LES CHIFFREMENTS AUTOMATIQUES DANS L'ENQUÊTE EMPLOI

Gérard Arrivault, Catherine Bergera, Michel Cézard, Nicole Roth

Une rénovation de l'enquête Emploi a été effectuée lors du recensement de la population de 1990. À cette occasion, outre les modifications du questionnaire, le mode de traitement de l'enquête a été rénové.

Pour la rénovation de 1990, un des objectifs à atteindre était "le raccourcissement des délais de disponibilité des résultats" (voir rapport d'étape de 1987). Pour diminuer la durée de traitement de l'enquête, l'option prise a consisté à automatiser autant que possible les chiffrements requis dans l'enquête, opération qui auparavant se faisait manuellement en DR.

Tous les chiffrements à effectuer n'ayant pas la même complexité, plusieurs solutions ont été apportées en fonction des données recueillies, du degré d'imprécision acceptable et des outils de chiffrement existant déjà. Des algorithmes "maison" ont été développés pour chiffrer la catégorie sociale et l'activité économique à deux chiffres. Le logiciel QUID a été choisi pour chiffrer la profession. L'activité économique, le statut public ou privé et la taille de l'entreprise sont chiffrés à partir des informations figurant dans le répertoire SIRÈNE.

Les outils de chiffrement automatique utilisés dans le traitement de l'enquête Emploi ne permettent pas d'effectuer 100% des chiffrements mais déchargent les DR selon les cas de 50% à 90% des chiffrements sans perte de qualité par rapport à un chiffrement manuel, tout en assurant une certaine homogénéité. La codification automatique est réalisée par un programme Batch. Des niveaux minimum de qualité ont été prévus. En cas de doute ou d'échec de ces procédures automatiques, les données sont traitées par un agent du Service enquêtes. Il complète les chiffrements manuellement et saisit les codes à partir d'un terminal, ainsi que les autres corrections éventuelles à apporter au questionnaire. Lors de cette opération, une aide automatique lui permet de vérifier les modalités du chiffrement effectué sur le terminal.

Certains des outils développés dans le cadre de l'enquête Emploi peuvent être réutilisés à l'identique ou moyennant quelques modifications pour les adapter à une nouvelle application.

Les outils de codification "maison" :

Pour coder les catégories sociales et pour coder les activités à un niveau regroupé, deux algorithmes ont été bâtis. Ils réalisent le codage à partir de données de base sommaires et conduisent à des codes à un niveau agrégé.

Le principe consiste à rechercher parmi une ou plusieurs listes de libellés de référence chiffrés celui que l'on peut identifier au libellé à chiffrer, soit parce qu'il sont exactement semblables, soit parce qu'ils ont un certain nombre de mots en commun (placés ou non dans le même ordre), soit parce qu'ils contiennent des mots ayant la même racine.

La recherche est donc fondée sur une logique de type dictionnaire, et passe successivement dans plusieurs listes.

Les deux algorithmes fonctionnent de la même facon :

- des listes de référence externes ;
- un programme de chiffrement automatique ;
- un module d'appel du programme de chiffrement automatique qui lui fournit les données utiles et gère le résultat du chiffrement.

Catégories sociales (CS) :

La codification est programmée à partir des consignes figurant dans l'Index alphabétique des catégories socio-professionnelles (1983), qui fournit, pour chaque intitulé fréquent, un ensemble de codes pour diverses valeurs de la classification et du statut. Lorsque aucune information n'est disponible à propos de la classification et/ou du statut, un code est cependant affecté par défaut. Dans l'algorithme, l'information sur la classification n'est pas directement prise en compte. Cinq cas sont considérés suivant le statut, pour chaque intitulé : salarié de l'État ou d'une collectivité locale, autre salarié, non-salarié avec moins de 10 salariés, non salarié avec 10 salariés ou plus, aucune information disponible sur le statut.

Lors de la phase de reconnaissance, quand un intitulé est trouvé dans une liste, le code correspondant à l'un des cinq cas est attribué à cet intitulé.

La recherche s'effectue d'abord dans une première liste où, pour être reconnu, un intitulé à chiffrer doit être exactement égal à l'un des intitulés de la liste. Dans cette liste, les intitulés sont classés par ordre alphabétique et la recherche est dichotomique.

