux consistent à traiter en priorité les anomalies des établissements de plus de 50 salariés (anomalies d'identification et anomalies statistiques), ce qui représente environ 31 000 établissements (soit un peu moins de 10 % des établissements en anomalie).

Tableau 1 : Quelques statistiques sur le 1er trimestre 2004

Période	**T1 2004**
Nombre total d'établissements	1 904 331
Nombre d'établissements à traiter	351 931
En % du total des établissements	18,5 %
Nombre d'établissements en anomalie d'identification	91 152
Nombre d'établissements en anomalie statistique	296 853

[6] Depuis 2001

Actuellement, la fin des traitements manuels est conditionnée par le nombre de semaines accordé à cette opération. Le pourcentage d'anomalies expertisées par les gestionnaires diffère fortement d'une DR à une autre (de plus de 100 %[7] dans certaines régions à moins de 30 % dans d'autres).

Une des composantes de la démarche « qualité » mise en place pour EPURE, consiste à proposer un critère d'arrêt de l'expertise des gestionnaires tout en garantissant la même « qualité » dans les principaux domaines de diffusion et en essayant de quantifier la « qualité » des données obtenues. Afin de concilier les différents niveaux de publications, national et régional, les domaines de diffusion retenus dans notre démarche sont définis comme le croisement de l'activité en Nes16 (voir annexe 1) et du département, soit 1368 domaines France entière. L'activité retenue est celle de début d'année (activité également retenue pour la diffusion des résultats issus d'EPURE), autrement dit aucun changement d'activité en cours d'année civile n'est pris en compte pour les publications. Notons que cette démarche ne remet pas en cause l'utilité des contrôles au niveau agrégés qui sont par conséquent maintenus.

3. La démarche envisagée

La problématique exprimée ci-dessus revient à trouver un ordre adéquat de traitement des établissements en anomalie et à fixer un seuil au-delà duquel l'expertise manuelle n'apporte rien, c'est-à-dire ne modifie plus qu'à la marge les évolutions obtenues par domaine de diffusion.

3.1. 1ère étape : la détermination d'un ordre de traitement des établissements

La première étape consiste à déterminer un ordre adéquat de traitement des établissements en anomalie. A priori, le classement des établissements en vue de leur traitement doit s'appuyer sur les éléments suivants :

- Notation du redresseur
- Taille de l'établissement
- Contribution de l'établissement à l'évolution du domaine de diffusion
- Taux de couverture en effectif des établissements en anomalie dans le domaine (en niveau).

On se placera dans le cadre d'un croisement de l'activité en Nes16 et du département que nous appellerons « case » par la suite. X_T^j et X_{T-1}^j représentent respectivement la valeur de la variable X relative au trimestre T pour l'unité j et la valeur de la variable X au trimestre T-1 pour la même unité j.

La contribution de l'établissement à l'évolution de sa case n'est pas une variable qui existe déjà dans les différents fichiers générés par la chaîne de production EPURE. Il a donc été nécessaire de la définir. Plusieurs indicateurs pour appréhender la contribution d'un établissement à l'évolution peuvent être envisagés.

Premier indicateur :

Le principe de ce premier indicateur consiste pour une unité j et pour une variable X donnée à calculer l'écart entre X_T^j et X_{T-1}^j et à le rapporter à la somme des valeurs absolues des différences individuelles entre 2 trimestres successifs[8], ce qui correspond à :

$$Cont1 = 100 * \frac{\left| X_T^j - X_{T-1}^j \right|}{\sum_i \left| X_T^i - X_{T-1}^i \right|}$$

[7] Cela signifie que les gestionnaires corrigent également des établissements qui ne sont pas en anomalie.
[8] Ou peut-être entre le même trimestre de deux années successives.

Cet indicateur a été mis en place dans la chaîne de traitement des enquêtes annuelles d'entreprise (EAE) afin de contrôler l'impact des unités prises individuellement sur l'évolution d'un agrégat cible.

Par construction, cet indicateur est facilement calculable pour les établissements « présents dans la case » en T et en T-1. Cependant, il peut arriver que pour quelques unités (créations ou non-réponse par exemple), il n'y ait aucune valeur pour l'année précédente. Ainsi, si on considère pour une unité non-répondante que X_{T-1}^{j} est nulle alors que la vraie valeur est non nulle, on va surestimer inutilement l'impact de cette unité. La solution retenue dans les EAE consiste à fixer pour ces unités une mesure d'écart fixée arbitrairement à 0,1 X_{T}^{j}. Plusieurs valeurs de coefficients peuvent être testées. Pour les unités créées, considérer que X_{T-1}^{j} est nulle a un sens.

Second indicateur :

Le principe de cet indicateur est très voisin de celui présenté précédemment ; il permet d'évaluer la croissance individuelle de l'unité j pour la variable X mais cette croissance est évaluée vis-à vis du total de la variable *X*, soit :

$$Cont2 = 100 * \frac{\left| X_{T}^{j} - X_{T-1}^{j} \right|}{\hat{X}_{T-1}}$$

où $\hat{X}_{T-1}$ est le total de la variable *X* estimé à partir des données du trimestre T-1.

Une fois que l'on a récupéré ou calculé pour chaque établissement les différentes variables utiles pour le classement, on réalise une régression linéaire expliquant l'écart entre la valeur de l'effectif issue du redresseur et celle considérée comme « vraie », en fonction de ces variables. On suppose que, si le gestionnaire contrôle un établissement, la valeur obtenue après expertise est la valeur « vraie ». La réalisation de cette étape nécessite de disposer pour tous les établissements de la valeur proposée par le redresseur ainsi que celle du gestionnaire. Autrement dit, toutes les anomalies détectées par le redresseur sur le champ considéré doivent être expertisées par les gestionnaires.

Pour les trimestres ultérieurs, les établissements pourraient alors être classés par ordre décroissant d'écart estimé par le modèle retenu.

3.2. 2ème étape : critère d'arrêt et mesure de la « qualité » du fichier obtenu

Nous présentons deux approches possibles : une approche par modélisation et une approche empirique. Ces deux approches nécessitent de disposer comme pour la première étape d'un fichier contenant en particulier les valeurs « vraies » pour un champ donné (qui sont obtenues suite à une vérification exhaustive des anomalies). Celles-ci permettent d'estimer les paramètres du modèle proposé dans le premier cas et de définir explicitement le critère dans le second cas.

