Sicore embarqué pour améliorer les libellés et raccourcir le questionnement : utilisation dans la future enquête emploi

25 janvier 2012

Sophie Destandau (*), Romain Warnan(*)

(*) Insee, Direction des statistiques démographiques et sociales, division Emploi

Introduction

Coder les réponses des enquêtés selon une nomenclature cible est une opération importante dans le traitement d'une enquête. En effet, cela concerne — par définition — des variables particulièrement utiles. Par exemple la profession, la catégorie sociale, le diplôme, l'activité de l'entreprise, etc.

Depuis longtemps, le codage est largement réalisé par des ordinateurs utilisant des logiciels spécialisés. Mais une partie des réponses doit toujours être codée « à la main » par des codeurs professionnels. Cette part d'échec de codage automatique varie entre 60 % pour les activités d'entreprise et 2 % pour les communes. Elle est de 15 % pour les professions.

Les échecs de codage sont dus la plupart du temps à la « mauvaise qualité » des libellés (trop imprécis, ou représentant une activité plutôt qu'une profession, ou encore correspondant à deux professions différentes, etc.). Or il est possible d'utiliser le logiciel de codage Sicore au moment de la collecte, sur le poste Capi [1] des enquêteurs. C'est ce qu'on appelle *Sicore embarqué*. C'est un outil précieux pour améliorer la qualité des libellés ainsi que le questionnement. Par exemple, en indiquant à l'enquêteur que le libellé ne permet pas le codage, on espère susciter une interaction entre l'enquêteur et l'enquêté pour trouver un meilleur libellé. De fait, Sicore embarqué peut être utilisé dans trois directions différentes :

- indiquer à l'enquêteur que le libellé ne permet pas le codage ;
- proposer une liste de possibilités lorsque le libellé est ambigu ;
- poser les questions correspondantes lorsque le logiciel Sicore indique avoir besoin de telle ou telle variable annexe pour coder.

Après avoir brièvement expliqué le fonctionnement de Sicore, cet article présente les différentes façons dont peut fonctionner Sicore embarqué. On verra le cas du tronc commun des enquêtes ménages (TCM). Puis on proposera une solution optimale qui utilise toute

1. *Computer assisted personal interviewing*, entretient assisté par ordinateur.

l'information disponible. Enfin on décrira plus en détail les solutions qui ont été retenues pour la future enquête Emploi (projet Reflee). L'article s'appuie souvent sur l'exemple de la variable profession, car c'est certainement pour cette variable que les enjeux sont les plus grands.

1 Coder des libellés en clair avec Sicore

Les enquêtes auprès des ménages comportent fréquemment des libellés en clair. Chacun de ces libellés doit être codé dans une certaine nomenclature. Les méthodes de codage ont évolué au cours du temps, en fonction des changements de logiciels de codage, des changements de nomenclatures et d'évolutions des questionnaires.

En outre, le passage à une collecte sur ordinateur portable, qui a été réalisé progressivement à partir de 1992, a permis de ne plus saisir *ex post* les réponses aux questions.

En 1993, l'Insee a débuté la construction de l'outil Sicore, remplaçant de Quid. Le système a été opérationnel dès 1995, et c'est toujours celui qu'on utilise aujourd'hui. Les principales exigences auxquelles satisfait Sicore sont :

- chiffrer *différentes variables* selon différentes nomenclatures ;
- chiffrer *rapidement* car les volumes sont parfois très importants (comme dans le recensement de la population) ;
- chiffrer *efficacement* pour limiter la reprise manuelle ;
- chiffrer *précisément* afin d'avoir des codes de qualité ;
- chiffrer *de manière homogène* pour assurer la comparabilité des données.

À l'heure actuelle, Sicore code de nombreuses variables :

- profession (PCS et PCS-ESE [2]) ;
- diplômes et niveaux de formation ;
- communes, pays et nationalités ;
- dépenses des ménages (produits de consommation et magasins) ;
- occupations (enquête emploi du temps) ;
- activités d'entreprise ;
- ...

1.1 Point de départ : le libellé

Lorsque Sicore doit coder une réponse libre selon une nomenclature cible, il commence toujours par tenter de reconnaître le libellé. Cette première phase se décompose en une classique étape de normalisation du libellé, puis en une seconde étape d'analyse du libellé.

1.1.1 Normalisation du libellé

D'abord la réponse en clair de l'enquêté est mise en majuscules. Puis les accents, les caractères spéciaux, les articles et les mots de liaison sont supprimés. Ensuite les mots vides de sens, répertoriés dans une liste, sont supprimés. Puis certains mots sont remplacés par un synonyme. Par exemple les expressions « mairie » , « commune » et « collectivité territoriale » sont toutes remplacées par « COLLECTERR [3] » . Enfin les mots en double sont supprimés et certains termes sont permutés.

2. Professions et catégories socioprofessionnelles des emplois salariés d'entreprise.
3. Contraction de « collectivité territoriale » qui n'utilise qu'un seul mot.

1.1.2 Analyse du libellé

Une fois le libellé normalisé, Sicore tente de le mettre en correspondance avec une liste de libellés de référence déjà normalisés : *le fichier d'apprentissage*. La comparaison ne se fait pas sur la totalité des deux chaînes de caractères, mais par groupe de deux lettres — des bigrammes — dans un ordre spécifique :

- d'abord le 2^{e} bigramme du 1er mot ;
- puis le 1er bigramme du 1er mot ;
- puis le 3^{e} bigramme du 1er mot ;
- *idem* pour le second mot ;
- ...

Dans la langue française, cet ordre est effet le plus discriminant entre deux libellés. Pour la profession, Sicore n'analyse que cinq mots de douze lettres au maximum. Cette analyse par bigrammes présente deux intérêts majeurs : d'une part elle est plus rapide qu'une comparaison de deux chaînes dans leur totalité, et d'autre part elle autorise des différences en fin de mot. Comme la marque du féminin est très souvent portée par la fin du mot, ce deuxième point est intéressant car il permet de ne lister que le masculin dans le fichier d'apprentissage. Pour Sicore, « boulanger » = « boulangère » . En plus, cela permet de passer outre certaines fautes d'orthographe. Par exemple, « statistissien » = « statisticien ». Les autres fautes d'orthographe plus fréquentes sont traitées par le biais des synonymes.

1.2 Qualité du libellé

Le libellé en clair est le point de départ de tout processus de codage, mais c'est aussi la variable qui contient le plus d'information. Un bon libellé est essentiel pour un chiffrement de qualité. Ainsi, Desrosières établit que la qualité du codage de la profession repose d'abord sur la qualité de la déclaration de la personne enquêtée concernant son occupation. Il note que *le dispositif d'interrogation lui-même a une influence sensible sur la forme des réponses concernant l'occupation* et que notamment *un dispositif administratif et officiel, dépourvu de la médiation d'un enquêteur*, comme le recensement n'aboutira pas aux mêmes libellés qu'une enquête menée par un enquêteur.

1.3 Affiner le codage avec les variables annexes

1.3.1 Tables et règles de décision

Bien souvent, le seul libellé ne suffit pas à coder précisément selon la nomenclature cible. Sicore gère donc un système de variables annexes associées à des règles de décision qui permette de préciser le codage. Exemples :

- l'environnement « commune » utilise la date comme variable annexe ;
- l'environnement « occupation » utilise entre autre l'heure, la durée et le lieu de l'activité ;
- l'environnement « profession » utilise entre autre le statut, la nature de l'employeur et l'activité de l'entreprise ;
- ...

Si le libellé déclaré ressemble à un libellé du fichier d'apprentissage, Sicore lui affecte un *précode*. Dans le fichier d'apprentissage, des groupes de libellés similaires sont mis en

regard du même précode. Cela signifie que ces libellés seront traités exactement de la même manière dans la suite du processus de chiffrement.

Le précode indique le nom d'une table de décision faisant appel à des variables annexes. Des combinaisons de modalités des différentes variables annexes sont prévues dans cette table, ce sont les *règles de décision*. Une règle qui s'applique dans tous les autres cas est généralement aussi prévue ; on l'appelle la *règle balai*. Chaque règle de décision est associée

- soit à un code de la nomenclature ; dans ce cas le codage prend fin ;
- soit au nom d'une autre table de décision ; dans ce cas le processus se poursuit jusqu'à aboutir à un code.

