Sociologie et sociétés

Note de recherche
Les données manquantes multiples : un problème. Une solution pour une recherche sur les communautés religieuses
Reaserch Note - Multiple Missing Data: a Problem
A Solution in the Context of Reaserch on Religious Communities

Lorraine DUCHESNE et Yves LEPAGE

La construction des données
Volume 25, numéro 2, automne 1993

URI : https://id.erudit.org/iderudit/001209ar
DOI : https://doi.org/10.7202/001209ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0038-030X (imprimé)
1492-1375 (numérique)

Découvrir la revue

Citer cette note

DUCHESNE, L. & LEPAGE, Y. (1993). Note de recherche : les données manquantes multiples : un problème. Une solution pour une recherche sur les communautés religieuses. *Sociologie et sociétés*, *25* (2), 47–51. https://doi.org/10.7202/001209ar

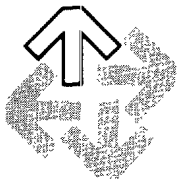
Résumé de l'article

Cet article propose une solution au problème lié aux données manquantes, dans le cadre d'une étude sur les communautés religieuses. Parmi les différentes approches présentées, la solution retenue préconise la réduction de la population visée. La portée des conclusions de cette étude est discutée en fonction de l'approche retenue.

Sociologie et sociétés, vol. XXV, nº 2, automne 1993

Les données manquantes multiples : un problème

Une solution pour une recherche sur les communautés religieuses



LORRAINE DUCHESNE et YVES LEPAGE

L'analyse quantitative pose à tout chercheur plusieurs problèmes complexes lorsque les données sont incomplètes. Une enquête effectuée par questionnaire administré soit au téléphone ou expédié par la poste exige que les chercheurs prennent toutes les mesures nécessaires pour obtenir les données les plus complètes possibles sur l'échantillon étudié. On sait qu'il est difficile de recueillir certaines informations sur des questions très personnelles comme le statut matrimonial, le revenu, etc. Plusieurs personnes refusent de répondre à certaines questions qu'elles jugent trop confidentielles. D'autres informations se perdent, pour des raisons techniques, au moment de la codification ou de l'informatisation des données.

Des statisticiens et des sociologues se sont penchés sur les problèmes liés aux données manquantes : dans les années soixante-dix, Goodman (1974) tenta de résoudre certains aspects de ce problème. Depuis, plusieurs auteurs dont Little & Rubin (1987) et Little & Su (1989) ont écrit sur le sujet.

Notre propos est de traiter le problème des données manquantes dans le cadre d'une recherche où les données sont tirées de documents d'archives. Dans ce type de recherche, les données concernent souvent des personnes décédées et elles représentent fréquemment les seules vestiges de notre histoire. D'une part, le chercheur ne dispose d'aucun moyen de retrouver ou de compléter les données manquantes par une autre enquête, un autre questionnaire ou une autre étude. D'autre part, le chercheur a aussi le souci de tirer toutes les informations possibles des données rassemblées, même si elles sont incomplètes.

La recherche dont il est question ici porte sur le cheminement de carrière des religieuses de 1920 à 1971[1], à partir des informations recueillies dans le cadre d'une vaste étude sur les communautés religieuses[2]. Les données de cette recherche sont tirées des archives de 24 communautés de femmes choisies au hasard sur le territoire québécois et touchent 3 700 religieuses[3]. L'échantillon a été construit à partir de la taille et de l'activité principale des communautés. Nous voulions savoir quelle est l'influence de

1. Pour les résultats de cette recherche, voir Duchesne (1992).
2. Pour un aperçu de cette recherche, voir Juteau et Laurin (1986).
3. Pour les détails de l'échantillon, voir Laurin, Juteau et Duchesne (1991).

l'origine sociale, mesurée par la profession du père, de la scolarité, du premier et du troisième emploi sur la trajectoire professionnelle des religieuses.

Les communautés religieuses conservent certaines informations personnelles concernant leurs sujets : date de naissance, date des vœux, profession du père, scolarité au moment de leur entrée au couvent, postes occupés durant leur carrière, etc. La façon de conserver ces données varie selon le temps, la taille et les types de communautés. Au début du siècle, les informations sont souvent dispersées dans différents documents qui n'ont aucun souci scientifique : registres d'entrée, nécrologies[4], listes d'obédiences, etc. Aussi les informations concernant la scolarité et la profession du père sont très souvent absentes. Toutefois, à partir des années trente ou quarante, la plupart des communautés créent un système de dossiers personnels. Ces dossiers regroupent les informations que les communautés jugent utiles de conserver sur chacun de leurs sujets. Certains de ces dossiers sont toutefois incomplets; la profession du père et le niveau d'études atteint avant l'entrée n'y sont quelquefois pas notés.

