

LE CONTRÔLE DE QUALITÉ DES PHASES DE SAISIE-CHIFFREMENT DU RP90 EN FRANCE

Gérard Badeyan

Le recensement général de la population et des logements réalisé en France en 1990 a été marqué par un effort important de prise en compte de l'exigence accrue de qualité du produit statistique, dans ses divers aspects : adéquation du contenu des questionnaires aux besoins d'information, qualité de la collecte et enquête post-censitaire de mesure du degré d'exhaustivité, rapidité de mise à disposition des résultats, optimisation des tirages d'échantillon pour le dépouillement au 1/4 et au 1/20, précision des données, mesure de cette précision...

La présente note décrit les systèmes de contrôle qualité mis en place pour les phases de saisie-chiffrement des données, systèmes qui ont constitué un des principaux développements méthodologiques du RP90 sur cette question de la qualité, à partir d'une première expérience de cette nature menée à l'occasion du RP82.

En préalable on rappelle en un mot le plan d'exploitation retenu et l'organisation de la production à l'Insee car cela conditionne en partie les solutions retenues.

Le contenu du bulletin individuel porte d'une part sur les principales caractéristiques démographiques de l'individu (date et lieu de naissance, sexe, nationalité...), d'autre part sur sa situation professionnelle (profession, secteur d'activité économique et lieu de travail...). La saisie des données s'effectue de façon simple pour un grand nombre de variables (cases cochées ou chiffrement simple), mais nécessite un chiffrement complexe pour la profession et l'activité économique (renseignements fournis en clair par les enquêtés). La feuille de logement ne comporte que de la saisie simple. La plupart des variables simples à saisir ou chiffrer sont l'objet d'un traitement exhaustif (exploitation dénommée "exhaustif léger") alors que les variables complexes donnent lieu à un sondage au quart (avec les variables simples résiduelles).

Pour assurer la rapidité de mise à disposition des premiers résultats, l'exhaustif léger est traité en premier et le quart lourd en second. L'ampleur de la saisie à effectuer pour l'exhaustif léger (84 millions de documents) conduit à la sous-traiter alors que le quart lourd, compte tenu de la technicité du chiffrement, est réalisé par les agents de l'Insee, selon une technique de chiffrement assisté par ordinateur (dénommée "Colibri"). Ces

agents sont répartis dans une vingtaine d'établissements régionaux aux effectifs assez variables. Pour l'exhaustif léger comme pour le quart lourd, il se pose donc des problèmes spécifiques d'organisation et d'homogénéité de traitement.

Finalités des systèmes de contrôle de qualité mis en place pour les phases de saisie-chiffrement.

Les systèmes de contrôle de qualité pour les deux phases de saisie-chiffrement du RP90 (exhaustif léger et Colibri) ont poursuivi deux finalités :

- assurer un niveau de qualité du travail de production correspondant à des normes choisies *a priori*.
- permettre une mesure de l'erreur issue de cette phase du processus de production et produire des études méthodologiques sur ce thème.

Ces systèmes se sont inspirés des expériences conduites à l'occasion du recensement de 1982, en innovant particulièrement sur le deuxième point dans une perspective de publication.

Le contrôle de qualité de l'exhaustif léger.

Principes et normes

Comme indiqué plus haut, la saisie-chiffrement de l'exhaustif léger a été sous-traitée à des façonniers (sept au total). Les documents [84 millions de Bulletins Individuels (BI) et de Feuilles de Logement (FL)] ont été préparés en 242 lots (dénommés "unités de traitements") de volume variable par les établissements régionaux de l'Insee : aux alentours de 150 000 BI pour les premiers, puis compris entre 350 000 et 500 000 BI par la suite. Commencer par des lots de taille moindre permettait d'approvisionner très tôt les façonniers et de minimiser le coût d'un refus du travail, refus dont le risque était plus fort au démarrage.

Le contrat de sous-traitance spécifiait un niveau de qualité minimal à respecter pour que le travail soit accepté, niveau fixé à partir des considérations suivantes :

- obtenir un niveau de qualité au moins égal à celui du recensement précédent;
- être conforme aux normes de qualité de saisie habituelles dans les entreprises de saisie informatique.

