

Statistique et société

Mai 2023

Volume 11, Numéro 1

Sondages et enquêtes

Sommaire

Statistique et société

Volume 11, Numéro 1

- 7 Éditorial**
Emmanuel DIDIER
Rédacteur en chef de Statistique et société
-
- 9 Introduction au dossier « Sondages et enquêtes »**
Jean CHICHE
Chercheur émérite au Cevipof (Sciences PO, CNRS)
- 11 Comment mesurer ? L'exemple de la rénovation énergétique des logements**
Thomas MERLY-ALPA
Institut National des Études Démographiques (Ined), Chef du Service des Enquêtes et Sondages
- 23 La qualité des enquêteurs importe dans la non-réponse partielle des répondants qui sont moins motivés par l'enquête**
Thierry BORNAND
Baptiste FERAUD
Institut Wallon de l'Évaluation, de la Prospective et de la Statistique (IWEPS), Belgique
- 39 Abandons dans une enquête sur internet : l'exemple de l'inclusion dans la cohorte Coset-MSA**
Noémie SOULLIER
Santé publique France, Chargée de projet organisation et suivi des enquêtes
Hugo ROGIE
Santé publique France, Stagiaire
Guilhem DESCHAMPS
Santé publique France, Data scientist
Jean-Luc MARCHAND
Santé publique France, Chef de projet COSET-Indépendants
Béatrice GEOFFROY-PEREZ
Santé publique France, Cheffe de projet COSET-MSA

Sommaire

Statistique et société

Volume 11, Numéro 1

- 53 **L'appariement statistique des bases de données SILC, HBS et HFCS : aspects méthodologiques et applications à l'étude de la pauvreté et des inégalités**
Huyen TRAN
Guillaume OSIER
STATEC Luxembourg et Luxembourg Income Study (LIS)
-
- 65 **Un entretien avec Marc Hallin**
Marc HALLIN
Jean-Jacques DROESBEKE
Université libre de Bruxelles
- 75 **Comment être élu à tous les coups ? Petit guide mathématique des modes de scrutin de Jean-Baptiste AUBIN et Antoine ROLLAND (2022)**
Antoine HOULOU-GARCIA
Chercheur associé au CESPRA (Centre d'études sociologiques et politiques Raymond-Aron)
- 79 **Adolphe Quetelet**
Passeur d'idées
de Jean-Jacques DROESBEKE (2021)
Emmanuel DIDIER
Directeur de recherche au CNRS, Membre du Centre Maurice-Halbwachs (ENS et EHESS)

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique. Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Directrice de publication

Anne Philippe, Présidente de la SFdS

Rédaction

Rédacteur en chef : Emmanuel Didier, CNRS, France

Rédacteurs en chef adjoints :

Thomas Amossé, Conservatoire national des arts et métiers, France

Jean Chiche, Institut d'études politiques de Paris, France

Quentin Dufour, Ecole Normale Supérieure, France

Jean-Jacques Dreesbeke, Université libre de Bruxelles, Belgique

Chloé Friguët, Université Bretagne-Sud, France

Pauline Hervois, Sorbonne Université & Muséum national d'histoire naturelle, France

Olivier Martin, Université Paris Cité, France

Antoine Rolland, Université Lyon 2, France

Jean-Christophe Thalabard, Université de Paris, France

Catherine Vermandele, Université libre de Bruxelles, Belgique

Comité éditorial

Représentants des groupes spécialisés de la SFdS :

AGRO : Nicolas Pineau (Nestlé)

Banque Finance Assurance : Idriss Tchabda-Djamen (BNP Paribas)

Biopharmacie et Santé : Emmanuel Pham (IPSEN)

Enquêtes : Alina Gabriela Matei (IRDP Université de Neuchâtel)

Enseignement : Catherine Vermandele (Université Libre de Bruxelles)

Environnement : Thomas Opitz (INRAE)

Fiabilité-Incertitudes : Vlad Stefan Barbu (Univ. Rouen)

Histoire de la Statistique : Jean-Jacques Dreesbeke (Université Libre de Bruxelles)

Jeunes Statisticiens : Vivien Goepp (CBIO, Mines ParisTech)

MALIA : Christine Keribin (Université Paris-Sud)

Stat&Sport : Christian Derquenne (EDF)

Statistique et Enjeux Publics : Chantal Cases (INSEE)

Autres membres :

Jose Maria Arribas Macho, revue Empiria (Espagne)

Assaël Adary (Occurrence)

Denise Britz do Nascimento Silva (IASS - International Association of Survey Statisticians)

Gwenaëlle Brihault (INSEE)

Yves Coppieters't Wallant (Ecole de santé publique ULB)

Christophe Ley (Société Luxembourgeoise de Statistique, Gent Universiteit)

Theodore M. Porter (UCLA)

Walter J. Radermacher (La Sapienza Università, Rome)

Design graphique

fastboil.net

ISSN 2269-0271

Éditorial



Emmanuel DIDIER

Rédacteur en chef de *Statistique et société*

Chère Lectrice, Cher Lecteur,

J'ai le plaisir d'annoncer que, pour célébrer les 10 ans de notre revue, nous venons avec Jean-Jacques Droebeke et Catherine Vermandele de faire paraître un ouvrage intitulé *Les Nombres, acteurs de changement* qui célèbre la décennie passée de notre revue¹. Vous y trouverez un florilège d'articles qui y ont été publiés au cours de son existence. Ils montrent ensemble à la fois la diversité de ton, d'objets, de formats de nos publications, et l'unité de notre propos qui consiste toujours à montrer que la quantification est un acteur à part entière des sociétés dans lesquelles nous vivons.

Le premier numéro de 2023 que je vous présente ici ouvre une nouvelle décennie de Statistique et Société. Il comporte un dossier sur un thème que nous avons déjà abordé mais qui se renouvelle sans cesse : les enquêtes et sondages. Le dossier a été construit par Jean Chiche, membre de notre rédaction, qui a rédigé une introduction. Les thèmes qui y sont abordés sont premièrement, les problèmes de mesure appliqués au cas particulier, mais d'une grande actualité, de la rénovation énergétique par Thomas Merly-Alpa ; deuxièmement, le rapport entre les caractéristiques des enquêteurs et le taux de non-réponse, par Thierry Bornand et Baptiste Feraud ; troisièmement, le serpent de mer de l'abandon dans les enquêtes par internet, par Noemie Soullier et une équipe de Santé Publique France et enfin, quatrièmement, l'appariement de bases de données, sur un exemple de questions liées à la pauvreté, par Huyen Tran et Guillaume Osier. On remarquera que ce dossier est particulièrement international puisque les auteurs viennent de Belgique, du Luxembourg et de France.

Vient ensuite un entretien de Jean-Jacques Droebeke avec Marc Hallin, professeur de statistique mathématique à l'Université Libre de Bruxelles, qui a reçu en 2022 le prix Simon de Laplace remis par la SFdS pour l'ensemble de son œuvre. Et enfin nous vous présentons deux recensions, l'une par Antoine Houlou-Garcia du dernier livre de Jean-Baptiste Aubin et Antoine Rolland sur les modes de scrutin ; et l'autre par votre serviteur à propos de la biographie d'Adolphe Quetelet écrite par Jean Jacques Droebeke.

Bonne lecture.
Emmanuel Didier

1. Emmanuel Didier, Jean-Jacques Droebeke et Catherine Vermandele (Eds) *Les Nombres, Acteurs de Changement*, Paris, Presses des Mines (2023) <https://www.mollat.com/livres/2781606/les-nombres-acteurs-de-changement>.

Introduction au dossier « Sondages et enquêtes »



Jean CHICHE¹

Chercheur émérite au Cevipof (Sciences PO, CNRS)

Les articles présentés dans ce dossier spécial « Sondages et enquêtes » ont été écrits à partir de quatre communications présentées et sélectionnées durant le 11^e Colloque Francophone sur les Sondages.

Les colloques francophones sur les sondages sont organisés, depuis 1997, tous les deux à trois ans, sous l'égide de la Société Française de Statistique (SFdS). Ils réunissent des chercheurs et praticiens, issus des instituts publics ou du secteur privé, qui réalisent ou utilisent des enquêtes par sondage. Ces rencontres scientifiques internationales sont des événements importants pour les statisticiens d'enquêtes. Elles permettent de faire le point sur les avancées théoriques comme le partage des poids, l'estimation de variance, la robustesse et les fonctions d'influence, les algorithmes accélérés de calage, l'utilisation des parodonnées... Mais elles permettent aussi de partager les pratiques des concepteurs et des utilisateurs d'enquêtes par sondage. Les conférences et tables rondes posent et tentent de répondre aux questions brûlantes du moment. Voici quelques exemples de problèmes souvent discutés aux derniers colloques : comment endiguer la baisse de la participation aux enquêtes, la pratique des nouveaux modes de collecte, la construction et l'utilisation des panels et access panels internet, le rôle de l'enquêteur, les enquêtes en milieu difficile, la confrontation aux méthodes informatiques qui se sont développées avec le stockage et l'analyse de données massives, comme l'apprentissage profond, les forêts aléatoires, les réseaux de neurones.

L'Université libre de Bruxelles (ULB) a accueilli du mercredi 6 au vendredi 8 octobre 2021, sur le campus du Solbosch, le 11^e Colloque International Francophone sur les Sondages. Sous la houlette du comité d'organisation présidé par Catherine Vermandele et du comité scientifique présidé par Alina Matei, cela a été pour la discipline un grand succès après les années de la pandémie. Pour beaucoup, ce fut, sur un plan personnel, une première occasion et une très belle opportunité de se revoir physiquement et, sur un plan scientifique, de faire un point sur les évolutions de la discipline.

Ce dossier propose quatre articles très différents qui permettent au lecteur non spécialiste des sondages de se faire une idée des problématiques actuelles.

Thomas Merly-Alpa de l'Ined se pose la question de *comment mesurer* à l'aide de nouvelles enquêtes de la statistique publique. Il prend pour exemple la mesure de la rénovation énergétique des logements. Il propose « la construction d'indicateurs statistiques robustes et pertinents à partir de sources administratives, multiples et hétérogènes ».

Thierry Bornand et Baptiste Feraud de l'IWEPS, Belgique, nous proposent de limiter la non-réponse partielle en mesurant la performance des enquêteurs dans les entretiens par questionnaires en face à face. Ils avancent qu'améliorer la qualité des enquêteurs réduira les abandons en cours d'enquête. Pour ce faire, ils proposent une panoplie d'outils opérationnels qu'ils mesurent et testent avec des modèles logistiques. Ils prennent appui sur des enquêtes d'opinion dans le domaine politique et électoral. Si le sujet a été souvent abordé, leur méthodologie est intéressante.

1. jean.chiche@sciencespo.fr

Noémie Soullier, Hugo Rogie, Guilhem Deschamps, Jean-Luc Marchand et Béatrice Geoffroy-Perez de Santé publique France utilisent et analysent une enquête administrée par Internet. Il s'agit de la cohorte française Coset-MSA qui a pour objectif de décrire et de surveiller l'état de santé de la population selon l'activité professionnelle en France. Les auteurs proposent d'explorer les abandons en cours de questionnaire lors du recrutement de la cohorte, à la fois en fonction des profils des répondants mais également de leurs données de connexion. Cette approche novatrice explore des variables peu utilisées, les métas ou paradonnées.

Huyen Tran et Guillaume Osier de STATEC, Luxembourg, nous présentent une étude basée sur l'appariement statistique de bases issues de différentes sources. Ils montrent quelques aspects méthodologiques ainsi que quelques applications dans l'étude de la pauvreté. Ces techniques leur permettent de construire par exemple un indicateur de taux de pauvreté multidimensionnelle basé sur le revenu, la consommation et le patrimoine, un taux d'épargne et un taux de précarité énergétique. Mais ils montrent bien que ces indicateurs construits sont le fruit de modélisation.

Ces articles sélectionnés pour correspondre à l'esprit de la revue ne comportent pas de développements formels. Ils témoignent de l'évolution des techniques d'enquêtes par sondages appliqués à des domaines très divers, mais qui touchent au cœur de la société et sont des aides précieuses pour les politiques publiques et essentielles à notre compréhension du présent.

Comment mesurer ? L'exemple de la rénovation énergétique des logements



Thomas MERLY-ALPA¹

Institut National des Études Démographiques (Ined), Chef du Service des Enquêtes et Sondages

TITLE

How to measure? The example of energy-efficient housing renovation

RÉSUMÉ

La mesure d'un phénomène social et économique demande de choisir des indicateurs et une méthodologie permettant de les calculer. Cet article vise à donner un exemple d'une telle construction statistique sur le domaine de la rénovation énergétique des logements, sujet majeur des politiques publiques actuelles.