3.2.1. Approche par modélisation

On notera Y la variable effectif salarié et on supposera que $y_i = y_i^* + \varepsilon_i$ où y_i est la valeur obtenue pour l'établissement i pour la variable Y dans le fichier final, y_i^* la « vraie » valeur de la variable pour l'établissement i et ε_i un aléa qui modélise l'erreur de mesure. La valeur retenue dans le fichier final correspond soit à la valeur vérifiée et/ou corrigée par le gestionnaire soit à la

valeur proposée par le redresseur en absence de vérification du gestionnaire. De plus, on postule que, si le gestionnaire contrôle un établissement, la valeur obtenue après contrôle est la valeur vraie et que si le redresseur ne détecte aucune anomalie la valeur est vraie. Par conséquent, les établissements vérifiés conduisent à une erreur nulle ; seuls les établissements non vérifiés contribuent à « créer » de l'erreur.

L'aléa ε_i est supposé suivre une loi normale $N(0,\sigma^2_{X_i})$ où $\sigma^2_{X_i}$ est la variance des aléas qui dépend d'une variable notée X_i définissant l'ordre de classement des établissements au sens de la première étape (cette variable peut éventuellement correspondre à une combinaison linéaire d'autres variables).

Une règle d'arrêt inspirée de cette approche par modélisation peut être la suivante :
En notant E l'erreur totale[9], on obtient :

$$E = \sum_{i\in NV}\left(Y_i - Y_i^*\right) \rightarrow N\left(0, \sum_{i\in NV}\sigma^2_{X_i}\right)$$

où NV représente l'ensemble des établissements non vérifiés.

Si on se fixe un seuil S de "tolérance" pour E, on peut se donner comme contrainte :

$$\text{Proba}\left[|E| \le S\right] \ge 0{,}95$$

ce qui correspond à $S \ge 2\cdot\sqrt{\sum_{i\in NV}\sigma^2_{X_i}}$

Il convient donc de fixer le domaine NV de telle sorte que

$$\sum_{i\in NV}\sigma^2_{X_i} \le \frac{S^2}{4}$$

Auparavant, il aura fallu expliciter et estimer $\sigma^2_{X_i}$. On peut par exemple tester une modélisation du type

$$\sigma^2_{X_i} = \sigma^2 \cdot X_i^{\alpha}$$

où α est un réel (probablement positif -on commencera par tester $\alpha = 0$ et $\alpha = 1$). On estime σ^2 par $\hat{\sigma}^2$ à partir des données disponibles, et la règle devient :

$$\boxed{\sum_{i\in NV} X_i^{\alpha} \le \frac{S^2}{4\hat{\sigma}^2}}$$

En pratique, la démarche se résume de la façon suivante après avoir fixé le seuil[10] S et estimé α sur des données disponibles :

- Classement les établissements dans l'ordre décroissant de traitement (selon la variable X_i)
- Calcul de la taille limite T telle que :

$$\text{si } NV = \{i / X_i \le T\} \text{ alors } \sum_{i\in NV} X_i^{\alpha} \le \frac{S^2}{4\hat{\sigma}^2}$$

- Arrêter le contrôle dès que $X_i \le T$.

[9] L'erreur est ici définie de façon absolue. Il est aussi possible de développer la même approche théorique en raisonnant en valeurs relatives.
[10] Le seuil S est fixé par le responsable EPURE. Celui-ci sera très certainement différent selon le croisement département*activité considéré.

3.2.2. Approche empirique

Une autre façon de procéder est de raisonner de manière empirique (ou encore de façon "non paramétrique", c'est-à-dire sans supposer que l'erreur suit une loi normale). Le critère est déterminé sur des données disponibles puis ensuite appliqué en mode production sur des données courantes.

L'idée serait alors de se fixer un seuil S* pour quantifier l'erreur admissible, selon :

$$\frac{\left|\sum_{NV}\left(Y_i - Y_i^*\right)\right|}{\sum_{i=1}^{N} Y_i^*} \leq S^* \qquad (1)$$

où N représente la taille totale de la population. Comme on dispose des Y_i et des Y_i^* sur toute la population[11], si on est en mesure de déterminer une variable X_i qui classe les établissements dans un ordre de risque décroissant, alors en posant :

$$NV = \left\{i / X_i \leq T\right\}$$

on arrêtera le contrôle lorsque l'inégalité (1) sera satisfaite. Autrement dit, on arrête le contrôle quand on considère que, conditionnellement aux informations historiques dont on dispose, les corrections à venir ne sont pas susceptibles d'avoir a priori un impact numérique sur le résultat diffusé.

Il est aussi possible de choisir comme dénominateur pour le critère (1) le total de la variable Y pour le trimestre précédent (c'est cette approche qui sera retenue par la suite).

4. La réalisation

4.1. Les jeux de tests

Pour mettre en œuvre la démarche, plusieurs régions[12] ont accepté de participer à un test qualité sur les données des 3ème et 4ème trimestres 2003. Le travail demandé aux équipes EPURE consistait à vérifier l'ensemble des anomalies de leur région.

Les premières observations ont porté sur la modification du nombre de salariés et sur le suivi de l'indice d'évolution de l'emploi entre deux trimestres successifs T-1 et T, appelé par la suite indice T/T-1, aux différentes phases de l'exploitation des données (état brut URSSAF, données après passage du redresseur - semaine 0 et données après travail des gestionnaires - semaine 10). Le champ observé est celui de la diffusion d'EPURE[13]. Ces travaux ont permis de comparer la valeur redresseur à la valeur gestionnaire ce qui est impossible en période courante car seule la dernière valeur est conservée dans les fichiers sans que l'on sache si elle est différente de la valeur redresseur.

4.1.1. Premières constatations

Le premier constat est que le travail des gestionnaires impacte fortement l'indice d'évolution de l'emploi, celui-ci évoluant surtout dans les tranches de taille les plus élevées. Les plus fortes corrections de l'indice T/ T-1 sont apportées pour les notes du redresseur indiquant les anomalies

[11] Comme dans l'approche a.
[12] Bretagne, Haute-Normandie, Picardie et Midi-Pyrénées pour le 3ème trimestre 2003 plus le Limousin, le Poitou-Charentes et les Pays de la Loire pour le 4ème trimestre 2003.
[13] Établissements présents dans les bases de diffusion hors secteur 00, EA et ER de la Nes16 (cf. annexe 1).

les plus graves et pour les établissements ayant un nombre de salariés important. Une exception pourtant pour les établissements de moins de 50 salariés pour lesquels les gestionnaires modifient fortement l'indice en traitant ceux qui ont des notes relevant essentiellement des anomalies sur les créations-cessations. Il ne semble pas y avoir d'effet département marqué pour les régions ayant participé au test.