1.3.2 Écho de codage

Lorsqu'il reconnaît un libellé, Sicore va presque toujours trouver un code dans la nomenclature. Pour la profession, on vérifie cette remarque dans 98 % des cas. Cela provient de l'existence des règles de décision par défaut (règles balais), et cela satisfait à l'exigence d'efficacité de Sicore. Autrement dit, l'existence des règles balais augmente le taux de codage automatique. Ces règles balais indique à Sicore la situation la plus probable. Exemple pour la profession :

En l'absence d'information sur le statut, Sicore considère que l'individu est salarié car il n'y que 10 % d'indépendants environ. Cependant si l'enquêté s'est déclaré architecte, Sicore choisira plutôt un statut libéral par défaut, car les architectes libéraux sont deux fois plus nombreux que les architectes salariés.

Ce système de choix par défaut basé sur la situation la plus probable augment le taux de codage automatique mais peut nuire à la qualité du codage. Sicore retourne donc en même temps que le code un *écho de codage*. Cet écho nous renseigne sur la qualité du codage. Il y a quatre classes d'échos :

1. **Codage de qualité** : les variables requises étaient toutes bien remplies
2. **Codage réussi sous réserve de qualité** : il a manqué au moins une variable annexe
3. **Libellé reconnu mais non codé** : une information importante a manqué
4. **Libellé non reconnu**

Dans les cas où il a manqué des variables annexes, Sicore indique leurs noms et l'ordre dans lequel elles ont manqué. *C'est grâce à ces informations qu'on peut optimiser Sicore embarqué.*

1.4 Codage de la profession avec Sicore

La profession est une variable particulièrement lourde à coder. Dans ce paragraphe on décrit dans les grandes lignes comment se déroule son chiffrement avec Sicore.

Depuis 2003, on utilise la nomenclature des PCS révisée pour classer les professions. Cette nomenclature compte 486 rubriques qui sont agrégées en 31 catégories socioprofessionnelles. La nomenclature sépare les indépendants des salariés et les fonctionnaires des salariés du public. Elle hiérarchise les indépendants suivant la taille de l'entreprise et les salariés selon leur niveau de qualification. Par ailleurs l'activité de l'entreprise est déterminante pour classer au niveau détaillé.

On comprend donc que le nombre de variables annexes — et donc de règles de décision — est particulièrement important dans « l'environnement PCS ».

1.4.1 Fichier d'apprentissage de la profession

Le fichier d'apprentissage de l'environnement PCS compte 27 000 lignes. Bien que certaines lignes soient presque des doublons, il faut garder en mémoire les principes de l'analyse par bigramme et des synonymes qui augmente le nombre de libellés reconnus. En plus, dans le fichier d'apprentissage, le caractère « $ » prend la place de n'importe quel mot et permet de reconnaître encore plus de libellés.

Ces libellés de référence sont associés à 3 000 précodes différents qui conduisent aux règles de décision [4].

1.4.2 Variables annexes utilisées

Au total il y a 13 variables annexes que Sicore PCS peut utiliser :
- STATUT : statut dans l'emploi (salariés, indépendants) ;
- PUB : statut de l'employeur (public, privé) ;
- CPF : position professionnelle ou qualification (ouvriers, ingénieurs...) ;
- NAF2 : division d'activité de l'établissement employeur sur deux positions ;
- NBS : nombre de salariés employés (distinction entre artisans, commerçants et chefs d'entreprise) ;
- NAF : sous-classe d'activité de l'établissement employeur sur cinq positions ;
- FN : fonction professionnelle (fabrication, installation, vente, secrétariat...) ;
- T : taille de l'entreprise (petite, moyenne, grande) ;
- S : sexe ;
- SP : statut d'apprenti ;
- DEP : département ;
- OPA : orientation principale agricole ;
- SAU : surface agricole utilisée.

Si on veut coder précisément une profession au hasard, il faudrait demander sept questions aux salariés et huit aux indépendants [5]. Cela n'est pas envisageable dans la plupart des enquêtes, en particulier dans l'enquête Emploi, car elle comporte de nombreuses professions (principale, secondaires, un an avant, des parents...).

On peut nuancer ceci en considérant la figure 1. On peut en effet y lire qu'avec les seules variables statut et nature de l'employeur [6], 53 % des individus sont correctement codés au niveau profession (quatre positions). Ce pourcentage augmente rapidement : 65 % avec une variable supplémentaire, 74 % avec deux variables en plus. Sachant que le graphique ne tient pas compte des 13 % de libellés non reconnus et donc le pourcentage de « convergents » plafonne à 87 %.

1.4.3 L'activité de l'établissement employeur

L'une des variables utiles au codage de la profession est l'activité de l'établissement employeur. Elle n'est utilisable par Sicore PCS qu'à condition d'avoir été au préalable codée selon la nomenclature d'activité en vigueur en France : la NAF rév. 2. La règle est la suivante : avec un code d'activité sur deux positions (division) on obtient un bon

4. Les 3 000 premières règles sont appelées *règles mères*. Elles conduisent pour la plus part à d'autre règles : les *règles filles*

5. Certaines questions sont communes aux deux statuts.

6. La variable STRE est fixée à 7 et n'alourdit donc jamais le questionnement.

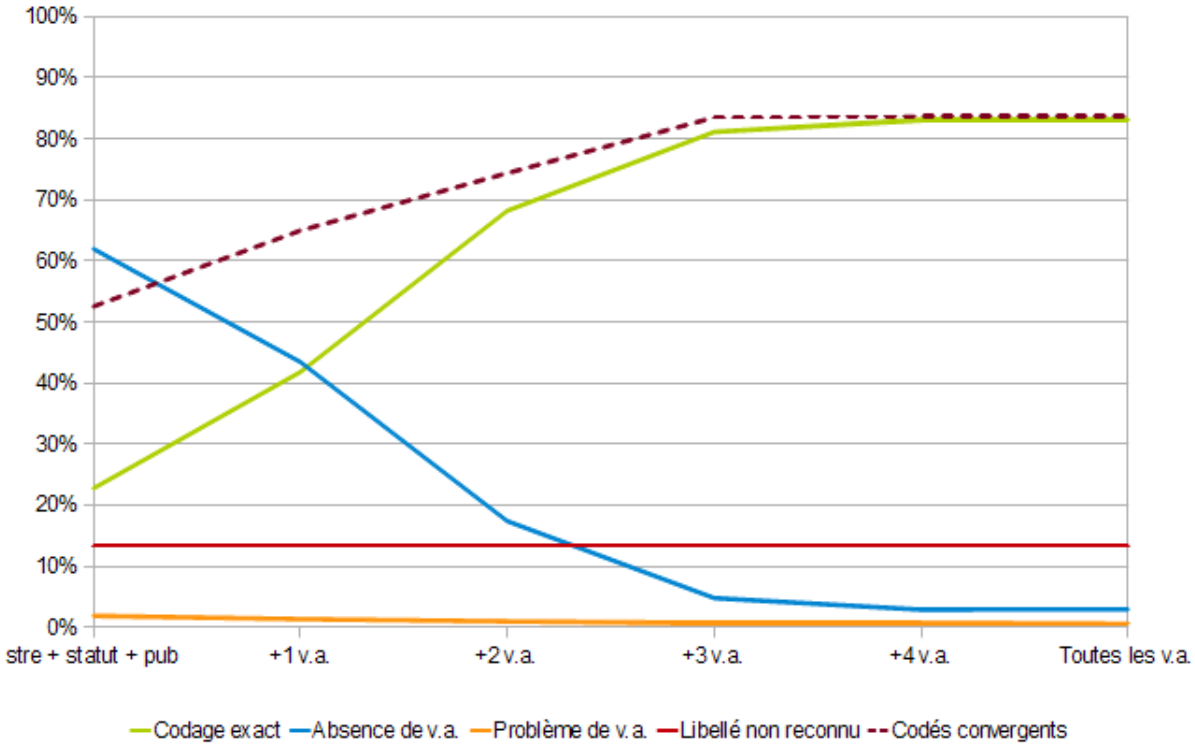

FIGURE 1 – La qualité du codage augmente rapidement avec le nombre de variables annexes utilisées

code de catégorie sociale sur deux positions. Avec un code d'activité sur cinq positions (sous-classe), on obtient un bon code de profession sur quatre positions.

Pour cette raison, l'environnement Sicore PCS est en fait constitué de deux environnements : l'un utilisant la NAF2 qui donne une catégorie sociale de qualité et l'autre utilisant la NAF5 qui donne une profession de qualité [7]. La plupart des enquêtes ménage utilisent l'environnement en NAF2. Certaines autres (enquête Emploi, recensement de la population, enquête FQP) utilisent dès que possible l'environnement en NAF5.

