Sciences de l'homme et de la Société | **ISSN: 2325-0798**
Vol : 07 Num: 29 | **Décembre,2018, pp 3-19**

Evaluation de la Qualité des Donnéesde l'Enquête Algérienne(PAPFAM 2002) : Application de la Concaténation comme Méthode de Traitement des Réponses Multiples

Mesli Redhouane, Saadi Rebah,Madani Salima

Département des Sciences Sociales, Université BLIDA2, redlimadz@yahoo.fr.
Département des Sciences Sociales, Université BLIDA2,rsaadi56@hotmail.com.
Département des Sciences Sociales, Université BLIDA2, selredamin@gmail.com.

Reçu le:03/10/2018 *Accepté le:15/11/ 2018*

Résumé :

L'exploitation et la mise en profit des données démographiques sont conditionnées par la qualité de ceux-là. Cette qualité peut faire défaut suite aux manquements à certaines règles afférant aux différentes phases et traitements. Parmi les erreurs pouvant entachées les données on trouve les erreurs de saisie et de codage, et qui peuvent être détectées grâce aux tests de champs et aux tests de vraisemblance.

L'article proposée c'est donné pour objectifs principal, l'évaluation de la qualité des données de l'enquête PAPFAM 2002, et plus précisément le fichier ménage, au travers de ces tests, en déployant une procédure nouvelle, reposant sur la concaténation des variables à contrôler. Les résultats de l'expertise montreront comme bien cet artifice est efficace dans la détection de ce type d'erreurs.

Mots clés : Test de champs, Test de cohérence interne, concaténation, Données manquantes.

Abstract:

The profitable utilization of population data is conditioned by the quality of this one. This quality can be altered by the bad application of certain rules relating to the different phases of collection and processing.Some of the errors that may affect the data include typing and coding errors, which can be detected by field tests and plausibility tests.

The main purpose of this paper is to evaluate the quality of the 2002 PAPFAM survey data and, more specifically, the household file, through these tests, by deploying a new procedure based on the concatenation of variables to control.The results of the survey will show that this device is effective in detecting such errors and will provide insight into the quality of the above-mentioned survey data.

Key words:field tests, plausibility tests, concatenation, missing values.

I INTRODUCTION

La qualité des données d'une opération de collecte dépendra dela préparation de cette opération, le dénombrement et le traitement des données récupérées.Le présent travail s'inscrit dans la troisième étape. En effet, cette recherche tente de mesurer la fiabilité des données de l'enquête PAPFAM 2002. Plus précisément, le but est de valider les données en question, au moyen de tests de contrôle et de cohérence internes.

Plusieurs aspects viennent justifier l'intérêt de cette recherche. Pour ne pas dire inexistants, les travaux sur les aspects ayant trait à la collecte des données en démographie sont très rares en Algérie.

KOUAOUCI (1992) a consacré quelques paragraphes aux sources traditionnelles des données sur la population dans une publication intitulée familles, femmes et contraception en Algérie (Kouaouci, 1992, p 12).

Sur le plan qualité des données, LAZAAR (1996) a publié un article sur les données démographiques algériennes. Il s'agit en fait d'une simple correction de la structure par âge et par sexe de la population, basée sur l'indice combiné des Nations Unies. En 1987, cet indice était de 16.39 (Lazaar, 1996, pp 23-29) attestant d'une bonne structure.

Les travaux de fin d'études en démographie, très mal diffusés en Algérie, sont le plus souvent axés sur des thèmes très particuliers et surtout initiés sur un nombre réduit d'observations. Ils sont, par conséquent, discutables sur le plan de la représentativité et de la signification statistique des résultats. Celui qui, en 2000, avait été consacré à l'état civil serait le plus important (Oucief .M, 2000).

Cette réalité n'est pas due à un manque d'intérêt, elle est surtout conséquente au problème d'accessibilité aux données. Les données des enquêtes ne sont mises à la disposition des chercheurs que très rarement. Aucun accès aux fichiers des opérations de collecte de données n'est autorisé. Cette recherche serait presque l'a seule exception à cette règle générale.

Plusieurs aspects relatifs à la collecte et aux résultats des opérations de collecte de données peuvent faire l'objet d'une problématique et justifient l'opportunité de conduire différentes recherches. Quelle qu'en soit l'optique d'intervention, ces recherches apporteraient des réponses aux interrogations qu'on est en droit de se poser face à la collecte et aux données démographiques algériennes.

