Recensement de la population du Canada en 1991. Expérience avec un système de codification automatique

Jocelyn Tourigny,
Statistique Canada

1. INTRODUCTION

La codification des libellés du recensement canadien de la population de 1991 constitue la première utilisation massive du système généralisé CART (Codification Automatique par Reconnaissance de Texte), logiciel développé à Statistique Canada. Durant le traitement du recensement plus de 16 millions de libellés en clair provenant de 10 questions touchant la langue, la religion, le lieu de naissance, l'origine ethnique, l'éducation et la mobilité ont été traités par le logiciel CART. Un taux de succès de 92% a été obtenu avec un taux d'erreur inférieur à 1%. Pour ces questions, les coûts de l'opération de codification ont été réduits de moitié par rapport à la procédure de codification traditionnelle.

Le présent document est divisé en deux parties. Dans la première partie nous décrivons la méthodologie du système de codification automatique CART. Dans la seconde partie, nous présentons l'application de codification automatique du recensement canadien de 1991 et les résultats obtenus. La conclusion décrit les projets de codification pour le recensement de 1996.

2. SYSTÈME DE CODIFICATION AUTOMATIQUE (CART)

2.1 PROBLÉMATIQUE DU CODAGE

Dans le contexte d'une enquête, les libellés en clair sont très utiles lorsque la variable étudiée a un ensemble de réponses possibles très vaste ou lorsque certaines réponses ne peuvent être prédites. Ce type de libellé dans une question permet:

- *une économie dans la formulation de la question en offrant au répondant moins de choix à cocher (parfois le nombre de questions sur le sujet peut être réduit afin de laisser de l'espace pour des questions sur d'autres sujets);*
- *d'être objectif en réduisant ou éliminant la structure artificielle des choix proposés (et l'ordre de ces choix) et de contrer la tendance du répondant à sélectionner le premier choix approprié;*
- *d'obtenir une variété de réponses permettant une revue de la classification et parfois sa mise à jour; et*
- *d'être plus simple pour le répondant: ses réponses sont du même médium que la question.*

Toutefois, afin de faciliter la synthèse et l'analyse statistique, il est nécessaire de grouper ensemble les libellés en clair qui, suivant une

nomenclature existante (par exemple: la nomenclature des langues), ont essentiellement le même sens. Cette opération est dénommée codification.

Traditionnellement, la codification est une opération effectuée par des commis sans aucun support informatique. Utilisant un libellé en clair (parfois des informations annexes soumises par le répondant) et les instructions de codification produites par un spécialiste de la nomenclature, un commis cherche un libellé dans un manuel de nomenclature. Le code associé au libellé est inscrit sur le questionnaire. C'est ce code qui, au lieu du libellé, est saisi avec les autres réponses du répondant.

Il peut y avoir des variations à cette approche, telles l'interprétation de la réponse du répondant, l'utilisation de procédures spéciales et complexes et la référence du problème à un expert en codification.

Les problèmes rencontrés lors de la codification par des commis se situent à plusieurs niveaux.

> *La codification est sujette à erreur. Il est difficile de chercher dans un manuel de nomenclature qui a parfois plus de 50,000 entrées. Les instructions peuvent être inadéquates ou être parfois appliquées incorrectement par le commis. Les libellés sont parfois vagues et leur interprétation est très subjective, d'où la possibilité d'un mauvais chiffrement en des codes statistiques. Cependant seul un commis peut repérer et solutionner adéquatement un cas "difficile".*
>
> *Bien contrôler l'opération de codification est un défi. Codifier précisément nécessite beaucoup de jugement et il est parfois très difficile de choisir le bon code numérique. Il n'est pas surprenant de trouver beaucoup de variation entre le chiffrement de différents commis, et même, à l'intérieur du travail d'un même commis. Il faut donc développer un programme de formation étoffé, obtenir le support continu d'experts et développer un contrôle qualitatif approprié.*
>
> *L'opération de codification est difficile à administrer. Il s'agit d'une opération qui exige beaucoup de temps et de ressources. La courte durée de l'opération peut difficilement être réduite sans affecter les coûts et la qualité. Il faut donc engager et motiver un groupe important d'employés temporaires pour effectuer un travail relativement monotone et espérer une rotation de personnel minimale.*

Pour remédier aux désavantages énumérés, plusieurs pays ont développé et utilisent avec succès des systèmes de codification automatique, notamment la France, la Suède et les États-Unis. Statistique Canada a aussi mis au point un système de codification automatique pouvant répondre aux besoins de plusieurs enquêtes. Ce système généralisé, connu sous le sigle de CART (pour Codification Automatique par Reconnaissance de Texte) est utilisé par quelques enquêtes dont la plus importante fut le recensement de 1991.

2.2 MÉTHODOLOGIE DU CODAGE AUTOMATISÉ (CART version 1.06)

Cette section décrit les éléments principaux de la méthodologie de la codification automatique; ces éléments sont détaillés pour mieux comprendre ce qui constitue la force et parfois la faiblesse du système CART selon la question à chiffrer.

2.2.1 Généralités

Les méthodes utilisées par le système CART s'inspirent de méthodes qui ont été élaborées à l'origine au Bureau du Recensement américain (Hellerman, 1982) et de l'expérience de Statistique Canada dans le développement d'algorithmes et de systèmes d'appariement des dossiers administratifs. Essentiellement, la méthode consiste à examiner une série de libellés préalablement codés. Si le libellé à chiffrer est repéré, le code correspondant est enregistré et l'opération prend fin. Dans le cas contraire, l'examen se poursuit en faisant intervenir un algorithme pour repérer le libellé le plus comparable; une fois cette opération réalisée, le système attribue le code correspondant.

Ce repérage est rendu complexe par le fait que le langage humain a plusieurs façons d'exprimer la même chose. Les mots ne sont pas toujours dans le bon ordre, un mot important peut être absent, un mot non pertinent peut être présent, un mot peut être un synonyme ou une abréviation d'une expression, ou les règles de ponctuation et de syntaxe peuvent ne pas avoir été respectées. CART tente de contourner ces difficultés grâce à un traitement préalable des libellés et à ses deux techniques d'appariements.

La figure 1 représente les différents modules du système CART que nous décrirons.