Tableau 2 : T3 2003 - Impact de la reprise manuelle - régions ayant participé au test qualité

Nombre d'établis-sements	Effectifs T-1	Effectifs Bruts T	Effectifs après redressement automatique T	Effectifs après 10 semaines de travail manuel T	Évolution T/T-1 brute	Évolution T/T-1 après redressement automatique	Évolution T/T-1 après reprise manuelle
204 957	2 175 518	2 134 470	2 173 447	2 168 317	-1,9	-0,1	-0,3

En pratique, les gestionnaires « modifient » peu d'établissements. En effet, la valeur proposée par le redresseur automatique est remplacée pour seulement **12 % des établissements déclarés en anomalie pour le contrôle automatique.**

Lorsque la valeur brute issue des URSSAF est contrôlée par le redresseur automatique et que ce dernier décèle une anomalie, celui-ci propose une nouvelle valeur ou, faute de mieux, reconduit la valeur brute. A son tour, le gestionnaire, lorsqu'il vérifie les établissements en anomalie, peut confirmer la valeur retenue à l'issue du redressement automatique ou en proposer une autre.

Lorsque le gestionnaire confirme l'effectif issu du redressement automatique, il confirme également l'effectif brut (issu des URSSAF) dans 75 % des cas (ainsi que l'effectif T-1 pour la moitié). Lorsque l'effectif issu du redresseur confirmé par le gestionnaire est différent de l'effectif brut, il est presque toujours égal à l'effectif T-1.

Lorsque le gestionnaire ne valide pas la proposition du redresseur, il reprend l'effectif brut dans 14 % des cas, l'effectif T-1 dans 38 % des cas et pour le reste, il choisit une autre valeur qui est souvent proche de l'un des effectifs (brut, T-1 ou redresseur).

Tableau 3 : Choix de correction ou non des effectifs des établissements à vérifier par les gestionnaires

	Le gestionnaire confirme la valeur issue du redresseur						
Établissements à vérifier	Valeur différente de l'effectif Brut			Valeur égale à l'effectif brut			Ensemble
	Valeur = effectif T-1	Valeur différente de l'effectif T-1	**Total**	Effectif brut= effectif T-1	Effectif brut différent de l'effectif T-1	**Total**	
T3-2003 (1)	6 350	654	**7 004**	9 737	10 352	**20 089**	**27 093**
T4-2003 (1)	7 763	688	**8 451**	15425	15 592	**31 017**	**39 468**

	Le gestionnaire modifie la valeur issue du redresseur								**Ensemble**
Établissements à traiter	Val. Gest et Redr. différents de Brut			Val. Gest = brut	Valeur redresseur égale à valeur brute			**Ensemble**	
	Val Gest.= effectif T-1	Val. Gest. autre	**Total**		Val. Gest.= effectif T-1	Val. Gest. autre	**Total**		
T3-2003 (1)	167	317	**484**	527	1 221	1 456	**2 677**	**3 688**	**30 781**
T4-2003 (1)	161	495	**656**	698	2 016	2 111	**4 127**	**5 481**	**44 949**

(1) régions ayant participé au test qualité

4.1.2. Mise en œuvre de la démarche

In fine, on dispose pour chaque trimestre de trois effectifs : la valeur déclarée par l'établissement, la valeur redressée par le redresseur et la valeur corrigée par le gestionnaire. En supposant que la valeur 'vraie' est la valeur proposée par le gestionnaire, la démarche a consisté à modéliser l'écart entre la valeur 'redressée' et la valeur 'vraie', de manière à être capable, par la suite, de prédire au mieux cet écart. Une fois un modèle retenu et ajusté, il faut vérifier qu'il est pertinent pour toutes les régions et quel que soit le trimestre considéré. Pour la validité de la démarche, la

relation estimée doit être largement indépendante du **trimestre et de la région considérés**. En effet, en raison de la lourdeur du travail demandé aux directions régionales, il n'est possible de calculer la valeur « vraie » que sur un nombre de trimestres et de régions limités.

Le modèle retenu consiste à régresser au niveau de l'établissement la différence entre les effectifs issus du redresseur et les effectifs obtenus après reprise manuelle sur les établissements en anomalie au 3ème trimestre 2003[14] selon les variables ci-dessous dont les modalités sont explicitées en annexe 3 :

- Tranche de note ;
- Tranche de taille ;
- Contribution de l'établissement à l'indice d'évolution de la case Département*Nes16[15] (comme défini selon le 1er indicateur - partie III.1) ;
- Taux de couverture en effectif des établissements en anomalie sur la case Département*Nes16.

La différence estimée entre les effectifs à 0 semaine de travail manuel et les effectifs à 10 semaines de travail manuel est ensuite calculée à l'aide des coefficients estimés issus de la régression. Les établissements sont ensuite classés par département, Nes16 et différence estimée décroissante.

4.2. Les difficultés rencontrées

La première difficulté a consisté à construire la base de travail. En effet, les fichiers EPURE sont constitués de bases départementales et les variables nécessaires aux calculs sont disponibles soit dans les bases brutes, soit dans les bases redressées ou encore dans les bases de diffusion de chaque trimestre ; certaines variables n'existent pas et doivent être purement et simplement construites (comme par exemple la variable qui indique si l'établissement est à vérifier). A titre d'exemple, pour la validation France entière, construire la base de travail regroupant les 2 millions d'établissements et toutes les variables calculées nécessite d'agréger 349 fichiers.

Plusieurs modèles ont été testés (cf. annexe 4-résultat de la régression retenue). Pour mesurer l'efficacité du classement issu de chaque modèle, l'indice d'évolution de l'emploi est recalculé pour chaque classement en simulant le traitement de 25 %[16] des établissements en anomalie dans chaque case Département*Nes16 (au lieu de 100 % des anomalies). Le classement jugé le plus pertinent est celui pour lequel l'indice recalculé pour 25 % d'anomalies vérifiées est le plus proche de l'indice obtenu en traitant toutes les anomalies.

