Sociologie

N° 2, vol. 9 | 2018
Théories et Méthodes

Échantillonner des populations rares

Une expérimentation du *Respondent Driven Sampling* en milieu musical

Sample rare populations. A Respondent Driven Sampling experimentation among musicians

PIERRE BATAILLE, MARC PERRENOUD AND KAREN BRÄNDLE

Abstracts

L'échantillonnage de populations dont il est impossible de connaître *a priori* les contours avec précision constitue un défit majeur en sciences sociales. Certaines solutions ont néanmoins été proposées pour tenter de produire des données fiables dans ce type de cas. Parmi les procédures d'échantillonnage alternatives, c'est indubitablement la méthode du *Respondant Driven Sampling* (RDS) – « échantillonnage guidé par les répondants » –, apparue à la fin des années 1990 qui a connu la plus forte audience. L'ambition de cet article sera de présenter les points forts et les faiblesses de cette procédure, à travers un retour sur expérience dans le cadre d'une recherche sur des « musicien-ne-s ordinaires » suisses romand-e-s. Il s'agira premièrement de présenter l'outillage théorique, méthodologique et statistique qui structure cette procédure d'échantillonnage. Dans un deuxième temps, on montrera comment le RDS permet de construire des données originales et de qualité par rapport à d'autres procédures plus souvent utilisées – notamment l'enquête « en ligne » – pour des recherches sur des populations « rares ». Enfin, on reviendra sur les limites de l'échantillon recruté, du point de vue de la structuration du réseau complet des collaborations musicales sur une année de référence.

Sample rare populations. A Respondent Driven Sampling experimentation among musicians

Sampling populations without clear administrative boundaries is a major issue in social sciences. Some solutions have, however, been proposed in order to gather reliable data on this kind of populations. The Respondant Driven Sampling (RDS) undoubtedly figures as the most widespread of alternative sampling techniques. The main goal of this article is to discuss the strengths and the weaknesses of this procedure through our field work experience on Swiss "ordinary musicians." The theoretical, methodological and statistical tools involved in this kind of procedure will be presented first. Secondly, the article turns to discuss the originality and robustness of data gathered by the RDS method as compared with other more common procedure

(like online surveys) in research on “rare” but not necessarily “hard to reach” populations. Finally, the article evaluates the limits of the sample by analysing the structuration of the whole network of musical collaborations.

Index terms

Mots-clés : échantillonnage, réseaux, populations rares, musiciens, Suisse romande
Keywords : Switzerland, sampling, network, rare populations, musicians

Full text

1 L’échantillonnage de populations dont il est impossible de connaître *a priori* les contours avec précision constitue un défi majeur en sciences sociales. Ces populations « difficiles à atteindre » du fait de certaines de leurs caractéristiques potentiellement stigmatisantes (comme les usagers de drogues, les personnes LGBTQ…), ou plus simplement « rares » car trop spécifiques pour être saisies par les outils statistiques officiels sont en effet au cœur de nombreuses enquêtes. Les chercheur-e-s sont dans les faits bien souvent obligé-e-s de « bricoler » en croisant sources secondaires et données de première main au risque d’abandonner tout espoir de voir leurs conclusions « prétendre à l’assise du raisonnement probabiliste » (Schiltz, 2005, p. 30).

2 Certaines solutions ont néanmoins été proposées pour tenter de produire des données fiables sur ce type de population dans le cadre d’un raisonnement inférentiel[1]. Le champ de recherche autour des méthodes d’échantillonnage « adaptatives » (Thompson & Collins, 2002), où l’on découvre les contours de la population visée au cours du processus de recrutement des enquêté-e-s, a longtemps été relativement segmenté (Spreen, 1992). Les approches par dépistage de liens (*link tracing*), plus ou moins directement dérivées de l’échantillonnage par « boule de neige » (Goodman, 1961)[2], ont fait l’objet d’une attention particulière en comparaison d’autres propositions restées plus confidentielles, comme l’échantillonnage basé sur la proximité géographique ou la construction de cluster (Thompson & Collins, 2002). Parmi les méthodes utilisant le dépistage de liens, c’est indubitablement la méthode du *Respondent Driven Sampling* (RDS) – « échantillonnage guidé par les répondants » –, apparue à la fin des années 1990 (Heckathorn, 1997), qui a connu la plus forte audience. En 2013, on recensait ainsi dans la littérature disponible près de 460 recherches menées dans 69 pays ayant utilisé le procédé du RDS, la plupart du temps dans le cadre d’enquêtes menées en épidémiologie autour de problématiques de santé publique (White *et al.*, 2015).