La construction d'indicateurs statistiques robustes et pertinents à partir de sources administratives, multiples et hétérogènes, présente de nombreux défis pour la statistique publique. De plus, certaines informations, par exemple les gestes de rénovation réalisés en dehors de tout cadre incitatif, nécessitent la mise en place d'enquêtes statistiques dédiées. La combinaison de l'ensemble des sources offre alors la possibilité de quantifier les efforts de rénovation faits et restant à faire.

Mots-clés : énergie, logement, rénovation, quantification, indicateurs.

ABSTRACT

The measurement of a social and economic phenomenon requires choosing indicators and a methodology to calculate them. This article aims to give an example of such a statistical construction on the field of energy renovation of housing, a major subject of current public policies.

Designing the official indicators mixing heterogeneous, administrative data, is a challenge. Moreover, these data cannot cover the whole scope. For instance, some people might repair their homes without asking for financial help from the government. Combining information from all data sources allows us to produce an estimation of the dynamics of renovation.

Keywords: energy, housing, renovation, quantification, indicators.

Mesurer des phénomènes sociaux et économiques dans la population française est la base du métier de statisticien public. Pour cela, il est nécessaire de concevoir des indicateurs et des méthodologies permettant de les calculer à partir des sources disponibles. Nous proposons dans cet article de détailler ce processus en prenant l'exemple de la rénovation énergétique des logements. La première partie donne une introduction des enjeux de ce domaine. La seconde partie détaille la question du nombre de logements à rénover, en se demandant comment caractériser ces logements, puis comment compter le nombre de logements relevant de chaque catégorie. La troisième et dernière partie se concentre sur la question du nombre de rénovations réalisées chaque année ; après avoir discuté sa définition, nous évoquerons une

1. thomas.merly-alpa@ined.fr

piste d'approximation à partir de sources préexistantes (les aides publiques à la réalisation de travaux) puis à partir d'enquêtes, permettant de compléter et d'enrichir les résultats obtenus.

1. La rénovation énergétique en France

Le gouvernement français s'est engagé à la réduction de ses consommations énergétiques et de ses émissions de gaz à effet de serre, depuis 1997 et la signature du protocole de Kyoto ; cet engagement a été renouvelé et amplifié à l'occasion des accords de Paris, issus de la COP 21 en 2015. Le secteur du bâtiment fait partie des secteurs consommant le plus d'énergie et émettant le plus de gaz à effet de serre ; il a été identifié par les acteurs politiques et économiques comme un secteur sur lequel des économies étaient possibles. Au sein de celui-ci, la rénovation énergétique des logements est identifiée comme un des piliers de la transition énergétique et un gisement de bénéfices environnementaux, économiques et sociaux.

Ces ambitions se sont concrétisées dans plusieurs plans et lois. Après le plan de rénovation énergétique de l'habitat (PREH) promulgué le 21 mars 2013, puis la loi n° 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte (TECV), le plan de rénovation énergétique des bâtiments (PREB), promulgué le 26 avril 2018, définit les ambitions de la France en matière de rénovation énergétique. Il fixe pour objectifs, entre autres (cf. action n° 1), d'éliminer les « passoires thermiques » du parc privé d'ici 2025 (logements classés en étiquettes F et G du Diagnostic de Performance Energétique), et de rénover 500 000 logements par an – ce qui constitue un objectif ambitieux, car il se construit moins de 400 000 logements chaque année en France. Depuis, la loi « Climat et Résilience » du 22 août 2021, qui reprend une partie des 146 propositions de la Convention citoyenne pour le climat, a introduit des mesures d'interdiction de mise sur le marché locatif de certains logements, en fonction de leur étiquette énergétique. En effet, depuis le 1er janvier 2023, les propriétaires des logements les plus énergivores (logements classés G) devront réaliser des travaux de rénovation énergétique s'ils veulent augmenter le loyer de leur location ; ces logements seront interdits à la location dès 2025, et dès 2028 pour ceux classés F.

Plusieurs aides à la rénovation énergétique ont été mises en place en 2005. Les deux principales sont le crédit d'impôt pour la transition écologique (CITE), qui a remplacé en 2014 le crédit d'impôt développement durable (CIDD), et les certificats d'économie d'énergie (CEE). Ces aides sont complétées par l'Anah (agence nationale de l'habitat), qui offrait jusqu'en 2019 des prestations ciblées notamment pour les ménages les plus précaires, et centralise depuis 2020 le dispositif « Ma Prime Renov' », qui a remplacé le CITE. De nombreuses aides locales (régionales, départementales, voire municipales) existent également.

Suivre de tels objectifs et comprendre l'impact des politiques publiques nécessite d'être en mesure de qualifier le parc de logements et d'évaluer finement les dynamiques de rénovation. À cette fin, le commissariat général au développement durable (CGDD) a été missionné pour la mise en place d'un Observatoire de la rénovation énergétique, visant à produire un tableau de bord d'indicateurs assurant le suivi des dynamiques de rénovation des logements, individuels et collectifs. Dans cet article, nous reviendrons ainsi sur la façon dont les agents de la statistique publique ont proposé des indicateurs permettant de répondre aux deux questions principales de la rénovation énergétique des logements : d'une part, combien de logements sont à rénover – ou, plus prosaïquement, combien y a-t-il de passoires thermiques en France ? Et d'autre part, combien de logements sont rénovés chaque année ?

2. Combien de logements à rénover ?

Le premier indicateur à construire est relatif à l'état du parc. Il vise à synthétiser une information

sur la performance énergétique des bâtiments existants, pour permettre aux pouvoirs publics de mieux évaluer le nombre de logements ayant une forte consommation énergétique, qui seront interdits à la location dans un futur proche. Pour cela, il repose principalement sur le diagnostic de performance énergétique (DPE), qui associe à chaque logement une étiquette selon la quantité d'énergie qu'un ménage consommerait pour habiter ce logement (et notamment pour le chauffer, le ventiler et chauffer son eau chaude sanitaire – nécessaire aux douches, etc.). Cette consommation est ensuite associée à une étiquette, selon une grille prédéfinie (Figure 1), conduisant les logements les plus consommateurs – appelés usuellement « passoires thermiques » – à avoir des étiquettes élevées (F voire G).

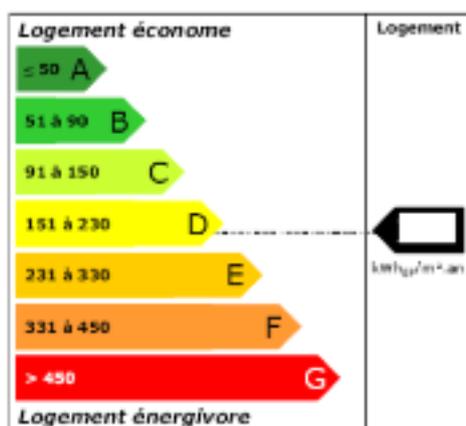


Figure 1 – Les étiquettes « énergie » du diagnostic de performance énergétique en vigueur en 2020, attribuées une fois la consommation connue. Par exemple, une maison bien isolée de 100m² peut nécessiter 10 MWh/an pour son chauffage et son eau chaude sanitaire, ce qui correspond à une consommation surfacique de 100kWh/an*m², qu'on associe à une étiquette C.

Le DPE est un outil à destination des professionnels et du grand public pour informer sur la performance énergétique et climatique d'un logement. Le DPE existe depuis 2006 et a connu plusieurs évolutions, notamment celle relative à la loi ELAN (Évolution du Logement, de l'Aménagement et du Numérique) de 2018, rentrée en vigueur au 1er juillet 2021, qui, outre des ajustements méthodologiques, a rendu le diagnostic de performance énergétique opposable ; cela signifie que, au même titre que pour les autres diagnostics immobiliers (comme le plomb ou l'amiante), en cas d'erreur du diagnostic, la responsabilité du vendeur ou du bailleur peut être engagée par l'acquéreur ou par le locataire qui a subi un préjudice.

Le diagnostic de performance énergétique est un processus mené par un spécialiste (diagnostiqueur certifié par l'État ou l'un de ses opérateurs) qui analyse un bien en parcourant chacune de ses pièces pour établir les éléments physico-techniques du bâti (murs donnant sur l'extérieur ou non, matériaux, etc.), compiler les informations sur les systèmes de chauffage, de ventilation et d'eau chaude sanitaire, et évaluer les flux et les déperditions thermiques. Il conduit à établir une consommation d'énergie (et les conséquences environnementales associées, notamment en termes d'émissions de CO₂) nécessaire à un usage « standard » du bâtiment, c'est-à-dire une période d'occupation moyenne (par jour et sur l'année), une température de chauffage standardisée ainsi qu'un nombre d'habitants moyen en fonction de la superficie du logement.

Réaliser une telle analyse pour chaque logement demanderait un trop fort investissement des diagnostiqueurs. Pour éviter cela, dès la conception du DPE, des méthodes (et des logiciels permettant de les mettre en œuvre) ont été élaborées afin de rendre comparables tous les diagnostics. La méthode usuelle est la méthode dite 3CL (pour « calcul de la consommation conventionnelle des logements »), qui recense l'ensemble des caractéristiques du logement

(forme, mitoyenneté, nombre d'ouvertures, matériaux...) et, en combinaison avec une localisation (zone climatique, altitude), en déduit une consommation moyenne annuelle associée au logement pour le chauffer, le ventiler et pour l'eau chaude sanitaire.

Il est à noter que, jusqu'à l'entrée en vigueur de la loi ELAN, cette méthode n'était pas applicable pour les logements les plus anciens, pour lesquels les méthodes de construction et les matériaux étaient plus hétérogènes, ce qui réduisait fortement la pertinence du calcul ; un jeu d'hypothèses plus complet a désormais été intégré à la méthode pour l'étendre à ces logements. Pour permettre l'édition d'un DPE pour les logements construits avant 1948, une estimation à partir des factures de consommation d'énergie était réalisée : c'est ce que l'on appelle la méthode « factures ». L'utilisation de cette méthode conduit à des biais importants par rapport à une méthode de type 3CL :

- Cette méthode repose sur des factures énergétiques observées, et la performance énergétique peut être sur- ou sous-évaluée selon le comportement du ménage (température de confort, présence dans le logement, etc.).
- Elle conduit à une estimation plus proche du concept de consommation réelle d'énergie que l'autre méthode, qui conduit à des consommations conventionnelles ; si l'objectif est bien sûr de faire concorder les deux concepts, il demeure toujours des écarts, liés notamment aux hypothèses faites dans les méthodes de calcul du DPE 3CL.
- Enfin, certains logements étant inoccupés lors de la réalisation d'un DPE (cas d'un logement remis en vente ou en location), la réalisation du diagnostic sur ce mode était impossible.

Ainsi, la même étiquette DPE obtenue pour deux logements avec deux méthodes différentes ne recouvre pas la même réalité, mais conduit pourtant à un affichage et un classement identique. Par exemple, un logement occupé par un ménage utilisant une température de chauffage élevée (21 ou 22°C) sera considéré comme très énergivore par la méthode « factures » en raison de factures élevées, même si une analyse de type 3CL pourrait conclure à une consommation conventionnelle beaucoup plus limitée, car elle suppose une température de chauffage de 19°C.

Malgré ses limites, le DPE est un indicateur qui permet de construire des statistiques sur l'état du parc des logements. En 2012, la réalisation d'une enquête dite PHEBUS (voir Rouquette, 2015, par exemple) a permis la réalisation d'un diagnostic auprès d'un échantillon de quelques milliers de logements. Cette étude concluait à un nombre de passoires thermiques (étiquettes F et G du DPE) d'environ 7 millions. Depuis 2020, une autre analyse (Merly-Alpa *et al.*, 2020) a été effectuée sur la base de l'ensemble des diagnostics réalisés en France et collectés par l'Ademe (Agence de la transition écologique), qui les diffuse en accès libre sur un site Internet dédié. Cette masse de données, d'environ un million de diagnostics par an, offre une vision très large de l'état du parc français. Cependant, ces données souffrent de certains biais, liés aux événements de la vie d'un logement pour lesquels la réalisation d'un DPE est obligatoire : mise en location, vente, construction. Or, tous les logements anciens ne sont pas mis en location ou vendus chaque année, et ne bénéficient donc pas tous d'un DPE ; alors que tous les logements nouvellement construits sont concernés par un tel diagnostic. La base est donc représentative d'un parc sensiblement plus jeune que le parc français de référence, et des opérations de correction par appariement avec les sources fiscales et le recensement de la population (détaillées dans Merly-Alpa *et al.*, 2020) sont nécessaires. Cette estimation conduit à un chiffre d'environ 5 millions de passoires thermiques, en corrigeant les limites liées aux DPE « factures » évoqués plus haut.

