ODSEF Observatoire démographique et statistique de l'espace francophone

Note de recherche

Codage de l'activité principale
Recensement de 2013 au Sénégal

Par
Jean Pierre Diamane BAHOUM

Avec la collaboration de
Laurent RICHARD

Agence universitaire de la Francophonie
Organisation internationale de la francophonie
Québec
Université Laval

Codage de l'activité principale

Recensement de 2013 au Sénégal

Note de recherche réalisée par :

Jean Pierre Diamane Bahoum
Chef de la Division des Opérations de Terrain
Agence Nationale de la Statistique et de la Démographie
Sénégal

Avec la collaboration de :
Laurent Richard
Professionnel de recherche, Université Laval

Note de recherche de l'ODSEF

Québec, mars 2021

Éléments de référence pour citer ce document :

BAHOUM, Jean Pierre Diamane, avec la collaboration de Laurent RICHARD (2021). *Codage de l'activité principale : recensement de 2013 au Sénégal.* Québec, Observatoire démographique et statistique de l'espace francophone, Université Laval, Note de recherche de l'ODSEF, 11 p.

Note à propos des auteurs :

Jean Pierre Diamane BAHOUM est Chef de la Division des Opérations de Terrain (DOT), à la Division des Statistiques Démographiques et Sociales (DSDS) de l'Agence Nationale de la Statistique et de la Démographie du Sénégal.

Laurent RICHARD est professionnel de recherche à l'Université Laval, Québec, Canada.

ISBN : 978-2-924698-29-7 (PDF)

Remerciements

En 2019, Jean Pierre Diamane Bahoum a eu l'opportunité de séjourner à Québec, à l'Observatoire démographique et statistique de l'espace francophone (ODSEF) de l'Université Laval. Lors de ce séjour, Jean Pierre a travaillé au codage des réponses de type « autre à préciser » du recensement de 2013 au Sénégal. C'est grâce au troisième Accord - Cadre liant l'ANSD à l'ODSEF et favorisant la valorisation des données du recensement sénégalais de 2013 que cette activité a été réalisée. Les auteurs remercient l'ANSD et l'ODSEF qui ont permis d'approfondir le traitement des données du recensement sénégalais de 2013. Ce fastidieux travail de codage s'est poursuivi après le séjour de Jean Pierre Diamane Bahoum à Québec, dans un contexte difficile, devenu de plus en plus complexe avec la pandémie de COVID-19 et la planification du prochain recensement sénégalais. Nous espérons que la diffusion de ce travail permettra d'améliorer les processus de collecte et offrira de nouvelles possibilités de traitement des informations pour de nombreux pays africains.

Résumé

Dans le cadre d'une opération de collecte d'envergure comme celle du recensement sénégalais de 2013, le codage des réponses constitue un défi. L'utilisation de formulaires électroniques permet de traiter instantanément plusieurs réponses. Toutefois, les listes de réponses préétablies sont limitées et une part importante de répondants fournissent des informations qui sont colligées à l'aide de l'item « autre à préciser ». C'est notamment le cas des questions au sujet de la profession et de l'activité principale. Cette note de recherche montre que le traitement des informations contenues dans les modalités « autre à préciser » permet d'obtenir un portrait plus précis de ces phénomènes. Les travaux de codification réalisés uniquement pour la région de Dakar ont ainsi permis d'attribuer un code spécifique d'activité pour plus de 210 000 individus parmi ceux qui se trouvaient dans la modalité « autre à préciser », soit un peu plus des deux tiers des individus concernés (67,9%). Les activités liées au secteur commercial sont celles qui ont été les plus touchées par ce travail minutieux: près de 35 000 Dakarois et Dakaroises déclarant une activité commerciale s'ajoutent ainsi aux 27 000 initialement estimés.

Mots-clés

Codification, recensement, Sénégal, activité principale.