Ce travail a pour but de traiter les données de l'enquête PAPFAM -2002- de manière critique, en se limitant à quelques aspects qui touchent à la

collecte et au traitement des données. Nous nous sommes fixés un objectif principal. Il s'agit de mesurer de manière grossière la qualité des données de l'enquête susmentionnée, en essayant de détecter les erreurs de collecte, de saisie et de traitement.

Nous examinons, une à une, toutes les variables du fichier informatique ménage (118 variables). Cette recherche a, par ailleurs, un objectif d'ordre secondaire. En effet, ce travail nous permettra de présenter et de testé encore une fois une procédure de contrôle de la qualité et de traitement des données (cette technique a été tester une première fois sur les données du recensement algérien de 1998) (Saadi, 2007).

Ce travail est divisé en deux parties. La première examinera la qualité des données du fichier en question sur la base de tests de validité des champs.

La deuxième partie est une application de la procédure proposée aux données examinées. Elle sera retenue comme:

- Outil de détection des erreurs de cohérence ;
- Et comme procédure plus générale de tabulation.

la confection des tests et la tabulation ont étés réalisés en utilisant le logiciel SAS 9.01.

II METHODES ET MATERIELS

2.1. Source de données

Le projet arabe pour la santé de la famille (PAPFAM), connu communément sous l'appellation enquête PAPFAM, est conçu pour mettre à

la disposition des décideurs des données et des indicateurs permettant de suivre et d'évaluer la politique de santé de la population. Ce projet est la continuité du projet PAPCHILD, également initié par la Ligue des Etats Arabes.

Quatrième pays s'engageant dans la réalisation de ce type d'enquête, L'Algérie est représentée par Le Ministère de la Santé, de la Population et de la Réforme Hospitalière. L'Office National des Statistiques a été chargé de l'exécution de l'enquête.

L'enquête compte quatre principaux objectifs :

- Produire des données nécessaires à la mise en œuvre, au suivi et à l'évaluation des politiques relatives à la santé de la population ;
- Améliorer et consolider les programmes de santé (la santé reproductive, en particulier), en enrichissant les bases de données existantes ;
- Produire des données en mesure d'initier des politiques régionales de population ;
- Disposer de données comparables entre les différents pays, pour évaluer les efforts à fournir en santé de population.

L'enquête compte des questionnaires communs : les caractéristiques du ménage et la santé reproductive et des questionnaires facultatifs. L'Algérie a optée pour les questionnaires facultatifs suivants:

- Les femmes en âge de ménopause ;
- La population des jeunes (15- 19 ans) ;

- Les personnes âgées.

L'enquête Algérienne s'est dotée ainsi de trois questionnaires, en plus du questionnaire ménage

La base de sondage est constituée de l'ensemble des ménages ordinaires et collectifs, issues du recensement général de la population et de l'habitat de 1998. La notion urbain/rural est adoptée comme critère de stratification.

On a opté pour un échantillon stratifié à deux degrés. Avant de procéder au tirage des ménages, on a classé la population des wilayas en quatre zones sanitaires (Centre, Est, Ouest et Sud).

Le tirage des unités primaires (510 districts) est effectué à probabilités égales, dans les différentes zones constituées.

Après mise à jour des districts-échantillon, les ménages à observer sont tirés également à probabilités égales (20 ménages par district-échantillon). A ce deuxième degré de tirage, on a procédé au tirage de deux échantillons. Un échantillon de base de 10200 ménages et un échantillon complémentaire de 10200 autres ménages en vue d'étudier la mortalité infantile (40 ménages par district-échantillon) (Ministère de la Santé, de la Population et de la Réformes Hospitalières, 2004).

Le premier échantillon est appelé 'échantillon principal'. Le deuxième est connu sous l'appellation 'échantillon élargi'.

2.2. Technique d'analyse : une alternative à la méthode de traitement des réponses multiples

La procédure proposée est décisive dans le traitement des réponses multiples, en définissant la totalité des profils (combinaisons de réponses). En effet, la fréquence d'une chaîne de caractères créée par concaténation définit tous les profils possibles. Cette procédure peut être généralisée à n'importe quel type de variables (unique, multiple, numérique ou caractère). On peut également recourir à cette technique, lorsque les données sont rapportées selon le rang de réponse.

En présentant l'ensemble des modalités de réponse de manière combinatoire, la technique que nous avons proposée permet de sélectionner des positions de lecture variées. Dans ces cas, on utilisera le nombre de réponses pour calculer les proportions. Cet artifice présente un avantage décisif, en offrant plusieurs possibilités de lecture. La lecture horizontale permet de repérer les spécificités rares et les profils dominants dans la population soumise à l'observation. Verticalement, on est en mesure de sélectionner une position de lecture. On peut également combiner les possibilités de lecture.