Comme tout indicateur, le nombre de passoires thermiques est une construction statistique mobilisant un outil utilisé par le grand public à d'autres fins : choix d'un logement à louer ou acheter, décision de réalisation de travaux... La situation correspond ainsi à une application classique de la loi de Goodhart, qui s'énonce de la façon suivante : « lorsqu'une mesure devient

un objectif, elle cesse d'être une bonne mesure ». En effet, compter le nombre de passoires thermiques a pour objectif d'atteindre un nombre le plus faible possible de tels logements, mais repose aussi, par définition, sur la réalisation de diagnostics de performance thermique dont le résultat n'est pas neutre pour le propriétaire du logement. Il y a en effet un intérêt notamment économique, car on peut mettre en évidence la « valeur verte » d'un logement (Laurenceau et Giraudet, 2011), mais aussi réglementaire avec la mise en place à horizon 2025 de l'interdiction de location de certains logements trop énergivores (étiquettes G, étendue aux étiquettes F en 2028). La mesure de l'état du parc de logements obtenue en mobilisant l'ensemble des DPE entre ainsi en contradiction avec l'intérêt d'un ménage réalisant un diagnostic pour son logement. Ainsi, un propriétaire pourra être tenté de mettre en concurrence plusieurs diagnostiqueurs, ou inversement un diagnostiqueur vérifiera à deux fois avant de passer le logement à une étiquette inférieure.

Nous pouvons trouver un reflet de cet écueil dans les données : l'analyse de la structure des consommations d'énergie dans les DPE (Figure 2) montre des « pics » de la distribution au moment du changement d'étiquette, et ce uniquement à gauche du passage. En d'autres termes, il y a une accumulation, incohérente d'un point de vue statistique, des consommations avant chaque passage à une étiquette plus mauvaise. Sur la base de ce constat, Girard (2011) procède à une correction dans l'estimation du nombre de passoires thermiques en lissant ces pics, ce qui conduit à faire basculer certains logements dans l'étiquette suivante. Le bien-fondé de cette correction est, au-delà d'une question statistique, une question de mesure : cherche-t-on à mesurer le nombre de logements ayant officiellement une étiquette donnée, ou cherche-t-on à mesurer le nombre de logements dont la consommation, éventuellement corrigée, dépasse un certain seuil ? Selon Girard (2011), choisir parmi ces deux alternatives conduit à des estimations très différentes : il aboutit en effet à une estimation d'environ 7 millions de passoires.

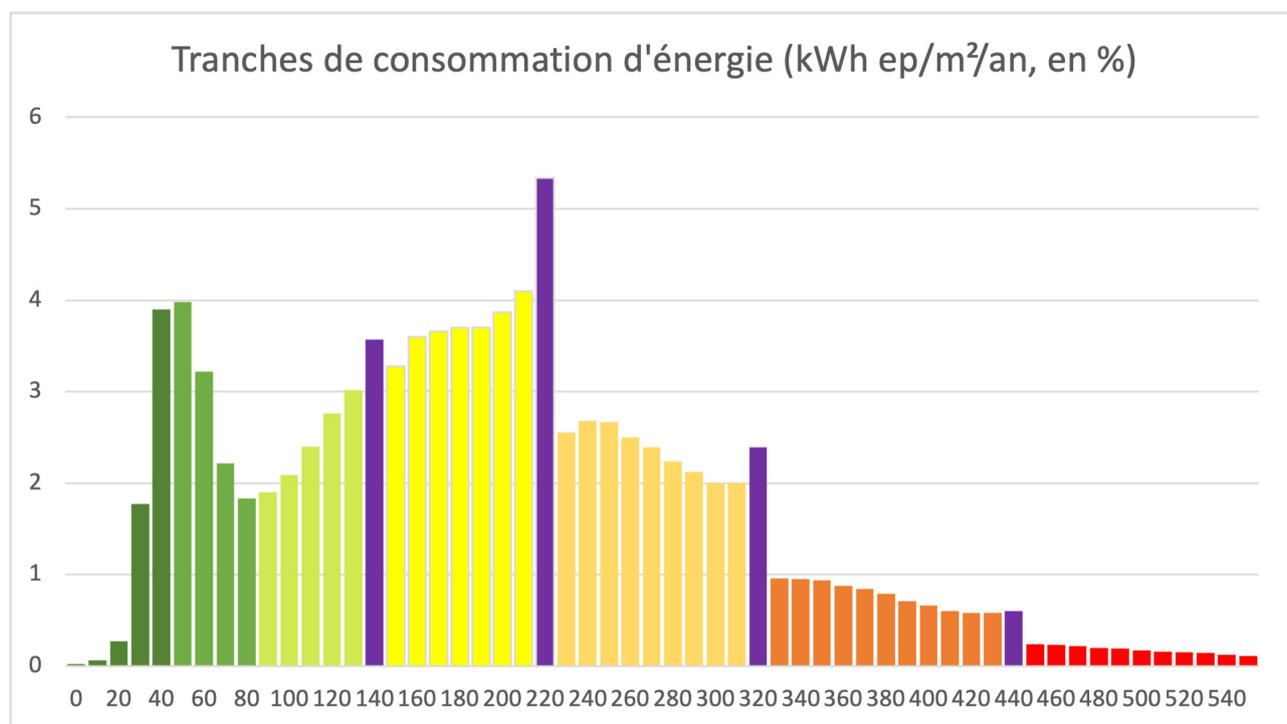


Figure 2 – Répartition des consommations énergétiques calculées pour l'ensemble des DPE de la base Ademe

3. Combien de rénovations chaque année ?

3.1 Un concept statistique à construire

Une fois estimé le stock de logements à rénover (5 millions de « passoires thermiques », selon l'Observatoire national de la rénovation énergétique), se pose la question de la mesure du flux : combien de logements sont rénovés chaque année ? Ou, plus précisément, combien de rénovations sont effectuées ? Répondre à cette question demande certes des données et des informations sur les travaux réalisés en France, mais surtout une définition claire de la notion de rénovation, d'un point de vue statistique. Or, les rénovations sont multiples et peuvent concerner des logements divers : on peut rénover un petit studio ou une grande maison, et on peut changer une seule fenêtre ou refaire entièrement l'isolation et le chauffage.

Pour résoudre ce problème, il faut revenir à la logique derrière la comptabilité des rénovations. Comme indiqué en Partie 1, les objectifs du PREB sont de rénover 500 000 logements par an ; ce chiffre a été choisi comme objectif permettant d'atteindre des cibles de consommation d'énergie qui soient compatibles avec les engagements issus des accords de Paris. Or, une rénovation peut recouvrir de nombreuses situations : entre la rénovation complète d'une grande maison qui avait une étiquette G et le remplacement d'une fenêtre dans un appartement moderne, l'effet sur la consommation d'énergie totale sera sensiblement différent. Il est donc nécessaire de trouver une définition de la rénovation qui permette une cohérence entre les 500 000 logements rénovés et l'atteinte de faibles niveaux de consommation d'énergie. Pour cela, il ne faut pas compter en nombre de logements rénovés (ce qui comporte une forte hétérogénéité), mais considérer une rénovation comme une unité mesurant une économie d'énergie. Ainsi, rénover entièrement une passoire thermique pourrait correspondre à plusieurs « rénovations », tandis qu'un simple changement de fenêtre ne correspondrait qu'à 0,1 rénovation. Par ailleurs, cette vision permet de sommer des gestes hétérogènes par nature (changer un radiateur, isoler des combles) pour arriver à un unique indicateur.

Il ne reste plus alors qu'à déterminer la quantité d'énergie à laquelle doit correspondre une rénovation. Pour cela, il faut calculer l'écart entre la consommation actuelle du parc de logements et les prévisions en 2050 de la stratégie nationale bas-carbone (SNBC), en prenant en compte les évolutions de structure du parc de logements (notamment sa surface moyenne). Une fois la rénovation considérée comme une unité d'énergie, le suivi de cette politique pourrait sembler direct : la simple mesure de la consommation d'énergie pour les besoins de logement, corrigée éventuellement de la consommation d'électroménager, des destructions et constructions de logements, serait à même de nous donner un chiffre comparable à l'objectif de 500 000 rénovations.

Cependant, comme évoqué en Partie 2 sur le DPE, la mesure d'énergie associée à une rénovation est en fait une consommation d'énergie conventionnelle ; on peut l'obtenir en réalisant deux diagnostics de performance énergétique, avant et après les travaux. Or, ce gain de consommation « conventionnelle » ne correspond pas nécessairement à une baisse réelle de consommation d'énergie : cela peut s'expliquer par un comportement des ménages occupants qui n'est pas conforme aux standards (occupation plus importante, température de chauffage différente), mais aussi pour ce que la littérature en économie de l'énergie appelle un « effet rebond » (ou, en anglais, *Energy Efficiency Gap* ; voir Jaffe et Stavins, 1994, et Allcott et Greenstone, 2012), qui signifie ici que les ménages ayant investi pour la rénovation énergétique de leur logement peuvent voir *in fine* leur consommation et donc leur facture augmenter car ils peuvent enfin chauffer leur logement à un prix raisonnable, ce qui n'était pas le cas avant les travaux ; si l'effet est positif sur le bien-être et la santé des ménages (voir, par exemple, Dervaux *et al.* (2022) sur le lien entre santé et précarité énergétique), il peut conduire à une baisse de la consommation d'énergie plus faible qu'attendue.

Pour compter le nombre de rénovations en France, il ne suffit donc pas de connaître la consommation d'énergie du secteur résidentiel, il est aussi nécessaire de connaître le nombre de gestes de rénovation qui sont réalisés, de connaître leurs caractéristiques et celles des logements rénovés, puis de définir des règles de calcul associant à une famille de rénovation un gain énergétique. Nous n'entrerons pas ici dans le détail de ce dernier calcul, qui est explicité dans les annexes de Kraszewski et Le Jeannic (2021), mais qui ouvre lui-aussi un vaste éventail de questions de conception et de pertinence statistique.

3.2 Des rénovations accompagnées par les aides publiques

Si aucun décompte complet des gestes de rénovation n'existe en France à notre connaissance, il existe cependant de nombreuses aides à la rénovation, offrant un aperçu partiel des gestes de rénovation réalisés. Cet aperçu est partiel car ces aides sont parfois complexes : en effet, si leur objectif est de massifier et d'encourager à la rénovation énergétique des logements, de nombreuses aides à la rénovation différentes existent, avec des conditions spécifiques pour en bénéficier, certaines pouvant parfois être cumulées. Certains ménages ne savent donc pas qu'ils peuvent en bénéficier, ou y renoncent face à la complexité des dispositifs.

Pour les statisticiens publics, cette source de données est néanmoins très utile car elle permet d'obtenir, avec un léger retard et malgré les limites imposées, un premier aperçu du rythme des rénovations. L'Observatoire national de la rénovation énergétique a mobilisé les données relatives aux deux aides principales : le crédit d'impôt pour la transition écologique (CITE) et les certificats d'économie d'énergie (CEE).

Le CITE (crédit d'impôt pour la transition écologique) vise à encourager les ménages à réaliser des travaux de rénovation énergétique. Cette prime est associée à la déclaration de travaux précis dans le cadre des déclarations de revenus des ménages ; elle ne concerne que la réalisation de travaux dans la résidence principale du ménage. Elle correspond à de nombreux travaux tels que le changement d'une chaudière, l'installation d'une pompe à chaleur, l'isolation de parois extérieures ou intérieures... Le CITE est progressivement remplacé par « Ma Prime Renov' », rentrée en vigueur en 2020. Les données du crédit d'impôt sont centralisées par la direction générale des finances publiques (DGFiP), combinées à des informations sur les foyers fiscaux les ayant réalisés (ainsi que le logement rénové), ce qui permet des analyses directes dès lors que l'on peut accéder aux données, ce qui est le cas des statisticiens travaillant dans le système statistique public (SSP).

Le dispositif des CEE (Certificats d'économie d'énergie) repose sur une obligation de réalisation d'économies d'énergie imposée par les pouvoirs publics aux vendeurs d'énergie (électricité, gaz, GPL, chaleur et froid, fioul domestique et carburants pour automobiles). Ceux-ci sont ainsi incités à promouvoir activement l'efficacité énergétique auprès de leurs clients et des autres consommateurs d'énergie : ménages, collectivités territoriales ou professionnels. Les CEE sont par définition un dispositif très décentralisé, mais qui est piloté par la Direction générale de l'énergie et du climat (DGEC), qui dispose d'un registre compilant tous les certificats réalisés chaque année. Chaque certificat est renseigné suivant une fiche spécifique, qui détaille le geste technique (par exemple, épaisseur de l'isolant installé) ainsi que l'adresse du logement rénové.