TABLE DES MATIÈRES

Liste des tableaux

Sigles et abréviations

ANSD	Agence Nationale de la Statistique et de la Démographie, Sénégal
CITP	Classification internationale type des professions
NAEMA	Nomenclature d'activités (États membres d'AFRISTAT)
DOT	Division des Opérations de Terrain
DSDS	Division des Statistiques Démographiques et Sociales
ODSEF	Observatoire démographique et statistique de l'espace francophone
RGPHAE	Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Élevage, 2013, Sénégal

Introduction

L'utilisation de la technologie numérique lors du RGPHAE de 2013 au Sénégal a été une innovation majeure qui a permis de réduire les coûts de stockage des questionnaires, d'accroitre la fiabilité des données, de réduire les délais de diffusion et, surtout, de limiter les erreurs d'écriture. Par ailleurs, cette façon de faire a permis de collecter des données avec des règles de contrôle de cohérence préétablie pour éviter les erreurs d'observations et faciliter la transmission de manière instantanée vers le niveau central.

Cependant, quelle que soit la qualité des différents outils technologiques utilisés, il existe des avantages et des inconvénients associés à cette méthode de collecte. D'aucuns sont liés à l'outil ou aux procédures de recueil sur le « terrain », et d'autres émanent de la responsabilité des agents recenseurs dans la transcription des réponses données par les interviewés. Par conséquent, le choix technologique ne saurait être le seul facteur de succès dans un recensement. En effet, certaines données issues du RGPHAE, notamment celles recueillies à partir des modalités « autre à préciser », pour les variables relatives à l'activité économique et à la profession, nécessitent d'être codifiées.

La codification est une opération fondamentale du recensement qui permet de transcrire en données numériques par l'affectation d'un code, les données alphabétiques recueillies sur le « terrain ». En effet, si la collecte des données sur le terrain est une opération primordiale, une mauvaise codification des réponses ouvertes pourrait conduire à des résultats erronés. Par conséquent, les données codifiées doivent refléter le mieux possible la réalité mesurée lors de l'opération de collecte. Pour ce faire, il est nécessaire d'avoir une bonne connaissance de la méthode de codification, ainsi qu'une excellente maîtrise du système de codes en vigueur. Ainsi, afin de traiter les chaînes de textes saisies par les agents recenseurs dans les modalités « autre à préciser » des deux

variables considérées ici, la NAEMA[1]-rév1 et la CITP-08[2] ont été retenues comme systèmes de référence.

L'objectif de cette note de recherche est de :

1) Proposer une démarche méthodologique pour la codification des occurrences de la modalité « autre à préciser » pour les variables « activité principale et profession ».
2) Présenter quelques résultats des travaux de codification concernant les modalités « autre à préciser ».
3) Dégager des pistes d'amélioration pour les prochaines opérations de recensement, comme une meilleure prise en charge des variables relatives à l'activité économique au moment de la conception et durant les travaux de traitement des données.

La présente note technique s'inscrit dans les activités consignées dans le troisième Accord - Cadre liant l'ANSD à l'ODSEF et portant sur la valorisation des données du recensement sénégalais de 2013. Bien que des travaux aient été entrepris en vue de codifier l'ensemble des réponses ouvertes concernant la profession et l'activité principale, seuls les résultats ayant trait à l'activité principale sont présentés ici. De plus, les résultats portent uniquement sur la région de Dakar. La note de recherche est structurée en quatre sections. La première concerne principalement le processus de collecte de données. Ensuite, la seconde section porte sur la méthodologie mis en œuvre afin de coder les réponses ouvertes. Enfin, les deux dernières sections présentent les principaux résultats et des réflexions au sujet d'améliorations potentielles dans le cadre de futures collectes de données.

---

[1] www.afristat.org/contenu/.../NAEMA_NOPEMArev1_def_vers_12mai11.

[2] https://www.ccss.lu/fileadmin/file/ccss/PDF/SECUline/CITP-08.pdf

1.1. La production des données : agents recenseurs et équipe « terrain »

L'étape dite du « terrain » est fondamentale dans la réalisation d'un recensement. Le « terrain » recoupe plusieurs acceptions terminologiques. Il désigne déjà une phase dans la production des données : tout ce qui se rapporte au recueil de données, tout le travail de production qui n'est effectué ni par un directeur technique ni par un chargé d'études relève de la compétence de l'équipe du terrain. Ainsi, l'administration du questionnaire par les agents recenseurs et le codage des réponses ouvertes correspondent à des travaux menés lors de la phase dite de terrain. Par ailleurs, ce terme fait également référence à un lieu géographique, à des locaux où se déroulent ces différentes phases de la collecte de données, aussi bien pour les questions ouvertes que pour les questions fermées.