Lorsque les données sont entachées d'erreurs d'observation et de saisie, la technique proposée permet de détecter les enregistrements erronés.

En effet, l'artifice que nous avons proposé, est doublement intéressant. Il présente une alternative à la procédure de traitement des réponses multiples en application de nos jours. Il offre, par ailleurs, la possibilité d'engager plusieurs variables d'analyse pour dresser les profils des populations soumises à l'observation. Ce sera une possibilité d'analyse accessible à un large public. L'analyse multivariée proprement dite est peu commode pour les non spécialistes.

Cette méthodologie présente, toutefois, une limite importante dans la présentation des données, lorsqu'on est en présence de données volumineuses en nombre d'observations et de variables à concaténer. Cette limite pourrait être contrebalancée par un regroupement des profils les plus rares. On peut, en effet, limiter les combinaisons de réponses (dans ce type de tableaux) aux profils les plus dominants. Les profils rares pourraient être regroupés dans une seule catégorie (autres combinaisons de réponses, par exemple).

Cette technique, vu les possibilités de lecture, pourrait être engagée dans les plans de tabulation.

Elle présente, par ailleurs, une force décisive dans les travaux de contrôle de qualité des données, lorsque la vraisemblance entre variables est difficile à établir. L'expérience a montré que les profils rares correspondent très généralement à des erreurs de collecte et de codification des données. Dans ces conditions, une simple lecture horizontale des profils permet la mise en œuvre de tests de vraisemblance, utiles pour la correction des données brutes.

Cet artifice pourrait être appliqué sous n'importe quel logiciel statistique de traitement des données (SAS, SPSS, STATA, …).

III RESULTATS

3.1. **Les tests de champs**

Les données manquantes relatives au mois de naissance représentent 9% environ de l'information collectée. Ce constat fragilise la possibilité de

dater les évènements d'une manière satisfaisante, lorsqu'il s'agit de calculer des âges et des durées exacts (tableau n°1 en Annexe).

Les données manquantes concernant la variable 'raison de l'interruption de la scolarité' (tableau n°2) semblent poser de sérieux problèmes quand on parle de qualité de données (47%). Cette variable mérite d'être traitée pour réduire l'importance des données manquantes. Le rapport méthodologique n'a pas apporté une précision à cette question.

La variable 'prise du tabac' (tableau n° 3) est codée d'une manière différente, en comparant le questionnaire et le fichier de données. Sur le questionnaire, la modalité '8' identifie la catégorie des personnes ayant répondu 'ne sait pas'. Cette même population est reconnue par le code '4' sur le fichier de données. Dans les deux cas, cette catégorie de population ne devrait pas exister. La justification de ce constat est validée par le fait que ces personnes ne sont pas interrogées individuellement.

Selon les données de l'enquête, l'affection d'un handicap ne dépend pas de l'âge (0 - 93 ans révolus, dans notre cas). Or, la codification mise en œuvre limite cette variable à 95 ans. Autrement dit, on ne devrait pas rencontrer de personnes devenues handicapées au-delà de 95 ans. Les modalités 96, 98 et 99 identifient des populations spécifiques non libellés(tableau n° 4).

En réalité la question sur 'le type de réservoirs' ne devrait être posée qu'aux ménages ne disposant pas d'une source d'eau potable. Malheureusement, l'exclusion ne concerne que les ménages utilisant l'eau minérale en bouteilles.

La codification de la variable ‘traitement d’eau’ n’est pas continue (c’est le cas de toutes les variables du fichier ou presque). Cette manière d’opérer n’est pas très opérationnelle(Tableau n°5).

Deux valeurs (99,7 et 99,9 : (tableau 6)) de la variable ‘poids de l’enfant’ posent un problème. Elles n’appartiennent pas à l’intervalle de variation de la variable. Aucune précision n’est apportée à ce constat, ni sur le questionnaire ni sur le fichier de données. Elles sont certainement réservées à des situations particulières.

La variable ‘taille de l’enfant en cm’ varie entre 38 et 138,5 cm (tableau 7). Deux valeurs renvoient à des situations particulières. Aucune indication n’est apportée par l’enquête à ce constat.

Les valeurs de la variable ‘poids de la femme éligible’ s’étendent entre 30 et 155,0 kg. Deux valeurs abusives sortent du champ de variation possible. Elles ne sont pas, par ailleurs, libellées (999,7 et 999,9).