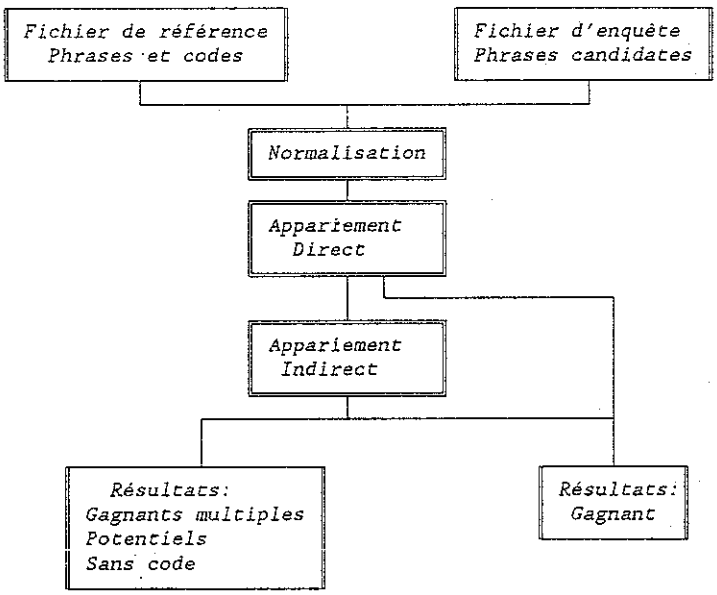


figure 1: système CART

2.2.2 Fichier de référence

Pour chaque question à chiffrer, il faut en premier lieu créer un fichier de référence contenant les libellés en clair typiques (dites phrases) pour une question donnée. Ce fichier comprend les phrases et leur code numérique associé. Il est construit à partir de manuels de nomenclature standard, de phrases codifiées par des experts provenant d'une enquête similaire conduite antérieurement ou d'une combinaison de ces deux sources comme dans le cas du recensement de la population de 1991. Idéalement les phrases choisies sont représentatives des phrases les plus fréquemment observées lors d'une opération d'appariement. Il est recommandé de conserver les phrases dans leur forme originale, avec les erreurs d'orthographe, de grammaire et de syntaxe. Ce fichier de phrases et de code numérique est intégré à une base de données servant à faciliter les opérations d'appariement.

2.2.3 Normalisation

Les phrases du fichier de référence et celles devant être chiffrées sont mises sous une forme normalisée (terme anglais: parsing) afin de permettre à l'ordinateur de reconnaître comme identique les réponses qui sont sémantiquement équivalentes. CART fournit à l'usager un module de normalisation très flexible. Dans un premier temps, les phrases sont considérées comme une suite ininterrompue de caractères; on ne reconnait pas que la phrase contient des mots, des espaces et des signes de ponctuation. Cette suite de caractères est analysée par le système afin de déterminer les mots distincts. Les mots distincts sont ensuite scrutés et mis sous forme normalisée; cette dernière étape réduit le problème des synonymes, des mots doubles, des mots vides, des suffixes différents, etc. L'annexe A donne la liste des fonctions de normalisation offertes par CART.

2.2.4 Appariement direct

Les mots normalisés de la phrase candidate sont placés en ordre alphabétique et la phrase est comprimée pour former une "clé d'expression condensée" dont la longueur équivaut en moyenne à 35% de la longueur initiale de la phrase. En pratique, cette clé est construite par l'élimination des espaces entre les mots normalisés et en convertissant en des codes de 8 bits les caractères (lettres et chiffres) individuels et les combinaisons fréquentes de caractères (digrammes et trigrammes seulement). La clé est utilisée pour chercher un appariement "exact" dans le fichier de référence où chaque phrase possède déjà sa clé.

2.2.5 Appariement indirect

Cette méthode consiste à chercher l'appariement le plus "comparable" de la phrase candidate dans le fichier de référence. Toutes les phrases qui possèdent un ou plusieurs mots normalisés en commun avec la phrase candidate sont extraites du fichier de référence. Le système évalue chacune de ces phrases et leur attribue un "pointage". Ce pointage, combiné à certains paramètres établis a priori, permet de déterminer s'il existe un appariement "gagnant", des appariements

"gagnants multiples" ou "potentiels" dans le fichier de référence. Cette méthode est inspirée des travaux de Hellerman (1982) et de Knaus (1981).

2.2.5.1 Calcul d'un poids pour chaque mot normalisé du fichier de référence

Le système calcule un poids pour chaque mot normalisé contenu dans le fichier de référence. Ce poids donne une indication du pouvoir de discrimination du mot, c'est-à-dire si le mot peut conduire à un seul code numérique.

Le poids heuristique d'un mot est construit de telle façon que le poids diminue lorsque le nombre de codes auquel il est associé augmente. Le poids H d'un mot a la forme:

$$H=\frac{E_U-E_M+\epsilon}{E_M+\epsilon}$$

où:

$$E_M=-\sum_{i=1}^{n}(p_i*log_2\ p_i)\ \text{ et }\ E_U=-\sum_{i=1}^{k}\frac{1}{k}*log_2(\frac{1}{k})$$

E_M *est l'entropie du mot. L'entropie est une mesure de l'uniformité d'une distribution. Lorsqu'un mot est particulier à un seul code, l'entropie est nulle; elle atteint son maximum lorsque le mot est associé à tous les postes (soit les n codes) de la nomenclature.*

p_i *est la proportion d'occurrences du mot dans le fichier pour le* $i^{ème}$ *code; cette quantité représente donc une mesure de la probabilité qu'étant donné le mot, le code approprié est le code i.*

$$p_i=\frac{x_i}{k}\ ,\ \sum_{i=1}^{n}x_i=k\ \text{ et }\ \sum_{i=1}^{n}p_i=1$$

x_i *est le nombre d'occurrences du mot considéré parmi les phrases qui ont le code i*