À l'examen de la première liste, si la recherche n'a pas abouti, elle se poursuit dans d'autres listes basées sur un principe différent : un code est attribué si l'intitulé à chiffrer est trouvé dans la liste, mais cet intitulé à chiffrer peut comporter d'autres mots. La recherche se termine dès que le premier groupe de mots accordé à l'intitulé en examen est trouvé dans une liste. De ce fait, l'ordre des items dans ces listes est très important et doit être l'objet de soins attentifs. Par exemple, le libellé de référence "clerc avocat" doit être placé avant le libellé "avocat" pour éviter que tous les clercs d'avocat ne soient chiffrés avocats.

Certains mots d'un libellé apportent une information sur la classification (secrétaire, directeur, ouvrier...). À ces mots, appelés mots clés, sont associées des listes particulières vers lesquelles s'oriente la recherche lorsqu'elle n'a pas abouti dans la première liste et qu'un mot clé est détecté dans le libellé à chiffrer. Si ce dernier ne contient pas de mot clé, la recherche se poursuit dans une deuxième liste générale.

Pour certains types de chiffrement, les libellés sont plus imprécis et on se contente d'un chiffrement approximatif (CS secondaire, CS recherchée par les personnes à la recherche d'un emploi). Nous avons alors choisi de privilégier l'efficacité : une liste spécifique, dite liste "balai", dont les libellés de référence sont de plus en plus imprécis, rattrape les échecs de chiffrement dans les premières listes. Dans ce cas, l'information "chiffrée par la liste balai" est transmise au module d'appel qui accepte ou refuse le chiffrement.

Avant d'effectuer la recherche dans les diverses listes de référence, le libellé à chiffrer est formaté par le programme de chiffrement afin de pouvoir être comparé au format des intitulés de référence :

- les mots sont tronqués à 10 caractères ;

- les S terminaux sont enlevés ;

- les mots vides (de, la, un,...) et les caractères spéciaux (- ("" ...) sont éliminés.

Cette normalisation des libellés permet de regrouper sous une même forme plusieurs libellés se ressemblant, et donc de ne pas trop augmenter la taille des listes de références. D'autres manipulations du libellé à chiffrer permettent d'accroître le rendement et la qualité du chiffrement, par exemple : on ne cherche pas à chiffrer les libellés contenant "hors" ou "sauf" qui peuvent en inverser le sens, on regroupe tous les synonymes de "qualifié" (OP, OP1, OS ...) sous le même terme.

Dans le traitement de l'enquête Emploi, cet algorithme est utilisé pour chiffrer la CS du père, la CS recherchée par les personnes à la recherche d'un emploi, la CS trouvée par les personnes ayant trouvé un emploi qui commence plus tard et la CS secondaire des personnes ayant une double activité. Le chiffrement automatique effectué par la liste balai est retenu dans tous les cas sauf pour la CS du père. Dans ce dernier cas, le chiffrement automatique doit être validé par le service enquêtes. À l'enquête de mars 1991, les taux de chiffrement automatiques étaient les suivants :

	Total	Dont balai
CS père	91,8	12,9*
CS recherchée	78,6	17,0
CS ultérieure	81,3	21,9
CS secondaire	70,2	15,2

** Pour la CS du père, les chiffrements effectués par la liste balai sont considérés comme des échecs.*

L'outil de codification de la CS peut facilement être utilisé dans le cadre d'autres applications où seul le module d'appel devra être développé. Les listes de libellés de référence peuvent facilement être enrichies mais cette opération doit être effectuée par un expert et testée avec soin sous peine de provoquer une détérioration du chiffrement automatique.

Activités (à deux chiffres) :

Pour coder l'activité économique, nous avons pris la nomenclature comme base, et nous avons construit des listes permettant de réaliser une codification automatique aussi souvent que possible, et à n'avoir que le moins possible de cas chiffrés de manière inexacte. Mais ce codage, pour traiter les cas difficiles, reste approximatif car aucune règle de décision claire n'émane de la nomenclature elle-même. Par exemple, avec le mot MÉTALLURGIE, il est impossible de savoir si cela concerne le fer ou un autre métal ; dans ce cas, le code attribué renferme donc une part d'arbitraire.