Pour comptabiliser les rénovations (cf. partie 3.1), il est désormais nécessaire d'associer à chaque geste une économie d'énergie envisagée, afin de pouvoir les sommer. Pour cela, Kraszewski et Le Jeannic (2021) ont mobilisé les informations disponibles dans les bases de données pour arriver à une estimation aussi précise que possible : localisation et superficie du bien, geste de rénovation réalisé, et, autant que possible, caractéristiques techniques. Pour le CITE, l'information est plus maigre, car l'on ne dispose pas, par exemple, de l'épaisseur de l'isolant déclaré, celle-ci ne devant pas être indiquée dans le cadre de la déclaration annuelle

de revenus. Pour l'estimer, Kraszewski et Le Jeannic (2021) proposent d'utiliser le montant de dépenses de travaux déclaré aux impôts comme indicateur de la performance énergétique des travaux réalisés, en se basant sur des résultats des enquêtes Tremi (détaillée ci-après en partie 3.3.1). Par exemple, un ménage pourrait déclarer avoir dépensé 1 300€ pour isoler la toiture de sa maison de 80m² ; or, nous savons qu'une isolation de toiture conduisant à un gain énergétique de 2MWh/an pour une telle maison coûte 2 600€. Nous pouvons donc supposer que cette isolation n'a pas concerné la totalité de la toiture, ou que l'isolant utilisé est de moindre qualité, ou d'une épaisseur plus faible. Une hypothèse raisonnable est alors de penser que le gain énergétique est moitié moindre, soit 1MWh/an.

Il reste une dernière étape pour conclure le calcul. En effet, additionner directement les gains totaux associés aux deux aides conduirait à surestimer la quantité d'énergie économisée, car il est possible pour un ménage de solliciter les deux aides pour les mêmes travaux. Autrement dit, compter deux fois le même geste conduirait ainsi à une surestimation de la rénovation énergétique en France. Il est donc nécessaire de chercher à identifier ces « doublons », ce qui est fait en rapprochant les adresses et les gestes de travaux entre les deux bases de données, afin de s'assurer de les retirer du calcul. On aboutit alors à une estimation du nombre de rénovations qui ont été réalisées grâce aux aides gouvernementales : 2,4 millions de ménages ont bénéficié d'aides à la rénovation énergétique en 2019.

3.3 Une estimation à consolider

Cependant, ces aides ne permettent pas de connaître l'ensemble des rénovations réalisées dans les logements français. En effet, environ une rénovation sur trois (en 2017) serait réalisée sans bénéficier d'aucun dispositif d'aide. Il est donc nécessaire de compléter ces informations par le biais d'enquêtes statistiques à échantillonnage aléatoire. Ces enquêtes, Tremi (Travaux de rénovation énergétique dans les maisons individuelles) et Treco (Travaux de rénovation énergétique dans le logement collectif), détaillées ci-après, nous offrent également des informations sur différents sujets :

- les motivations des ménages (confort, économies d'énergie) réalisant des rénovations ;
- à l'inverse, les freins à la rénovation (complexité, coût...) ;
- la connaissance et le recours éventuel à des services d'appui à la rénovation ;
- le coût des rénovations en rapport aux montants par les aides.

Elles permettent également de mieux appréhender les recours à des aides multiples (gestion des « doublons » évoquée en fin de partie 3.2) et améliorent ainsi les indicateurs construits à partir de sources administratives.

Enfin, les questionnaires des deux enquêtes permettent de récupérer des informations précises sur les gestes de rénovation dans les bâtiments (type et épaisseur d'isolants, par exemple) ; cela permet de mettre en place une méthode semblable à celle mise en place dans le cadre des diagnostics de performance énergétique pour évaluer plus précisément la performance des rénovations.

3.3.1 Description de l'enquête Tremi

L'enquête Tremi est la déclinaison de l'observatoire OPEN sur les maisons individuelles, créé en 2006 à l'initiative de l'Ademe. Réalisé tous les deux ans, cet observatoire avait pour objectif de suivre de manière qualitative et quantitative le marché de la rénovation énergétique dans les maisons individuelles et parties privatives des logements collectifs. L'Ademe a décidé de refondre l'observatoire qui est ainsi devenu en juillet 2018 l'enquête Tremi, qui présente une méthodologie davantage transparente et concertée, et qui se concentre désormais sur les maisons individuelles.

L'enquête 2020 concerne la France métropolitaine ainsi que les DOM, pour lesquels les problématiques de rénovation sont spécifiques (confort d'été principalement, le chauffage étant plus marginal). L'échantillon, tiré dans les sources fiscales par l'Insee, est de grande taille (100 000 logements) afin de disposer d'un nombre suffisant de ménages répondants ayant réalisé des travaux dans chacune des régions françaises. Le tirage de l'échantillon a également été implicitement stratifié sur des variables fortement corrélées à la réalisation de travaux et à la performance de ceux-ci (suite aux résultats des éditions précédentes, et comme montré par Charlier, 2013) : date de construction, statut d'occupation, tranche de revenus des occupants...

La collecte des données est réalisée par Internet (questionnaire en ligne), sur la base des informations de contact contenues dans l'échantillon, via l'envoi de lettres-avis avec des identifiants ainsi que de mails lorsque l'adresse électronique est disponible. Des questionnaires papier sont également mis à disposition des ménages lorsque la réponse par Internet est impossible. Cela conduit à un risque qui, s'il existe pour la grande majorité des enquêtes, pourrait être particulièrement important ici : le simple fait d'avoir réalisé ou non des travaux peut avoir un impact sur la motivation d'un ménage à répondre à l'enquête (par intérêt pour le sujet, par exemple), même si plus de la moitié des ménages enquêtés ont finalement répondu. C'est ce que l'on appelle le caractère « non-ignorable » de la non-réponse (ou *Non Missing at Random* ; Little et Rubin, 2002). Pour pallier ce problème, les traitements statistiques d'exploitation des données de l'enquête doivent mobiliser le plus d'informations possibles, et notamment des sources plus directement liées aux variables d'intérêt de l'enquête : les informations fiscales relatives au CITE et les CEE (tous deux présentés en partie 3.1). La méthode mise en œuvre, détaillée par Baudry *et al.* (2022), permet ainsi de limiter autant que possible ce risque et d'obtenir des résultats non biaisés.

Les résultats de l'enquête Tremi, présentés par Caumont *et al.* (2021), montrent que la rénovation non aidée est une part non négligeable de l'ensemble des rénovations énergétiques en France ; en 2019, intégrer les rénovations non aidées en maisons individuelles conduit à augmenter de près d'un tiers l'estimation de l'énergie économisée sur ce champ. Cela peut s'expliquer par une méconnaissance des aides ou une trop grande restriction des critères d'application, même si les résultats de l'enquête montrent assez logiquement que les ménages ayant rénové connaissent mieux les aides à la rénovation que les autres.

3.3.2 Treco – un projet d'enquête sur un nouveau champ

En combinant les résultats sur les aides à la rénovation et ceux de l'enquête Tremi, un indicateur composite sur la rénovation peut ainsi être construit. Cependant, celui-ci ne couvre que le champ des maisons individuelles ; il est donc nécessaire de réaliser un travail sur le champ du logement collectif.

Celui-ci se répartit en deux problématiques très différentes. Premièrement, il concerne l'ensemble du parc social (HLM), géré par des opérateurs publics ou privés qui ont la charge de la rénovation (y compris énergétique) des bâtiments et logements, et qui en rendent compte à l'État, notamment via le répertoire des logements locatifs des bailleurs sociaux (RPLS). Deuxièmement, il faut créer un dispositif statistique pour mesurer la dynamique de rénovation énergétique des copropriétés privées, encore trop mal connue. Cela est dû à la plus grande difficulté à collecter de l'information sur ce champ, car les rénovations peuvent concerner, d'une part, les parties communes (isolation lors d'un ravalement de façade, changement de la chaudière collective), d'autre part les appartements eux-mêmes (changement d'un radiateur électrique individuel, remplacement de fenêtres). Pour cela, l'Observatoire national de la rénovation énergétique et ses partenaires proposent de mettre en place l'enquête Treco, comme un complément à Tremi sur le champ du logement collectif.

Cette enquête, encore en projet, est bien plus complexe. En effet, apprécier l'effort de rénovation énergétique demande de connaître les deux versants de l'histoire, ce qui nécessite d'interroger plusieurs interlocuteurs :

- les occupants, pour leur ressenti ;
- les propriétaires, pour les travaux réalisés dans leurs appartements ;
- le syndic de copropriété, pour les détails des travaux sur les parties communes.

Par ailleurs, la prise de décision est bien plus complexe qu'en maison individuelle (voir, par exemple, BrisePierre, 2014) : réaliser une opération d'isolation est bien plus coûteux et concerne l'ensemble des copropriétaires, qui doivent financer l'opération et donc décider de sa réalisation lors d'une assemblée générale (AG).

Réaliser une enquête sur le sujet suppose donc de réaliser plusieurs volets et plusieurs questionnaires. Une piste privilégiée est d'enquêter les syndics par rapport aux copropriétés qu'ils gèrent (identifiées via le registre des copropriétés de l'Anah) sur la réalisation ou non de travaux, les discussions en assemblée générale, l'état de l'immeuble ; elle s'intéresse particulièrement au coût des travaux et aux aides mobilisées. Une enquête complémentaire auprès de plusieurs occupants de ces mêmes copropriétés permet de capter les motivations ainsi que d'éventuels travaux de rénovation des parties privatives.

Le premier volet de l'enquête, auprès des syndics, serait basé sur un échantillon de copropriétés suivant différentes caractéristiques (nombre de lots, typologie de copropriétés – revenus, type d'occupation, date de construction –, type de chauffage, localisation géographique...) dans le registre des copropriétés. Ces caractéristiques permettent de segmenter efficacement le parc des copropriétés, que ce soit en termes de gestion (le syndic peut être professionnel ou bénévole), de réalisation de travaux (les copropriétaires occupants étant souvent plus disposés à réaliser des travaux que lorsqu'ils sont bailleurs) et de nature des rénovations (une copropriété récente n'aura pas le même potentiel qu'une copropriété construite dans les années 1960, en particulier si cette dernière dispose d'un chauffage collectif).

De façon simultanée, un échantillon d'au plus une dizaine d'occupants dans les copropriétés enquêtées serait sélectionné. Un questionnaire court leur serait alors adressé pour connaître leurs informations personnelles (composition du ménage, statut d'occupation, date d'emménagement, revenus, âge), la position du logement dans l'immeuble (quel étage, quelle orientation...), le confort thermique du ménage, ses votes aux dernières AG, et les éventuels travaux réalisés dans les parties privatives.

Ce projet est à l'étude, mais il comporte des risques. En effet, il est possible que les connaissances techniques soient perdues si des changements de syndics ont eu lieu. Pour ce dernier problème, une solution pourrait être de récupérer des documents tels que les audits énergétiques. Par ailleurs, le faible échantillon sur les occupants réduit la qualité des informations que l'on peut récupérer, en particulier sur les travaux dans les parties privatives. Enfin, la structure de l'échantillon conduirait à disperser les poids de sondage associés aux questionnaires des occupants, ce qui réduirait la précision de ces résultats. Cela pourrait être corrigé par un échantillonnage adapté ou en interrogeant l'ensemble des occupants de chaque copropriété, mais ces deux solutions pourraient dégrader la qualité de l'enquête (soit parce que l'échantillon devient moins adapté pour les syndics, soit parce que cela crée une forte charge de réponse sur une même copropriété).

4. Conclusion

Tout au long de cet article, nous avons suivi la démarche du statisticien public pour construire des indicateurs permettant de mesurer et mieux comprendre la rénovation énergétique des

logements. La mobilisation de sources dites administratives (ici sources fiscales, répertoires de diagnostics énergétiques, registres d'aides comme les certificats d'économie d'énergie...) permet de fournir des estimations sur l'état du parc de logements français : nous estimons qu'il existe 5 millions de passoires thermiques sur le territoire français, mais ce chiffre repose sur les hypothèses simplificatrices du diagnostic de performance énergétique. Elle offre également une vision partielle du nombre de rénovations réalisées chaque année : 2,4 millions de ménages ont bénéficié d'aides à la rénovation énergétique en 2019. Mais certains ménages ne sollicitent pas ou ne peuvent pas bénéficier de ces aides. L'utilisation de ces données administratives doit donc s'accompagner d'enquêtes statistiques, certes beaucoup plus coûteuses, mais permettant d'enrichir et de compléter les premiers indicateurs produits : l'enquête Tremi 2020 permet de savoir que près de 700 000 ménages ont réalisé des travaux de rénovation énergétique dans leur maison sans bénéficier d'une aide de l'État, ces travaux conduisant à près d'un quart de l'énergie totale économisée grâce à la rénovation énergétique des maisons.