Le nombre d'anomalies et leur difficulté diffèrent selon le département et selon la Nes16. Des critères d'arrêts plus précis seront définis dans la suite du document.

Pour le test qualité, le classement issu de la régression permet avec seulement 25 % des anomalies traitées au lieu de 100 % d'approcher l'indice total à 0,1 point près, soit à 2 131 salariés près. Ce résultat est très différent lorsqu'il est ventilé par département et Nes16. Pour la moitié des cases, l'indice s'approche de 0,06 point, mais pour 10 % des cases, l'indice diffère de plus de 0,3 point, ce qui reste insatisfaisant. Néanmoins, il faut considérer que le critère d'arrêt uniforme de 25 % n'est sans doute pas optimal et par conséquent ne conduit pas aux meilleurs résultats. En effet, il est très certainement bien trop élevé pour certaines cases et au contraire bien insuffisant pour d'autres.

[14] Champ restreint aux bases de diffusion EPURE hors secteurs 00, EA et ER (Nes16)
[15] France entière, il existe 1368 cases Département*Nes16 au 1er trimestre 2004 , 138 cases pour les régions du test au 3ème trimestre 2003 et 265 cases pour les régions du test au 4ème trimestre 2003.
[16] Ce chiffre a été choisi arbitrairement. Il correspond à l'économie de 2 semaines de travail manuel, les gestionnaires traitant 30% des établissements en anomalie en régime courant en 10 semaines de travail manuel.

Tableau 4 : Résultat du traitement de 25 % des anomalies dans l'ordre du classement issu de la régression pour les régions du test

T3-2003						
	Nes16	**Indice initial avant travail manuel**	**Indice final avec 100 % des anomalies traitées**	**Indice avec 25 % des anomalies traitées**	**Écart**	**Écart en effectif**
	EB	3,31	3,14	3,18	0,05	44
	EC	0,11	-1,71	-1,67	0,04	25
	ED	-1,33	-1,33	-1,33	0,00	0
	EE	-1,36	-0,67	-0,60	0,07	76
	EF	-1,45	-1,14	-1,09	0,06	114
	EG	-9,49	1,88	2,01	0,13	34
	EH	2,94	0,49	0,54	0,05	82
	EJ	0,91	0,21	0,37	0,15	513
	EK	0,81	0,73	0,68	0,04	59
	EL	6,90	-0,66	-0,64	0,02	12
	EM	2,13	2,55	2,57	0,02	8
	EN	0,21	-1,02	-1,01	0,01	47
	EP	0,19	-1,63	-1,11	0,51	732
	EQ	-3,15	-0,57	-0,43	0,13	503
Ensemble des régions du test		**-0,09**	**-0,31**	**-0,21**	**0,10**	**2 131**

Tableau 5 : Dispersion des écarts entre les effectifs et les indices à 100 % et à 25 % de traitement des anomalies dans l'ordre du classement issu de la régression sur les cases Département*Nes16.

pour les régions du test

	T3-2003		
	Indice	Effectif	Nb de cases
Max	1,32	494	138
90%	0,37	81	
75%	0,15	27	
50%	0,06	5	
25%	0,01	1	
5%	0	0	
Min	0	0	

Le modèle ne propose donc pas un ordre de classement suffisamment efficace des établissements. La raison en est, **que pour 88 % des établissements, le gestionnaire se borne à confirmer la proposition du redresseur**, ce qui perturbe le modèle linéaire en introduisant beaucoup de valeurs nulles pour la variable à expliquer. Un deuxième enseignement de la modélisation est la très grande importance de la variable « contribution à l'indice d'évolution de l'emploi ». En effet, cette variable écrase toutes les autres dans le modèle. Ces enseignements nous amènent à poursuivre les travaux selon différents axes :

- Classer les établissements principalement selon la variable contribution à l'indice d'évolution (voir partie IV.3.)
- Si on crée une variable de type binaire 0/1 modélisant le fait que la valeur retenue par le gestionnaire corresponde ou non à la valeur proposée par le redresseur, une autre méthode est de réaliser une régression logistique expliquant cette variable en fonction des différentes variables disponibles. Les établissements pourraient alors être classés selon la probabilité estimée que le gestionnaire modifie la valeur proposée par le redresseur. Cette méthode a été testée et conduit à des résultats similaires à ceux présentés. Elle n'a pu être totalement menée à son terme et comme les résultats obtenus jusqu'ici sont moins pertinents que les autres méthodes ceux-ci ne sont pas détaillés dans ce document.

- La méthode exposée avec la régression logistique a l'inconvénient de ne pas prendre en compte l'ampleur de la correction sur les effectifs parmi les établissements pour lesquels le gestionnaire modifie la valeur du redresseur. Une solution serait de considérer que l'on est en présence de données censurées ou tronquées et d'utiliser par conséquent un modèle Tobit ; la variable latente associée reflèterait l'amplitude de la correction réalisée conditionnellement au fait que le gestionnaire ait pris la décision de modifier la valeur proposée par le redresseur. Cette piste non explorée jusqu'à présent devrait l'être prochainement.

4.3. Le classement direct

Cette approche consiste à classer directement les établissements à traiter dans chaque domaine de diffusion (croisement département et Nes16) selon leur contribution à l'indice d'évolution de la case à laquelle ils appartiennent. Plusieurs classements ont été testés et le plus pertinent est celui qui consiste à retenir dans l'ordre les variables suivantes :

- Département
- Nes16
- Contribution de l'établissement à l'indice d'évolution de la case Département*Nes16
- Évolution T/T-1 en valeur absolue (à 0 semaine de travail manuel) de l'établissement
- Tranche de note
- Tranche de taille
- Indicateur d'anomalie statistique sur les effectifs inscrits du 3ème mois du trimestre

Par construction, ce classement permet d'avoir en tête de liste les établissements dont la contribution à l'indice d'évolution est la plus importante et ensuite les établissements dont la contribution est nulle. Ces derniers sont classés à l'aide des trois dernières variables (tranche de note, tranche de taille et indicateur d'anomalie statistique sur les effectifs du 3ème mois du trimestre).

Les établissements dont la contribution est nulle ne sont pas à négliger dans les traitements des gestionnaires. En effet, une contribution nulle signifie que l'effectif à l'issue du redressement automatique est identique à l'effectif du trimestre précédent. Cependant, cela ne signifie pas que la contribution réelle de l'établissement est nulle et par conséquent que celui-ci n'a aucun impact sur l'indice. En fait, il arrive que les gestionnaires corrigent, parfois de façon importante, des établissements en anomalie ayant une contribution nulle à l'indice d'évolution.