Avant d'effectuer différentes analyses statistiques pour étudier le cheminement de carrière des religieuses, il fallait donc résoudre le problème posé par les valeurs manquantes. La première étape a été d'effectuer une étude exhaustive des valeurs manquantes pour l'ensemble des périodes de temps. Combien y a-t-il de valeurs manquantes pour chacune des variables étudiées et comment se distribuent-elles selon les activités principales des communautés, selon leur taille et selon les périodes ? Il faut préciser qu'il n'y a aucune donnée manquante pour les variables qui décrivent les communautés : activité principale, taille et période puisque ces variables étaient connues avant de commencer la recherche et ont servi à construire l'échantillon. Les données manquantes portent sur les informations décrivant la vie de chaque religieuse. Les archives des communautés contenaient presque toutes les informations sur les dates importantes qui marquent la vie de leurs sujets : date de naissance, date des vœux temporaires et perpétuels, date du décès ou de sortie, etc. Les emplois que les religieuses occupaient au cours des différentes décennies étaient aussi généralement bien conservés; il n'y avait que 3,5 % de valeurs manquantes pour le premier emploi et 1,4 % pour le troisième emploi occupé par les religieuses. Ce sont les informations plus personnelles, comme la profession du père et la scolarité des religieuses avant leur entrée en communauté, qui ont posé le plus de problèmes. Pour l'ensemble des périodes, seulement 81,5 % des informations concernant la profession du père et 65,3 % de celles décrivant le niveau de scolarité atteint avant l'entrée étaient connues. Or, rien ne nous autorise à croire que les religieuses pour lesquelles nous n'avions aucune information étaient plus ou moins scolarisées que les autres religieuses.

Le chercheur se trouve alors devant un problème qu'il doit solutionner avant de poursuivre l'analyse. Doit-il renoncer à inclure dans son étude les informations concernant la scolarité des religieuses, variable pour laquelle nous ne connaissons l'information que dans les deux tiers des cas environ ? Cependant étudier le cheminement de carrière des religieuses sans tenir compte de leur scolarité équivaudrait à amputer notre objet de ce qu'il possède de plus pertinent du point de vue sociologique, à savoir les bases sur lesquelles les postes sont attribués. Nous serions obligés d'éviter les questions les plus intéressantes. Les communautés sont-elles des organisations méritocratiques ? Est-ce que les communautés tiennent compte de la formation académique dans les nominations ? Il aurait toujours été possible d'étudier le cheminement de carrière selon les différentes activités des communautés et, selon les périodes et d'élaborer une analyse en ne tenant pas compte de la scolarité. Mais, du point de vue sociologique, nous aurions été insatisfaits de ce choix. Nous croyions que la scolarité des religieuses est un facteur explicatif de la trajectoire professionnelle de ces femmes trop important pour le laisser dans l'ombre. Nous devions, toutefois, respecter certains principes statistiques. Voici comment nous avons procédé pour minimiser l'effet des valeurs manquantes dans notre étude.

---

4. Une nécrologie est une notice biographique écrite à la suite du décès d'une religieuse.

Plusieurs raisons expliquent le fait qu'il y ait beaucoup de données manquantes : le refus de certaines communautés de transmettre les informations, le fait que les informations soient trop éloignées dans le temps et n'aient pas été conservées, l'inutilité, pour certains types de communautés, de conserver ces informations, etc. L'analyse détaillée des valeurs manquantes a permis de cerner des catégories de variables contenant une grande proportion de données manquantes. Nous avons éliminé ces catégories de notre étude.

Une très grande communauté responsable de services sociaux-hospitaliers a refusé de fournir toute information sur la scolarité de ses sujets; nous avons retiré cette communauté de notre analyse. Nos conclusions ne s'appliquent donc pas à elle.

Il y a un lien entre notre connaissance de la scolarité des recrues et la période d'entrée. Il est conforme à ce qu'on s'attendait; plus on avance dans le temps, plus la proportion de valeurs manquantes diminue. Il n'est pas étonnant que les communautés religieuses n'aient pas conservé dans leurs archives les données concernant la scolarité de filles qui entraient dans leur communauté, surtout à la fin du XIX^{e} siècle et au début du XX^{e} siècle. Les données du recensement canadien ne sont d'ailleurs pas très explicites sur le sujet; en effet, jusqu'en 1931, les catégories de l'information sur le niveau de scolarité atteint par les Canadiens sont : « sait lire/ne sait pas lire, sait écrire/ne sait pas écrire ». Il se peut aussi que les communautés aient gardé à jour des informations sur la scolarité de leurs sujets à l'époque, mais que celles-ci n'aient pas été conservées; elles n'existeraient plus. Nous avons donc retiré de notre étude les religieuses admises avant 1922. À partir de cette date, les informations sur la scolarité et la profession du père sont suffisamment complètes pour que notre analyse soit valide.