Se contenter de prendre comme norme le niveau de qualité observé au recensement précédent peut paraître timide comme objectif. Ce n'était pas le cas car la priorité donnée au traitement de l'exhaustif léger qui permettait une publication beaucoup plus rapide des résultats constituait une nouveauté par rapport aux recensements précédents. Or cette priorité conduisait à un défi redoutable en terme de rythme, défi qui fut tenu : la saisie-chiffrement des 84 millions de documents fut achevée mi-janvier 1992, soit en huit mois et demi. D'un point de vue multidimensionnel (délais et qualité de saisie-chiffrement) on assurait ainsi une amélioration globale importante de qualité.

Le contrat spécifiait la clause de qualité de la façon suivante : dans le fichier livré, chaque document (BI, FL) correspondait à un enregistrement. Un enregistrement était réputé erroné dès lors qu'il contenait au moins un caractère erroné. Le fichier livré ne devait pas contenir plus de 3 % d'enregistrements de type BI ni plus de 1 % d'enregistrements de type FL erronés (conditions à remplir simultanément), les pourcentages d'enregistrements erronés étant estimés par sondage.

L'établissement de taux d'erreurs par enregistrements entiers et non code par code représentait une perte de précision acceptée pour simplifier la règle de décision pour les sous-traitants. En effet on ne tenait pas compte de ce fait de la plus ou moins grande difficulté des enregistrements, en particulier de leur longueur, ni d'éventuelles erreurs multiples sur un même enregistrement.

On a contrebalancé en partie ce fait (déjà présent au recensement précédent) en fondant la règle de décision sur deux sous-échantillons (sous-échantillon de FL, sous-échantillon de BI) : le sous-échantillon FL permettait de tester la qualité de la saisie pure, le sous-échantillon BI la qualité du chiffrement (si la qualité de saisie était bonne, ce qui était attesté par le résultat sur les FL, les erreurs sur les BI allaient alors provenir pour l'essentiel du chiffrement).

Pour les FL, le taux contractuel retenu (1 %) visait à imposer une double saisie aux façonniers. En effet en fonction du taux considéré comme habituel d'erreurs de saisie dans la profession (2 caractères erronés sur mille) et du nombre de caractères à saisir en moyenne par enregistrement, une saisie simple conduisait à des taux prévisibles d'enregistrements erronés qui ne pouvaient qu'induire des refus nombreux du travail, même compte-tenu de la marge prise pour limiter le risque de rejet à tort.

Pour les BI, le taux retenu se calait sur celui de 1982 (4 %), diminué à 3 % pour tenir compte de l'absence en 1990 d'un chiffrage qui avait causé beaucoup d'erreurs en 1982 (celui du type d'activité).

Le risque de rejet à tort fut pris faible (1/40), ce qui conduisait à des seuils de décision effectifs relativement élevés (2 % et 4,5 %). Ces seuils n'étant pas contractuels, on se réservait cependant la possibilité de décider au cas par cas pour tenir compte empiriquement de l'information apportée par les résultats constatés sur les unités de traitement précédentes du même façonnier.

Pour les BI, en plus du comptage des enregistrements erronés destiné à la prise de décision concernant l'acceptation ou le rejet du fichier concerné, les établissements régionaux comptabilisaient les erreurs par variable afin d'obtenir des mesures de qualité sur celles-ci prises isolément.

Mode opératoire du contrôle :

Une fois une unité de traitement constituée et avant son envoi chez le sous-traitant, l'établissement régional saisissait un échantillon de contrôle, en laissant les documents vierges de toute annotation, de sorte que le sous-traitant ne puisse pas repérer les documents qui en faisaient partie. Au retour du fichier réalisé par le sous-traitant, l'échantillon de contrôle préparé par l'établissement régional était apparié à la partie correspondante du fichier sous-traitant. Les enregistrements en divergence étaient alors édités et, par examen des documents correspondants, l'établissement régional déterminait si la divergence constatée était due ou non à une erreur du sous-traitant. On déterminait ainsi les pourcentages d'enregistrements erronés sur les sous-échantillons constitués par les enregistrements de type BI et de type FL.

Simple dans le principe un tel système pose dans le concret la question de l'homogénéité de traitement des divergences par les différents établissements. Dans de nombreux cas, en effet, on peut hésiter entre imputer une erreur au façonnier ou estimer que l'information contenue sur le bulletin est trop floue pour appeler un traitement univoque.