Le terrain est également le lieu où l'on commet certaines erreurs liées à l'observation directe ou au contenu. Dans le cadre de l'utilisation de la technologie mobile, les sources d'erreurs viennent en partie de la manipulation des terminaux utilisés pour la collecte. En général, elles sont imputables au fait que chaque interface correspond, à une question avec des instructions et des réponses y afférant. Ainsi, la validation d'une réponse entraîne le passage automatique à la question suivante, donc à l'interface suivante, avec le risque de répéter l'erreur commise sur toute la chaine de production.

De façon concrète, pour les questions ouvertes, comme c'est le cas des questions portant sur la profession et l'activité économique, les agents ont la latitude de saisir directement la réponse donnée par un répondant au niveau de la modalité « autre à préciser ». Autrement dit, bien que cette situation ne soit pas très courante, il se peut que la réponse inscrite sous cette modalité corresponde sensiblement à un item de la liste préétablie. Évidemment, les agents recenseurs n'ont pas de formation de sténotypiste et l'information transmise par le répondant par débit oral ne peut généralement être saisie à la même vitesse via un clavier. La saisie intégrale des informations est donc difficile à obtenir. À cela s'ajoutent, les déficiences en orthographe, de même que l'utilisation d'abréviations et de certains sigles.

En définitive, la prise en charge de la codification des réponses ouvertes et des manquements liés à la saisie automatique et à l'abréviation des données, pour l'activité économique et la profession, requiert une démarche méthodologique assez soutenue. Rappelons que le traitement des réponses à une question ou à une modalité ouverte pose des défis plus complexes qu'un enchaînement de questions à modalités fermées.

1.2. Méthodologie de codage

Le codage des réponses ouvertes relatives à l'activité principale et à la profession a été effectué par une affectation manuelle des codes correspondant (NAEMA-rév1 ou CITP-08) pour les individus dont la réponse se trouve à l'item « autre à préciser ». Afin de procéder efficacement à la codification, il a été résolu de considérer uniquement les réponses comptant minimalement cinq occurrences. En priorisant la codification des valeurs verbatim les plus fréquentes, on maximise l'utilisation des ressources humaines requises pour attribuer les codes tout en s'intéressant à une très large part des individus concernés. Évidemment, plusieurs réponses à la modalité « autre à préciser » concernent un nombre de répondants inférieur à cinq ou corresponde à une valeur unique. Le traitement de ces réponses aurait requis un effort supplémentaire considérable. Par exemple, traiter l'ensemble des occurrences uniques pourrait représenter une somme de travail quasi équivalente aux efforts consentis pour coder les réponses comptant minimalement cinq occurrences. Ce choix méthodologique a donc été guidé par une utilisation maximale souhaitée de nos ressources. De même, les réponses illisibles à cause d'une déficience de l'orthographe ou de l'usage d'une abréviation, dont il est difficile de connaître la signification, ont été exclues du processus de codification.

La démarche méthodologique est spécifique pour chacune des deux variables. Pour la profession des individus qui se retrouvent dans une catégorie résiduelle « autre à préciser », les codes de la CITP-08 à quatre positions leur ont été appliqués. En ce qui a trait à l'activité principale, les codes de la NAEMA-rév1 ont été utilisés en procédant ainsi :

1. L’utilisation du code des grandes divisions de la « NAEMA-rév1 » pour les modalités de réponse assez générales et imprécises.
2. Le recours aux codes des sous-divisions de la « NAEMA-rév1 » pour les modalités de réponse assez précises et plus détaillées.
3. La recherche ciblée des modalités de réponse fournies dans « l’univers » d’activités de la « NAEMA-rév1 » par le repérage de l’univers approprié.