3 Ce succès s’explique par la simplicité de la méthode mise en place par Douglas Heckathorn, mais également par la formalisation d’outils statistiques destinés à réduire les potentiels effets des biais de sélection sur les estimations de la structure de la population étudiée, alors que la plupart des autres méthodes d’échantillonnage par dépistage de liens ne permettent pas de prétendre à ce niveau de fiabilité (Van Meter, 1990). Les récents travaux menés sur les résultats obtenus par les méthodes d’estimations forgées par D. Heckatorn et ses collaborateurs montrent que ces estimations restent néanmoins bien souvent sujet à caution du fait des divers postulats sur lesquels elles reposent (Léon *et al.*, 2016).

4 L’objectif de cet article est double. À partir des résultats d’une recherche menée sur les musicien-ne-s « ordinaires » suisses romands[3], il vise premièrement à montrer combien la méthode du RDS gagnerait à être plus largement diffusée chez les sociologues, tant elle permet de recueillir un matériau empirique original et riche sur

des populations « rares », c'est-à-dire sous-représentées dans les enquêtes statistiques nationales mais pas nécessairement stigmatisées. À partir des données recueillies grâce à notre dispositif d'enquête original, il ambitionne également de proposer une évaluation de cette technique de recrutement. Alors que les travaux critiques quant aux résultats obtenus grâce au RDS sont souvent fondés sur des simulations informatiques, nos analyses permettent de donner une appréciation empirique des potentiels biais induits par ce processus d'échantillonnage.

5 Dans un premier temps, nous détaillerons les principales caractéristiques de la méthode du RDS. Ensuite, nous montrerons comment nous l'avons concrètement mise en place dans le cadre de notre recherche, en essayant de tenir compte des critiques déjà formulées dans d'autres travaux. Puis, nous comparerons les résultats obtenus *via* le RDS et les données récoltées sur la même population au moyen d'une technique d'échantillonnage plus souple. Enfin, nous analyserons la qualité du recrutement RDS à l'aune de la structuration de la population visée.

Le *Respondent Driven Sampling* : principes, limites et perspectives d'amélioration

Une méthode d'échantillonnage et d'estimation

6 La technique du RDS est basée sur une idée assez simple et bien connue depuis les travaux de Stuart Milgram sur le phénomène du « *small world* » (Milgram, 1967) : tous les individus d'un même groupe social sont liés entre eux par un nombre limité de liens[4]. Autrement dit, quel que soit le point de départ des chaînes de relations, on peut potentiellement atteindre n'importe quel individu du groupe en question au bout de *x* mises en relation successives.

7 Fort de ce constat, dans le premier article qu'il consacre au RDS en 1997, D. Heckathorn défend donc l'idée suivante : lorsque la sélection des informateurs et informatrices privilégié-e-s choisi-e-s pour nouer des contacts dans la population cible est non-aléatoire et arbitraire, si l'on multiplie les vagues de recrutement à partir de ce petit groupe initial, la sélection des personnes intégrées dans l'échantillon dépendra de moins en moins des graines initiales et va tendre à devenir, tout de même, aléatoire. La méthode exposée dans cet article séminal et dans les principaux travaux qui ont été depuis consacrés à son développement (Gile & Handcock, 2010 ; Salganik & Heckathorn, 2004 ; Volz & Heckathorn, 2008 ; Wejnert, 2009) recouvrent en fait deux aspects bien distincts : la procédure de récolte concrète des données, d'une part ; les traitements statistiques particuliers à appliquer à ces données pour contrecarrer les biais de recrutement, d'autre part.