La taille des femmes éligibles (15-54 ans) s’étale sur un intervalle allant de 100 jusqu'à 198,8 cm. Dans ce cas, on observe également deux valeurs extrêmes (999,7 et 999,9), sans aucune précision.

3.2. Les tests de vraisemblance

Dans le tableau 52 (non présenté), la combinaison ‘. 1’ fait référence aux femmes n’ayant pas été enquêtées. Le rapport PAPFAM final ne donne pas le taux de non réponse en ce qui concerne l’enquête femme 15-49 ans.

Les données sont de bonne qualité puisque toutes les femmes enquêtées ont un ménage d’appartenance (deuxième ligne du tableau). Le taux de non réponse est de 3,5%.

Dans le tableau 51(non présenté)on devrait avoir une combinaison basée sur la valeur '1' pour les deux variables concaténées. La première fait référence au fichier femmes. La deuxième ce rapporte au fichier ménage

Le taux de non réponses est de 5% environ. Il s'agit de femmes non observées, dans le cadre de l'enquête femmes ménopausées.

Selon la même logique, l'enquête personnes âgées n'a pu être réalisée auprès de 390 personnes, traduisant ainsi un taux de non réponse relativement élevé (tableau 53)(non présenté).

Par concaténation, nous avons créé une chaîne de caractères à partir de six variables (le genre, l'âge, le numéro de ligne, la situation individuelle, le poids, la taille et le résultat des mesures). Deux hommes sont classés parmi la population des femmes au foyer. Les autres cas (tableau 17, (non présenté)) se rapportent à des femmes éligibles, pour lesquelles on n'a pas obtenu le poids, la taille et le résultat des mesures. Pour ces femmes, on devrait théoriquement coder les mesures selon les modalités discutées plus haut (code spécifique : 999,7, 999,9). Le nombre de données manquantes est important (plus de 8860 cas).

La variable âge intervient dans la formation de plusieurs catégories de personnes. On peut imaginer plusieurs tests de vraisemblance entre l'âge et autres variables (le statut par rapport à la scolarisation à l'enquête, le niveau d'instruction, la situation individuelle, certaines maladies chroniques,..).

En concaténant les variables âge, situation par rapport à la scolarisation, le niveau d'instruction et la situation individuelle on a remarqué quelques enregistrements qui méritent d'être vérifiés :

- Jeunes de moins de 20 ans en service national ;
- Jeunes en retraite ;
- Personnes âgées en scolarisation (tableau 18, annexe 1).

La technique proposée a permis de détecter quelques erreurs. En effet, le questionnaire adopté permet de confectionner des tests de vraisemblance entre les variables proposées (tableau numéro 54, (non présenté)). On peut signaler les erreurs relatives à l'aptitude à lire. Certaines personnes ont déclaré l'incapacité de lire alors qu'elles ont réussi un nombre d'années d'études. D'autres n'ont pas donné la raison d'interruption de la scolarité.

Les renvois utilisés par le questionnaire ont limité ce traitement au niveau primaire. En concaténant l'âge (en groupes), la situation individuelle et la situation dans l'emploi on saura que quelques enregistrements présentent certaines anomalies. On a interrogé une personne sur sa vie professionnelle alors qu'elle n'est pas concernée par le module occupation (Age en deçà de 16 ans). Le questionnaire réserve l'exploration de la vie active aux personnes âgées de 16 ans et plus. Certaines personnes ont été classées parmi la population des retraités avant l'âge requis. Le problème de classer les hommes parmi les femmes au foyer a été évoqué précédemment.

Dans le tableau 56 (non présenté) la quasi-totalité des erreurs relevées concerne l'affection par une maladie chronique (première et deuxième). En effet, on n'a pas précisé la situation de certaines personnes

tout en ignorant le reste des questions. Cette manière d'opérer suppose que ces personnes ne sont pas concernées, ce qui n'est pas vrai dans tous les cas (tableau 19, annexe1). Ce nombre de cas est reconnu par la modalité '9' en première ou en quatrième positions. Les modalités relatives à chaque variable sont précisées sur le questionnaire, en annexe.

L'axe handicap a été sommairement exploré. La sévérité de ces problèmes de santé a été posée de manière globale. On aurait obtenu des informations plus détaillées si l'on avait distingué la sévérité par handicap. La sévérité est exprimée en première position (tableau 57, (non présenté)). Les différents handicaps sont représentés dans les autres positions (note de bas de page).