Enfin, le travail de l'Observatoire national de la rénovation énergétique va s'enrichir dans les prochaines années en mobilisant de nouvelles sources : l'accès aux données de consommation réelle d'énergie via les distributeurs (Enedis, GRdF, entreprises locales de distribution...) va permettre d'enrichir les estimations avec l'effet de la rénovation énergétique sur la trajectoire de réduction des consommations observées d'énergie. Ces travaux vont poser des questions méthodologiques (variations climatiques, etc.), mais aussi d'autres natures, notamment autour du consentement à l'usage des données de consommation d'énergie (discutées lors des commissions « Environnement et Développement Durable » du Conseil national de l'information statistique – CNIS), dans le contexte des polémiques Linky mais aussi plus largement sur le contrôle de la consommation de l'énergie (comme détaillé par Danieli, 2021).

Références

Allcott H. and M. Greenstone (2012), « Is There an Energy Efficiency Gap? », *Journal of Economic Perspectives*, 26, pp. 3-28, DOI: 10.1257/jep.26.1.3.

Baudry M., T. Merly-Alpa et J.-P. Rathle (2022), « Observer la rénovation énergétique par des enquêtes statistiques », 14es Journées de Méthodologie Statistique de l'Insee.

Brisepierre G. (2014), « Comment se décide une rénovation thermique en copropriété ? Un nouveau mode d'organisation de l'habitat comme condition de l'innovation énergétique », *Flux*, 2, pp. 31-39.

Caumont R., T. Merly-Alpa et J.-P. Rathle (2021), « La rénovation énergétique des maisons individuelles, résultats de l'enquête Tremi 2020 », in « La rénovation énergétique des logements, bilan des travaux et aides entre 2016 et 2019 », Commissariat au développement durable, Observatoire national de la rénovation énergétique, <https://www.statistiques.developpement-durable.gouv.fr/la-renovation-energetique-des-logements-bilan-des-travaux-et-des-aides-entre-2016-et-2019-resultats>

Charlier D. (2013), « La rénovation énergétique en France dans le secteur résidentiel privé », *Revue de l'OFCE*, 2, pp. 395-419.

Danieli A. (2021), « Mesurer pour contrôler ? Les épreuves de la quantification domestique avec le compteur d'électricité (de 1880 à nos jours) », *Statistique et Société*, 9(3), pp. 31-61.

Dervaux B., L. Rochaix, B. Meurisse *et al.* (2022), « L'évaluation socioéconomique des effets de santé des projets d'investissement public », Rapport du groupe de travail. France Stratégie - Commissariat général à la stratégie et la prospective, <https://hal-lara.archives-ouvertes.fr/hal-03664885/>

Kraszewski M. et T. Le Jeannic (2021), « Les rénovations énergétiques aidées du secteur résidentiel entre 2016 et 2019 », *in* « La rénovation énergétique des logements, bilan des travaux et aides entre 2016 et 2019 », Commissariat au développement durable, Observatoire national de la rénovation énergétique, <https://www.statistiques.developpement-durable.gouv.fr/la-renovation-energetique-des-logements-bilan-des-travaux-et-des-aides-entre-2016-et-2019-resultats>

Jaffe A. B. and R. N. Stavins (1994), « The energy-efficiency gap: What does it mean? », *Energy Policy*, 22, pp. 804-810, DOI: 10.1016/0301-4215(94)90138-4.

Chotard D., M. Midori, C. Berthon, S. Laurenceau, L.-G. Giraudet (2011), « Analyse préliminaire de la valeur verte pour les logements », pp. 261, https://hal-enpc.archives-ouvertes.fr/file/index/docid/799764/filename/ADEME_11_ValeurVerte.pdf

Little R. J. and D. B. Rubin (2002), *Statistical analysis with missing data* (vol. 793), John Wiley & Sons.

Girard R. (2021), « Sur le nombre de passoires énergétiques en France »,

Merly-Alpa T., N. Riedinger et M. Baudry (2020), « Le parc de logements par classe de consommation énergétique », Commissariat au développement durable, Observatoire national de la rénovation énergétique, <https://www.statistiques.developpement-durable.gouv.fr/le-parc-de-logements-par-classe-de-consommation-energetique-au-1er-janvier-2018>

Rouquette C., D. François, G. Houriez et CGDD/SOeS (2015), « L'enquête Performance de l'habitat, équipements, besoins et usages de l'énergie (Phébus) : quel dispositif pour quels enseignements ? », *La Revue du CGDD*, pp. 47-56, <https://temis.documentation.developpement-durable.gouv.fr/docs/Temis/0081/Temis-0081488/21873.pdf>

26. Dans les provinces où il n'existe pas d'université, un des collèges communaux doit avoir pour but « de propager [...] le goût et les lumières parmi toutes les classes de la société, sans en excepter celles qui ne se destinent point aux études académiques ». Ces établissements privilégiés, désignés par le nom d'*athénée*, sont établis à Bruxelles, Maestricht, Bruges, Tournai, Namur, Anvers et Luxembourg (Mailly, 1875).

27. Verhulst (1845).

La qualité des enquêteurs importe dans la non-réponse partielle des répondants qui sont moins motivés par l'enquête



Thierry BORNAND¹

Institut Wallon de l'Évaluation, de la Prospective et de la Statistique (IWEPS), Belgique



Baptiste FERAUD²

Institut Wallon de l'Évaluation, de la Prospective et de la Statistique (IWEPS), Belgique

TITLE

The quality of the interviewers matters in the partial non-response of respondents who are less motivated by the survey

RÉSUMÉ

Le rôle des enquêteurs dans la qualité des données d'enquêtes en face-à-face est un élément important qui ne doit pas être sous-estimé. Bien qu'en raison du coût ce type d'enquête tend de plus en plus à être remplacé par de l'auto-administré (web + questionnaire papier), elle reste une référence en termes de fiabilité des données et peut se montrer particulièrement adéquate pour interroger certaines catégories de la population, moins enclines à répondre à des enquêtes. La qualité et le professionnalisme des enquêteurs peuvent se montrer alors déterminants pour le niveau de qualité des données récoltées. Dans cette étude, nous présentons une procédure mise en place par l'IWEPS, l'institut de statistique Wallon en Belgique, qui permet d'évaluer la qualité des enquêteurs au début de la collecte des données et utilisons cette évaluation pour voir dans quelle mesure la qualité des enquêteurs influence le nombre de non-réponses partielles. En nous basant sur le modèle du processus de réponse, nous formulons l'hypothèse de trois sources d'influence sur ce type de non-réponses. Une première source est liée aux caractéristiques socio-démographiques des répondants : leur âge, leur niveau d'éducation ou leur intérêt pour la thématique de l'enquête (dans le cas présent, l'intérêt pour la politique). Une deuxième source est liée aux caractéristiques des enquêteurs, que nous mesurons par l'évaluation que nous en avons fait. Enfin, une troisième source est liée à l'interaction entre les caractéristiques des répondants et les caractéristiques de l'enquêteur. Les résultats confirment l'effet des deux premières sources d'influence, comme cela a déjà été montré par ailleurs dans la littérature. Concernant l'interaction, les résultats montrent un effet lorsque les deux conditions non propices se trouvent réunies, à savoir : un faible intérêt des répondants pour la politique et une faible qualité des enquêteurs. Lorsque ces deux conditions sont réunies, le nombre de non-réponses partielles augmente de manière très importante. Les conséquences de ces résultats tant sur l'importance d'évaluer les enquêteurs en situation que sur la manière d'améliorer la procédure pour s'assurer de la qualité des données sont discutées en conclusion.

Mots-clés : effet de l'enquêteur, non-réponse partielle, enquête en face-à-face, évaluation des enquêteurs.

ABSTRACT

The role of interviewers in the quality of face-to-face survey data is an important element that should not be underestimated. Although this type of survey is increasingly being replaced by self-administered surveys (web + paper) because of the cost, it remains a reference in terms of data reliability and can be particularly suitable for interviewing certain categories of the population that are less inclined to respond to surveys. Thus, the quality and professionalism of the interviewers can be a determining factor in the quality of the data collected. In this study,

1. t.bornand@iweeps.be

2. b.feraud@iweeps.be

we present a procedure implemented by IWEPS, the Walloon statistical organization in Belgium, to evaluate the quality of interviewers at the beginning of data collection and to ensure that they correctly apply the instructions given to them during field preparation and use this assessment to see to what extent the quality of interviewers influences the number of partial non-responses. Based on the response process model, we hypothesize three sources of influence on this type of non-response. A first source is related to the socio-demographic characteristics of the respondents: their age, their level of education or their interest in the subject of the survey (in this case interest in politics). A second source is linked to the characteristics of the interviewers, which we measure by the evaluation we made of them. Finally, a third source is related to the interaction between the respondents' characteristics and the interviewer's characteristics. The results confirm the effect of the first two sources of influence, as this has already been shown elsewhere in the literature. Regarding the interaction, the results show an effect when the two non-conducive conditions are met, namely, low political interest of the respondents and low quality of the interviewers. When these two conditions are met, the number of non-responses increases significantly. The implications of these results both for the importance of assessing interviewers in the situation and for possible improvement of the procedure are discussed in the conclusion.

Keywords: *interviewer effect, partial non-response, face-to-face survey, interviewer assessment.*

1. Introduction

La non-réponse à une enquête est l'un des éléments les plus critiques pour la qualité des résultats issus d'une enquête (Groves, 2004). On parlera de non-réponse totale lorsque le répondant refuse de répondre ou qu'il n'a pas pu être contacté, alors qu'on parlera de non-réponse partielle lorsque le répondant ne répond pas à une partie de l'enquête. La compréhension de ce qui amène à la non-réponse est d'autant plus importante que celle-ci tend à être plus élevée dans les récentes enquêtes (de Leeuw *et al.*, 2018 ; Groves *et al.*, 2009). Une des conséquences à cela est alors l'augmentation des coûts d'enquête en raison de l'augmentation des efforts que cela implique (relance, rappels, ...) pour atteindre un taux de réponse similaire (Bethlehem *et al.*, 2011 ; Tourangeau *et al.*, 2013) ou un nombre final d'enquêtes exploitables identique. Il faut aussi distinguer la non-réponse partielle de la non-réponse totale car elle n'intervient pas au même moment dans le processus de réponse à l'enquête. La non-réponse partielle intervient une fois que les répondants ont accepté de répondre à l'enquête alors que la non-réponse totale intervient en amont, soit par le refus des répondants à participer soit par l'impossibilité à les contacter. Ainsi, la non-réponse partielle peut globalement être conçue comme le résultat de deux processus décisionnels de la part des répondants au cours de l'enquête (Beatty & Herrmann, 2002). Premièrement, sont-ils capables de produire une réponse valide (ce qui implique de comprendre la question mais aussi de formuler une réponse dans les modalités attendues) et, deuxièmement, ont-ils la volonté de communiquer cette réponse (ce qui implique cette fois la question de désirabilité sociale, de la confiance dans la confidentialité des réponses ou encore de la motivation à répondre à l'enquête) ? Ce processus décisionnel peut bien sûr être influencé par les caractéristiques propres des répondants (tel que leur âge, leur niveau d'éducation ou encore leur intérêt pour l'enquête), mais aussi par les caractéristiques du contexte telles que les compétences de l'enquêteur à mener l'enquête, ou encore par l'interaction entre les deux, à savoir des caractéristiques personnelles du répondant et les compétences de l'enquêteur.

Cette étude se focalise sur les éléments conduisant à la production de non-réponses partielles dans une enquête en face-à-face, en se basant sur le modèle cognitif du processus de réponse développé par Tourangeau *et al.* (2000) et Bradburn *et al.* (2004). Elle évalue dans quelle mesure la qualité des enquêteurs, telle que mesurée en début de terrain sur base de l'enregistrement d'une des premières enquêtes réalisées, est susceptible d'accentuer ou de réduire le recours des répondants à la non-réponse. La non-réponse partielle est ainsi comprise comme le résultat d'un « processus interactionnel » (Dykema *et al.*, 2020) entre les caractéristiques du répondant et celles de l'enquêteur.