À terme, une vérification de la comparabilité entre la structure des modalités issues de la codification et celle des modalités pré-codifiées, aussi bien pour l’activité principale que pour la profession, est faite, afin de statuer sur la qualité de l’information recueillie dans les modalités ouvertes (« autre à préciser »). Il s’y ajoute l’identification des nouvelles activités et professions listées par les répondants et qui ne figurent ni dans la NAEMA-rév1, ni dans la CITP-08. Ces nouvelles activités et professions identifiées, devraient être gardées et inscrites comme modalités nouvelles lors des prochaines opérations de collecte, si elles disposent d’un seuil minimum de cinq occurrences de réponses valides. Par contre, les catégories résiduelles qui ont des occurrences de réponse inférieures à cinq et qui ne figurent pas dans l’univers d’activités de la NAEMA-rév1, encore moins dans la liste des professions de la CITP-08, sont isolées et attributaires d’un code distinctif.

1.3. Quelques résultats issus du codage

De façon générale, les réponses fournies par le répondant entrainent trois situations principales. Premièrement, la réponse fournie correspond à un des codes prévus dans la NAEMA-rév1 ou la CITP-08, selon le cas. Deuxièmement, l’activité ou la profession déclarée par le répondant semble être une valeur tout-à-fait plausible, mais aucun des codes prévus ne correspond à cette valeur. Troisièmement, la réponse donnée n’est pas exploitable (défaut orthographique ou utilisation d’une abréviation dont la signification est inconnue).

Tel que nous l’avons mentionné précédemment, les travaux de codage concernant la variable « profession » sont très avancés mais ils ne sont pas entièrement complétés.

Bref, seuls les résultats portant sur la variable « activité principale » (B37) sont présentés ci-après. Il est important de mentionner que la base originale de données du recensement de 2013 compte 313 874 Dakarois ayant déclaré pratiquer une activité économique qui n'a pas été rattachée à une modalité préétablie du questionnaire électronique (Tableau 1). Les réponses fournies par ces personnes se trouvent donc dans la catégorie résiduelle « autre à préciser ». En fait, parmi les 1 224 934 individus ayant déclaré une activité principale, le quart d'entre eux (25,5%) n'ont pas obtenu de code spécifique d'activité dans la base originale de données.

Comme l'illustre également le sommaire au bas du tableau 1, les récents travaux de codification réalisés ont permis d'attribuer un code spécifique d'activité pour 213 193 individus parmi ceux qui se trouvaient dans la modalité « autre à préciser », soit un peu plus de deux individus concernés sur trois (67,9%). Ainsi, la base modifiée de microdonnées de recensement ne contient plus que 100 681 individus regroupés dans la modalité « autre à préciser », ce qui représente 8,1% de tous les répondants ayant déclaré une activité principale.

Certes, l'attribution d'un code spécifique d'activité principale pour près de 220 000 individus constitue le fait saillant du travail effectué ici. Il est facile d'imaginer comment cette modification de la base de données peut avoir comme impact dans le cadre d'analyses statistiques bivariées ou multivariées. Si ce résultat justifie à lui seul les efforts de codage consentis ici, le tableau 1 offre aussi une comparaison de la structure d'activité principale avant et après l'exercice de codification de la modalité « autre à préciser ».

Ainsi, ce sont les activités commerciales où le changement d'effectifs est le plus marquant (+33 614). À contrario, le travail de codage a permis d'identifier seulement 120 nouveaux individus dont la grande catégorie correspond aux « activités extractives ». En termes relatifs, ce sont les activités « spécialisées, scientifiques et techniques » qui ont connu le taux maximal de variation (3 625%), cette catégorie passant de 244 à 9 089 individus.