Autant la NAF2 peut se déterminer lors de l'enquête grâce à un Sicore embarqué pour l'activité, autant la NAF5 ne peut pas être déterminée sans appariement avec le répertoire Sirene. Cet appariement n'est actuellement pas possible au moment de l'interview pour des raisons techniques.

2 Sicore embarqué sur le portable des enquêteurs

Vu les éléments présentés avant, notamment concernant le recueil de la profession, l'Insee a décidé d'embarquer Sicore sur le poste de collecte des enquêteurs. Le codage, jusqu'alors réalisé à l'Institut, se rapproche de la source d'information que constitue l'enquêté. L'un des objectifs est de recueillir ainsi une information de meilleure qualité.

7. Attention cependant, Sicore retourne toujours des codes détaillés sur quatre positions, mais avec une NAF2, seuls les deux premiers chiffres sont significatifs.

2.1 Reconnaître le libellé déclaré

2.1.1 Premier avantage de Sicore embarqué

Un avantage immédiat d'avoir Sicore sur le poste des enquêteurs est de savoir au moment de la collecte si le libellé déclaré est reconnu ou pas. On rappelle ici que si le libellé n'est pas reconnu, il y a *nécessairement* échec de codage automatique, et donc reprise manuelle.

Au moment de l'enquête, l'enquêteur pose une question ouverte à l'enquêté. Par exemple : « Quel est le dernier diplôme que vous avez obtenu ? ». Et il saisit sa réponse dans le champ réservé. Sicore est alors appelé sur ce libellé et fournit instantanément une réponse. La partie de la réponse intéressante n'est pas tant le code de la nomenclature — calculé avec peu, voire aucune, variable annexe — mais l'écho de codage. Celui-ci nous indique en effet si le libellé a été reconnu ou pas. Au cas ou il n'a pas été reconnu par Sicore, l'enquêteur en est averti et il a la possibilité de faire préciser sa réponse à l'enquêté.

Ensuite le questionnaire se déroule comme prévu, ce qui laisse entier le problème des variables lourdes à collecter comme la profession. Pour un codage de qualité, il faut toujours poser de nombreuses questions.

Cette technique a été mise en place dans le tronc commun des enquêtes ménages pour coder la profession. Sur l'enquête SILC (*statistics on income and living conditions*, condition de vie des ménage), le taux de codage automatique est passé de 80 % à 90 %. Cela divise par deux la charge de reprise !

2.1.2 Limites du simple Sicore embarqué

Avec ce simple Sicore embarqué, l'objectif d'efficacité est largement rempli mais qu'en est-il de la précision ? Durant le test Capi du questionnaire de la future enquête Emploi, on a pu enregistrer deux libellés de profession. On regarde ce que l'enquêteur a saisi en premier et, si cela n'a pas été reconnu par Sicore, comment il a modifié le libellé. Quantitativement, on note une nette diminution du taux d'échec de codage automatique, conformément à ce qui a été observé dans SILC.

D'un point de vue qualitatif, le second libellé est :

- meilleur[8] que le premier dans 60 % des cas ;
- appauvri dans 25 % des cas (ex. : *technicien process* → *technicien*) ;
- identique ou vide dans 15 % des cas.

Parmi les libellés codés en deuxième instance, 40 % sont des libellés appauvris. En effet, avec le système des règles balais, Sicore parvient à coder des libellés assez vagues comme « technicien » alors qu'il ne reconnait pas certains libellés trop précis. Ce point n'est pas trop inquiétant dès lors qu'on dispose d'une bonne information annexe. Les enquêteurs devront être sensibilisés à ce sujet.

Force est de constater que ce simple Sicore embarqué affiche des résultats plutôt encourageants : une augmentation notable du taux de codage acquise au prix d'une faible baisse de qualité. La principale critique qu'on peut lui opposer c'est qu'il a plutôt tendance à rallonger le temps de questionnement (deux libellés sont parfois saisis au lieu d'un seul), même si cela est assez marginal. En tout cas, ce Sicore embarqué ne permet pas de diminuer le nombre de questions posées.

8. Par meilleur, on entend reformulé sans perte d'information ou simplifié à raison. Exemples : *assistante manager* → *secrétaire de direction* ou encore *ajusteur monteur en aéronautique* → *ajusteur monteur*.

2.2 Implémentation dans le tronc commun des enquêtes ménages

Vu les bons résultats obtenus par ce simple Sicore embarqué, il a été décidé de l'incorporer dans le tronc commun des enquêtes ménages. Mais pour les professions le projet est allé plus loin : on a voulu alléger le questionnement en ne posant pas de questions superflues lorsque le codage est déjà de qualité.

Dans les enquêtes à tronc commun, on ne code la PCS qu'au niveau catégorie sociale. On verra que dans ce cas, on peut se passer de certaines variables annexes. Malgré tout, la NAF2 est considérée comme une variable d'intérêt. Par ailleurs, les variables STATUT et PUB sont jugées indispensables, pour le codage de la profession mais aussi pour les filtres dans la suite du questionnaire.

2.2.1 Importance relative des variables annexes pour chiffrer la PCS

Suivant la précision et la qualité recherchées lors de la phase de chiffrement, on peut omettre certaines variables annexes prévues par Sicore PCS. En effet, elles n'ont pas toutes la même importance : certaines sont essentielles alors que d'autres sont d'utilisation plus marginale. En posant les questions les plus discriminantes en premier, on augmente la probabilité d'avoir un codage de qualité tôt dans le processus de codage. Cela peut permettre de poser moins de questions.

Importance théorique Certaines variables sont très clivantes par construction de la nomenclature des PCS. D'autres en revanche interviennent à un niveau plus détaillé et dans des cas plus particuliers. On peut dresser une liste — presque — ordonnée en fonction de l'importance des variables annexes :

1. STATUT : utilisée dans presque tous les cas pour séparer les indépendants des salariés. Cette variable joue au niveau du groupe social (1^{er} niveau).
2. CPF : utilisée pour beaucoup de salariés, elle permet de hiérarchiser les salariés au niveau groupe social.
3. PUB : utilisée pour beaucoup de salariés, elle permet de séparer les fonctionnaires des salariés du privé. Elle joue au niveau CS [9], et parfois au niveau groupe [10].
4. NBS : utilisée pour presque tous les indépendants, elle permet de les hiérarchiser au niveau CS.
5. NAF2 : utilisée pour tous, essentiellement pour situer le contexte d'exercice du métier, elle intervient au niveau CS. Elle est très importante pour les petits indépendants car on peut assimiler l'activité de l'entreprise à la profession.
6. OPA, SAU : ne concernent que les agriculteurs, elles permettent de les hiérarchiser au niveau CS.
7. FN : utilisée pour les salariés, elle intervient fréquemment au niveau profession mais aussi au niveau CS [11].
8. NAF : utilisée pour tous dans de nombreux cas pour situer le contexte d'exercice du métier, elle intervient au niveau profession.

9. Catégorie socioprofessionnelle.
10. Par exemple : une femme de ménage qui ne travaille pas chez des particuliers est classée parmi les ouvriers (groupe 6) sauf si elle est fonctionnaire : elle est alors classée parmi les employés (groupe 5).
11. Entre autre pour séparer les cadres techniques des cadres administratifs et commerciaux.

9. T : utilisée pour hiérarchiser les chefs d'entreprise, et séparer les cadres en grande entreprise, elle intervient uniquement au niveau profession.
10. SP : utilisée pour séparer certains apprentis parmi les ouvriers. Ils sont alors considérés comme non qualifiés, cette variable intervient donc au niveau CS.
11. DEP, S : utilisation tout à fait marginale pour les agriculteurs et les aides familiaux.

Cette hiérarchisation théorique des variables se vérifie assez bien empiriquement.

Ordre des variables dans Sicore PCS Plus que l'ordre d'importance théorique des variables annexes, c'est l'ordre dans lequel les utilise Sicore qui importe. En effet, dès qu'une variable annexe demandée par Sicore est manquante, elle est ajoutée à la liste des variable manquantes et le codage est signalé comme douteux.

L'objectif est donc de recueillir les premières variables demandées par Sicore dès le début du processus. De cette manière, le codage emprunte le chemin le plus rapide vers un code de qualité, c'est-à-dire un chemin qui évite les règles balais.