1.1 Processus cognitif de formation d'une réponse et rôle du répondant

Tourangeau *et al.* (2000) et Bradburn *et al.* (2004) ont proposé un modèle cognitif du processus de formation d'une réponse à une enquête. Ce modèle distingue quatre processus qui chacun peut conduire à la production d'une non-réponse (« refus » ou « ne sais pas ») si celui-ci ne se déroule pas correctement. Le premier processus cognitif implique que le répondant fasse un travail pour comprendre la question. Si la question est mal posée ou trop ambiguë, le répondant peut répondre par « ne sais pas » ou éventuellement refuser de répondre. Le deuxième processus nécessite que le répondant retrouve des informations qu'il a en mémoire pour se faire un avis ou pour répondre de manière factuelle à la question. Par exemple, si on lui demande pour quel parti il a voté aux dernières élections et qu'il ne s'en souvient plus, il répondra logiquement par « ne sais pas ». Le troisième et le quatrième processus sont fortement liés. Le troisième processus implique que le répondant fasse un travail de reformulation de la réponse qu'il a en tête dans les modalités proposées par l'enquête et le quatrième processus implique qu'il choisisse s'il va ou non exprimer sa réponse. Pour reprendre la proposition de Beatty et Herrmann (2002), il faut distinguer la possibilité qu'ont les répondants de répondre et la volonté qu'ils ont de le faire. L'item sur le positionnement politique gauche-droite qui reçoit toujours un nombre important de non-réponses permet d'illustrer ce phénomène. Dans ce cas, le problème n'est pas la compréhension ou la confidentialité de la réponse qui pose problème mais le fait que, pour un certain nombre de répondants, le positionnement gauche-droite n'est pas une manière pertinente de se situer alors que bon nombre d'acteurs politiques ne se revendiquent plus d'un tel positionnement (Purdam *et al.*, 2020).

Comme nous pouvons le voir, le processus de réponse à une enquête est un phénomène complexe et la non-réponse s'explique différemment selon la partie du processus qui est en jeu. La question qui se pose alors est de savoir quelles sont les caractéristiques du répondant qui amènent à davantage de non-réponses dans ce processus. Les deux principales caractéristiques associées à la non-réponse partielle dans les enquêtes sont le niveau d'éducation et l'âge des répondants (voir, par exemple, Bishop *et al.*, 1986 ; Colsher & Wallace, 1989 ; Schuman & Presser, 1996). Mais les explications du lien entre ces caractéristiques et la non-réponse peuvent être diverses. Une première explication serait que ces caractéristiques seraient associées à une moins bonne compréhension des questions. Une des stratégies mises en place par les personnes avec une faible compréhension serait alors de choisir soit une réponse au hasard, soit de répondre « ne sais pas ». Cette stratégie leur permettrait ainsi de répondre aux attentes de l'enquêteur tout en court-circuitant le processus de compréhension de la question (Krosnick, 1991). Mais cette explication n'est pas la seule possible. Une alternative est donnée par Laurison (2015) qui voit dans la non-réponse des personnes à faible statut socio-économique davantage un choix qu'une impossibilité de le faire. Pour lui, ces personnes sont moins habituées à être sollicitées pour émettre un avis personnel que les personnes de plus haut statut. Prises ainsi dans une situation d'enquête inhabituelle, elles se sentiraient moins légitimes et auraient donc davantage tendance à recourir à la non-réponse pour des questions où on leur demande leur avis personnel, comme les questions politiques. Enfin, une explication importante en lien avec ces caractéristiques individuelles est celle de la motivation. La motivation des répondants est généralement liée à l'aptitude des répondants à soutenir une attention mentale, à la pertinence du sujet de l'enquête pour les répondants, ou encore à la distraction et la fatigue croissante au cours d'un entretien (Krosnick, 1991).

1.2 Le rôle de l'enquêteur

L'effet de l'enquêteur sur la qualité des données est un effet bien connu et qui s'observe notamment par le fait qu'il existe une corrélation plus forte des réponses entre répondants d'un même enquêteur. En d'autres termes, les répondants interrogés par un même enquêteur auraient tendance à davantage se ressembler qu'avec ceux interrogés par d'autres enquêteurs.

Cet effet a été modélisé par certains auteurs dans le cadre de la *Total Survey Error (TSE) framework* (West *et al.*, 2013, 2018 ; West & Blom, 2017). Les auteurs parlent alors de deux biais qui expliquent cet effet. Le premier serait lié au fait que l'enquêteur procéderait à une sélection, bien que non-intentionnelle, des répondants qui répondent à l'enquête (*non-response error*). Ceci s'expliquerait notamment par le fait que les répondants acceptent plus facilement de participer si le profil socio-démographique de l'enquêteur correspond au leur (Vercruyssen *et al.*, 2017). Le deuxième biais interviendrait dans la relation entre l'enquêteur et le répondant. L'enquêteur aurait tendance à accentuer et à atténuer les réponses ou encore à favoriser les non-réponses de manière systématique pour tous les répondants qu'il interroge. En outre, cet effet de l'enquêteur serait encore plus prononcé lorsque la passation se fait en face-à-face plutôt que de manière téléphonique (West *et al.*, 2018).

Purdam *et al.* (2020), qui se sont intéressés aux données de collecte (paradata) de l'enquête *European Social Survey* qui bénéficie d'une forte réputation en matière de qualité des données, ont montré que le nombre de non-réponses partielles était très variable entre les pays mais aussi entre les enquêteurs. Ils ont montré notamment que la variabilité entre pays expliquait 10 % de la variance de ce nombre alors que la variabilité entre enquêteurs expliquait quant à elle 30 % de cette variance. Si les différences entre pays peuvent s'expliquer par des différences dans le design ou dans la manière de délivrer les consignes aux enquêteurs, au final c'est donc bien la manière d'appliquer les consignes de l'enquête qui influence le plus le nombre de non-réponses. Les enquêteurs peuvent ainsi avoir un rôle positif important dans la réduction des non-réponses (Fowler Jr, 1991). Même si la passation de l'enquête est fortement standardisée, ils peuvent expliquer les questions que les répondants ne comprennent pas, ils peuvent s'adapter au rythme de réponse et maintenir l'intérêt de la personne interrogée au fil de l'enquête. Enfin, il semble aussi que l'effet de l'enquêteur soit plus marqué lorsque les répondants ont besoin de davantage d'explications comme, par exemple, les personnes les plus âgées (Beullens *et al.*, 2019).

1.3 L'interaction entre le répondant et l'enquêteur

On l'a vu, la littérature scientifique a bien documenté les caractéristiques individuelles qui influencent la non-réponse partielle par le biais de la compréhension des questions, de la motivation ou encore dans la légitimité perçue des répondants pour y répondre. De même, il existe une vaste littérature sur l'influence systématique des enquêteurs comme biais d'erreur. Cependant, il existe encore peu de travaux qui se sont intéressés aux interactions entre l'enquêteur et le répondant, alors que ce type d'interactions, que ce soit sur le plan verbal ou non-verbal, apparaît comme fondamental dans le cas d'enquêtes en face-à-face. Dykema *et al.* (2020) proposent par exemple un modèle interactionnel dans la séquence question-réponse. Dans ce modèle, la réponse dépend de l'interaction entre les caractéristiques de l'enquête, le processus cognitif du répondant et l'interaction entre le répondant et l'enquêteur. Ceci signifie que l'interaction entre le répondant et l'enquêteur peut aussi bien fonctionner comme facilitateur que comme inhibiteur par rapport à la réponse et la non-réponse. Dans ce cadre, il s'agit donc d'évaluer si le niveau d'expertise des enquêteurs est susceptible d'augmenter ou de réduire le nombre de non-réponses partielles des répondants (qui sont susceptibles de donner des non-réponses).

1.4 Hypothèses

Comme nous l'avons vu plus haut, pour identifier le rôle de l'enquêteur sur l'occurrence des non-réponses, il faut d'abord distinguer les caractéristiques propres au répondant qui influencent les non-réponses, puis la qualité des enquêteurs et enfin l'interaction entre les caractéristiques du répondant et la qualité des enquêteurs. C'est seulement dans ce dernier cas de figure que

l'on pourra dire si l'enquêteur a un effet sur l'augmentation ou la réduction de la non-réponse. De manière plus concrète, nous formalisons ceci en trois hypothèses.

H1) Les caractéristiques des répondants telles que l'âge, le niveau d'éducation ou encore l'intérêt pour la politique sont susceptibles d'influencer le nombre de non-réponses partielles. Le nombre de non-réponses devrait être plus élevé chez les personnes les plus âgées, chez les personnes avec des niveaux d'éducation faibles et chez les personnes qui ont un faible intérêt pour la politique.

H2) La qualité des enquêteurs, déterminée par leur aptitude à mener l'enquête de manière fluide, et la qualité de la relation qu'ils sont capables d'instaurer, influence le nombre de non-réponses indépendamment des caractéristiques des répondants. Toutes choses égales par ailleurs, nous nous attendons à ce que les « meilleurs » enquêteurs obtiennent moins de non-réponses que les « moins bons » enquêteurs.

H3) La qualité des enquêteurs est d'autant plus importante lorsque les caractéristiques du répondant sont propices à produire plus de non-réponses. On s'attend donc à ce qu'il y ait un effet d'interaction entre les caractéristiques du répondant et la qualité de l'enquêteur. On s'attend à ce que cette interaction s'observe particulièrement chez les personnes peu intéressées par la politique. Soit leurs non-réponses augmentent encore lorsqu'elles sont en présence d'un enquêteur faiblement évalué, soit elles diminuent lorsqu'elles sont en présence d'un enquêteur très bien évalué. Dans le premier cas, il y a une accentuation de la caractéristique personnelle (le faible intérêt politique), dans le deuxième cas il y a une réduction de l'effet. Ces deux interactions ne sont pas mutuellement exclusives.

2. Méthode

Le Baromètre social de la Wallonie (BSW) est une enquête d'opinion, en partie politique, réalisée en face-à-face auprès d'un échantillon ($n = 1304$) de la population wallonne francophone. L'échantillon a été tiré à partir du Registre national, qui est la base de données belge qui recense toutes les personnes résidant sur son territoire. Il s'agit d'un tirage aléatoire à deux niveaux. Le premier niveau est celui du code postal. Un tirage au sort de 130 codes postaux est réalisé sur la base des codes postaux des 253 communes francophones wallonnes (9 communes sont germanophones). Ce tirage au sort est réalisé en tenant compte du poids relatif de la population pour chaque code postal. Le deuxième niveau est celui des adresses individuelles. Pour chaque code postal sélectionné, 20 adresses individuelles sont tirées au sort. L'IWEPS sous-traite alors la phase de collecte des données auprès d'un institut de sondage par le biais d'un marché public. L'IWEPS exige cependant que l'institut de sondage ne fasse appel qu'à des enquêteurs qui ont une expérience préalable avec la réalisation d'une enquête similaire. La gestion du déroulement du terrain est laissée à l'institut de sondage mais l'IWEPS se charge de la formation des enquêteurs avant le terrain. C'est aussi lui qui évalue chaque enquêteur en début de terrain sur la base de l'enregistrement d'un des trois premiers entretiens réalisés par l'enquêteur. C'est sur la base de cette évaluation, qui est décrite dans la section suivante, qu'une note est attribuée à chaque enquêteur sur sa capacité à mener correctement les entretiens.

La collecte des données de l'enquête a eu lieu de mars à juin 2018. Elle s'est faite à partir du tirage d'un échantillon d'adresses de 2600 individus (20 adresses x 130 codes postaux). Au total, 58 enquêteurs ont été retenus pour collecter les données. Le taux de réponse (50,15 %) de l'enquête est similaire à celui d'enquêtes antérieures. Le calage de l'échantillon avec la population cible s'est fait en tenant compte du sexe, du niveau d'éducation et de l'âge des répondants. La comparaison a, en effet, montré une légère sous-représentation des personnes avec le niveau d'éducation le plus faible ou le plus élevé dans l'échantillon par rapport à la population cible. Concernant les réponses partielles, aucune imputation n'a été réalisée.

2.1 Évaluation et classification des enquêteurs

Une fois que les enquêteurs ont suivi la formation dispensée par l'IWEPS, ils reçoivent un lot de vingt adresses pour commencer le travail de terrain. Cependant, une évaluation est faite après les trois premiers entretiens. Les enquêteurs doivent enregistrer (à l'aide d'une fonction sur leur laptop) l'une des trois premières enquêtes et le fichier d'enregistrement est transmis à l'IWEPS qui en fait l'évaluation. Cette évaluation permet de s'assurer que les consignes données lors de la formation ont bien été intégrées, elle permet aussi de donner un retour aux enquêteurs pour améliorer la qualité de leurs entretiens ou de demander éventuellement à l'institut de sondage de stopper un enquêteur dont la qualité des entretiens ne serait pas suffisante. C'est sur la base de cette évaluation qu'une note a été attribuée aux enquêteurs. Cette note ne leur était pas directement communiquée. Ils ne recevaient dans leur retour que les points à améliorer et, éventuellement, s'ils devaient soumettre un nouvel entretien enregistré pour poursuivre leur travail. Le tableau suivant présente de manière synthétique les critères attribués pour chaque note, ainsi que la répartition du nombre d'enquêteurs et du nombre d'enquêtes pour chacune d'entre elles.