Les enregistrements sélectionnés (tableau20, (non présenté)) se rapportent à des personnes dont l'état de santé n'a pas été précisé (code 9 en première position) alors qu'elles ont répondu aux questions relatives à la prise en charge et à l'âge à l'affection. Ces personnes appartiennent à différentes catégories d'âges.

Les combinaisons (tableau 58, (non présenté)) suspectes sont celles qui associent, aux appartements et aux maisons individuelles, un sol en terre battue (combinaisons : 1 2 et 2 1).

Les anomalies reprises dans le tableau 59 se rapportent :

- Aux logements dont le nombre de chambres à coucher n'a pas été déclaré ;
- Aux appartements dont le nombre de pièces semble être surestimé ;

- Aux habitations (maison individuelle, appartement, maison traditionnelle) dont ni le nombre de pièces ni le nombre de chambres à coucher n'ont été déclarés.

Les anomalies détectées (tableau 61, (non présenté)) se rapportent aux appartements disposant d'une bâche d'eau à l'intérieur du logement (combinaisons intégrant les modalités '2' en première position et '1' en troisième position). Ce nombre de cas est estimé à 11% environ de la totalité des ménages enquêtés.

L'erreur la plus singulière est représentée par la combinaison '2 3 1 2', signifiant l'existence d'un puits (deuxième position = 3) dans un appartement (première position = 2).

Les cinq premières combinaisons (tableau 62, (non présenté)) identifient des logements de type Villa/Maison individuelle utilisant des toilettes situées en dehors de l'habitation de manière collective (1ère position = 1, 2ème position = 2 et 4ème position = 1).

Les autres combinaisons se rapportent à des appartements utilisant des toilettes situées en dehors de l'habitation, à titre collectif (1ère position = 2, 2ème position = 2 et 4ème position = 1). L'erreur la plus évidente désigne les appartements sans toilettes.

Le nombre de cas présentés (tableau 63, (non présenté)) représente des habitations (villa, maison individuelle ou appartement) dépourvues de cuisine (position 2 =3) et des appartements où les cuisines sont situées à l'extérieur du logement.

Les combinaisons (tableau 64, (non présenté)) représentent un nombre de cas qui mérite d'être vérifié. En effet, ces anomalies possibles identifient des habitations (villa, maison individuelle ou appartement) utilisant du charbon ou du bois comme principale source de cuisson.

Dans le tableau (65)(non présenté)nous concaténons le poids, la taille de l'enfant, la position de la mesure (taille) et le résultat des mesures. Cette dernière variable introduit une imprécision. En effet, la question relative au résultat des mesures est ambiguë. Le questionnaire ne tranche pas la dépendance. Dépend-elle du poids, de la taille ou des deux à la fois ?

Les enregistrements sélectionnés sont pris par rapport aux valeurs critiques précédentes. Nous n'avons aucune réponse quant au sens de ces valeurs. Nous n'avons, par ailleurs, détecter aucune relation possible reliant la dernière question (résultat des mesures : position 4) à la première, la deuxième ou aux deux à la fois.

Les limites évoquées précédemment sont valables lorsqu'on construit une chaine de caractères à partir du poids, de la taille de la femme et des résultats des mesures. En effet, les valeurs abusives (poids et taille) restent sans précision même en intégrant le résultat de la mesure. Cette dernière variable ajoute une confusion supplémentaire.

IV CONCLUSIONS

En somme, les résultats de cette expertise plaident en faveur d'une bonne qualité des données étant données la rareté d'erreurs conséquentes. Les anomalies signalées sont peu nombreuses pour porter préjudice à la

qualité des données. Encore une fois, l'artifice proposé a fait ces preuves, en s'avérant indispensable dans l'application des tests de vraisemblance.

V REFERENCES

[1] KOUAOUCI A, Femme, famille et contraception ; FNUAP, CENEAP, ALGER : 1992.

[2] LAZAAR A, "Qualité des données démographiques algériennes". In EL TAWASSOL, n° 03, 1996, Université de BADJI MOKHTAR, ANNABA : 1996.

[3] OUCIEF M, Les données de l'état civil en Algérie, Mémoire de Magister en Démographie, Université SAAD DAHLEB, BLIDA : 2000.

[4] Ministère de la Santé, de la Population et de la Réforme Hospitalière, ONS, Enquête Algérienne sur la santé de la famille. Rapport Principal, ALGER : 2004.

[5] SAADI REBAH, La qualité des données démographiques en Algérie : le recensement de 1998, Thèse de doctorat en Démographie, Université SAAD DAHLEB, BLIDA : 2007.