Les graphiques suivants indiquent quelles sont les quatre premières variables demandées par Sicore PCS dans le cas des salariés puis des indépendants. Les variables STATUT et PUB sont supposées connues.

Lecture pour les salariés : *Avec le libellé, le statut et la nature de l'employeur, sur le champ des salariés :*

- *38 % sont parfaitement codés (*),*
- *la 1re variable manquante est la qualification (CPF) dans 42 % des cas,*
- *la 1re variable manquante est la fonction (FN) dans 6 % des cas,*
- *la 1re variable manquante est l'activité (NAF) dans 14 % des cas*

Avec le libellé, le statut, la nature de l'employeur et la 1re variable, sur le champ des salariés :

- *60 % sont parfaitement codés (*),*
- *la 2e variable manquante est l'activité (NAF) dans 25 % des cas*
- *...*

Ces graphiques confirment l'importance de l'activité de l'entreprise, le rôle de la qualification pour les salariés (figure 2) et du nombre de salariés pour les indépendants (figure 3).

2.2.2 Variables manquantes pour le codage de la profession

Sachant cela, on se fixe un nombre restreint de variables annexes. On choisit de conserver les plus importantes : les seules variables que l'on recueille sont les suivantes : STATUT, CPF, PUB, NBS, NAF2, OPA, SAU et FN. Et l'objectif est de recueillir CPF, FN, NBS, OPA et SAU uniquement pour les personnes concernées.

Le cas de OPA et SAU concerne uniquement les agriculteurs exploitants. Grâce à Sicore embarqué pour l'activité [12], on dispose toujours de l'activité sur deux positions. On peut donc ne demander la surface et l'orientation des production qu'aux individus concernés. C'est-à-dire les indépendants dont l'activité commence par 01 (culture et production animale, chasse et services annexes), 02 (sylviculture et exploitation forestière) ou 03 (pêche et aquaculture).

Comme cela a été précisé, Sicore retourne un code de la nomenclature, un écho de codage et aussi la liste ordonnée des variables qui ont manqué durant le codage. Le Sicore

12. Voir la section 4.

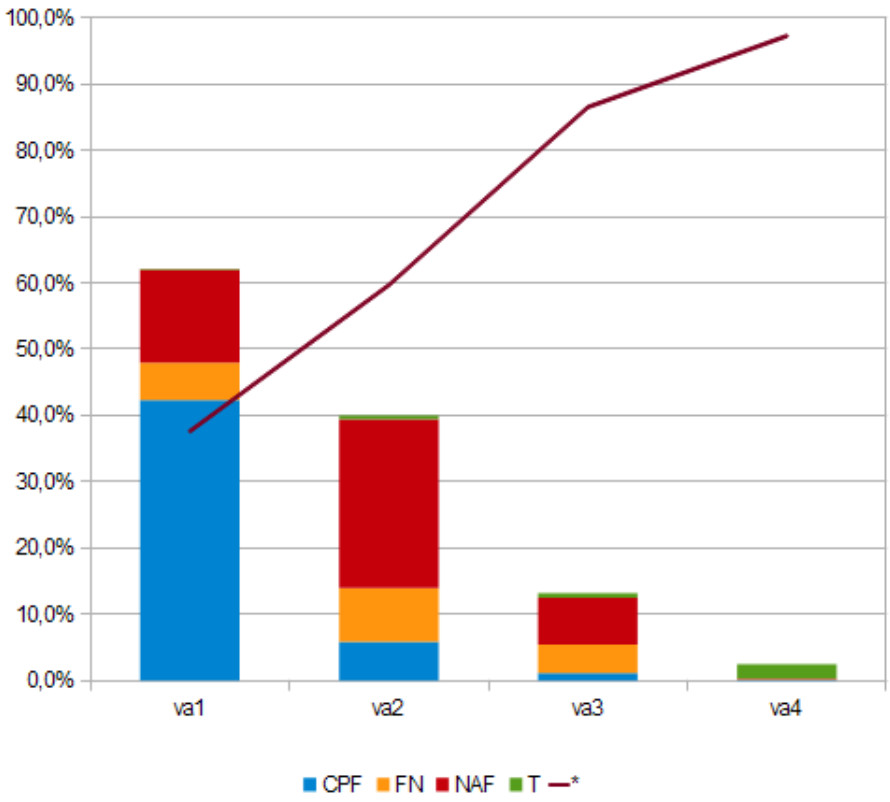

FIGURE 2 – Salariés : la qualification et l'activité de l'employeur en premier

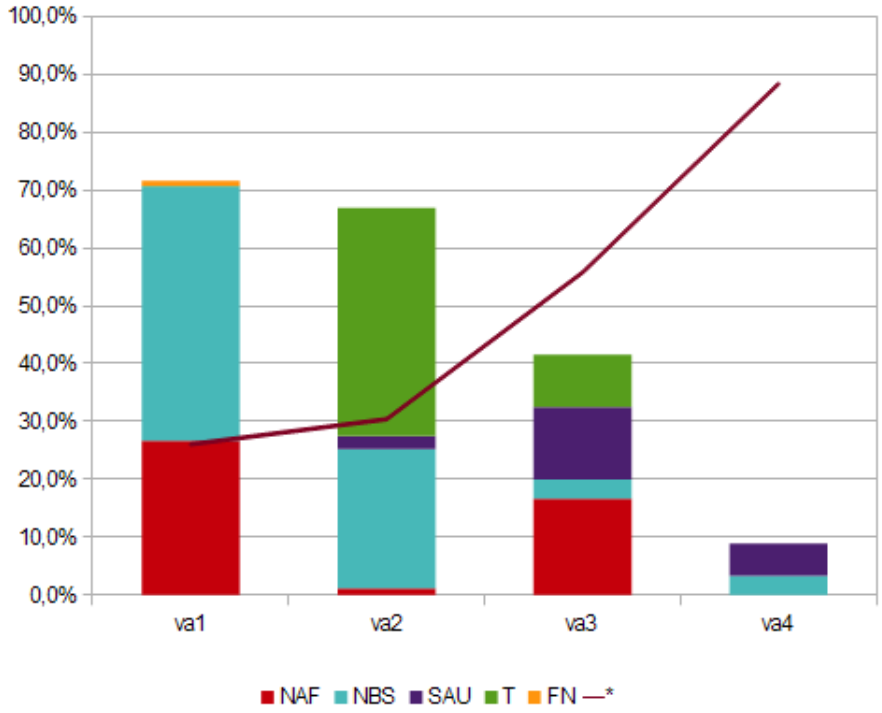

FIGURE 3 – Indépendants : la taille et l'activité de l'entreprise dirigée en premier

embarqué du TCM analyse cette liste et débloque les questions associées à CPF, FN ou NBS uniquement si ces variables font partie de la liste des variables manquantes.

Pour un salarié le codage se déroule ainsi :
- l'enquêteur demande d'abord le statut, la nature et l'activité de l'employeur ;
- ensuite il demande la profession et note le libellé ;
- Sicore PCS est appelé sur le seul libellé et les trois variables déjà collectées ;
- si le libellé est reconnu
  - Sicore PCS parcourt la liste des variables annexes,
  - si elle est vide ou si elle ne contient ni FN, ni CPF, le chiffrement de la profession est terminé,
  - si elle contient CPF ou FN, les filtres associés aux questions correspondantes sont activés et une ou deux questions sont posées ;
- si le libellé n'est pas reconnu
  - les filtres associés aux questions CPF et FN sont activés et les deux questions sont posées.

Ce Sicore embarqué conserve l'avantage de pouvoir reconnaître le libellé au moment de l'entretien. Il contribue donc à augmenter le taux de codage automatique. Et en plus il permet de diminuer le nombre de questions posées aux enquêtés. Cela fait gagner du temps sur certains questionnaires, ce qui est bénéfique pour l'Insee et pour les enquêtés.

2.2.3 Critique du système

Le système constitue une bonne solution alternative pour le codage au niveau catégorie sociale. Il présente cependant au moins un défaut.

Prenons l'exemple d'un salarié dont le libellé a été reconnu mais pour lequel la liste indique qu'il faut encore collecter CPF et FN. Les deux questions sont donc activées, à commencer par la question liée à CPF. Il est tout à fait possible — c'est même fréquent —, que le fait d'avoir CPF renseignée rende la question sur la fonction superflue. Or elle sera quand même posée parce qu'il n'y a pas d'appel Sicore après la collecte de CPF.