Tableau 1 – Notes et critères évaluatifs des enquêteurs ; répartition des enquêteurs et des enquêtes réalisées

Note	Critère	Conséquence	Nombre d'enquêteurs	Nombre d'enquêtes réalisées
A+ Très satisfaisant avec mention	Toutes les consignes sont respectées. L'enquêteur établit une très bonne relation de confiance avec la personne enquêtée. Il parvient bien à s'adapter au rythme de la personne interrogée et à son niveau de compréhension. Pas d'interprétation par l'enquêteur des réponses données par la personne interrogée.	L'enquêteur poursuit le terrain.	14	319
A Très satisfaisant	Toutes les consignes sont respectées. Pas d'interprétation par l'enquêteur des réponses données par la personne interrogée. Pas de remarques particulières.	L'enquêteur poursuit le terrain.	20	452
B+ Satisfaisant	Quelques remarques peuvent être formulées quant à la manière de conduire l'entretien et la fluidité de celui-ci (rythme de lecture, etc.). Mais les remarques ne concernent pas des problèmes d'encodage ou d'interprétation des réponses données.	L'enquêteur poursuit le terrain (généralement sans qu'une deuxième évaluation ne soit requise).	20	488
B Juste satisfaisant	Certaines consignes n'ont pas été bien respectées. Problème de risque d'interprétation des résultats ou de risque d'induction de réponse.	L'enquêteur doit tenir compte des remarques formulées et doit fournir un nouvel enregistrement qui sera à nouveau évalué. Il ne pourra poursuivre que si l'évaluation du deuxième enregistrement est meilleure.	4	45
Sous-total des enquêteurs ayant réalisé l'enquête			48	1304

C Non-satisfaisant	Ne tient pas compte des consignes. Trop d'erreurs ou interprétation abusive des réponses données par la personne interrogée.	L'enquêteur est arrêté. Les données récoltées sont supprimées. Éventuellement un autre enquêteur s'occupera de faire repasser l'enquête.	2	0
	Enquêteurs n'ayant pas respecté les délais pour les 3 premières enquêtes ou qui ont abandonné d'eux-mêmes.		7	0
Total des enquêteurs ayant suivi la formation			57	1304

2.2 La non réponse partielle

La non-réponse partielle par enquête est relativement faible dans notre sondage. La moyenne est de 1,9 non-réponse (« ne sait pas » ou « refus ») par enquête pour environ 200 questions que compte l'ensemble du questionnaire. Mais la dispersion, quant à elle, est relativement importante avec un écart-type de 7,59, un minimum de 0 et un maximum de 146. Le tableau suivant présente la distribution du nombre de non-réponses partielles par enquête.

Tableau 2 - Distribution de la non-réponse partielle par enquête

Nombre de non-réponses partielles	Nombre de répondants	%	% cumulés
0	743	57,0	57,0
1	248	19,0	76,0
2	114	8,7	84,7
3	65	5,0	89,7
4	34	2,6	92,3
5	18	1,4	93,7
6-10	36	2,8	96,5
11 et plus	46	3,5	100,0
Total	1304	100,0	100,0

Comme le montre le tableau 2, la majorité des répondants (57,0 %) n'ont aucune non-réponse dans l'enquête. Presque 90 % des répondants n'ont pas plus de 3 non-réponses et 96,5 % des répondants n'en ont pas plus de 10 ; seuls 3,5 % de l'échantillon en ont plus de 10. La distribution n'est donc pas normale mais bien asymétrique. Par la suite, nous effectuerons des régressions logistiques ordinales plutôt que des régressions linéaires afin de tenir compte méthodologiquement de cette asymétrie.

Enfin, afin d'éviter les scores extrêmes de non-réponses et étant donné que l'essentiel des répondants n'ont pas plus de 10 non-réponses, nous avons calculé un score de non-réponse en nous basant uniquement sur les 11 questions qui ont provoqué le plus de non-réponses. Le tableau suivant présente l'intitulé de ces 11 questions ainsi que la répartition des « ne sait pas » et des « refus » pour chacune de ces questions.

Tableau 3 – Liste des 11 questions ayant engendré le taux de non-réponse le plus élevé

Variables	Ne sait pas	Refus	Total non-réponses	% non-réponses (N = 1304)
1) Par le passé, les hommes et femmes politiques étaient davantage à l'écoute des citoyens	139	2	141	10,8
2) Selon vous, les comportements suivants sont-ils fréquents chez les hommes et les femmes politiques ? Payer une prestation en liquide pour ne pas payer de TVA	125	3	128	9,8
3) Selon vous, les comportements suivants sont-ils fréquents chez les hommes et les femmes politiques ? Utiliser ses relations pour conclure un marché en dépit des règles de concurrence	81	3	84	6,4
4) Pour quel parti avez-vous voté à la chambre des représentants aux élections législatives de 2014 ?	41	25	66	5,1
5) À propos de politique, les gens parlent souvent de « gauche » et de « droite ». Personnellement, vous diriez que vous vous situez ...	55	10	65	5,0
6) De nos jours, pensez-vous que les partis politiques sont suffisamment à l'écoute des citoyens... au niveau de la région ?	42	0	42	3,2
7) Pouvez-vous situer, sur l'échelle suivante, le montant mensuel net de votre revenu personnel ?	18	23	41	3,1
8) Les partis politiques font plus de problèmes qu'ils n'en résolvent ?	35	3	38	2,9
9) Le processus de décision démocratique est trop complexe, trop obscur et trop lent	35	3	38	2,9
10) Les partis politiques n'ont plus d'idéologies crédibles	36	2	38	2,9
11) Je pense que je ferais de l'aussi bon travail que la plupart des politiciens que nous avons élus	32	6	38	2,9

Le tableau précédent montre que les trois questions qui ont généré le plus de non-réponses concernent l'évaluation des comportements des hommes et des femmes politiques. Il s'agissait soit de se prononcer sur leurs comportements dans le passé, soit de se prononcer sur leur intégrité. Les répondants semblent avoir eu beaucoup de peine à se positionner pour ces questions en particulier. La quatrième question est plus factuelle puisqu'il s'agissait de se rappeler du vote aux dernières élections. Il y a là sans doute un problème de rappel mais aussi un problème lié à la confidentialité de la réponse puisque la part des « refus » est plus importante par rapport aux questions précédentes. En cinquième position, la question sur le positionnement politique gauche-droite récolte 5,0 % de non-réponses. Il s'agit d'une question qui a déjà été répertoriée par ailleurs comme posant problème à certaines personnes pour se positionner. On remarque enfin que la question sur le revenu n'est « que » à la septième position avec 3,1 % de non-réponses. Ce taux de non-réponses est plutôt faible pour ce type de question. Ici aussi la part de « refus » est plus importante, ce qui souligne le problème de la confidentialité de la réponse à cette question qui peut se poser pour un certain nombre de répondants.

2.3 Analyse croisée des taux de réponse partiels et totaux

Avant d'aborder les analyses statistiques de la Section 3, il est tout d'abord informatif de voir s'il existe un lien, une corrélation, entre le taux de réponse général obtenu par les interviewers et leurs performances au niveau des réponses aux items (taux de réponse partiel). En effet, certaines études ont montré qu'il pouvait exister un lien entre la non-réponse totale et la non-réponse partielle (voir, par exemple, Yan & Curtin, 2010). Les enquêteurs qui parviendraient

davantage à convaincre les répondants (non-réponse totale faible) auraient tendance à avoir plus de non-réponse partielle du fait de répondants moins motivés. Pour savoir si nous devons prendre en compte le taux de réponse total dans notre analyse, nous avons donc vérifié le lien entre les deux dans le cadre de notre enquête de référence BSW. Pour cela, nous avons repris les 11 questions qui ont engendré le plus de non-réponses (Tableau 3) et avons calculé une moyenne du nombre de non-réponses par enquête menée à bien pour chacun des interviewers. Notons que nous ne considérons ici que les interviewers qui ont été conservés, soit ceux qui sont notés de A+ à B. Pour obtenir un taux de réponse total, nous avons fait le rapport entre le nombre d'enquêtes effectivement menées à bien et le nombre d'adresses reçues en début de processus. Là encore, seuls les interviewers bien notés sont pris en compte (ces derniers ont par ailleurs récupéré des lots d'adresses attribués initialement à des interviewers qui ont été plus tard écartés). En mettant de côté les six observations les plus aberrantes, nous obtenons le graphique de la Figure 1.

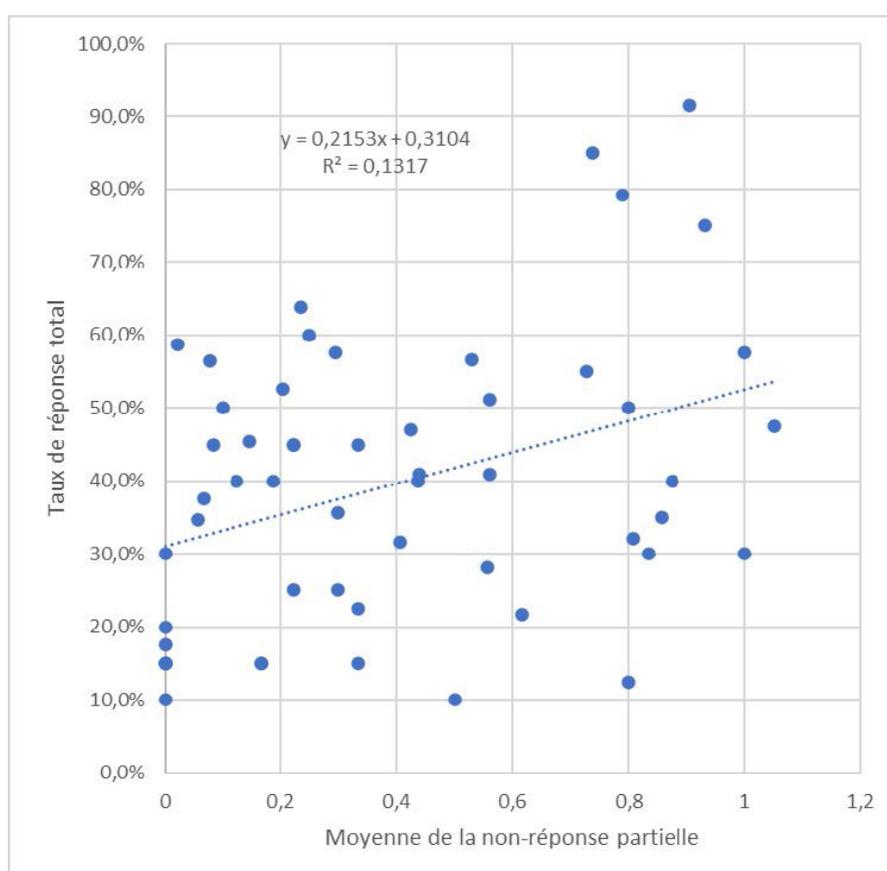


Figure 1 – Lien entre le taux de réponse total et la non-réponse partielle (données BSW)

Nous constatons dans ce graphique que le coefficient de détermination (R^2 , qui est équivalent au carré du coefficient de corrélation dans le cas d'une régression linéaire simple) entre les deux variables est de 0,1317. Il s'agit donc d'une relation positive mais très faible.

Par ailleurs, un test statistique de significativité du coefficient attribué à la variable explicative fournit une p-valeur égale à 0,4322. Une telle valeur implique un non-rejet de l'hypothèse nulle d'absence d'effet. Nous pouvons donc dire que, dans notre cas, il n'y a pas de lien statistiquement significatif entre la non-réponse partielle et la non-réponse totale. En d'autres termes, il n'est pas ici prouvé statistiquement qu'un interviewer qui obtient un meilleur taux de réponse total se retrouve avec plus de répondants qui ne répondent pas à certaines questions. Par conséquent, nous décidons de mettre de côté le taux de réponse total dans les analyses de la Section 3.