Résultats du contrôle :

Seules 4 unités de traitement parmi les 242 ont fait l'objet d'un refus. La qualité moyenne de la saisie-chiffrage s'est avérée bonne par rapport aux normes exigées : pour les BI 1,6 % d'enregistrements erronés pour une norme contractuelle de 3 %, pour les FL 0,5 % d'enregistrements erronés pour une norme contractuelle de 1 % ; 10 % seulement des unités de traitements dépassent sur l'échantillon de contrôle la norme de

1 % pour les FL, 60 % se trouvant en dessous du taux de 0,5 % et 6 % seulement dépassent la norme de 3 % sur les BI, 75 % ayant un taux inférieur à 2 %.

Ainsi non seulement l'exhaustif léger fut saisi en un temps très court, mais en plus avec une qualité nettement améliorée par rapport au RP précédent.

Cette réussite s'analyse par la conjonction de plusieurs facteurs : la précision des instructions de saisie-chiffrement fournies aux sous-traitants, l'effort de formation en direction du personnel recruté pour cette opération, l'efficacité de la gestion des documents dans les établissements régionaux qui a permis d'éviter toute rupture d'approvisionnement chez les façonniers, efficacité renforcée par la réduction de la taille de l'échantillon de contrôle grâce à l'optimisation du plan d'échantillonnage, ce qui limitait le coût du contrôle de qualité (saisie de l'échantillon de contrôle, analyse des divergences avec pour chaque opération des manipulations supplémentaires de documents) et évitait tout retard pour l'acceptation ou le refus des unités de traitements, la qualité de gestion de la sous-traitance (élimination rapide d'un sous-traitant défaillant, bonnes relations établies entre les sous-traitants et les établissements régionaux de l'Insee qui permettait une résolution rapide des difficultés rencontrées).

Par variable, les taux d'erreur constatés vont de 0,06 % pour les variables de saisie les plus simples (sexe) à 0,5 % pour certaines variables nécessitant un chiffrement (ilt : indicateur du lieu de travail) ou un quasi-chiffrement (cas du statut pour les bulletins renseignés de façon floue). La faiblesse, intéressante, de ces taux globaux ne doit pas masquer le fait que les erreurs peuvent être concentrées sur certaines modalités peu fréquentes et donc, pour ces modalités, représenter un taux relativement élevé. On ne peut qu'établir des bornes supérieures pour ces taux par modalités en raison du traitement manuel des divergences qui a conduit à limiter les comptabilisations demandées aux agents chargés du contrôle.

En relation avec la question de l'homogénéité de traitement des divergences signalée plus haut, on constate que le classement relatif des façonniers quant à la qualité du travail ne varie pas selon les DR, même si l'analyse est limitée du fait que certains façonniers se sont trouvés en relation presque exclusive avec un établissement pour des raisons de minimisation des coûts de gestion.

De fait les différences constatées dans les résultats sur la qualité du travail ont trouvé des explications dans les différences d'organisation et de gestion du travail des façonniers : renouvellement d'une partie du personnel en cours d'opération, importance de la formation préalable. En particulier les normes se sont révélées effectives pour la double saisie : les façonniers qui avaient opté pour une deuxième saisie limitée à une partie des documents à titre de test se sont trouvés en difficulté. Chez les façonniers ayant le mieux formé leur personnel, la deuxième saisie a même joué le rôle inattendu de contrôle supplémentaire du chiffrement, ce qui explique l'excellence étonnante de certains résultats.

Le contrôle qualité de Colibri¹

Principes

On rappelle en préalable que selon le plan d'exploitation retenu, le traitement du quart lourd ("Colibri") vient après celui de l'exhaustif léger. Colibri prend donc naturellement appui sur la saisie de l'exhaustif et la part de saisie simple se trouve très réduite par rapport au recensement précédent pour lequel Colibri précédait l'exhaustif. Les opérateurs avaient alors à saisir la totalité des données figurant sur les bulletins appartenant au quart.

La réalisation du travail de saisie-chiffrement complexe dans les ateliers de l'Insee donne l'opportunité de mettre en oeuvre un système de contrôle de qualité plus détaillé que pour l'exhaustif léger sous-traité.

Les codes à saisir et/ou chiffrer ont été répartis en sous-groupes :

- Les codes ne nécessitant que de la saisie ou du chiffrement simple;
- L'activité économique;
- La Profession et Catégorie Socioprofessionnelle (PCS) éclatée en :
 - PCS à un chiffre;
 - PCS à deux chiffres;
 - PCS à quatre chiffres;

Pour chacun de ces sous-groupes on se donnait comme objectif des taux d'erreur à ne pas dépasser sur toute la durée de l'opération.