ϵ *est une petite constante arbitraire pour éviter une division par 0 dans l'éventualité où* $E_M = 0$ *(qui correspond à la situation où un mot est particulier à un seul code).*

$$\epsilon=\frac{k}{k+1}log_2\frac{k}{k+1}$$

2.2.5.2 Calcul d'un pointage pour chaque phrase appariée

Chaque phrase du fichier de référence qui contient au moins un mot normalisé en commun avec la phrase candidate est considérée comme un appariement potentiel. Une méthode de pointage a été mise au point afin de déterminer la phrase la plus "comparable"; ce pointage est basé sur le nombre de mots contenus dans la phrase candidate qui sont "valides" dans le fichier de référence, le nombre de mots de la phrase du fichier de référence, et sur le poids des mots communs aux deux phrases. La formule utilisée est la suivante:

$$P = \frac{(\text{nombre de mots en commun})^3 * (\Sigma \text{ poids des mots en commun})}{(\text{nombre de mots valides dans la phrase candidate}) * (\text{nombre de mots dans la phrase du fichier de ref.})}$$

En présence de deux phrases identiques (donc d'un appariement exact), la formule devient:

$$P = (\text{nombre de mots en commun}) * (\Sigma \text{ poids des mots en commun})$$

2.2.5.3 Évaluation des appariements et choix d'un gagnant

Avant de procéder à un appariement indirect, l'usager fournit des valeurs aux trois paramètres suivants:

1. *MIN: borne inférieure du pointage*
2. *MAX: borne supérieure du pointage*
3. *PCNT: pourcentage de différence*

Supposons que m appariements potentiels existent dans le fichier de référence. Ordonnons les pointages obtenus par ces phrases en ordre décroissant:

$$P_1 > P_2 > \ldots > P_m$$

Quatre situations peuvent se produire:

i) Si $P_1 >= MAX$ *et* $\frac{P_1 - P_2}{P_1} >= PCNT$

alors la phrase ayant obtenu le pointage P_1 *est gagnante et son code numérique est assigné à la phrase candidate.*

ii) Si $P_1 >= MAX$ *et* $\frac{P_1 - P_2}{P_1} < PCNT$

alors toutes les phrases i telles que $P_i >= MAX$ *sont considérées comme étant gagnantes multiples.*

iii) $MIN <= P_1 < MAX$
alors toutes les phrases i telles que $MIN <= P_i < MAX$ *sont considérées comme des appariements potentiels.*

iv) Si $P_1 < MIN$
alors aucun appariement n'obtient un statut.

Toutes les phrases candidates se trouvant dans les situations ii, iii ou iv ainsi que celles qui ne sont pas appariées au fichier de référence doivent être codifiées par des commis. Durant les tests précédant la production, toutes ces phrases candidates disponibles sont étudiées dans le but d'améliorer le fichier de référence, les règles de standardisation et les paramètres d'évaluation des appariements.

2.2.6 Performance de CART

La technique d'appariement direct grâce à son utilisation de la clé d'expression condensée est très efficace même lorsque le fichier de référence est très volumineux.

Pour rendre l'appariement indirect plus efficace, CART identifie toutes les phrases du fichier de référence qui contiennent le mot de la phrase candidate ayant le plus haut poids et il établit leur pointage. Avant d'identifier les phrases additionnelles contenant le mot ayant le second poids en importance, un pointage potentiel est estimé. Lorsque ce pointage potentiel est inférieur au paramètre MIN la recherche est arrêté. Sinon l'identification et le calcul des pointages se poursuivent.

3. L'APPLICATION DE CODIFICATION DU RECENSEMENT 1991

3.1 Généralités

Le recensement canadien de la population et des logements utilise deux types de questionnaires auto-administrés pour recenser plus de 10 millions de logements. Durant l'établissement de la liste des logements de son secteur de dénombrement, le représentant du recensement distribue un questionnaire abrégé à 80% des logements et un questionnaire complet à 20% des logements suivant un échantillonnage systématique. Le répondant retourne par la poste le questionnaire complété. Le représentant du recensement vérifie les réponses et fait les suivis téléphoniques et en personne nécessaires pour corriger certaines réponses incohérentes ou incomplètes.

Le questionnaire complet est l'équivalent de la feuille de logement et de six bulletins individuels du recensement de la République Française; par contre beaucoup plus d'information sur les caractéristiques des personnes sont recueillies. Le questionnaire abrégé est une version réduite du questionnaire complet où seulement les questions de base sur le logement et les personnes sont incluses (e.g. type de logement, logement occupé par un propriétaire ou locataire; relation avec la personne de référence, sexe, date de naissance, état matrimonial légal, première langue apprise). Pour répondre à une question le répondant doit cocher un cercle, écrire un nombre ou imprimer un libellé en clair.

Quelques libellés sont codifiés par des commis à la préparation pour la saisie des données. Toute l'information des questionnaires abrégés et complets, à l'exception des libellés déjà codifiés, est saisie en une seule opération sur une période de 4 mois. Pour chaque variable à codifier de façon automatique, le

libellé en clair (dite phrase dans la terminologie de CART) ainsi que des variables annexes reliées à la personne et aux autres membres du logement sont transférées sur une base de données pour faciliter l'opération de codification.

L'application de la codification du recensement de 1991 est illustrée à la figure 2. L'application est hautement intégrée. Elle englobe la codification automatique par CART, la codification des commis assistée par ordinateur, le contrôle qualitatif des deux types de codification et la rectification des erreurs systématiques. Aucun retour au questionnaire n'est nécessaire et le système prend les décisions dans la majorité des situations.

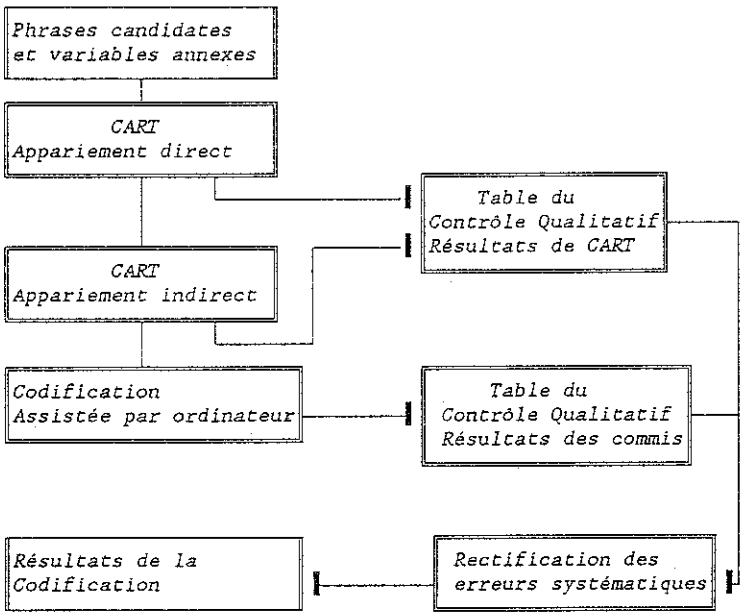


figure 2: module de l'application de la codification

Les 10 questions soumises à la codification automatique sont présentées à l'annexe B. De ces questions, 12 applications semblables mais personnalisées ont été établies (une application pour chaque question, une application pour une question qui a besoin d'un second fichier de référence et une dernière application pour une question qui apparait sur le questionnaire abrégé et complet avec des variations dans l'information annexe disponible).