Tableau 1 : Résultat du travail de codage de la modalité « autre à préciser », activité principale (B37), région de Dakar, 2013

Grande catégorie (NAEMA-rév1)	Avant		Après		Variation	
	(n)	(%)	(n)	(%)	(n)	(%)
A) AGRICULTURE, SYLVICULTURE, PÊCHE	37 082	3,99	41 699	3,65	4 617	12,45
B) ACTIVITÉS EXTRACTIVES	63 249	6,81	63 369	5,55	120	0,19
C) ACTIVITES DE FABRICATION	321 479	34,60	347 852	30,45	26 373	8,20
D) PRODUCTION ET DISTRIBUTION D'ÉLECTRICITÉ ET DE GAZ	15 826	1,70	15 990	1,40	164	1,04
E) PRODUCTION ET DISTRIBUTION D'EAU, ASSAINISSEMENT, TRAITEMENT DES DECHETS ET DE POLLUTION	313 885	33,79	315 695	27,64	1 810	0,58
F) CONSTRUCTION, INSTALLATION ET FINITION	9 327	1,00	32 205	2,82	22 878	245,29
G) COMMERCE	27 578	2,97	61 192	5,36	33 614	121,89
H) TRANSPORTS ET ENTREPOSAGE	4 196	0,45	14 455	1,27	10 259	244,49
I) HÉBERGEMENT, BAR ET RESTAURATION	1 706	0,18	11 706	1,02	10 000	586,17
J) INFORMATION ET COMMUNICATION	349	0,04	1 372	0,12	1 023	293,12
K) ACTIVITÉS FINANCIÈRES, COMPTABLES ET D'ASSURANCE	7 336	0,79	10 936	0,96	3 600	49,07
L) ACTIVITÉS ET PROMOTION IMMOBILIÈRES	769	0,08	1 254	0,11	485	63,07
M) ACTIVITÉS SPECIALISEES, SCIENTIFIQUES ET TECHNIQUES	244	0,03	9 089	0,80	8 845	3625,00
N) ACTIVITES DE SERVICES DE SOUTIEN ET DE BUREAU	78 461	8,45	100 949	8,84	22 488	28,66
O) ACTIVITES D'ADMINISTRATION PUBLIQUE	3 870	0,42	20 210	1,77	16 340	422,22
P) ENSEIGNEMENT	1 860	0,20	13 324	1,17	11 464	616,34
Q) ACTIVITÉS POUR LA SANTÉ HUMAINE ET L'ACTION SOCIALE	1 579	0,17	11 213	0,98	9 634	610,13
R) ACTIVITÉS ARTISTIQUES, SPORTIVES ET RECREATIVES	225	0,02	7 497	0,66	7 272	3232,00
S) AUTRES ACTIVITÉS DE SERVICES N.C.A.	423	0,05	730	0,06	307	72,65
T) ACTIVITÉS SPECIALES DES MÉNAGES	1 214	0,13	12 302	1,08	11 088	913,34
U) ACTIVITÉS DES ORGANISATIONS EXTRATERRITORIALES	26	0,00	420	0,04	394	1515,38
V) IMPRIMERIE ET REPRODUCTION D'ENREGISTREMENTS	37 777	4,07	39 541	3,46	1 764	4,67
W) ENTRETIEN ET REPARATION	347	0,04	6 060	0,53	5 713	1646,40
X) ACTIVITES D'ABATTAGE, DE TRANSFORMATION ET CONSERVATION DE VIANDE	252	0,03	3 193	0,28	2 941	1167,06
Sous-total	**929 060**	**100,0**	**1 142 253**	**100,0**	**213 193**	**22,95**
« Autre à préciser »	313 874	25,25	100 681	8,1	-213 193	-67,92
Total	**1 242 934**	**---**	**1 242 934**	**---**	**---**	**---**

Source : ANSD. Recensement général de la population et de l'habitat, de l'agriculture et de l'élevage, Sénégal, 2013.

De même, les activités « artistiques, sportives et récréatives » effectuent un saut quantique (3 232%) tout comme la grande catégorie « entretien et réparation » (1 646%). Évidemment, les taux de variation constituent de bons indicateurs des changements survenus dans chacune des catégories prises isolément, sans nécessairement impliquer des changements structuraux. Par exemple, la catégorie « activités des organisations extraterritoriales » regroupait initialement 26 individus (moins de 0,01% de la distribution) alors que les 420 individus de ce groupe après le processus de codage représentent seulement 0,04% de tous les individus disposant d'un code spécifique d'activité; il n'y a donc pas un réel changement structurel malgré un taux de variation de 1 515%.