8 Concernant la procédure d'échantillonnage, le but du protocole RDS est de favoriser la multiplication des vagues de recrutement – puisque c'est cet encastrement des recrutements successifs qui est *in fine* la clé d'un échantillonnage le moins biaisé possible. Le protocole commence donc par la sélection d'un nombre (limité, généralement moins de dix) d'informateurs ou informatrices initiaux. Ces *primo* participant-e-s – appelé-e-s « graines » – doivent présenter des profils aussi variés que possible du point de vue des variables d'intérêts de l'enquête, de manière à diversifier les points d'entrée dans la population cible. Ces graines sont ensuite invitées à recruter

deux ou trois de leurs contacts personnels correspondant aux critères de définition de la population. Pour ce faire, elles sont dotées de « coupons » que leurs contacts doivent retourner aux enquêteur-ice-s lors de leur interview – de manière à s'assurer qu'il y a bien eu un échange entre les individus recruteurs et les futurs recruté-e-s. Une autre recommandation pour s'assurer la participation active des recruteurs et recruteuses est de fournir une rétribution matérielle ou symbolique sous condition que les contacts participent effectivement à l'enquête et retournent les fameux « coupons » aux enquêt-eur-rice-s.

9 Le même procédé est répété à chaque vague avec les personnes nouvellement contactées, jusqu'à atteindre l'*equilibrium*, c'est à dire le stade où la structure globale de l'échantillon recruté se stabilise et ne varie plus d'une vague à l'autre sous l'angle des variables socio-démographiques classiques (âge, sexe, origine sociale, etc.) et des variables plus spécifiques à la population analysée. Cette procédure d'échantillonnage « dirigée par les répondants » fait le pari qu'en favorisant une implication des enquêté-e-s et la mobilisation de leurs réseaux personnels, on peut entretenir une dynamique de recrutement sur un nombre important de vagues.

10 Les expérimentations pratiques de ce protocole de recrutement et les problèmes rencontrés ont amené les chercheu-r-se-s investi-e-s dans ce champ à proposer un ensemble de techniques statistiques de pondération visant à limiter les effets des biais de sélection. Il ne s'agira pas ici de commenter dans le détail les formules derrières les estimateurs améliorés, fruits de ce travail de modélisation[5], mais de présenter brièvement les principes sur lesquels se basent les outils statistiques propres au RDS et les problèmes qu'ils sont censés permettre de contourner. Nous aborderons la présentation de ces « estimateurs » dans la sous-partie suivante, portant sur les débats qui ont animé la communauté des chercheurs et chercheuses s'intéressant au RDS.

Une approche en débat

11 De manière très compréhensible, c'est la question du biais induit par les propriétés sociales particulières des personnes recruteuses qui est généralement vu comme le principal problème des échantillonnages de type RDS. Notamment, il est très vite apparu que, dans un processus de recrutement par l'intermédiaire de pairs, les individu ayant des réseaux personnels plus réduits (ceux et celles infecté-e-s à un stade relativement avancé du VIH dans une enquête sur les usag-er-ère-s de drogues injectables par exemple) ont potentiellement moins de chance d'être connectés à de potentiel-le-s recruteu-r-se-s – et ont donc moins de chances d'être recrutés.

12 Partant de cette idée, une première série d'estimateurs pondérés par la taille du réseau personnel des recruté-e-s dans la population cible a été proposée : les estimateurs dits « RDS I » (Salganik & Heckathorn, 2004) et « RDS II » (Volz & Heckathorn, 2008). L'estimateur RDS II pouvant être considéré comme une amélioration du premier, c'est celui qui est aujourd'hui le plus souvent utilisé. S'il permet théoriquement d'améliorer la qualité des estimations effectuées sur l'échantillon RDS, sa validité est fondée sur d'importants présupposés difficilement vérifiables la plupart du temps. L'estimateur RDS II est considéré logiquement comme valable si les six propriétés suivantes sont vérifiées :

13 1. Le lien entre recruteu-r-se-s et recruté-e-s est réciproque (*i.e.* le réseau de recrutement n'est pas dirigé).

14 2. Tous les individus de la population cible sont d'une manière ou d'une autre liés entre eux (*i.e.* le réseau que forme la population cible n'a qu'une seule composante).

3. Le recrutement des répondant-e-s s'effectue avec remplacement (*i.e.* même une fois