Les prochains paragraphes détaillent chacun des modules de ces applications.

3.2 CART - appariement direct

Seule la phrase est utilisée pour la codification automatique. Les phrases sont ordonnées par ordre alphabétique et regroupées par phrase unique. C'est cette phrase unique qui est normalisée et appariée avec les phrases normalisées du fichier de référence. S'il y a appariement toutes les phrases correspondantes reçoivent le même code et le résultat est inscrit dans la table du contrôle qualitatif des résultats de CART.

Pour le recensement canadien, la codification automatique de 9 des 10 questions provient uniquement de cette méthode d'appariement. Seule la question Lieu de résidence, il y a cinq ans (libellé des villes et municipalités canadiennes) utilise également l'appariement indirect pour augmenter son taux de codification automatique.

3.3 CART - appariement indirect

Toutes les phrases uniques non chiffrées sont ensuite soumises à la méthode des appariements indirects. Pour augmenter son taux de codification automatique, seule la question Lieu de résidence, il y a 5 ans (libellé des villes et municipalités canadiennes) peut obtenir un appariement gagnant, c'est-à-dire une codification automatique; dans ce cas, toutes les phrases correspondantes reçoivent le même code et le résultat est inscrit dans la table du contrôle qualitatif des résultats de CART. L'information concernant les appariements "gagnants multiples" et "potentiels" (la phrase appariée, le code correspondant et le pointage) est inscrite au dossier de toutes les phrases correspondantes à cette phrase. Cette information sera utilisée à la codification assistée par ordinateur. S'il n'y a pas d'appariement ou seulement des appariements dont le pointage est inférieur au pointage minimal MIN, aucune information n'est consignée.

3.4 CART - notes sur l'exécution

Plusieurs applications ont partagé les mêmes fichiers de référence et les mêmes stratégies de normalisation. Ces fichiers bilingues furent bâtis à partir des entrées des manuels de nomenclature, d'un échantillon de libellés du recensement de 1986 et des libellés d'enquêtes-ménages courantes.

Puisque l'exécution était faite sur une base journalière, il fut possible d'analyser régulièrement les résultats de CART et les phrases non chiffrées. Les fichiers de référence furent mis à jour cinq fois afin d'augmenter le taux d'appariement automatique et la qualité des résultats. Aucune amélioration des stratégies de normalisation n'était permise parce leur impact sur la qualité des résultats était imprévisible.

3.5 Codification des commis assistée par ordinateur

L'ordinateur scrute le fichier original des phrases candidates (ordonnées alphabétiquement) et prépare des lots de 200 phrases non chiffrées pour les

commis. Le commis n'a pas accès au questionnaire original, mais l'information suivante apparait sur 2 écrans (voir figure 3 et 4, page 11). Sur le premier écran, il voit la phrase à codifier, les résultats de CART (phrases appariées et codes associés) et enfin les réponses des autres membres du ménage à la même question. Sur un deuxième écran, le commis peut obtenir les réponses de la personne à des variables annexes. Le commis peut soit choisir un des résultats de CART, soit inscrire un code basé sur un manuel de nomenclature ou référer la codification à un expert. Chaque fois que le commis inscrit un code, le système imprime au bas de l'écran l'énoncé officiel du manuel de nomenclature; le commis doit lire et confirmer le code. Le résultat de la codification est inscrit dans la table du contrôle qualitatif des résultats du commis.

L'ordinateur transfère électroniquement les phrases référées à l'expert de service. L'expert a accès, sur écran, à plus d'information tel les pointages de CART et l'information annexe pour tous les autres membres du ménage. De plus, il peut consulter des manuels de référence plus spécialisés.

3.6. Table du contrôle qualitatif des résultats de CART

Le contrôle qualitatif pour la codification automatique a les mêmes objectifs que celui pour la codification traditionnelle. Cependant il diffère en étendue car beaucoup plus d'information sur l'opération est disponible et cette information peut facilement être modifiée.

Chaque aspect du contrôle qualitatif exploite la nature systématique de la codification automatique car une phrase reçoit toujours le même code s'il n'y a pas d'intervention humaine. Donc l'examen d'une seule occurrence d'une phrase suffit pour établir sa qualité. Les conclusions sur la qualité s'étendent à toutes les répliques de cette phrase.

La table du contrôle qualitatif contient une entrée pour chaque couple phrase-code. Un indicateur de statut est associé au couple. Sa valeur est 1 pour un couple approuvé a priori, 2 pour un couple vérifié et valide, 3 pour un couple vérifié et invalide et 4 pour un couple non vérifié. Durant la production, chaque nouveau couple phrase-code codifié automatiquement est ajouté à la table tandis que la fréquence d'occurrence est augmentée pour chaque couple répété.

Puisque les entrées initiales du fichier de référence ont fait l'objet de tests intensifs, tous les couples appartenant à ce fichier sont inscrits dans la table du contrôle qualitatif avec le statut approuvé a priori et ils ne sont pas vérifiés. Ceci rend plus efficace le contrôle qualitatif.

Les autres couples sont échantillonnés sur une base de priorité. Dès que le couple phrase-code a une fréquence de trois ou plus, une des répliques est sélectionnée et regroupée avec d'autres couples par lot de 200 pour être codifié par un commis de première ligne.