Ce classement donne un résultat toutes régions et Département*Nes16 proche du classement issu de la régression. Néanmoins, par Nes16 (toutes régions), il est plus pertinent que le classement issu de la régression pour tous les secteurs sauf pour deux d'entre eux : EH et EN.

En poursuivant l'analyse du comportement des gestionnaires sur les données du test, on se rend compte qu'ils corrigent de façon importante, des établissements en anomalie ayant une contribution nulle à l'indice d'évolution dans 7 secteurs dont les deux signalés ci-dessus (EE, EH, EJ, EK, EL, EM et EN). Il semble alors plus judicieux d'en tenir compte et de proposer pour les 7 secteurs concernés le classement suivant (les autres secteurs restant soumis au classement direct -cf.ci-dessus) :

Case Département*Nes16	**15 %** des établissements en anomalies classés par : o Contribution **non nulle** de l'établissement à l'indice d'évolution de la case Département*Nes16 ; o Évolution T/T-1 en valeur absolue de l'établissement ; o Tranche de note ; o Tranche de taille ; o Indicateur d'anomalie statistique sur les effectifs inscrits du 3ème mois du trimestre.
	10 % des établissements en anomalies, pour lesquels la **contribution est nulle** classés par : o Tranche de note ; o Tranche de taille ; o Indicateur d'anomalie statistique sur les effectifs inscrits du 3ème mois du trimestre.
	75 % des établissements en anomalies classés par : o Contribution de l'établissement à l'indice d'évolution de la case Département*Nes16 ; o Évolution T/T-1 en valeur absolue de l'établissement ; o Tranche de note ; o Tranche de taille ; o Indicateur d'anomalie statistique sur les effectifs inscrits du 3ème mois du trimestre.

Cela permet d'améliorer très nettement les résultats, puisqu'on passe d'un écart en valeur absolue de 0,1 point (sur l'ensemble des régions du test) avec le classement par la régression à 0,04 point. Par case Département*Nes16, la qualité s'améliore également comme on peut le constater dans le tableau 6.

Tableau 6 : Résultat du traitement de 25 % des anomalies dans l'ordre du classement issu du classement direct avec reclassement des secteurs EE, EH, EJ, EK, EL, EM et EN pour les régions du test

T3-2003					
Nes16	**Indice initial avant travail manuel**	**Indice final avec 100 % des anomalies traitées**	**Indice avec 25 % des anomalies traitées**	**Écart**	**Écart en effectif**
EB	3,31	3,14	3,16	0,02	21
EC	0,11	-1,71	-1,68	0,03	20
ED	-1,33	-1,33	-1,33	0,00	0
EE	-1,36	-0,67	-0,64	0,03	28
EF	-1,45	-1,14	-1,12	0,03	51
EG	-9,49	1,88	2,01	0,12	32
EH	2,94	0,49	0,55	0,07	106
EJ	0,91	0,21	0,28	0,07	232
EK	0,81	0,73	0,80	0,07	95
EL	6,90	-0,66	-0,65	0,01	5
EM	2,13	2,55	2,49	0,05	17
EN	0,21	-1,02	-1,02	0,00	1
EP	0,19	-1,63	-1,49	0,14	198
EQ	-3,15	-0,57	-0,55	0,01	46
Ensemble des régions du test	**-0,09**	**-0,31**	**-0,27**	**0,04**	**816**

Une autre solution pour répartir les anomalies sur les cases consiste à classer les établissements sans tenir compte dans un premier temps du département et de la Nes16 selon les variables :

- Contribution de l'établissement à l'indice d'évolution de la case Département*Nes16
- Évolution T/T-1 en valeur absolue (à 0 semaine de travail manuel) de l'établissement
- Tranche de note
- Tranche de taille
- Indicateur d'anomalie statistique sur les effectifs inscrits du 3ème mois du trimestre

On « coupe » ensuite globalement à 25 % des anomalies puis on les répartit par case Département*Nes16 pour les traiter et recalculer l'indice après traitement des anomalies. Les résultats sont légèrement meilleurs sauf pour les secteurs EN et EP. Notons que pour ces 2 secteurs on traite seulement 22 % et 18 % des établissements contre 25 % selon l'autre méthode.

Au vu de ces résultats et des hypothèses choisies, le « meilleur » classement parmi ceux testés jusqu'à présent est le classement direct avec un traitement différencié pour les 7 secteurs EE, EH, EJ, EK, EL, EM et EN. Cependant, le fait de prendre arbitrairement 25 % comme seuil de traitement des anomalies dans chaque case peut avoir un impact sur la performance des classements. Par conséquent, le classement à retenir in fine sera choisi après la construction des critères d'arrêt.

4.4. Construction des critères d'arrêt

L'exercice inverse au précédent consiste à définir un écart maximal (en valeur absolue) d'indice souhaité par rapport à l'indice obtenu en traitant 100 % des anomalies et regarder à quel pourcentage d'anomalies traitées cela conduit. Ceci correspond à la démarche présentée dans la partie III.2. dans l'approche empirique pour la construction d'un seuil d'arrêt.

En effet, le critère retenu sur l'écart entre les indices se modélise pour une case donnée sous la forme suivante :

$$\left| \frac{\left(\sum_{j \in case} \left(y_T^j - y_{T-1}^j \right) \right)}{\sum_{j \in case} y_{T-1}^j} - \frac{\left(\sum_{j \in case} \left(y_T^{j\,*} - y_{T-1}^j \right) \right)}{\sum_{j \in case} y_{T-1}^j} \right| * 100 \leq S$$

où :

- $y_T^{j\,*}$ est l'effectif à l'issue de la vérification au trimestre T pour l'établissement j.
- y_T^j est l'effectif au trimestre T pour l'établissement j dans les fichiers de publication. Si l'établissement a été vérifié, c'est l'effectif à l'issue de la vérification (soit $y_T^{j\,*}$), mais s'il ne l'a pas été, on garde la valeur de l'effectif proposé par le redressement automatique.
- y_{T-1}^j est l'effectif au trimestre T-1 pour l'établissement j dans les fichiers de publication.