La situation inverse (plus grave) peut aussi se produire : la liste des variables manquantes ne comporte que CPF. Seule la question correspondante est activée, alors qu'il peut se produire que pour une valeur de CPF donnée et non vide, le processus de codage requiert FN. Et cela alors que la règle balai en l'absence de CPF empruntait un chemin qui ne nécessitait pas FN.

Plus généralement, ce système ne s'adapte au codage détaillé (niveau profession) qu'à condition de faire un appel Sicore supplémentaire après chaque nouvelle variable collectée.

3 Un Sicore embarqué optimal pour la profession

Ici on décrit le fonctionnement d'un Sicore embarqué pour la profession. L'objectif est d'améliorer la qualité du codage tout en diminuant le nombre de questions posées. Pour cela, il convient d'exploiter toute l'information que nous donne Sicore après un codage.

La description se veut assez générale de manière à pouvoir être reprise dans n'importe quelle enquête. Par ailleurs, ce qui est dit pour la profession peut très bien s'étendre à d'autres variables qui mobilise un minimum d'information annexe.

3.1 Principe général

On va exploiter l'écho de codage (figure 4) et la liste des variables annexes manquantes afin de poser uniquement les questions effectivement nécessaires au codage de la profession.

Écho de codage	Libellé	Qualité	Remarque
CCS, RCS	Reconnu	Codage de qualité	Le code CCS est impossible
RC*	Reconnu	Il manque des variables annexes	On a un code PCS dans 98 % des cas
CCM	Incomplet		Le code CCM est impossible
C_R, C_C, R_V, R_R, R_B	Non reconnu		

FIGURE 4 – Les échos de codage de Sicore PCS

Lecture du tableau : *un écho de codage égal à CCS ou RCS indique que le libellé de profession est reconnu et que le codage est de qualité. En pratique, le code CCS n'est jamais retourné.*

Certaines variables sont collectées avant même que l'on ne recueille le libellé de profession, ou toujours connues, comme le sexe.

Certaines variables, comme l'activité de l'établissement employeur, peuvent avoir un intérêt propre en dehors du codage de la profession. D'autres, comme le statut, sont trop importantes dans le chiffrement de la PCS pour être négligées. On les demandera donc systématiquement.

Certaines variables sont d'utilisation très marginale, comme le statut d'apprenti. Enfin certaines, comme la sous-classe d'activité ne sont pas toujours possibles à collecter. On ne les demandera donc jamais.

Toutes les autres variables seront facultatives : elles seront recueillies ou pas suivant le libellé de profession. Une variable sera recueillie si et seulement si elle est effectivement utile dans le codage dudit libellé.

3.2 Algorithme détaillé

Initialisation Pour le codage d'une profession donnée, on définit d'abord quatre listes de variables :

- celles qu'on connaît déjà (L_c) ;
- celles qu'on demande forcément (L_f) ;
- celles qu'on demande si nécessaire (L_o) ;
- celles qu'on ne demande pas (L_p).

Les variables de L_f doivent toujours être collectées avant celles de L_o. À chaque variable de L_f et L_o, on associe un nom de question (nom de la variable de l'enquête) et un questionnement.

Reconnaissance du libellé

1. D'abord on collecte les variables de L_f, puis on demande le libellé de profession.
2. Une fois que le libellé a été recueilli, on lance un appel Sicore avec les variables de L_c afin de déterminer si il est reconnu ou pas.
3. S'il n'est pas reconnu, on recueille un second libellé et on fait un second appel Sicore avec les variables de L_c.

4. Si ce second libellé n'est pas reconnu (CCM, C_R, C_C, R_V, R_R, R_B) , on pose toutes les questions associées à L_o.
5. Sinon on procède comme suit, on notant L_m la liste des variables annexes manquantes :

Sélection des variables annexes utiles

1. Si l'écho de codage est CCS ou RCS, le codage de la profession est terminé.
2. Sinon l'écho est RC* :
   (a) On cherche la 1re variable de L_m qui appartient à L_o et qui n'a pas déjà été collecté ;
      i. Si il n'y a plus de variable vérifiant cette condition, le codage de la profession est terminé.
      ii. Sinon on va à l'étape (b) :
   (b) On collecte cette variable, en respectant les filtres et les redirections ;
   (c) On refait un appel Sicore avec cette variable supplémentaire ;
   (d) On retourne à l'étape 1.

3.3 Problème de l'ordre des questions

Les questionnaires réalisés en Blaise sont linéaires, c'est-à-dire que l'ordre des questions est préétabli. On peut en passer certaines grâce aux filtres, mais on ne peut pas intervertir deux questions pendant le questionnement. Or, l'algorithme décrit dans le paragraphe précédent suppose ce genre de manipulations. On ne peut en effet pas savoir une fois pour toutes dans quel ordre vont être recueillies les variables annexes. Certes, il existe un ordre plus fréquent mais aucune possibilité n'est à exclure.

On peut malgré tout contourner cette contrainte. Il suffit en effet de répéter le bloc optionnel définit par L_o autant de fois qu'il y a de variables dans L_o. Ces blocs optionnels sont tous identiques (à un numéro d'ordre près) et on place un appel à Sicore entre chacun d'eux. La liste des variables manquantes retournée indique quelle unique question du prochain bloc doit être activée. Le programme Blaise sera assez volumineux — $(\text{card}L_o)^2$ questions pour seulement $\text{card}L_o$ questions posées au maximum — mais cela n'a pas vraiment d'importance.

Grâce à cet artifice, on obtient un programme linéaire qui donne l'impression que les questions sont posées dans un ordre indéterminé *a priori*.

3.4 Gains escomptés grâce à la nouvelle méthode

On se place dans le cas raisonnable suivant :

$$\begin{cases} L_c = (\text{S}, \text{DEP}, \text{SP}) \\ L_f = (\text{STATUT}, \text{NAF2}) \\ L_o = (\text{PUB}, \text{CPF}, \text{FN}, \text{NBS}, \text{T}, \text{SAU}) \\ L_p = (\text{NAF}, \text{OPA}) \end{cases}$$

Autrement dit :
– on connait toujours le sexe et le département de l'individu ;
– on sait aussi s'il est apprenti, par exemple parce qu'on connait son type de contrat ;

– on ne veut jamais se passer du statut et on considère l'activité comme une variable d'intérêt (sans avoir les moyens de récupérer une NAF détaillée) ;
– on ne demande pas l'orientation des productions agricoles car on pense pouvoir l'approximer avec l'activité ;
– enfin, on s'autorise à poser ou pas toutes les autres questions.

Sur des données de l'enquête Emploi, on a simulé le fonctionnement du Sicore embarqué optimal. Ensuite on a compté le nombre de questions posées au total pour en déduire un nombre moyen de question posées. Pour les salariés, on pose 3,42 questions en moyenne et pour les indépendants, 3,90. Dans l'ensemble, **on pose 3,47 questions pour coder la profession** (dont deux questions obligatoires) en plus du libellé.

Sans le Sicore embarqué, et pour atteindre la même qualité de codage, **il faudrait poser 5,89 questions**. En effet, pour un salarié, on demande : STATUT, NAF2, PUB, CPF, FN et T — 6 questions ; et pour un indépendant : STATUT, NAF2, NBS, SAU et T — 5 questions [13].

Si l'activité n'est pas une variable d'intérêt de l'enquête, **Sicore embarqué permet de ne poser que 3,07 questions** par individu.

Un autre avantage d'utiliser ce Sicore embarqué est qu'il ne pose que les questions qui ont un sens vis à vis de la profession de l'enquêté. Cela évite aux enquêteurs de se retrouver dans des situations embarrassantes, comme par exemple poser une question à laquelle l'enquêté a déjà répondu. Il peut être gênant de demander sa position professionnelle à quelqu'un qui l'a déjà indiqué dans sa profession comme : « ingénieur en informatique ». De même, on sait que la variable de fonction laisse perplexe de nombreux enquêtés : ils ne parviennent pas à se classer parmi les dix modalités de réponse possibles. Cette situation ne pourra plus se produire puisque la fonction ne sera demandée qu'aux personnes que cela concerne et qui savent donc se classer parmi les réponses prévues.

4 Application dans la future enquête Emploi

4.1 Présentation de la future enquête Emploi

L'enquête Emploi de l'Insee est un élément central de l'observation structurelle et conjoncturelle du marché du travail et de la situation des individus vis-à-vis de ce marché. Actuellement, elle est collectée par les enquêteurs auprès des ménages, en continu en métropole et annuellement dans les quatre Dom suivants : Guadeloupe, Martinique, Guyane et Réunion.