Le système compare le code assigné par CART avec celui fourni par le commis. Si les codes correspondent, le couple est dit valide. Sinon, la codification est soumise à un autre commis de première ligne. Si le nouveau code correspond à celui de CART, alors le couple est jugé valide. S'il correspond à celui du

Figure 3: exemple d'un premier écran - codification par commis

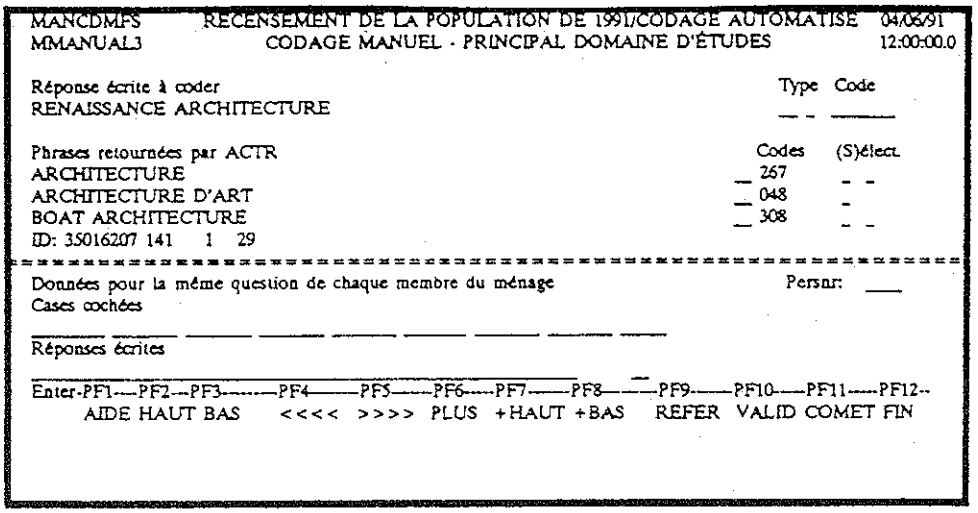
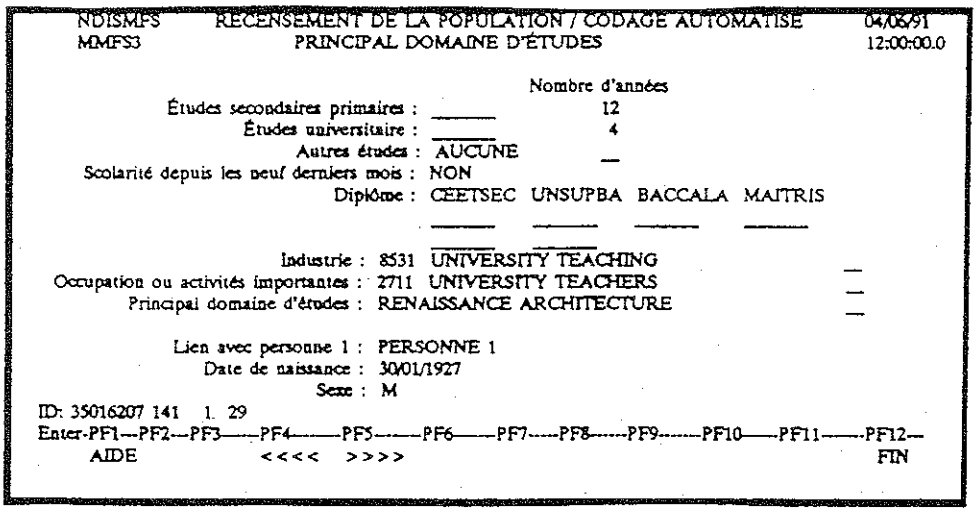


Figure 4: exemple d'un deuxième écran - codification par commis



premier commis, le couple est jugé invalide. Finalement, s'il ne correspond pas à aucun des deux codes, le cas est référé à un expert.

Ce type de contrôle qualitatif identifie les différences entre le code établi par un commis et celui de CART et aide à repérer les problèmes opérationnels dans les deux types de codification. Le responsable de la variable qui est aussi un spécialiste dans la nomenclature doit éventuellement faire une revue des résultats et établir ce qui est vraiment en erreur. Celui-ci aura la possibilité de rectifier les erreurs systématiques.

En plus de faciliter l'échantillonnage pour le contrôle qualitatif, la table du contrôle qualitatif sert à calculer régulièrement des taux d'erreur. Le responsable de la variable peut aussi scruter les couples phrase-code dont les fréquences sont inférieures à 3 et établir la qualité de la codification.

3.7 Table du contrôle qualitatif des résultats du commis

La table du contrôle qualitatif pour les résultats de la codification par les commis contient une entrée pour chaque phrase candidate traitée. Cette phrase est accompagnée du code assigné par le commis, un numéro de lot, du numéro du commis et du code final lorsque la phrase a subit un contrôle qualitatif.

L'objectif du contrôle qualitatif est de déterminer la performance des commis, d'identifier les zones à problèmes, de s'assurer que les objectifs de qualité sont atteints, de donner une rétroaction à l'opération et de prévenir la répétition d'erreur.

La méthode de contrôle qualitatif utilisé est la méthode d'échantillonnage par attribut avec une rectification à 100% des lots rejetés. En pratique 5 phrases d'un lot de 200 sont vérifiées par un commis de première ligne. Comme pour le contrôle qualitatif des résultats de CART, il n'y a pas de vérification supplémentaire lorsque les codes correspondent. Dans le cas contraire, on fait appel à un deuxième commis de première ligne et finalement à un expert pour déterminer le code exact.

Un lot est rejeté et recodifié dès qu'une phrase a un code en erreur.

Le code qui apparaît sur le fichier du recensement est le code établi lors de la vérification ou le code original s'il n'a pas été vérifié. Des taux d'erreur sont régulièrement produits. Le responsable de la variable a accès à toute l'information de la table et peut apporter les correctifs qui s'imposent.

3.8 Rectification des erreurs systématiques

Les deux tables du contrôle qualitatif contiennent l'histoire de la codification automatique et de la codification par des commis. Durant son analyse de l'information de ces tables, le responsable de la variable identifie les erreurs (de préférence systématiques) qui doivent être corrigées. L'analyse peut mener à une modification de la nomenclature pour refléter une nouvelle réalité. L'application du recensement renferme un module de rectification qui est utilisé à la fin de la production immédiatement avant l'intégration des résultats à la base de données principale du recensement.

Le module de rectification des erreurs systématiques agit globalement sur les couples phrase-code en erreur et étend son action sur toutes les répliques du couple. Des rapports détaillés des actions prises sont produits afin de bien contrôler cette opération.