Autrement dit, pour mesurer les changements structuraux, il vaut mieux comparer les pourcentages avant et après codification plutôt que d'utiliser uniquement les taux de variation. Ainsi, avec un taux de variation de 8,2% des « activités de fabrication », on observe une diminution du poids relatif de ce groupe qui passe de 34,6% à 30,5%, avant-après la codification. De même, les diverses activités liées à la gestion de l'eau et des déchets, qui regroupaient le tiers (33,8%) des répondants ayant un code spécifique d'activité avant le codage supplémentaire, forment désormais 27,6% du corpus. Globalement, la structure des activités principales avant et après codification est assez stable. Cet état de fait est plutôt rassurant; l'exercice de codage des valeurs verbatim saisies à l'item « autre à préciser » semble avoir généré des codes spécifiques dans des proportions semblables à celles de la distribution statistique originale.

Au total, environ 200 codes spécifiques d'activité principale ont été requis afin de procéder au codage des individus initialement inscrits à la modalité « autre à préciser ». Les codes utilisés le plus fréquemment et qui ne se trouvaient pas parmi la liste suggérée dans le formulaire électronique pourraient être ajoutés à cette liste. Il s'agit déjà d'un élément susceptible d'améliorer de futures collectes de données. La section suivante énonce justement quelques pistes d'amélioration à envisager.

1.4. Réflexion liminaire au sujet de pistes d'amélioration

Tel que nous venons de le mentionner, la présente expérience de codage permet d'envisager la modification des listes d'items pré-encodés fournis à même les questionnaires électroniques. Les codes utilisés fréquemment dans l'actuel exercice pourraient être considérés de même que certaines occurrences qui ne se retrouvent pas nécessairement dans les nomenclatures officielles. En fait, la liste suggérée automatiquement gagne à être relativement courte toute en reflétant bien les spécificités locales.

De façon plus générale, l'ensemble des questions dites ouvertes, soient celles disposant généralement d'une modalité « autre à préciser » en supplément d'une liste de choix préétablis, pourraient faire l'objet de nouvelles conventions de codification, prises en compte dans l'élaboration du manuel des spécifications techniques de collecte et de traitement. Par exemple, au départ, un code comme « 96 » pourrait être attribué aux réponses fournies à l'item « autre ». Au moment du traitement initial de ces informations, il pourrait être intéressant de ventiler et regrouper les valeurs résiduelles de manière à isoler les mentions qui ne pourront être codées à l'aide des nomenclatures. Ainsi, les réponses incohérentes ou inexploitables seraient codées « 97 »; les réponses de type « ne sait pas » se verraient attribuer un code « 98 » et les réponses ouvertes ne correspondant à aucun item de la nomenclature obtiendrait la valeur «99 ». Ces regroupements faciliteraient les révisions subséquentes et toute tentative de codage supplémentaire.

Enfin, étant donné les limites imposées par l'utilisation de questionnaires électroniques et des difficultés éventuelles de certains agents à saisir des informations fournies oralement par les interviewés, il serait possible de recommander aux agents d'inscrire les réponses aux questions ouvertes au stylo puis de les saisir en fin de passation du questionnaire. Toutefois, cela pose aussi certains risques d'erreurs ou d'omissions. Bref, peu importe les solutions envisagées, celles-ci devront faire l'objet de prétests.

Annexe

Annexe A : Programmation SPSS de codage du contenu de la modalité « autre à préciser » de la variable « Activité principale », RGPHAE, Sénégal, 2013

Le programme rédigé afin de convertir les expressions alphabétiques en codes numériques selon la nomenclature NAEMA-rév1 est disponible en téléchargement gratuit à cette adresse :

https://www.odsef.fss.ulaval.ca/sites/odsef.fss.ulaval.ca/files/uploads/Pgm_SPSS_B37_Codage.pdf

Ce programme peut être lu à l'aide d'un simple éditeur de texte. Le logiciel SPSS n'est pas requis afin de consulter les règles de codage créées.