Ces taux ont été établis en fonction des résultats du contrôle de qualité de Colibri 82 et d'une étude méthodologique consistant en la comparaison sur un échantillon de bulletins du chiffrement issu de Colibri 82 et d'un chiffrement manuel pour la PCS.

L'éclatement de la PCS se justifie par le fait qu'une erreur n'a pas la même portée selon le niveau de la nomenclature sur lequel elle porte.

(1) Colibri est le système de chiffrement en ligne des bulletins du recensement de la population et du logement.

Les taux correspondent aux taux moyens d'erreur constatés au milieu de l'opération 82. On assurait ainsi une qualité moyenne au moins égale à celle du recensement précédent. Par ailleurs en 1982 la norme était unique. En ce sens le contrôle 90 est plus sévère puisqu'il ne permet plus de compensations entre les principales variables. Il assure une qualité du fichier final plus homogène.

Période de croisière

Le principe de base du système a consisté à contrôler sur un échantillon la conformité du travail effectué en une semaine dans chaque établissement régional aux normes de qualité retenues pour chacun des 5 groupes de variables définis précédemment avec comme conséquence l'acceptation ou le refus global de ce travail.

En cas de refus le travail devait être repris en première saisie comme s'il n'avait jamais été traité (et non pas sous forme de mise à jour). Cette solution a été retenue pour éviter de créer une chaîne informatique supplémentaire pour un cas que l'on escomptait bien ne voir se présenter que rarement.

La semaine DR (semaine par direction régionale) a été retenue comme unité à laquelle s'appliquait le contrôle en raison du compromis à réaliser entre les deux contraintes suivantes :

- nécessité d'avoir un échantillon de taille suffisante pour obtenir une estimation de précision acceptable. Cette contrainte conduit à 1000 BI environ comme volume minimum;
- limiter le coût du contrôle. Ce coût correspond au taux de sondage. Une fois atteinte une limite raisonnable à la diminution du volume de l'échantillon, il était donc nécessaire d'augmenter le volume de l'unité contrôlée pour abaisser le coût. Dans ces conditions, avec 1000 comme taille d'échantillon, on était conduit à la semaine-DR comme unité contrôlée.

Les résultats des premiers contrôles s'étant révélés tout à fait satisfaisants, il fut décidé d'alléger pour les petites DR le coût du contrôle en prenant deux semaines consécutives comme nouvelle unité.

Pour la détermination des seuils effectifs d'acceptation du travail, le risque de rejet à tort d'un travail correct a été pris assez faible (1/20).

Période d'apprentissage

L'exigence de normes de qualité à respecter et la nouveauté du poste de travail pour les agents imposaient une période d'apprentissage.

Pendant cette période le travail de chaque agent a fait l'objet d'un contrôle exhaustif. Chaque fin de semaine on déterminait ainsi des indicateurs exacts de qualité pour chaque agent, à condition toutefois que le volume de documents sur lequel le contrôle fût suffisant pour être significatif (seuil fixé à 300 BI).

Un agent sortait de la période d'apprentissage lorsqu'il avait satisfait à l'ensemble des normes de qualité deux fois dans un ensemble de trois résultats consécutifs. Après deux mois et demi de travail, alors que les deux tiers des agents se trouvaient déjà en phase de croisière, le système a été légèrement assoupli pour permettre la sortie d'apprentissage d'agents qui avaient obtenu des résultats tangents (taux d'erreurs inférieurs à 110 % de la norme). Cet assouplissement ne risquait pas en effet de provoquer un rejet collectif ni des zones localisées de moins bonne qualité sur le fichier et il était nécessaire de diminuer le coût très lourd que représentait cette phase pour les équipes en termes d'avancement (chaque bulletin étant saisi deux fois) et d'encadrement (pour l'analyse des divergences, voir paragraphe suivant).

Mode opératoire concret

On a retenu un mode opératoire commun aux deux périodes, la seule différence consistant en ce que pendant la période d'apprentissage l'ensemble des bulletins était saisi deux fois par deux agents différents alors que pendant la période de croisière la deuxième saisie ne portait que sur l'échantillon de contrôle.

La deuxième saisie était effectuée en aveugle, c'est-à-dire sans que l'agent qui l'effectuait ait la moindre connaissance du contenu de la première. Les divergences entre les deux saisies étaient l'objet d'un arbitrage à l'écran. L'arbitre avait trois possibilités : confirmer une des deux saisies, ce qui classait l'autre en erreur ; valider les deux saisies, quoiqu'elles fussent différentes, en utilisant un indicateur de flou (si l'information contenue sur le bulletin, par son ambiguïté, faisait droit aux deux solutions adoptées par les colibristes) ; entrer une troisième valeur, ce qui classait les deux saisies comme fausses.