3.9 Résultats et observations
3.9.1 Volume de codification et taux d'appariement

Pour la présentation des résultats, les libellés des 10 questions soumises à la codification automatique ont été regroupés sous 7 variables qui employaient des fichiers de référence et des stratégies de normalisation distincts. Le Tableau 1 présente ces variables et des statistiques opérationnelles.

Tableau 1: Codification Automatique - variables et statistiques

Variable	Traités	Appariés par CART	Taux CART	Codifiés par commis
Origine ethnique	1,160,491	1,062,015	91.51%	98,476
Langage	5,998,021	5,741,294	95.72%	256,727
Indien(ne) inscrit(e)	236,501	169,675	71.74%	66,826
Lieu de résidence - 5 ans (ville/muni.)	1,042,951	793,425	76.08%	249,526
Principal domaine d'études	1,905,959	1,485,196	77.92%	420,763
Province - Pays - Territoire	880,077	821,510	93.35%	58,576
Religion	4,859,569	4,752,021	97.79%	107,548
Total	16,083,569	14,825,136	92.18%	1,258,433

Des 16 millions de libellés soumis à la codification automatique 14.8 millions ou 92.18% ont été chiffrés par CART (taux d'appariement). Les autres 1.2 millions ont été résolus par une codification assistée par ordinateur.

Les taux d'appariement sont regroupés en deux groupes principaux; dans l'intervalle 71% à 78 % et dans l'intervalle 91% à 98%. La différence des taux par variable s'explique par le volume traité, la variation des réponses, la longueur des libellés, l'utilisation d'abréviation par le répondant, le changement dans les frontières dû à l'effondrement du bloc communiste et le fait que certains libellés (par exemple: un nom de municipalité qui est associé à plusieurs codes) étaient délibérément envoyés à la codification par commis où l'information annexe pouvait être utilisée pour établir le code exact.

La question sur les Indiens inscrits était nouvelle et il était difficile de prévoir les réponses surtout parce que plusieurs noms ont récemment subi de nombreux changements. La variable Lieu de résidence, il y a cinq ans évitait l'utilisation des noms de lieu répétés en ne les incluant pas dans le fichier de

référence. Les noms de lieu répétés incluent les lieux géographiques qui ont le même nom à l'intérieur d'une province ou, si la province n'est pas identifiée, le même nom dans plus d'une province. De plus, on excluait un nom comme "Québec" puisque celui-ci pouvait référer à la province ou à la ville. La variable ***Principal domaine d'études*** *avait un nombre de réponses très variées, une nomenclature diverse et l'utilisation d'abréviations ou de libellés très longs. Le problème avec les longs libellés est qu'une erreur dans seulement un des mots peut empêcher un appariement direct, seul appariement permis pour cette variable. De plus, il n'était pas possible de répertorier toutes les variations d'épellation et les abréviations de ces libellés. Finalement, les longs libellés sont plus sujets à des erreurs à l'opération de saisie des libellés.*

3.9.2 Mise à jour des fichiers de référence

Durant la production, il y a eu 5 mises à jour des fichiers de référence. On estime qu'elles ont augmenté le taux d'appariement de 2 points de pourcentage, ou alternativement, qu'elles ont réduit le volume de codification par des commis d'environ 25%. Dans certains cas des phrases ont été enlevées car elles étaient ambiguës et elles généraient des erreurs.

3.9.3 Analyse de la Table du contrôle qualitatif des résultats de CART

Tel que mentionné précédemment, tous les couples uniques phrase-code avaient un des statuts suivants: approuvé a priori, vérifié et valide, vérifié et invalide, non vérifié.

Le terme "invalide" indique ici qu'il y a différence entre le code de CART et celui établi au contrôle qualitatif. Les différences proviennent parfois de codes erronés dans le fichier de référence, de phrases trop normalisées, de commis qui n'avaient pas les instructions les plus récentes ou qui ont fait des erreurs de jugement ou d'inattention. Une autre cause de différence est la possibilité que le libellé est associé à plusieurs codes. Donc ce que nous mesurons ici est une différence brute qui doit être analysée avant d'initier une rectification. C'est aussi la fonction de l'analyste de repérer les quelques erreurs qui ont été manquées au contrôle qualitatif.

Le tableau 2 reflète le volume des phrases selon les différents statuts. Plus de 87% des phrases codifiées par CART étaient approuvées a priori. Moins de 1% des phrases ont été identifiées comme ayant un code invalide.

Tableau 2: Résultat du contrôle qualitatif - toutes les variables

STATUT	COUPLES UNIQUES	FRÉQUENCE	FREQ. (%)	CONTRÔLE (%) total
Approuvé a priori	14,787	12,898,773	87.01%	
Vérifié et invalide	2,705	89,743	0.61%	0.018%
Vérifié et valide	34,499	1,735,931	11.71%	0.233%
Non vérifié	82,128	100,689	0.67%	
Total codifié par CART		14,825,136	100.0%	

3.9.4 Ressources pour le contrôle qualitatif

Les ressources planifiées pour le contrôle qualitatif visaient à vérifier 3.0% des libellés codifiés par CART et 10.0% des libellés codifiés par les commis. Ce dernier pourcentage était réparti comme suit: 2.5% pour l'échantillon et 7.5% pour recodifier les lots rejetés.

Les taux finaux furent de 0.251% (tableau 2: [2,705 + 34,499]/ 14,825,136) pour la codification automatique et de 10.02% pour la codification par les commis.

Le taux de 0.251% est attribuable à la haute fréquence d'occurrences des couples phrase-code approuvés a priori et au fait que chaque couple unique était sélectionné et vérifié seulement une fois. Cette stratégie de vérification est impossible dans une opération traditionnelle de contrôle qualitatif. Ce taux indique donc que l'exploitation de toute l'information produite par les systèmes peut augmenter l'efficacité de la vérification sans compromettre la qualité.

Le tableau 3 illustre, par variable, la fréquence moyenne d'occurrences des couples uniques phrase-code codifié par CART.

La fréquence moyenne des couples phrase-code approuvés a priori est de 872. La fréquence la plus intéressante est celle des couples vérifiés et invalides avec une moyenne de 33. Ceci signifie que la correction d'un de ces couples corrige en moyenne 33 erreurs.