Elle permet :
– de mesurer directement les grandes catégories d'activité retenues par le BIT (chômage, population active, sous emploi...), ce qui permet notamment de comparer le niveau et l'évolution des taux d'activité et de chômage de la France à ceux des autres pays européens ainsi que ceux des pays membres de l'organisation internationale du travail ;
– de préciser les caractéristiques (durée du travail, temps partiel, multiactivité...) et la structure des emplois ;
– de caractériser les trajectoires individuelles ;

13. Compter SAU est un peu exagéré, c'est une question qu'on pourrait assez facilement ne poser qu'aux agriculteurs exploitants. Cependant, cela ne change pas grand chose puisque les indépendants représentent environ 10 % des actifs occupés.

– de constituer une base de données permettant la réalisation d'études approfondies sur les différentes approches de l'emploi.

Les données de l'enquête Emploi sont souvent utilisées par les autres enquêtes de l'Insee ou par les partenaires du système statistique public. La plupart des instituts de sondage et organismes d'études utilisent pour leurs travaux des données de cadrage issues de l'enquête Emploi.

En décembre 2007, un projet de refonte de l'enquête Emploi à horizon 2013 a été lancé, impliquant à la fois

– la rénovation du questionnaire et
– le développement d'une nouvelle application de gestion.

Ce projet inclut notamment une interrogation en continu dans les Dom comme en métropole. L'objectif principal de cette refonte est de gagner en qualité à tous les niveaux.

4.2 La rénovation du questionnaire de l'enquête Emploi

Le questionnaire de l'enquête Emploi comporte un certain nombre de questions que l'enquêteur doit remplir en saisissant un libellé. Trois types de libellés à coder coexistent au sein du questionnaire de l'enquête Emploi : les libellés d'activité, de formation et de profession.

Plus précisément, dans le cadre du projet de refonte de l'enquête Emploi, le questionnaire comprend :

– 9 libellés d'activité : activité principale de l'établissement employeur actuel, autres activités liés à la profession principale (jusqu'à 2) et activités secondaires (jusqu'à 3) si l'individu est actif, activité à l'entrée dans l'entreprise et activité un an auparavant si elles sont différentes de la première ; activité antérieure si l'individu n'est pas actif.
– les libellés de formation : niveaux d'études actuel ou atteint, diplôme obtenu et plus haut diplôme atteint, et spécialités de formation ;
– 9 libellés de profession : profession principale et professions secondaires (jusqu'à 3) si l'individu est actif, profession à l'entrée dans l'entreprise et profession un an auparavant si elles sont différentes de la première ; profession antérieure si l'individu n'est pas actif au moment de l'enquête ; professions des parents.

Dans le cadre de la rénovation du questionnaire de l'enquête Emploi, Sicore sera embarqué sur le poste Capi de collecte de l'enquêteur pour traiter ces trois types de libellés.

Un des objectifs à l'utilisation de Sicore embarqué dans le questionnaire est d'améliorer la codification du libellé — en jouant sur la relation entre l'enquêteur et l'enquêté — lorsque le libellé saisi ne permet pas la codification via deux mécanismes possibles :

– indiquer à l'enquêteur que le libellé est ambigu et ouvrir une liste de codifications possibles ;
– indiquer à l'enquêteur que le libellé ne permet pas la codification et demander un deuxième libellé.

L'environnement Sicore APE, qui n'utilise pas de variable annexe, est construit selon le premier mécanisme : lorsque le libellé ne permet pas de coder sans ambigüité une NAF2, une liste déroulante de possibilités est proposée. Ce mécanisme est donc utilisé dans la future enquête Emploi pour coder les activités d'établissement.

Les environnements PCS, diplômes et niveaux Sicore n'ont pas été construits de façon à offrir la première possibilité. Ce serait trop lourd en raison des multiples variables annexes et de leur poids dans la codification. C'est pourquoi, pour coder les libellés de profession et de formation, la future enquête Emploi utilise le second mécanisme, avec une gestion

des variables annexes différentes selon les cas et la précision de la codification visée.

4.2.1 Sicore embarqué sur les libellés d'activité

Les libellés d'activité pour lesquels on ne cherche pas une grande précision et dont on pense que les enquêtés peuvent à peu près comprendre les intitulés de la nomenclature NAF sur 2 positions, seront codés entièrement en interactif sur le poste de collecte Capi des enquêteurs.

Après la saisie unique du libellé d'activité, l'enquêteur va au bout de la démarche du codage durant l'entretien. Ainsi, après avoir collecté le libellé d'activité, Sicore lui indique un code sur 2 positions de la nomenclature NAF ou des propositions de nomenclature de code NAF au niveau division. L'enquêteur, en accord avec l'enquêté, choisit parmi ces propositions. S'il n'est pas satisfait des propositions, il choisit un item dans une liste déroulante de toutes les activités.

Les libellés d'activité — mise à part celui de l'activité principale de l'établissement employeur — sont donc entièrement codés en interactif et l'application informatique ne revient pas sur ces codes contrairement à ce qui se fait dans l'application de l'enquête Emploi actuelle.

Plus précisément, l'enquêteur demande le libellé de l'activité économique principale de l'établissement employeur (module B sur les activités professionnelles) ou de l'entreprise de l'enquêté lorsqu'il est actif. Sicore APE est lancé sans aucune variable annexe. En retour, l'enquêteur :

- obtient un seul code ou
- propose une liste d'une dizaine maximum de codes en rapport avec le libellé ou
- propose une liste déroulante de toutes les activités (classification hiérarchique).

Par exemple, le libellé d'activité « Conseil en informatique » est reconnu en collecte par Sicore et codé en code 65. Le libellé d'activité « Bâtiment » après le codage avec Sicore aboutit à une proposition de deux libellés de nomenclature de code :

- « Travaux de construction spécialisés » correspondant au code 43 ;
- « Construction de bâtiments » correspondant au code 41 ;
- et une autre possibilité « autres cas » qui aboutit à la liste complète des codes.

En revanche, il y a échec complet du codage avec Sicore pour le libellé « Tabac PMU » ; l'enquêteur proposera directement la liste déroulante complète à l'enquêté.

4.2.2 Sicore embarqué sur les libellés de formation

Pour les libellés de formation, l'enquêteur recueille dans un premier champ, le libellé déclaré par l'enquêté le plus compréhensible et adapté au type de libellé demandé. il ne se contente pas, par exemple, d'une déclaration de l'enquêté du type « baccalauréat » ; l'enquêteur doit affiner cette déclaration dès le premier libellé sans toutefois formater le libellé pour qu'il soit codé ! Ce premier libellé est alors traité avec Sicore avec les variables annexes disponibles à ce moment là de l'entretien. S'il n'est pas reconnu, un 2ème libellé est demandé à l'enquêté. Pour savoir si le libellé est reconnu ou pas, le rôle de l'écho de codage Sicore est essentiel car il renseigne justement sur la qualité du codage.

Dans Sicore, il y a deux variables annexes pour coder les libellés de formation : type d'enseignement et de l'année d'obtention du diplôme ou d'atteinte du niveau. À l'Insee, il y a deux nomenclature pour coder le libellé de formation une pour les niveaux et l'autre pour les diplômes.

Pour les libellés de formation, en plus de la qualité du libellé, la cohérence du libellé est également assurée avec Sicore embarqué grâce à l'interaction de l'enquêté avec l'enquêteur. Un certain nombre de libellés de formation sont donc reconnus en collecte et codés. Dans ces cas de réussite, le code trouvé est confronté à l'année d'obtention du diplôme ou année d'atteinte du niveau d'études. Si le libellé n'est pas cohérent avec cette variable annexe, l'enquêteur redemande à l'enquêté le libellé de formation ou corrige la variable annexe.

Une « roue de secours » a été également ajoutée pour les libellés de formation non reconnus à la fin du processus de Sicore embarqué : une liste fermée de niveau ou de diplômes est proposée à l'enquêté. Elle est utilisée par les codeurs en reprise sur le poste de reprise dédié.

4.2.3 Sicore embarqué sur les libellés de profession

Les libellés de professions sont traités de manière différente aux 2 autres types de libellés dans le futur questionnaire de l'enquête Emploi. L'objectif à l'utilisation de Sicore embarqué sur ces libellés est bien d'obtenir des libellés de qualité et « codables ».