L'indicateur de flou permettait de traiter de façon équivalente la première et la deuxième saisies en période d'apprentissage. En période de croisière, seule la première saisie était conservée. L'indicateur de flou était alors utile pour les études de qualité du chiffrage.

De manière générale un grand nombre de procédures de cette phase d'arbitrage ont été stockées, ce qui a permis des études détaillées des erreurs de saisie-chiffrement au niveau des modalités elles-mêmes des codes les plus importants, ce qui représente une différence substantielle avec ce qui a pu être mis en place pour l'exhaustif léger.

Résultats du contrôle

Les contrôles négatifs se sont limités à deux sur l'ensemble de l'opération (environ 12 mois de contrôles collectifs pour chaque établissement).

Ce très bon résultat est à mettre en relation avec la rigueur de la phase d'apprentissage et avec une bonne utilisation du système par les équipes.

Des contraintes fortes avaient été établies pour la sortie de la phase d'apprentissage car le rejet d'une semaine entière de travail d'un établissement est très pénalisant d'un point de vue psychologique et en terme d'avancement du travail et l'on voulait se donner *a priori* le maximum d'atouts pour diminuer ce risque. Il faut dire que si, de ce point de vue, le système a atteint ses objectifs, en contrepartie on a sous-estimé les difficultés psychologiques induites par l'aspect "examen" de l'entrée en phase de croisière.

La bonne compréhension du système par les équipes lui a permis de jouer de façon satisfaisante un rôle avertisseur. En effet au cours d'une opération de cette longueur, une dégradation des résultats du contrôle qualité pouvait survenir dans un établissement pour des raisons diverses : esprit de routine induit par la répétition de résultats positifs, accent mis sur l'accélération du rythme d'avancement... Or on note que de manière générale un résultat inquiétant était bien pris en compte par les équipes, ce qui se traduisait par une amélioration sensible lors des contrôles suivants.

La bonne qualité du traitement se retrouve dans l'examen des taux d'erreurs par codes.

Pour les variables de saisie, les taux correspondent aux taux habituels pour ce type de travail (0,38 %).

Pour les chiffrements complexes les taux obtenus sont particulièrement bons et assez nettement inférieurs aux normes exigées.

Le taux d'erreur sur l'activité économique dépend de façon essentielle de l'aboutissement ou non de la recherche dans Sirene de l'établissement de travail (0,61 % en cas d'aboutissement, 2,82 % dans le cas contraire, ce qui donne un taux global de 1,17 %).

Pour la PCS les taux globaux sont particulièrement faibles : 0,72 % pour la PCS à un chiffre, 1,34 % pour la PCS à deux chiffres, 2,34 % pour la PCS à quatre chiffres. On

retrouve les différences habituelles de difficultés de chiffrage entre catégories : un taux d'erreurs très faible pour les agriculteurs (0,22 %), faible pour les ouvriers (0,48 %) et les employés (0,60 %), plus fort pour les cadres (1,39 %) et les professions intermédiaires (0,89 %) et, dans une moindre mesure, pour les commerçants, artisans, chefs d'entreprise (0,71 %). La répartition des taux d'erreurs entre catégories respecte ainsi le même ordre qu'en 1982 mais dans une fourchette beaucoup plus étroite.

On peut expliquer ce résultat par une bonne connaissance des nomenclatures par les agents, par l'efficacité du système d'assistance au chiffrage par l'ordinateur et aussi par la rigueur de la phase d'apprentissage.

ANNEXE

Les principales normes de qualité retenues correspondent aux taux d'erreurs suivants :

Codes de saisie ou de chiffrage simple

*sauf pour les communes suivantes (codes : commune de résidence antérieure,
commune de lieu de travail) :*

	6 10⁻³
– Paris	48 10 ⁻³
– Marseille	44 10 ⁻³
– Lyon	56 10 ⁻³
Activité économique	39 10⁻³
PCS à un chiffre	
– agriculteurs	8 10 ⁻³
– commerçants	35 10 ⁻³
– cadres	79 10 ⁻³
– professions intermédiaires	62 10 ⁻³
– employés	36 10 ⁻³
– ouvriers	33 10 ⁻³
PCS à deux chiffres	61 10⁻³
PCS à quatre chiffres	110 10⁻³