Tableau 3: Fréquence moyenne des couples phrase-code par variable et statut

VARIABLE / STATUT	APPROUVÉ A PRIORI	VÉRIFIÉ ET INVALIDE	VÉRIFIÉ ET VALIDE	NON VÉRIFIÉ
Origine ethnique	528	12	27	1
Langage	1,906	167	128	1
Indien inscrit	103	13	37	1
Lieu de résidence il y a 5 ans (villes)	---	19	44	1
Principal domaine d'études	180	16	29	1
Province - Pays - Territoire	588	393	38	1
Religion	4,252	25	105	1
Toutes les variables	872	33	50	1

Pour le prochain recensement, le but sera d'approuver a priori le plus de couples possibles afin de minimiser les ressources consacrées au contrôle qualitatif. Les ressources dégagées pourront être utilisées pour mieux analyser les deux tables du contrôle qualitatif.

3.9.5 Rectification des erreurs systématiques

Environ 94,000 codes furent rectifiés par le module de rectification. Les codes provenaient des deux types de codification (automatique et par commis). La plupart des rectifications ont amélioré la qualité. Pour les variables **Origine Ethnique**, **Langage** et **Province-Pays-Territoire**, quelques codes furent changés pour refléter la nouvelle réalité mondiale, réalité qui changea beaucoup entre la production du questionnaire et la fin du traitement des données du recensement.

Notre estimation de la qualité finale pour les deux types de codification est un taux combiné d'erreur inférieur à 1%: la codification par les commis est la source principale des erreurs. Cependant le taux atteint est remarquablement bas puisque dans les recensements précédents le taux d'erreur se situait dans l'intervalle de 4% à 8% dépendant de la question.

3.9.6 Coût de l'opération de codification

Le coût de l'opération de codification est estimé à 2.5 millions de dollars soit 60% de moins que si la codification avait été faite par des commis uniquement. Le coût ne comprend pas le coût de développement de CART et le coût supplémentaire de saisie des libellés (0.9 million de dollars) mais il reflète les coûts associés au développement des fichiers de référence, des stratégies de normalisation, et au développement des systèmes pour la codification assistée par ordinateur, le contrôle qualitatif et la rectification des erreurs. La réduction de coût provient de la réduction du nombre de commis de 600 à 25 et de leur efficacité accrue.

4. CONCLUSION

L'utilisation de la codification automatique pour le recensement de 1991 a été un franc succès sur lequel nous voulons capitaliser pour le recensement de 1996.

Nos intentions pour le recensement de 1996 sont les suivantes:

Le logiciel CART sera utilisé de nouveau mais il subira certaines modifications afin d'augmenter sa polyvalence. Il aura la capacité - de spécifier l'ordre des fonctions lors de la normalisation des libellés; de conserver l'ordre original des mots lors de la création de la clé d'expression condensée utilisée par l'appariement direct; et de calculer le poids des mots et les pointages suivant un choix de méthodes.

Les applications de codification de 1991 seront légèrement modifiées pour les rendre plus performantes. Les fichiers de référence et stratégies de normalisation seront mis à jour. Un nouveau module localisé au début de l'application est à l'étude; il décidera si un libellé doit être soumis à la codification automatique, être envoyé directement à la codification par les commis ou recevoir un code intérimaire indiquant qu'il n'y pas suffisamment d'information pour chiffrer. Finalement le manuel de nomenclature sera disponible à l'écran afin de faciliter la codification par les commis.

Deux nouvelles questions seront codifiées en 1996: Relation avec la personne de

référence et Lieu de travail (codifié au niveau du pâté de maisons). Pour ces questions, l'application de codification sera plus complexe et fera appel à CART et à d'autres logiciels d'appariement de dossiers (voir Tourigny, Moloney, Miller (1993).

Le défi pour le recensement de 2001 sera de codifier de façon automatique les deux dernières questions ayant des libellés en clair, soit l'Activité économique de l'entreprise et la Profession. Ironiquement, l'intention première lors du développement de CART était de codifier ces deux questions.

BIBLIOGRAPHIE

Ciok R. (1993). "The results of automated coding in the 1991 Canadian Census of Population". Document présenté à "1993 Annual Research Conference", conférence organisée par le Bureau du Recensement des États-Unis.

Hellerman E. (1982). "Overview of the Hellerman I&O Coding System". Document interne. Bureau du recensement des États-Unis.

Knaus R. (1981). "Pattern-based Semantic Decision Making". Texte du livre "Empirical Semantics", édité par Rieger B., Bochum, Allemagne de l'Ouest.

Tourigny J., Moloney J., Miller D. (1983). "The 1991 Canadian Census of Population experience with automated coding". Document présenté à la session de travail sur la vérification statistique des données. Session organisée par la Conférence des statisticiens européens. Stockholm, Suède.

Wenzowski, M.J. (1988). "ACTR - Un système généralisé de codage automatique". Techniques d'enquête, vol 14, pp. 317-326.

ANNEXE A

NORMALISATION DES PHRASES

Le logiciel de codification automatique CART contient un module qui permet la normalisation des phrases du fichier de référence et du fichier d'enquête. Il s'agit d'une suite fixe de 14 fonctions qui, suivant l'application de codification, peuvent ou non être utilisées. Les quatre premières fonctions identifient les mots de la phrase; les 10 autres fonctions normalisent ces mots. Pour chaque fonction utilisée, le responsable de la variable doit fournir une liste de caractères valides, de mots, de mots de remplacement ou de suffixes.

Traitement de texte:

La phrase est traitée comme une chaîne ininterrompue de caractères afin de pouvoir éventuellement identifier des mots distincts:

Fonction 1: clauses d'exclusion - pour les phrases du fichier de référence, le texte qui indique une clause d'exclusion (par exemple,"commis (sauf dans l'armée)") doit être exclu car un répondant ne s'exprime pas de cette façon. Le résultat sera des phrases normalisées identiques dans le fichier de référence qui conduiront à des appariements "gagnants multiples". CART n'assignera pas un code mais sur ces appariements sera acheminé à un commis qui devra décider du code approprié.

Fonction 2: élimination de caractère - permet d'éliminer les caractères inutiles, tels les apostrophes dans la langue anglaise, qui seraient interprétés comme des indicateurs du début ou de la fin d'un mot par la fonction 4.