Il peut s'écrire aussi $\left| \frac{\sum_{j \in case} \left(y_T^j - y_T^{j\,*} \right)}{\sum_{j \in case} y_{T-1}^j} \right| * 100 \leq S$ soit encore avec les notations de la partie III $\left| \frac{\sum_{NV} \left(y_T^j - y_T^{j\,*} \right)}{\sum_{j \in case} y_{T-1}^j} \right| * 100 \leq S$. On retrouve donc la formule présentée dans la partie III.2.

L'écart maximal que nous retenons est de 0,05 point (soit S = 0,05). Cette valeur s'explique par le fait que les principes de diffusion actuels conduisent à la publication de l'indice avec un seul chiffre après la virgule. Un écart inférieur strictement à 0,05 point permet donc d'arrondir l'indice de la même façon que l'indice final obtenu avec 100 % des anomalies traitées.

Si on considère le graphique 1 qui représente l'évolution de l'indice au fur et à mesure du traitement des 50 anomalies de la case concernée, on constate que l'indice final (0,6) est approché à moins de 0,05 point deux fois lors du traitement, la première fois à l'anomalie n°14, et à partir de l'anomalie n°26. Avec le critère d'arrêt retenu, les traitements manuels sont arrêtés après l'anomalie n°14 c'est-à-dire à la première fois où l'on approche l'indice final. La fluctuation de l'indice qui suit l'anomalie 14 est due à la correction d'un établissement ayant une contribution jugée nulle à tort par le redresseur à l'indice d'évolution de l'emploi ou encore au traitement d'une création ou d'une cessation.

Graphique 1 : Simulation de l'évolution de l'indice lors du traitement des anomalies d'une case selon l'ordre d'un classement donnée.

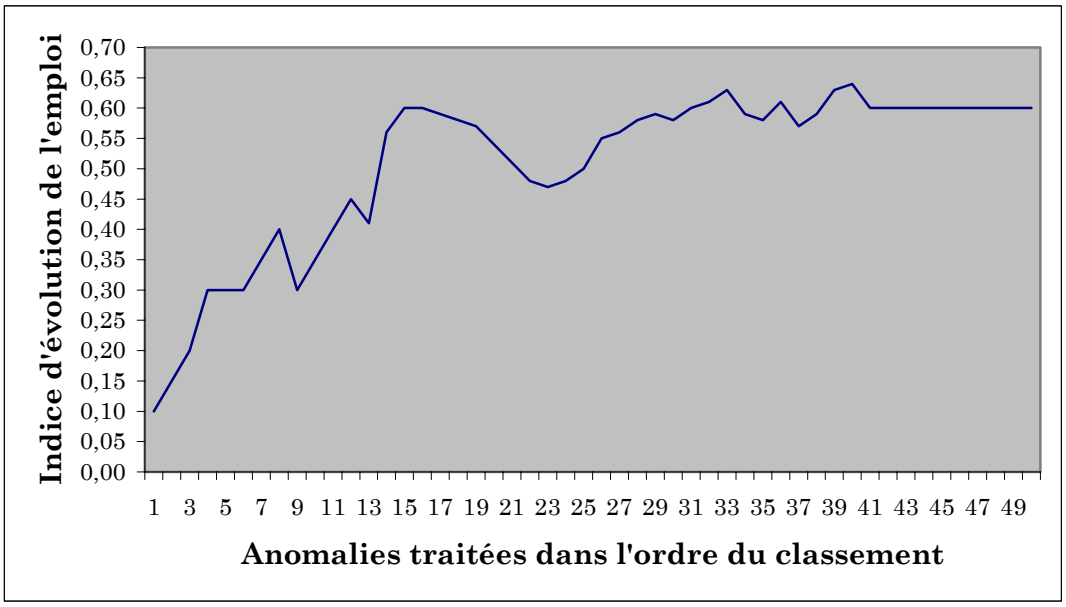

Tableau 7 : Nombre d'anomalies à traiter pour les régions du test selon le classement retenu

Classement issu :	**de la régression**		**du classement direct**		**du classement direct avec reclassement des secteurs EE, EH, EJ, EK, EL, EM et EN**	
T3 2003	% d'anomalies à traiter pour un indice identique	% d'anomalies à traiter pour un indice proche à moins de 0,05 point	% d'anomalies à traiter pour un indice identique	% d'anomalies à traiter pour un indice proche à moins de 0,05 point	% d'anomalies à traiter pour un indice identique	% d'anomalies à traiter pour un indice proche à moins de 0,05 point
Total régions du test	90,4	37,5	80,8	27,4	80,9	25,2

Lecture du tableau : En traitant 90,4 % des anomalies dans l'ordre du classement issu de la régression on obtient un indice identique à l'indice final pour chaque case Département*Nes16. En traitant 37,5 % des anomalies dans l'ordre du classement on obtient un indice approchant à moins de 0,05 point de l'indice final pour chaque case Département*Nes16.

En appliquant les critères d'arrêt décrits précédemment pour améliorer l'efficacité des gestionnaires, le classement direct avec reclassement des secteurs EE, EH, EJ, EK, EL, EM et EN conduit aux meilleurs résultats toutes régions confondues : **25,2 %** d'anomalies à traiter au lieu de 100 % pour approcher l'indice sur chaque case à moins de 0,05 points (voir tableau 7).

Ce classement ne fournit a priori que peu de gain par rapport au classement direct (25,2% d'anomalies à traiter contre 27,4%). Cependant, étant donné les volumes concernés, 3 % d'anomalies traitées en plus « représente » une semaine de travail manuel pour les gestionnaires[17]. La poursuite des travaux sur les données du 4ème trimestre 2003 permettra de conclure si le classement direct est finalement le plus pertinent à mettre en place.

Quel que soit le classement retenu, un quart des cases environ présente des écarts importants entre les effectifs des établissements non traités (après arrêt de la vérification selon le critère retenu) et les effectifs de ces même établissements traités (dans le cas de la vérification à 100 % des anomalies). Dans ces cases l'indice fluctue beaucoup au cours du traitement des établissements en anomalies comme le suggère le graphique 1. On peut alors se demander s'il est pertinent pour ces cases de retenir comme critère d'arrêt le rang du premier établissement permettant d'approcher l'indice final.

Tableau 8 : Dispersion du pourcentage d'anomalies à traiter par case selon le classement retenu pour obtenir un indice approchant l'indice final à moins de 0,05 point.