De même, les petites communautés de moins de 500 sujets ont conservé très peu d'informations sur la scolarité de leurs recrues. Il est probable qu'à l'intérieur de ces communautés à peu près toutes les religieuses se connaissent; elles n'ont pas à conserver dans les archives des renseignements connus. Nous avons exclu les petites communautés de notre analyse.

Les informations sont conservées de façon inégale selon les types de communauté. Les religieuses vouées à la protection — qui œuvrent dans les prisons ou les maisons pour jeunes filles délinquantes — ne constituent qu'une seule communauté. Lors de la cueillette de données, nous avons constaté que plusieurs informations étaient absentes des archives parce que des dossiers avaient été transférés dans une autre province canadienne pour des raisons d'ordre administratif. Nous avons dû exclure cette communauté. De même, les communautés de service au clergé n'ont gardé que peu d'informations sur la scolarité et la profession du père de leurs sujets. Ces religieuses sont responsables de l'entretien matériel et du service domestique dans les presbytères, les collèges, les évêchés, etc. Il est possible que les jeunes filles qui sont entrées dans ces communautés aient été moins instruites que les recrues des autres types de communautés.

Après avoir étudié tous les tableaux concernant les données manquantes, nous étions en mesure de découper dans notre échantillon un sous-échantillon de religieuses au sujet desquelles nous avions suffisamment de données pour que notre étude soit fiable. Après avoir retiré toutes les catégories de religieuses au sujet desquelles nous n'avions pas suffisamment d'informations, il nous reste quand même un échantillon de taille 1 482; notre analyse porte sur une population représentant 70 % de toutes les religieuses entrées dans les communautés au Québec, de 1922 à 1971. Évidemment, les résultats de notre étude ne s'appliqueront qu'aux religieuses présentes dans cette sous-population. Nous ne saurons jamais comment se fait le cheminement de la carrière dans les petites communautés, ni quel est le cheminement professionnel des religieuses entrées avant 1922, etc. Les résultats de notre étude sont certes partiels, mais il est préférable de ne considérer qu'une partie de la population religieuse et d'inclure dans notre analyse la profession du père, laquelle est un indice important du milieu socio-économique des futures religieuses, de ce qu'elles ont hérité de leur famille et la scolarité, laquelle représente ce que les religieuses apportent comme bagage personnel à leur entrée.

Même en retirant de notre analyse certaines périodes ou certains types de communauté, au sujet desquels il y avait trop de données manquantes, il reste des valeurs manquantes dans notre sous-échantillon; il manque 12,8 % des informations sur la profession du père et 20,8 % de celles sur la scolarité.

Nous avons étudié comment se distribuent les valeurs manquantes dans notre sous-échantillon. Nous avons constaté, à l'aide de la mesure d'association du khi-deux, que les valeurs manquantes pour la scolarité n'ont aucun lien avec la profession du père, ni avec l'activité principale des communautés, ni avec le premier emploi et ni avec le troisième emploi. Le niveau de signification tient compte évidemment du nombre de tests statistiques effectués. Les données manquantes sur l'occupation du père n'ont aucun lien avec l'activité principale des communautés, le premier emploi, le troisième emploi, ou avec l'occupation du père. Nous pouvons affirmer que les valeurs manquantes se répartissent au hasard dans le sous-échantillon [5].

Parce qu'il en est ainsi, nous pouvons analyser directement les données; cette méthode a été décrite par Little & Rubin (1989-1990)[6].

Nous n'aurions pas pu utiliser cette méthode si nous n'avions pas au préalable découpé notre échantillon global en un sous-échantillon parce que, dans l'échantillon global, les valeurs manquantes n'étaient pas dues au hasard; faire l'analyse directe des données dans ce cas aurait probablement créé des biais trop importants pour que les résultats soient statistiquement valables.

Nous avons utilisé les modèles loglinéaires pour décomposer l'effet de deux variables explicatives sur une variable expliquée. Winship & Mare (1989, p. 342) ont discuté de l'utilisation de ces modèles quand on omettait de tenir compte des valeurs manquantes, quand celles-ci se distribuaient au hasard .