Fonction 3: remplacement de caractères - permet de remplacer une abréviation par un ou des mots sinon le sens de l'abréviation sera détruit par la fonction 4. Par exemple télévision remplace "T.V." -

Fonction 4: bris du texte en mots - si un caractère n'est pas dans la liste des caractères valides pour un mot, il indique le début ou la fin d'un mot; par exemple si seulement les chiffres, les lettres et le trait d'union sont valides, les deux phrases suivantes seront divisées en 2 mots "T.V." = T V, "anglais/français" = anglais français, et la phrase "Electrician's Apprentice" en 3 mots.

Traitement des mots

La phrase est traitée comme une collection de mots. Par conséquent, les fonctions suivantes s'appliquent à chacun des mots pris individuellement.

Fonction 5: mots à trait d'union - permet de préserver en un mot deux mots qui ensemble ont un sens spécifique par exemple "post-secondaire". Si le mot à trait d'union n'est pas dans la liste, il est brisé en deux mots; autrement il est remplacé par un nouveau mot.

Fonction 6: caractères alphanumériques non valides - si un mot est formé d'une chaîne de caractères qui le rend inintelligible, ce mot est supprimé sans autre considération. Dans certaines applications, on utilise cette fonction pour supprimer des mots qui renferment des caractères numériques.

Fonction 7: mots de remplacement - cette fonction agit de la même façon que la fonction 3; la différence majeure est que la recherche est limitée à des mots entiers et non à une partie de mot. Cette fonction fait en sorte que deux mots synonymes soient reconnus comme pareils pour les fins d'appariement. Cette fonction peut aussi être utile pour corriger les fautes d'orthographe courantes.

Fonction 8: mots doubles - si deux mots, lorsque pris ensemble dans un certain ordre, ont un sens particulier, cette fonction permet de les remplacer par un seul mot. Par exemple les deux mots "radio" "active" sont remplacés par "radioactive" et "garde" "malade" par "infirmier". Cette fonction peut résoudre des incohérences dans l'orthographe et contrer une modification de l'ordre des mots qui aurait lieu lors de la construction de la "clé d'expression condensée" pour un appariement direct.

Fonction 9: mots sans importance - un mot sans importance, tel un article, un pronom, ne contribue pas au contenu sémantique de la phrase; il peut être supprimé sans autre considération.

Fonction 10: mots racine - les fonctions 11, 12 et 13 peuvent faire en sorte que deux mots sémantiquement différents peuvent être réduits à la même racine. Cette fonction examine les mots pour y déceler des mots racines. S'il en trouve un, le mot entier est remplacé par un mot substitut et les trois fonctions suivantes ne sont pas activées.

Fonction 11: remplacement de suffixes - un mot est scruté de droite à gauche pour y trouver la plus longue forme de suffixe se trouvant dans la liste. Si un tel suffixe est repéré, il est remplacé par le substitut prévu. Par exemple, la marque du pluriel peut être éliminé de telle manière que le suffixe est reconnu par la fonction 12. En anglais on peut remplacer "ies" par "y".

Fonction 12: suffixes - habituellement un suffixe ne change pas le contenu sémantique d'un mot. Cette fonction scrute un mot de droite à gauche pour y trouver la plus longue forme de suffixe se trouvant dans une liste, de telle sorte qu'une fois le suffixe enlevé, le mot contienne au moins cinq caractères. Si une forme définie de suffixe est repéré, elle est supprimée. Des exemples de suffixes sont able, aliste, icienne, trice.

Fonction 13: consonnes ou voyelles doubles - l'élimination des consonnes ou voyelles doubles ne change habituellement pas le contenu sémantique du mot. Cette élimination peut annuler des erreurs d'orthographe ou de saisie de données.

Fonction 14: mots répétés - seulement une occurrence de chaque mot normalisé est conservé dans la phrase normalisée.

ANNEXE B
Questions soumises à la codification automatique

Première langue apprise

Quelle est la langue que cette personne a apprise en premier lieu à la maison dans son enfance et qu'elle comprend encore?

Réponse: si la langue est autre que l'anglais ou le français, la personne précise celle-ci.

Note: Cette question apparait sur le questionnaire abrégé et complet.

Langue parlée à la maison

Quelle langue cette personne parle-t-elle le plus souvent à la maison?

Réponse: si la langue est autre que l'anglais ou le français, la personne précise celle-ci.

Langues non-officielles

Quelle(s) langue(s), autre(s) que l'anglais ou le français, cette personne connait-elle assez bien pour soutenir une conversation?

Réponse: la personne peut préciser jusqu'à trois langues.

Lieu de naissance

Où cette personne est-elle née?

Réponse: si la personne est née dans un pays autre que les 6 pays proposés, elle doit préciser ce pays.

Origine ethnique - ancêtres

À quel(s) groupe(s) ethnique(s) ou culturel(s) les ancêtres de cette personne appartenaient-ils?

Réponse: si la personne appartient à un groupe autre que les 15 groupes proposés, elle peut préciser jusqu'à deux autres groupes.

Indien(ne) inscrit(e)

Cette personne est-elle un(e) Indien(ne) inscrit(e) aux termes de la Loi sur les Indiens du Canada?

Réponse: si la case oui est coché, la personne précise la bande indienne ou première nation

Religion

Quelle est la religion de cette personne?

Réponse: la personne précise une seule confession ou une seule religion, ou coche la case "Aucune religion".

Lieu de résidence, il y a 1 an

Où cette personne habitait-elle il y a 1 an, c'est-à-dire le 4 juin 1990?

Réponse: si la personne n'habitait pas à une adresse dans la même province/territoire, elle doit préciser soit l'autre province/territoire ou le nom d'un autre pays.

Lieu de résidence, il y a cinq ans

Où cette personne habitait-elle il y a 5 ans, c'est-à-dire le 4 juin 1986?

Réponse: si la personne n'habitait pas à une adresse dans la même ville, elle doit préciser soit le nom de l'autre ville ou le nom d'un autre pays.

Principal domaine d'études

Quel était le principal domaine d'études ou de formation du plus haut grade, certificat ou diplôme de cette personne (sans compter les certificats d'études secondaires)?

Réponse: la personne indique que le plus haut diplôme est un certificat d'études secondaires ou précise un principal domaine d'études ou de formation.