Le processus de codage est similaire à celui des catégories sociales. Les principes de recherche sont les mêmes. Pourtant, il y a une différence : alors que les mots étaient tronqués à 10 caractères dans une étape préliminaire, puis entièrement examinés, pour les activités, on opère à partir d'une logique basée sur les racines des mots. Au début de chaque intitulé de la liste, le nombre de mots et leur longueur sont indiqués. Il est ainsi possible de traiter ensemble tous les mots appartenant à la même famille, ou racine, c'est à dire commençant par les mêmes caractères.

Après avoir formaté le libellé à chiffrer, la recherche s'effectue séquentiellement et successivement dans cinq listes de libellés de références où les intitulés sont classés selon leur nombre de mots puis, en règle générale, par ordre alphabétique. Dès qu'un libellé à chiffrer a pu être identifié à un libellé de référence le chiffrement automatique s'effectue sans examiner les autres libellés de référence.

Comme pour le chiffrement de la CS, la recherche s'effectue d'abord dans une liste pour laquelle l'intitulé à chiffrer doit être identique à l'un des libellés de référence. Dans les autres listes, il suffit de retrouver dans un libellé de référence l'ensemble des mots du libellé à chiffrer dans le même ordre.

La deuxième étape consiste à rechercher un éventuel mot clé dans le libellé à chiffrer et à poursuivre la recherche dans la liste associée au mot clé. Si la recherche n'a pas abouti précédemment (donc y compris pour les libellés à chiffrer contenant un mot clé), elle se poursuit dans une liste générale.

L'étape suivante détecte certains mots de l'intitulé qui orientent vers une recherche dans le libellé de profession. Par exemple, si le libellé d'activité à chiffrer contient "assurance", le chiffrement sera 88 si le libellé de CS contient "agent de bureau" ou "agent de maitrise" ou "agent technique" et 78 s'il contient "courtier" ou "courtage".

La dernière étape consiste également à rechercher dans une liste balai aux intitulés moins précis. Pour chiffrer, il suffit de retrouver dans le libellé à chiffrer l'ensemble des mots d'un libellé de référence quel que soit leur ordre. Les trois quarts des libellés de référence de cette liste sont constitués d'un seul mot. L'information "chiffré par la liste balai" est transmise au module d'appel.

Dans le traitement de l'enquête Emploi, cet algorithme est utilisé pour chiffrer l'activité antérieure des inactifs, l'activité secondaire des personnes ayant une double activité et, plus marginalement, les agriculteurs et personnels de service pour lesquels il n'y a pas d'identification dans le répertoire SIRENE. Dans tous les cas, les chiffrements effectués à l'aide de la liste balai sont retenus. En 1991, les taux de réussite étaient les suivants :

	Total	Dont balai
Activité antérieure	79,7	21,5
Activité secondaire	69,7	25,7

La codification de la profession (actuelle ou 1 an auparavant)

L'outil principal utilisé pour chiffrer la profession au niveau fin est QUID, un système de codification automatique réalisé et mis en oeuvre à l'Insee (Lorigny, 1988). La première étape du système Quid consiste à construire un gros fichier d'appellations de professions, avec les chiffrements correspondants vérifiés par un expert. Ce fichier, mémoire de tous les chiffrements manuels réalisés sur une enquête, et contrôlés, est appelé Fichier d'Apprentissage (FA). La gestion du FA est complètement indépendante de l'opération de chiffrement proprement dite. Année après année, le FA s'améliore : de nouvelles appellations et les chiffrements correspondants - supposés exacts - sont ajoutés au FA.

Le FA utilisé pour l'enquête Emploi à été constitué à partir de libellés profession de l'enquête Formation et Qualification Professionnelle (FQP) de 1985 et enrichi manuellement à partir des échecs de chiffrement rencontrés lors de sa constitution. Il contient 33 936 enregistrements.