En consultant les nombreux articles écrits sur les données incomplètes, nous constatons l'existence de plusieurs méthodes pour minimiser les biais qui peuvent s'introduire dans l'interprétation de ces données. Nous avons choisi de réduire notre analyse à un sous-échantillon dans lequel les valeurs manquantes se distribuent au hasard et la population visée par l'étude est donc aussi réduite. C'était une voie possible. Nous aurions pu utiliser d'autres méthodes comme la pondération, l'imputation multiple, etc. Il est important de constater qu'il existe plusieurs solutions aux problèmes des données manquantes et que chaque solution offre ses avantages et ses limites, qui doivent toujours être précisés.

Tous les choix doivent être faits pour répondre le mieux possible aux objectifs sociologiques de la recherche. Le rejet d'une variable importante, comme la scolarité dans notre étude, ne doit être envisagé qu'en dernier recours; cela est particulièrement important quand les données sont tirées de documents d'archives : elles représentent souvent les seules traces de notre passé.

Lorraine DUCHESNE
Institut national d'études démographiques
27, rue du Commandeur
75675 Paris – cedex 14
France

Yves LEPAGE
Département de mathématiques
et de statistique
Université de Montréal
C.P. 6128, Succ. « A »
Montréal (Québec)
Canada H3C 3J7

5. Ce que Little & Rubin (1989-1990, p. 297) nomment « *missing at random* » (MAR).

6. Little & Rubin (1989-1990, p. 295) nomment cette méthode « *the available-case analysis* ». Voici comment ces auteurs la décrivent : « *Available-case analysis replaces components in complete-data statistics by corresponding quantities calculable from available data. For example, univariate statistics such as means and variances are calculated using the set of cases for which each variable is observed, and covariances (or correlations) are computed using the set of cases for which both variables in the pair are observed.* »

RÉSUMÉ

Cet article propose une solution au problème lié aux données manquantes, dans le cadre d'une étude sur les communautés religieuses. Parmi les différentes approches présentées, la solution retenue préconise la réduction de la population visée. La portée des conclusions de cette étude est discutée en fonction de l'approche retenue.

SUMMARY

This paper proposes a solution to the problem of missing data, in the context of a study carried out on religious communities. From among the different approaches presented, the best solution is the reduction of research population size. The scope of this study's conclusions in relation to this approach is discussed.

RESUMEN

Este artículo propone une solución al problema relativo a los datos ausentes, en el marco de un estudio sobre las comunidades religiosas. Entre los diferentes puntos de vista presentados, la solución que se retiene preconiza la reducción de la población aludida. El alcance de las conclusiones de este estudio es discutido en función del punto de vista retenido

BIBLIOGRAPHIE

GOODMAN, Leo A. (1974) « The Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part 1 : A Modified Path Analysis Approach », *American Journal of Sociology* 79, pp. 1179-1259.

DUCHESNE, Lorraine (1992), *Le cheminement de carrière dans les communautés religieuses de femmes, au Québec, de 1920 à 1971*, thèse de doctorat, Département de sociologie, Université de Montréal.

JUTEAU, Danielle et Nicole LAURIN (1986), « Les communautés religieuses de femmes au Québec, Une recherche en cours » dans *Question de culture*, 9, Institut québécois de recherche sur la culture, pp. 145-157.

LAURIN, Nicole, JUTEAU, Danielle et Lorraine DUCHESNE (1991), *À la recherche d'un monde oublié. Les communautés religieuses de femmes au Québec de 1900 à 1971*, Montréal, Éditions Le Jour.

LITTLE, Roderick J. A. et Donald B. RUBIN (1987), *Statistical Analysis with Missing Data*, New York, John Wiley & Sons.

LITTLE, Roderick J. A. et Donald B. RUBIN (novembre 1989/févier 1990), « The Analysis of Social Science Data with Missing Values », *Sociological Methods and Research*, vol. 18, n^{os} 2 et 3, Sage Publications Inc, pp. 292-326.

LITTLE, Roderick J. A. et H. L. SU (1989) « Item Nonresponse in Panel Survey", dans D. Kasprzyk, G. Duncan et M. P. Singh (éds), *Panel Survey*, New York, John Wiley & Sons, pp. 400-425.

WINSHIP, Christopher et Robert D. MARE (1989), « Loglinear Models with Missing Data : A Latent Approach », *Sociological Methodology*, vol. 19, Basil Blackwell Ltd, Oxford & Cambridge MA for the American Sociological Association, pp. 331-367.