Il n'y a pas de codage complet en collecte comme pour les libellés d'activité — 13 variables annexes sont nécessaires au codage de la profession avec Sicore dont la NAF sur 5 positions — et il n'y a pas non plus de contrôle de cohérence sur le libellé, ni de « roue de secours » prévue en collecte comme pour les libellés de formation.

Il n'a pas été prévu de modifier les bases d'apprentissage de l'outil Sicore, ni le choix des variables annexes. Sicore embarqué fonctionne à l'identique de celui utilisé en batch sur le poste de collecte de l'enquêteur, et ceci afin d'inciter à une réelle interactivité entre l'enquêteur et l'enquêté pour trouver le meilleur libellé (au sens de la codification Sicore). L'objectif est d'améliorer la qualité de la codification tout en diminuant le volume de reprise manuelle des échecs de codage.

Ce mécanisme de Sicore embarqué présente un double avantage par rapport au Sicore embarqué utilisé pour les libellés d'activité. D'une part, on conserve une trace du premier libellé saisi. Il servira notamment à la reprise. Il est aussi utile pour comparer les seconds libellés avec les premiers. D'autre part, on diminue le nombre de questions posées aux enquêtés. Cela fait gagner du temps, ce qui est bénéfique pour les enquêtés.

Plus précisément, la profession principale bénéficie d'une attention la plus importante par rapport aux autres professions. Trois passages de Sicore au maximum sont programmés sur le libellé de profession avec un nombre croissant de variables annexes.

Pour les autres libellés de profession, deux passages de Sicore au maximum sur deux libellés saisis sont programmés avec également un nombre croissant de variables annexes.

L'idée du Sicore embarqué de Reflee est fort simple : si l'écho retourné appartient à la première classe, on cesse de récolter des variables annexes. Dans le cas contraire, on continue.

Concrètement, pour une profession donnée, on suit les étapes suivantes :

1. recueil des variables indispensables (essentiellement le statut) ;
2. saisie du libellé de profession ;
3. appel Sicore embarqué pour savoir si le libellé est reconnu ;
4. s'il ne l'est pas, saisie d'un 2^{e} libellé de profession ;
5. recueil des variables d'intérêt pour l'enquête (comme l'activité de l'entreprise) ;
6. autre appel Sicore avec le libellé et toutes les variables récoltées ;
7. analyse de l'écho de codage ;

8. s'il indique un codage réussi et de qualité, fin du codage de cette profession ;
9. sinon, recueil de la variable restante la plus importante dans le chiffrement de la PCS et retour à l'étape 6.

Exemple pour la profession antérieure :
- *Vous étiez : à votre compte ? salarié ? chef d'entreprise salarié ? aide familial non-salarié ?*
- *Quelle était la nature de votre employeur ?*
- *Dans cet emploi, quelle était votre profession ?*
- Appel à Sicore PCS avec le libellé, STATUT et PUB
- En fonction de l'écho de codage : *Pouvez-vous reformuler cette profession ?*
- *Quelle était était l'activité principale de votre entreprise ?*
- Appel à Sicore PCS avec le libellé, STATUT, PUB et NAF2
- En fonction de l'écho de codage et du statut : *Combien y avait-il de salariés dans l'entreprise dans laquelle vous travailliez ?* Ou bien : *Êtes-vous classé comme manœuvre, ouvrier qualifié, etc. ?*

Remarque : si le deuxième libellé n'est toujours pas reconnu, on recueille le maximum de variables prévues pour cette profession.

4.2.4 Les risques de l'utilisation de Sicore embarqué

Le risque le plus évident est que l'enquêteur veuille saisir dans le premier libellé, un libellé qu'il sait que Sicore Embarqué reconnaît pour ne pas avoir à saisir un second libellé. Or, un libellé n'est pas forcément faux quand il n'est pas reconnu par Sicore : il peut s'agir d'un nouveau libellé à intégrer éventuellement (si l'expert variable le juge) dans la base de connaissances de Sicore.

En formation des gestionnaires de l'enquête Emploi et des enquêteurs, il a été et sera bien précisé que l'objectif de l'utilisation de Sicore embarqué sur les libellés de professions et les formations dans le questionnaire n'est pas de coder les libellés en collecte ; il s'agit bien de collecter des libellés de qualité.

Le premier libellé saisi correspondant à une déclaration *spontanée et explicite* du libellé par l'enquêté constitue une variable d'intérêt de l'enquête. Elle sera étudiée en tant que telle par les chercheurs.

Les deux saisies du libellé feront également l'objet régulièrement de comparaisons à des fins méthodologiques pour surveiller une éventuelle dérive de saisie « formatée » des libellés.

4.3 Refonte de l'application informatique

L'application Emploi sert en amont à constituer l'échantillon trimestriel de l'enquête qui est utilisé par Capi et en aval à fournir les fichiers de collecte issus de Capi sous forme de tables SAS. Entre les deux, elle traite automatiquement certaines données, puis code manuellement les libellés de formation, et d'emploi.

La refonte de l'application de l'enquête Emploi porte notamment sur :
- les traitements automatiques des données,
- les postes de reprise des variables de formation et de profession ;

et a permis la création de postes de travail permettant de visualiser l'avancement et les résultats de la future enquête Emploi par les différents acteurs.

Outre le fait que les libellés de formation et de professions sont censés être de meilleure qualité grâce à Sicore embarqué et le fait que les libellés d'activité sont codés en collecte, les résultats de Sicore embarqué servent par ailleurs.

4.3.1 Utilisation des résultats de Sicore embarqué dans le traitement automatique des données

Le traitement automatique des données dans l'application Emploi consiste en deux types d'opération :

- appariement avec Sirène des coordonnées de l'établissement employeur saisies lors de l'interview afin d'obtenir un code NAF de l'activité d'établissement ;
- codification automatique avec Sicore des libellés de formation et de profession.

Dans la future application Emploi, ces deux types d'opération ont été maintenus et rénovés tous deux pour plusieurs raisons.

Appariement avec Sirène des coordonnées de l'établissement employeur Le codage de la profession principale par Sicore embarqué lors de la passation de l'enquête n'est pas aussi précis que celui réalisé dans l'application car il repose sur une variable d'activité plus regroupée (deux positions seulement). Dans l'enquête Emploi, le codage de l'activité principale doit être le plus détaillé possible : un code de la nomenclature NAF sur 5 positions est donc attendu. Pour cela, les coordonnées de l'établissement employeur saisies en collecte doivent être appariées avec un sous-fichier Sirène (dit appariement SIAM). Or cet appariement n'est pas possible sur le poste de collecte Capi de l'enquêteur. Il n'est possible que dans l'application informatique.

Outre le codage en NAF sur 5 positions, cet appariement permet d'obtenir d'autres informations sur l'établissement employeur s'il est identifié. Par exemple, le SIRET, la catégorie juridique de l'entreprise, les tranches d'effectif de l'établissement et de l'entreprise, le nom et la commune de l'établissement employeur déclarés dans le répertoire Sirene.

Le codage sur 5 positions de la NAF permet également d'améliorer le codage de la profession à l'entrée dans l'entreprise (ou dans la fonction publique) et de la profession un an auparavant lorsque l'établissement employeur n'a pas changé.

Le code sur deux positions de l'activité principale issu de Sicore embarqué intervient dans le choix des échos fournis par l'appariement avec Siam des coordonnées de l'établissement employeur en batch.

La codification automatique avec Sicore des libellés de formation et de profession Cette deuxième codification des libellés de formation et de profession avec Sicore est maintenue car un bug de Sicore embarqué sur le portable des enquêteurs n'est pas à exclure, une codification automatique dans l'application est donc programmée même si dans la grande majorité des cas, elle donne le même résultat.

D'autre part, lors du codage automatique avec Sicore, on compile le libellé de profession avec le libellé de grade ce qui facilite le codage de certaines professions (agents de la fonction publique d'État, territoriale ou hospitalière notamment) avant de lancer Sicore sur ce libellé combiné. La majorité du temps, le grade empêche le codage automatique du libellé de profession. Il était donc plus simple en collecte de ne pas présenter un libellé combiné de profession et de grade à Sicore embarqué. Le codage en batch prévoit lui de coder d'abord avec le grade puis, en cas d'échec, sans le grade.

Les codes d'activité (sauf celui de l'activité principale) sur deux positions obtenus en collecte sont utilisés dans l'application pour le codage des professions qui ne nécessitent pas un codage détaillé en PCS : professions antérieures, secondaires mais également pour les professions à l'entrée dans l'entreprise ou un an auparavant lorsque l'enquêté a changé d'entreprise ou d'établissement.