Principe de QUID

Afin d'éviter un temps de recherche trop long dans ce fichier, les appellations sont organisées sous la forme d'un "arbre" optimisé. L'unité d'information élémentaire est le bigramme (deux caractères consécutifs). Une suite de "questions" successives détermine les bigrammes les plus signifiants pour discriminer les appellations. Ainsi, après une standardisation préalable des intitulés (élimination des mots vides, compactage des sigles, troncature des mots selon un paramétrage modifiable), un intitulé est "représenté" par plusieurs "questions" aux bons bigrammes. Dans le paramétrage actuel, on retient cinq mots de dix caractères, les sigles sont resserrés et les mots vides éliminés.

Lors de la phase de codification, les intitulés sont traités comme lors de la construction du fichier d'apprentissage. Ainsi, seuls les bigrammes les plus efficaces sont appariés pour associer un nouvel intitulé avec un libellé déjà rencontré. Si le système reconnaît un cas déjà rencontré, il y a un écho, ou une décision; un code est proposé. Le fichier d'apprentissage utilisé pour l'enquête Emploi ne contient pas d'indécision (plusieurs codes associés à un même libellé du FA), le système chiffre donc de manière unique (douteuse ou non douteuse) ou bien ne chiffre pas. Pour confirmer l'identification, une procédure de contrôle confronte aussi d'autres bigrammes, appelés bigrammes de redondance. Le nombre et la place de ces bigrammes de redondance sont un paramètre du système et sont choisis selon le niveau de qualité désiré. Ainsi, l'écho initial peut être certain (non douteux), ou incertain (douteux). En cas de doute, ou lorsque aucun écho n'a été trouvé (inconnu), la codification doit être complétée par des méthodes manuelles. Après centralisation et validation par un expert, de tels cas pouront être intégrés dans le fichier d'apprentissage.

Avant l'appel des procédures de chiffrement QUID, les libellés de profession des salariés de l'Etat et des collectivités locales sont examinés et éventuellement reformulés pour accroître l'efficacité du système; ainsi, par exemple, la mention d'un indice ou d'un échelon est effacée; la mention du grade ou de l'échelle est placée au début du libellé.

Variables annexes
Variable de synthèse

Coder l'occupation actuelle déclarée dans l'enquête selon la nomenclature française des PCS est une tâche longue et délicate. Elle requiert à la fois un intitulé de profession (la réponse en clair à la question directe "quelle est votre profession?") et de nombreuses variables annexes, activité économique, statut, nombre de salariés, classification, fonction, etc. Ces variables sont plus ou moins utiles selon la zone en cause de l'espace social. Certaines sont essentielles, non-salarié ou salarié par exemple, d'autres sont moins utilisées (par exemple la distinction des spécialités selon l'orientation des productions de l'exploitation agricole). Aussi, en plus de l'intitulé, avons-nous construit une variable unique, appelée variable de synthèse, qui est traitée en même temps que les autres bigrammes appartenant à l'intitulé. Cette variable de synthèse est composée de 9 items qui délimitent les principaux découpages de la nomenclature, 10 à 60 pour les salariés du secteur privé, 70 pour les salariés du secteur public, 80 et 90 pour les personnes à leur compte.

10 ouvrier non qualifié

20 ouvrier qualifié

30 employé

40 technicien

50 agent de maîtrise

60 cadre

70 salarié de l'État ou d'une collectivité locale

80 non salarié, occupant moins de 10 salariés

90 non salarié, occupant 10 salariés ou plus

Par exemple l'intitulé TAXI, utilisé pour CHAUFFEUR DE TAXI, peut être associé avec trois variables de synthèse (au moins) :

TAXI--10 : salarié, non qualifié -- 6413

TAXI--80 : non salarié, moins de 10 -- 2171

TAXI--90 : non salarié, 10 ou plus -- 2332

Tables de décision

Cependant, la variable de synthèse ne conduit pas à une conclusion quand une information plus précise est requise. Dans ces cas, qui représentent environ 30%, un code intermédiaire est proposé. Environ 250 tables de décision ont été construites à partir des variables additionnelles pour achever la détermination automatique du code. Pour cela, nous avons pris comme référence les "tables Colibri" réalisées pour le chiffrement de la profession au recensement de la population et nous les avons adaptées à nos besoins. Ceci devrait fournir une codification de la profession à l'enquête cohérente avec celle obtenue à partir du recensement.