	T3-2003			
Classement issu :	**de la régression**	**du classement direct**	**du classement direct avec reclassement des secteurs EE, EH, EJ, EK, EL, EM et EN**	
	% d'anomalies à traiter	% d'anomalies à traiter	% d'anomalies à traiter	Nb de cases
Max	100,0	100,0	100,0	138
90%	86,0	72,9	73,1	
75%	62,3	51,1	42,9	
50%	15,6	10,6	10,6	
25%	3,9	4,1	4,1	
5%	0,0	0,0	0,4	
Min	0,0	0,0	0,0	

On constate que pour certaines cases 100 % des anomalies sont vérifiées alors que pour d'autres 0 % le sont (voir tableau 8). La dispersion du pourcentage d'anomalies vérifiées est relativement importante ce qui indique que choisir un seuil uniforme constant dans chaque case n'est pas la solution optimale. Cependant, choisir un critère d'arrêt propre à chaque case pose également le problème de sa validité temporelle. En effet, les résultats obtenus pour un trimestre ne seront pas forcément valables pour le trimestre suivant. Dans ce cas, le critère que l'on retiendra case à case est difficile à choisir : un des critères ou un critère composite ou encore un seul critère associé à d'autres méthodes à définir.

Les critères d'arrêt obtenu ci-dessus sont ensuite réappliqués dans chacune des cases afin de mesurer l'impact sur le total obtenu toutes régions confondues et par secteur. L'indice obtenu est quasiment identique à l'indice obtenu en traitant 100 % des anomalies sur l'ensemble des régions du test qualité (voir tableaux 9 et 10).

[17] Puisque les gestionnaires traitent 30 % des anomalies en 10 semaines.

Tableau 9 : Indice obtenu en appliquant les critères d'arrêt au classement direct

T3-2003					
Nes16	**Indice initial avant travail manuel**	**Indice final avec 100 % des anomalies traitées**	**Indice avec les anomalies traitées selon les critères d'arrêt**	**Écart**	**Écart en effectif**
EB	3,31	3,14	3,15	0,01	13
EC	0,11	-1,71	-1,72	0,01	6
ED	-1,33	-1,33	-1,33	0,00	0
EE	-1,36	-0,67	-0,68	0,01	11
EF	-1,45	-1,14	-1,13	0,01	24
EG	-9,49	1,88	1,87	0,01	2
EH	2,94	0,49	0,49	0,00	0
EJ	0,91	0,21	0,22	0,01	31
EK	0,81	0,73	0,73	0,00	3
EL	6,90	-0,66	-0,67	0,01	4
EM	2,13	2,55	2,55	0,00	1
EN	0,21	-1,02	-1,02	0,00	43
EP	0,19	-1,63	-1,62	0,01	29
EQ	-3,15	-0,57	-0,57	0,00	31
Ensemble des régions du test	**-0,09**	**-0,31**	**-0,30**	**0,01**	**84**

Tableau 10 : Dispersion des écarts entre les effectifs et les indices à 100 % de traitement des anomalies et les indices obtenus en traitant les anomalies selon les critères d'arrêt (classement direct).

	T3-2003		
	Indice	Effectif	Nb de cases
Max	0,05	23	138
90%	0,05	12	
75%	0,04	6	
50%	0,02	1	
25%	0,00	0	
5%	0,00	0	
Min	0,00	0	

La pertinence des différents classements les uns par rapport aux autres sur les données test du 3ème trimestre 2003 est confirmée à la fois pour les régions ayant participé au test qualité sur les données du 4ème trimestre 2003 et sur la France entière pour les données du 1er et du 2ème trimestre 2004 (travail en régime courant[18]). Il reste à déterminer quels critères d'arrêt seront appliqués en régime courant de production France entière pour chaque trimestre (moyenne des critères observés sur plusieurs trimestres ? ou maximum ?). De nouveaux tests qualité prévus d'ici la fin de l'année 2005 permettront de les valider pour quelques régions volontaires.

La méthodologie exposée dans ce papier avec en particulier la définition de critères d'arrêt sera mise en production en 2006 sur la chaîne EPURE. D'ici quelques années, il sera nécessaire de vérifier si les hypothèses retenues pour le classement des établissements en anomalies ainsi que celles relatives à la définition des critères d'arrêt eux-mêmes sont toujours valables. En effet, si ce n'est plus le cas, les résultats exposés dans ce papier concernant la proximité de l'indice issu de la chaîne avec celui obtenu dans le cas idéal où toutes les données sont vérifiées par les gestionnaires n'est plus forcément vraie. De nouveaux tests qualité devront être réalisés pour s'en assurer.

[18] Seules 30 % des anomalies ont été vérifiées. Le gain obtenu par les critères d'arrêt conduit donc dans ce cas à traiter moins de 30 % des anomalies pour obtenir un indice proche à moins de 0,05 point de l'indice final.

Bibliographie

[1] Lawrence D., McKenzie R., "The general application of significance editing", *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253, 2000

[2] Mauguin J., "Les procédures automatiques de contrôles de données dans les enquêtes annuelles d'entreprise », in *Echantillonnage et méthodes d'enquêtes*, Dunod, 2004

[3] Rivière P., "Enquêtes annuelles d'entreprise : à la rencontre du 4e type", *Courrier des statistiques*, n° 78, août 1996

[4] Un deuxième trimestre prometteur pour l'emploi salarié - Informations statistiques n°134-Septembre2004 - Direction régionale des Pays de la Loire :
http://www.insee.fr/fr/insee_regions/pays-de-la-loire/rfc/docs/Infostat134.pdf

ANNEXES

Annexe 1 : Nomenclature Economique de Synthèse en 16 postes

Nes16	
EA	Agriculture, sylviculture, pêche
EB	Industrie agroalimentaire
EC	Industrie des biens de consommation
ED	Industrie automobile
EE	Industrie des biens d'équipement
EF	Industrie des biens intermédiaires
EG	Énergie
EH	Construction
EJ	Commerce
EK	Transport
EL	Activités financières
EM	Activités immobilières
EN	Services aux entreprises
EP	Services aux particuliers
EQ	Éducation, santé, action sociale
ER	Administration
00	Indéterminé

<u>Annexe 2</u> : Signification de la note du redresseur

Cette note est attribuée aux données statistiques du cotisant par le redresseur automatique. Elle tient compte des anomalies détectées et de la taille de l'établissement.