Le futur traitement en automatique des libellés de profession inclut également une normalisation supplémentaire des libellés par rapport à celle que fait Sicore et une correction orthographique des libellés normalisés en échec de Sicore embarqué. Pour cela, on utilise à la fois l'écho de codage de Sicore embarqué et le libellé normalisé par Sicore embarqué.

4.3.2 Utilisation des résultats de Sicore embarqué lors de la phase d'apurement du questionnaire

Un indicateur de codage avec Sicore embarqué des libellés collectés a été créé et affiché sur le poste de gestion Capi afin que lors de la phase d'apurement, les gestionnaires dans les DEM repèrent les questionnaires dont au moins un questionnaire individuel comporte un libellé non reconnu par Sicore embarqué. Ils pourront alors lire ces questionnaires et éventuellement contacter les enquêteurs pour en parler.

4.3.3 Utilisation des résultats de Sicore embarqué sur les postes de reprise

Les postes de reprise des variables de formations et des variables d'emploi servent à coder manuellement les échecs des traitements automatiques et des cas particuliers (codage des spécialités ou coordonnées d'établissement employeur partant directement en reprise) ; ils ont été entièrement rénovés. Ils affichent notamment les 2 saisies de libellé (lorsqu'il y a eu échec de codage du 1er libellé saisi).

D'autres nouveautés ont été développées sur ces postes de reprise afin d'améliorer le codage final des libellés :

- Des variables contextuelles (revenus et diplômes pour les professions ou activités et professions pour les formations) peuvent être consultées par les codeurs ;
- Les remarques des enquêteurs faites lors de la collecte Capi mais actuellement uniquement consultables par les gestionnaires en DR sur les questions de formation et d'emploi le seront sur ces postes ;
- Les codeurs peuvent coder le type d'erreur de saisie du libellé par les enquêteurs. Ces codes seront retransmis aux enquêteurs via les DEM.

À noter quelques nouveautés notamment sur le poste de reprise des variables d'emploi :

- il ne permet pas de reprendre une activité en NAF2 ;
- il offre la possibilité de coder les libellés de professions dans une nouvelle nomenclature Isco une fois le code PCS fourni ;
- pour aider au codage de l'activité en NAF5, les codeurs visualisent le code de l'activité obtenu avec Sicore embarqué (NAF2), peuvent consulter Sirene en direct et ont accès à la nomenclature PCS sur le site *insee.fr* ;
- pour aider à la reprise des libellés de profession en échec de codage en automatique, il affiche le fait notamment que le libellé est particulier soit parce qu'il n'est pas passé par le codage automatique, soit parce qu'il y a eu changement d'établissement depuis la précédente interrogation soit parce que l'établissement ou l'entreprise ou la profession n'est pas la même qu'actuellement pour les professions un an auparavant et à l'entre dans l'entreprise.

4.3.4 Utilisation des résultats de Sicore embarqué sur le nouveau poste de travail

Un poste de travail accessible à tous les acteurs de l'enquête Emploi a été conçu dans le cadre du projet Reflee. Il permet notamment de visualiser des indicateurs d'avancement ou de bilans trimestriels de différentes phases de 'enquête : échantillon, collecte, codifications automatiques, reprise :

- Trois tableaux de résultats nationaux trimestriels des codages avec Sicore embarqué seront visibles sur ce poste de travail : sur les libellés d'activité, les libellés de formation, et les libellés de profession ;
- Un tableau de bilan national trimestriel de la correction d'orthographe du libellé de profession principale non reconnu par Sicore embarqué est également prévu sur ce poste.

4.4 De bons résultats de codage lors des tests Capi 2011

Les tests CAPI les plus récents sur la future enquête Emploi ont eu lieu en juin et septembre 2011 dans les Directions régionales de Bretagne, Midi-Pyrénées, Nord-Pas-de-Calais et pour la Martinique et la Guyane en juillet et octobre 2011 Ils se sont décomposés en deux : un test de première interrogation et un autre — 13 semaines plus tard — de réinterrogation. L'échantillon dans chaque région et Dom comportait 200 logements répartis en cinq enquêteurs. Les résultats concernant la codification sont encourageants.

4.4.1 Libellés de formation

En métropole, et en première interrogation, le taux de codage automatique était de 88 % pour le plus haut diplôme, 90 % pour le plus haut niveau de formation et 92 % pour la formation en cours. La plupart des formations ont été codées dès le premier intitulé (81 à 87 % selon les variables). Comme dans l'enquête Emploi actuelle, le diplôme est moins souvent codé que le niveau de formation et les formations de l'enseignement supérieur moins bien codées que celles du secondaire (car beaucoup plus nombreuses et complexes). Dans les Dom, et en 1ère interrogation, les taux de codage automatique sont proches de ceux enregistrés en métropole, ce qui est une surprise. Ils atteignent 86 % pour le plus haut diplôme, 95 % pour le plus haut niveau de formation et 89 % pour la formation en cours.

Comme en métropole, le mauvais codage des seconds intitulés provient essentiellement de trois problèmes :

- les enquêteurs concernés remettent souvent le même intitulé lors de la seconde saisie ;
- certains pensent que le second intitulé est la suite du premier et saisissent par exemple la spécialité après avoir saisi le diplôme au premier intitulé ;
- d'autres enquêteurs simplifient l'intitulé pour qu'il passe : première 1nnée bac pro (faute de frappe) devient bac pro, par exemple ; l'intitulé est codé mais en terminale bac pro au lieu d'une seconde.

4.4.2 Libellés de profession

Le taux de réussite dans l'appariement des coordonnées de l'établissement en métropole a quasi doublé (82 %) . Les taux de réussite de codage des professions avec Sicore sont autour de 80 % en PCS (principale : 84 % ; un an auparavant : 73 % ; à l'entrée dans

l'entreprise : 77 %) et au dessus de 95 % en CS (antérieure et parents) sauf pour les professions secondaires (89 %). On constate également qu'un deuxième libellé de profession principale est demandé dans 8 % des cas.

La correction d'orthographe a concerné moins de dix libellés de professions non reconnus par Sicore embarqué dans chaque type de profession.

Conclusion

L'introduction de Sicore embarqué sur le poste des enquêteurs permet de diminuer largement la reprise manuelle. Cela se fait au prix d'une légère baisse de qualité dans les libellés, ce qui n'a pas trop d'importance quand on dispose d'une bonne information annexe.

S'il y a échec de codage (car le libellé n'est pas reconnu), les gestionnaires de reprise disposent de deux libellés, et cela les aidera beaucoup dans leur travail.

Des marges d'amélioration dans l'utilisation de Sicore embarqué existent encore par rapport à celle du TCM et de la future enquête Emploi.

En effet, quand on exploite entièrement les retours Sicore, on parvient à diminuer le nombre de questions posées. Cela libère du temps pour collecter plus d'informations dans certains cas. En plus le questionnement est plus adapté au libellé déclaré par l'enquêté. Il en résulte un meilleur questionnaire, plus court et qui permet de coder plus précisément.

Références

[1] Annie Chanut : Codification de la profession avec Sicore – les variables nécessaires et leurs modalités. 2005.

[2] Alain Chenu et Francis Guglielmetti : Coder la profession : nouvelles procédures, vieux problèmes. *Insee Méthodes*, n° 102, 2002.

[3] Alain Desrosières : Les catégories socioprofessionnelles. *Courrier des statistiques*, n° 125, 2008.

[4] Insee, éditeur. *Nomenclatures des professions et catégories socioprofessionnelles 2003*. 2003.

[5] Pascal Rivière : *Glossaire Sicore*. Insee, 1995.

[6] Pascal Rivière : Sicore : système général de chiffrement automatique. *Journées de la méthodologie statistique*, 1995.

[7] Pascal Rivière : Sicore, un outil et une méthode pour le chiffrement automatique à l'Insee. *Courrier des statistiques*, n° 74, 1995.

[8] Pascal Rivière : *Sicore : Documentation méthodologique*. Insee, 1996.

[9] Pascal Rivière et Pierrette Schuhl : *L'outil Sicore*. 1996.

[10] Pierrette Schuhl : *Sicore : Fonctions complexes de synonymie*. Insee, 1997.

[11] Romain Warnan : Analyse du fonctionnement de Sicore embarqué et préconisation pour le TCM. 2010.

[12] Romain Warnan : Influence des variables annexes dans le codage de la profession selon la nomenclature des PCS. 2011.