Afin d'éviter de prendre en compte des changements fictifs, une comparaison est menée entre la codification de l'année en cours et celle de l'année précédente. La codification en cours est validée si elle est identique; sinon, elle doit être confirmée lors de la phase interactive. Tous les changements de profession sont de ce fait vérifiés.

Une codification complète et correcte de la profession, au niveau 4 chiffres, requiert certaines variables fournies par la codification de l'établissement, notamment le statut public ou privé de l'établissement, l'activité, la taille de l'entreprise. Aussi l'interrogation du fichier des établissements doit-elle avoir lieu avant la codification des professions. Si cette étape échoue, une tentative de codification automatique de la profession est cependant faite, sur la base des données "établissement" de l'année précédente.

Lorsque la profession a été ainsi codée, le code peut ensuite être invalidé pendant la phase conversationnelle si les données concernant l'établissement pour l'année en cours diffèrent finalement des données de l'année précédente.

À l'enquête de 1991, 71,4% des codes profession ont été chiffrés automatiquement soit par QUID (40%), soit par les tables de décision aval (28,9%). Les 2,2% restant concernent les aides familiaux traités différemment. Dans 15% des cas de chiffrement automatique, la comparaison avec le code de l'année précédente a fait apparaître une divergence que les services enquêtes ont dû valider à l'écran.

Le fichier d'apprentissage constitué pour l'enquête Emploi peut être utilisé pour d'autres applications. Un module spécifique devra être développé pour calculer la variable de synthèse, préparer les libellés des salariés de l'État et des collectivités locales et traiter les aides familiaux.

Les tables de décision utilisées en aval de QUID devront être revues avec l'introduction de la nouvelle nomenclature d'activité. Leur utilisation nécessite de développer un module spécifique de recherche des chiffrements intermédiaires effectué par QUID et des modalités des différents codes utilisés dans ces tables (catégorie juridique, tranche d'effectif salarié, activité...).

Codification de l'activité économique (actuelle ou 1 an auparavant)

L'activité détaillée de l'établissement ainsi que la taille de l'entreprise sont des données souvent sujettes à imprécision dans une enquête auprès des ménages : confusion avec l'activité individuelle, méconnaissance de l'unité "entreprise", imprécision des libellés. Nous avons donc choisi d'utiliser plutôt les informations du répertoire des établissements SIRÈNE. Ceci requiert une phase d'identification qui est d'abord réalisée automatiquement, puis ensuite par une interrogation au terminal en cas d'échec ou de doute.

Identification de l'établissement

D'un point de vue opérationnel, des outils d'identification ont été mis au point par les gestionnaires du répertoire afin de repérer une unité quelconque, soit par programme Batch, soit de manière interactive. Un appariement automatique repère les unités par leur nom et leur adresse, ou bien à partir de leur numéro d'identité. L'efficacité de cet appariement dépend d'abord de la précision et de la standardisation du nom et de l'adresse. Bien entendu, lorsqu'un numéro d'identification peut être utilisé pour la recherche, l'efficacité est bien meilleure. Nous avons essayé d'améliorer la qualité des

informations collectées concernant l'identification du lieu de travail (présentation plus standardisée du nom et de l'adresse, numéro SIRET demandé aux personnes interrogées pour la première fois). Enfin, nous utilisons l'identification de l'année précédente pour les personnes déjà enquêtées et qui n'ont pas changé de lieu de travail.

Pour chaque demande d'identification SIRÈNE, une réponse est fournie indiquant soit les raisons de l'échec de la recherche d'identification (SIRET non compatible avec le nom ou la raison sociale, adresse incomplète...), soit un ou plusieurs établissements susceptibles de répondre à l'identification recherchée affectés de codes qualitatifs permettant de les noter. Un programme Batch gère les retours des demandes d'identification en tenant compte des notes attribuées par le système et de la multiplicité des "échos" : un écho jugé "bon" sert de base au chiffrement automatique, un écho jugé "mauvais" est rejeté, les échos ayant une note intermédiaire sont, selon les cas, rejetés ou édités sur un listing pour servir de base au chiffrement à l'écran par les services enquêtes.