Note	Anomalie sur les effectifs	Anomalie sur les masses salariales	Création	Cessation
'.'	Pas d'anomalie	Pas d'anomalie		
20 (*)	Pas d'anomalie	Pas d'anomalie		
19	- cotisant trimestriel : pas d'ano en T, ano en T-1. - cotisant mensuel: don manquante en m1 et/ou m2 et/ou ano en m3(T-1)	eff rémunérés redressés		
18		MS red de 0% à -10% & eff etab < 10		
17	- cotisant mensuel: - don manquante en m3 si eff<10 - 3 anos (ano ponctuelle ou don manquante) en m3(T-1), m1, m2 - ano ponctuelle en m2 et ano sur m3(T-1)			
16	- cotisant mensuel eff < 50: m0 et/ou m1 et/ou m2 en forte évolution et évolution faible entre m3 et m3(T-1)			
15		MS red. de 10% à -30% & eff etab < 10		
14	cotisant mensuel eff ≥ 50: m0 et/ou m1 et/ou m2 en forte évolution et évolution faible entre m3 et m3(T-1)			
13		MS red de 0% à -10% & 10 ≤eff etab < 100		
12	- cotisant mensuel eff < 50: m3(T-1) et/ou m1 et/ou m2 en forte évolution et évolution modérée entre m3 et m3(T-1)		sans anomalie statistique (eff > seuil**) ou valeur manquante redressée	Sans anomalie statistique (eff > seuil**) ou valeur manquante redressée ou anomalie ponctuelle ou pseudo-cessation
11		MS red de 0% à -10% & eff etab≥ 100		
10	- cotisant mensuel eff ≥ 50: m3(T-1) et/ou m1 et/ou m2 en forte évolution et évolution modérée entre m3 et m3(T-1)			
9	- cotisant trimestriel: en forte évolution, correct en T-1 (eff <10)			
8	- cotisant mensuel: m3(T-1) et/ou m1 et/ou m2 en forte évolution et forte évolution entre m3 et m3(T-1)	-MS red. de +30% & eff etab < 10 -MS brute nulle cot trimestriel		
7		MS red. de 10% à -30% & 10 ≤eff etab < 100		
6			Autre ano stat (cotisant trim eff ≤ 20)	Autre ano stat ou valeur nulle attendue, red à zéro (cotisant trim eff ≤ 20)
5	- cotisant mensuel: m3 en forte évolution, red val brute conservée (proche de m3(t-1))			

4		MS red. de 10% à -30% & eff etab ≥ 100		
3		MS red. de +30% & 10 ≤eff etab < 100		
2		MS red. de +30% & eff etab ≥ 100		
1		MS brute du trimestre nulle-cotisant mensuel actif		
0	- cotisant mensuel: m3 =0 ou en forte évolution, red en val inexploitable . - cotisant trimestriel: =0 ou en forte évolution, déjà en erreur en T-1, red val inexploitable.		Autre ano stat (cotisant mensuel).	Autre ano stat, valeur nulle attendue, redressée à zéro (cotisant mensuel), Restructuration

(*) les cotisants n'ayant qu'une anomalie d'identification auront également une note de 20.

(**) seuil variable selon les régions

Quand il est indiqué « mensuel » lire « mensuel ou trimestriel >10 »
En sachant que l'information portée par la note rend directement compte de l'anomalie la plus forte rencontrée sur le cotisant.

Exemple un cotisant noté 2 est un cotisant ayant 100 salariés ou plus pour lequel on a redressé sa masse salariale du trimestre d'au moins 30%. Il peut également être en anomalie sur m3 correspondant à la note 5; la note globale masque cette anomalie, mais le gestionnaire la découvrira s'il traite le cotisant.

Ce tableau permet de déterminer le seuil de traitement des données par les gestionnaires.

Annexe 3 : Description des variables utilisées pour la régression

Tranche de taille

- Moins de 10 salariés
- De 10 à 49 salariés
- De 50 à 199 salariés
- De 200 à 499 salariés
- 500 salariés et plus

Tranche de note

- **1 : Notes à '.', 1 à 4, 7, 11, 13, 15, 18 et 20** : notes pour lesquelles il n'y a pas, selon le redresseur, d'anomalies sur les effectifs ;
- **2 : Notes à 14, 16, 17 et 19** : anomalies sur les mois 1 ou 2 du trimestre et évolution faible sur le mois 3 entre T et T-1, anomalies sur le T-1, donnée manquante ponctuelle, … ;
- **3 : Notes à 6, 10, 12** : anomalies sur les mois 1 ou 2 du trimestre et évolution modérée sur le mois 3 entre T et T-1 - anomalies dues aux cessations et aux créations ;
- **4 : Notes à 5, 8 et 9 :** fortes évolutions sur le mois 3 ou sur le trimestre ;
- **5 : Note à 0 :** fortes évolutions sur le mois 3 ou sur le trimestre avec redressement en valeur inexploitable.

Taux de couverture en effectif des établissements en anomalie sur une case Département*Nes16 donnée

$$TXcouv = 100 * \frac{\sum_{j} \left| X_T^i \right|}{\sum_{i} \left| X_T^i \right|}$$

avec j : établissements en anomalies en S0 sur les effectifs inscrits du 3ème mois du trimestre
et i : ensemble des établissements de la case (dep*nes16)

Annexe 4 : Résultat de la régression (sortie de la procédure proc REG de SAS)

The REG Procedure
Model: MODEL1
Dependent Variable: diffs0s10

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	38268809	3478983	1211.47	<.0001
Error	30770	88362073	2871.69559		
Uncorrected Total	30781	126630882			

Root MSE	53.58820	R-Square	0.3022
Dependent Mean	2.81641	Adj R-Sq	0.3020
Coeff Va	1902.71133		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
trnot1	1	-2.54592	0.78871	-3.23	0.0012
trnot2	1	-3.53946	1.21015	-2.92	0.0034
trnot3	1	-8.75950	0.94911	-9.23	<.0001
trnot4	1	-3.79704	0.75894	-5.00	<.0001
trnot5	1	-4.73108	0.99624	-4.75	<.0001
contrib	1	2471.25767	23.19329	106.55	<.0001
couvertureff	1	37.11920	5.07116	7.32	<.0001
taill12	1	-3.92197	0.85172	-4.60	<.0001
taill13	1	-12.20818	1.29049	-9.46	<.0001
taill14	1	-15.35555	2.66862	-5.75	<.0001
taill15	1	53.29146	4.75524	11.21	<.0001