Contrôles a posteriori

Certains contrôles supplémentaires ont été rajoutés sur certains champs de la nomenclature (redondance avec le code CHPUB, statut public ou privé déclaré par l'enquêté, ou numéros à risque, par exemple les cantines sont systématiquement traitées en conversationnel par les services enquêtes pour ne pas être confondues avec l'établissement se situant à la même adresse).

À l'enquête de mars 1991, 50,4% des identifications concernant l'activité actuelle de l'enquêté ont été jugées bonnes et intégrées dans les questionnaires. Dans la moitié restante, 17,5% des réponses ont été proposées à la validation des Services Enquêtes qui ont donc dû traiter complètement 32,1% des identifications.

Codification de la catégorie sociale antérieure

La question est posée à tous les inactifs ayant eu dans le passé une activité professionnelle. Les données brutes sont plutôt de bonne qualité, surtout lorsque l'activité passée se réfère à une période relativement récente. Le choix de l'algorithme utilisé reflète cette exigence de qualité ; une méthode mixte a été mise en place, combinant différents outils : QUID, la recherche dans les tables-aval type Colibri et l'algorithme "maison" de codification de la CS (sans module balai). En cas de succès de QUID ou des tables-aval, la CS est chiffrée en retenant les 2 premiers chiffres du code profession obtenu. Lorsque l'outil "maison" est utilisé, un contrôle supplémentaire est effectué : seules les CS en cohérence avec la qualification déclarée sont acceptées.

En 1991, ce "montage" a permis de chiffrer automatiquement 89,8% des codes profession antérieure, 70,3% par QUID ou par les tables aval et 19,5% par l'outil "maison" contrôlé.

L'automatisation des chiffrements a permis en 1991 de traiter 41,6% des questionnaires sans intervention des services enquêtes (questionnaires sans erreur et complètement chiffrés automatiquement). Le délai de production des résultats est identique à celui des précédentes enquêtes mais devrait être réduit au cours de la série avec l'appropriation croissante du nouveau mode de traitement par les services enquêtes.

Au niveau des chiffrements, la plus forte charge de travail concerne l'activité et la profession de l'ensemble des actifs, respectivement 50% et 43,8% des codes à chiffrer ou à confirmer à l'écran. L'introduction de la nouvelle nomenclature d'activité à partir de 1993 devrait quelque peu perturber les traitements de l'enquête, jusqu'à ce que les services enquêtes se familiarisent avec cette nouvelle nomenclature. Une importante formation leur sera nécessaire avant la première utilisation.

Cependant, l'enrichissement du fichier d'apprentissage au cours de la série et une formation appropriée des enquêteurs devraient avoir une influence positive sur le rendement qualitatif et quantitatif des chiffrements automatiques.

BIBLIOGRAPHIE

Rapport d'étape enquête Emploi, note interne n° 1367 / 472, 22 avril 1987.

Fiches SYDOSI.

Processing the French Labor Force Survey in the 1990's : M.CÉZARD, J.L.HELLER, N.ROTH ; ARC 1990, Bureau of the Census (1990).

QUID , une méthode générale de chiffrement automatique : J. LORIGNY ; Techniques d'enquêtes , vol 10, n° 2, 1988.

- Communication sur le chiffrement automatique : L.VIGLINO ; ISI 1991 , Le Caire

- *Outils de Codification automatique utilisés dans le traitement de l'enquête Emploi, Note de présentation, Note de référence*, G. ARRIVAULT, C. BERGERA, M. CÉZARD, N. ROTH.

ANNEXE

Performances des chiffrements automatiques principaux (enquête emploi mars 1991)

CS du père	**91,8**
dont : module balai	12,9
CS recherchée	**78,6**
dont : module balai	17,0
Profession actuelle	**71,4**
dont : QUID	40,0
tables aval (Colibri)	28,9
aides familiaux	2,2
dont à valider	15,0
Secteur d'activité économique actuelle (SIRÈNE)	**67,9**
dont : écho accepté	50,4
échos proposés	17,5
CS antérieure	**89,8**
dont : QUID	57,2
tables Aval	13,1
Module "maison" avec contrôle sur la qualification	19,5
Secteur d'activité économique antérieure	**79,7**