3. Résultats

Afin de tester nos hypothèses de départ, nous avons réalisé une analyse statistique multivariée pour expliquer le score de non-réponses (basé sur les 11 questions identifiées plus haut). Les variables explicatives ont été introduites dans l'ordre suivant : les variables socio-démographiques (sexe, niveau d'éducation, âge, lieu d'habitation), l'intérêt pour la politique (mesuré selon une échelle de Likert en quatre points), l'évaluation des enquêteurs, puis finalement les interactions entre l'intérêt pour la politique et l'évaluation des enquêteurs. La variable dépendante étant discrète ordinale avec 12 valeurs possibles différentes (allant de 0 à 11 non-réponses) et les variables explicatives étant des facteurs catégoriels, le choix de l'outil statistique approprié s'est porté sur une régression logistique ordinale.

Pour chaque variable, une modalité de référence a été choisie et est représentée en premier dans le graphique ci-dessous. Pour chaque variable, les modalités de référence correspondent aux modalités les plus fréquemment rencontrées. Le graphique représente visuellement les résultats de la régression logistique en présentant la valeur de l'Odd Ratio (OR) de chacune des modalités de la variable.

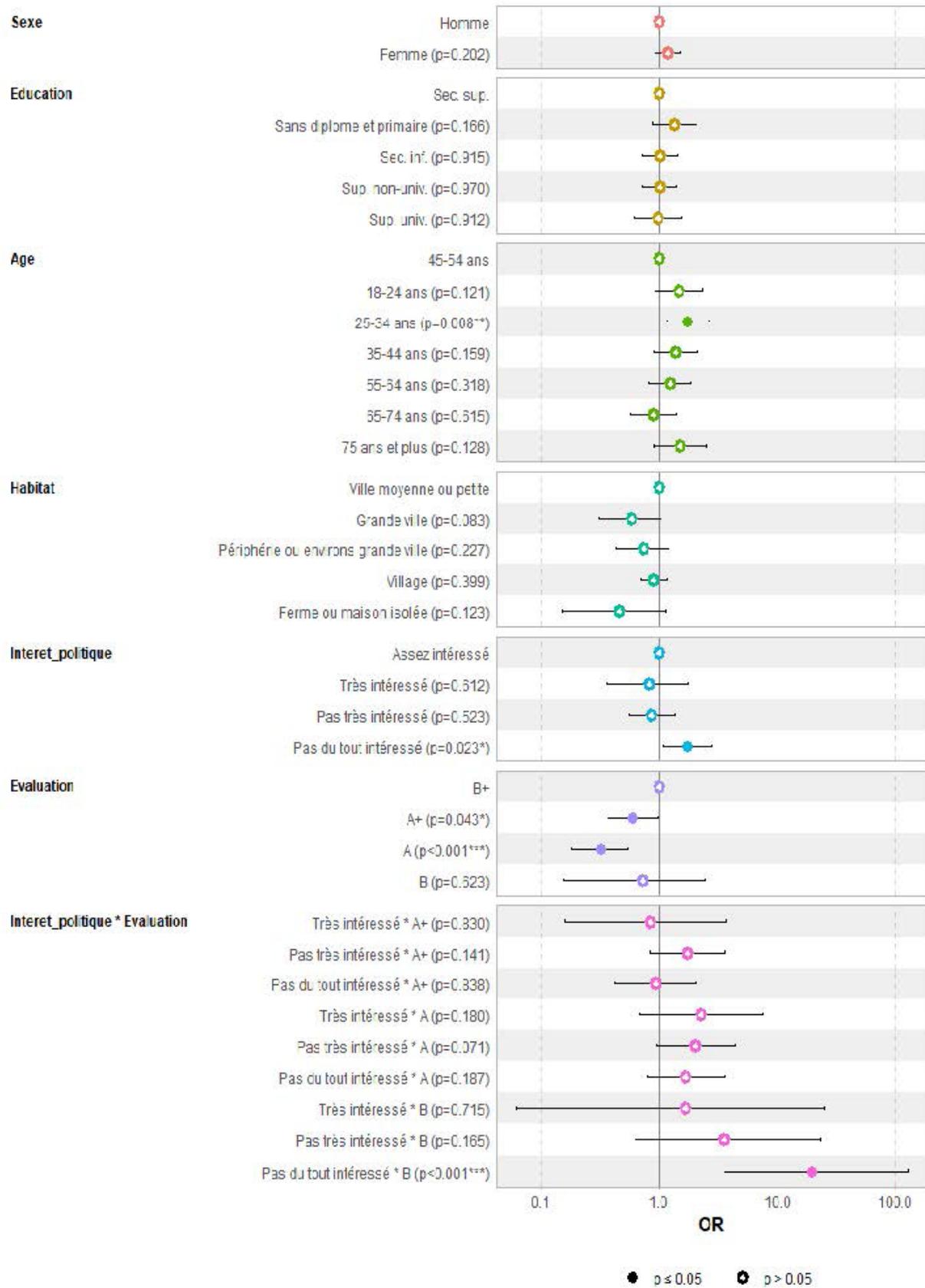


Figure 2 – Résultats graphiques de la régression logistique ordinaire pour le score des non-réponses

Comme on peut le voir dans la Figure 2, les variables socio-démographiques ont finalement peu d'effet sur les non-réponses. Concernant le sexe, on n'observe pas de différence. Concernant le niveau d'éducation, on observe que les personnes sans diplôme ou de niveau primaire présentent une légère tendance à avoir plus recours aux non-réponses (cependant, la différence n'est pas statistiquement significative ; $p=.166$) par rapport au secondaire supérieur (pris en référence). À l'opposé, un diplôme de niveau supérieur ne permet pas non plus de réduire les non-réponses. Concernant l'âge, il est intéressant de voir que la catégorie 25-34 ans produit plus de non-réponses que les 45-54 ans, pris en référence ($p=.008$). Cela est-il dû à une moins grande disponibilité de ces répondants pour réaliser l'enquête ? En revanche, la catégorie qui a légèrement tendance à moins produire de non-réponses est celle des 65-74 ans. On pourrait postuler ici, a contrario, que cette catégorie d'âge a une plus grande disponibilité pour répondre à l'enquête. Enfin, alors que l'on s'attendait dans notre hypothèse de base à ce que les personnes les plus âgées produisent plus de non-réponses, on observe ici que la différence pour les 75 ans et plus n'est pas significative ($p=.128$). Enfin, concernant la dimension rural-urbain du lieu d'habitation, on n'observe pas non plus de différences.

À propos de l'intérêt pour la politique, qui est aussi une caractéristique propre au répondant, on observe un effet de la variable qui va dans le sens de notre première hypothèse. En effet, les personnes « pas du tout intéressées » produisent statistiquement plus de non-réponses que les personnes assez intéressées qui sont la modalité de référence ($p=.023$). En revanche, les très intéressés ne produisent pas moins de non-réponses par rapport à la modalité de référence ($p=.612$). Les personnes les moins intéressées par la politique (sujet de l'étude) ont donc bien tendance à produire davantage de non-réponses ce qui va dans le sens d'un effet de la motivation lié au sujet même de l'enquête sur le nombre de non-réponses partielles.

Au niveau de la qualité des enquêteurs, indépendamment des caractéristiques des répondants, on observe là aussi un effet. Les enquêteurs évalués comme très satisfaisants (A) ou très satisfaisants avec mention (A+) induisent globalement statistiquement moins de non-réponses que les enquêteurs satisfaisants (B+) (pris en référence). On peut donc dire que l'effet de l'enquêteur a pour résultat de réduire le nombre de non-réponses partielles. En revanche, les enquêteurs juste satisfaisants (B) ne produisent pas plus de non-réponses que les satisfaisants. L'effet enquêteur est donc davantage lié aux bons enquêteurs qui parviennent à réduire les non-réponses plutôt qu'aux mauvais qui les augmentent.

Au niveau de l'interaction entre l'intérêt pour la politique du répondant et la qualité de l'enquêteur, on observe un effet d'accentuation de la qualité des enquêteurs mais pas un effet de réduction. En effet, en situation où la qualité de l'enquêteur est faible, les non-réponses augmentent encore chez les personnes dont l'intérêt politique est faible (l'effet est largement significatif avec $p<.001$). Par contre, lorsque la qualité de l'enquêteur est élevée, cela ne réduit pas les non-réponses des personnes avec un faible intérêt politique. On peut donc dire que lorsque le répondant manifeste un faible intérêt pour la politique, un bon enquêteur ne permettra pas de réduire les non-réponses mais un mauvais enquêteur, quant à lui, les augmentera de manière très significative !

4. Discussion et conclusion

Cette étude vise à présenter l'effet des enquêteurs sur la qualité des données d'une enquête en se basant sur le nombre de non-réponses partielles produites par les répondants. Elle propose une manière relativement innovante pour s'assurer de la qualité des enquêteurs qui consiste à évaluer la passation d'une enquête en situation grâce à l'enregistrement préalable de celle-ci. Il s'agit d'une procédure relativement coûteuse en ressources (il s'agit d'écouter et d'évaluer chacun des enquêteurs en début de terrain) mais qui permet de s'assurer que les consignes données lors de la formation sont bien prises en compte et respectées.

Les résultats de cette étude encore exploratoire méritent d'être étendus à plus grande échelle afin de vérifier si ceux-ci sont généralisables. L'enquête internationale European Social Survey est, à ce titre, peut-être une bonne source puisque certaines vagues ont fait l'objet de données collectées sur la passation.

Au niveau des résultats, on retiendra les éléments suivants, à mettre en lien avec les hypothèses de départ formulées dans la Section 1.3. On observe clairement un effet des enquêteurs sur la qualité des données récoltées. Les enquêteurs qui suivent bien toutes les consignes et pour lesquels l'enquête se déroule de manière fluide (évaluation A+ et A) ont moins de non-réponses partielles que les enquêteurs dont la fluidité des premières enquêtes n'était pas optimale, mais acceptable (B+). En revanche, on n'observe pas de différence entre les bons enquêteurs (A) et les très bons enquêteurs (A+).

Nous pensons que certaines caractéristiques propres aux enquêtés pouvaient faire augmenter, de façon directe et indépendamment de la qualité des interviewers, le nombre de non-réponses. L'effet est prouvé pour ce qui est de l'intérêt politique : un désintérêt total augmente effectivement les non-réponses (toutes choses égales par ailleurs). Par contre, l'âge et le niveau d'éducation n'entraînent que des tendances et non un effet statistiquement significatif. Ainsi, un âge élevé (75 ans et plus) et un niveau d'éducation faible (sans diplôme et primaire) sont des caractéristiques qui ont simplement tendance à provoquer plus de non-réponses partielles. Nous observons par contre un effet significatif à la hausse des non-réponses pour la tranche d'âge des 25-34 ans.

On observe un effet d'interaction entre la qualité des enquêteurs et des caractéristiques propres au répondant tel que l'intérêt pour la politique qui est le sujet principal de l'enquête. Ces effets s'observent surtout lorsque l'intérêt est très faible et que l'enquêteur est évalué comme juste satisfaisant (B). Dans ce cas, l'interaction entre le désintérêt du participant et probablement une forme de désinvestissement de l'enquêteur conduisent à une augmentation importante de la non-réponse partielle. Dans cette situation, le répondant semble davantage faire le choix de ne pas répondre aux questions plutôt que de donner son avis. À l'opposé, l'excellence des enquêteurs (A ou A+) ne permet pas de réduire l'effet du désintérêt sur les non-réponses partielles comme cela avait aussi été supposé. On observe donc que la faible qualité des enquêteurs aggrave l'effet du désintérêt des participants sur la non-réponse mais qu'à contrario l'excellence des enquêteurs ne permet pas de le réduire.

Ce dernier point, mettant en avant un effet d'interaction évident, montre à quel point la réussite d'une enquête s'apparente à un processus co-construit (Dykema *et al.*, 2020) qui fait appel aussi bien à la qualité des interviewers qu'à la motivation des personnes contactées. Ce processus co-construit doit idéalement permettre la rencontre entre la motivation et le devoir de réponse d'une part (Japéc, 2008), et la conscience professionnelle d'autre part.

Les conséquences concrètes pour la collecte des données sont les suivantes. Nous pensons qu'il est utile d'évaluer la passation de chaque enquêteur en début de terrain. C'est un moyen très informatif pour s'assurer qu'ils suivent bien les consignes qui ont été données et donc pour voir s'ils peuvent poursuivre ou non le terrain. Dans le cadre de cette étude, deux enquêteurs sur les 57 qui avaient suivi la formation ont été exclus du terrain suite à des passations d'enquête insuffisantes. Ce chiffre peut être considéré comme faible mais si cette procédure n'avait pas été mise en place ils auraient échappé à l'évaluation de la qualité de données.

Dans la majeure partie des cas, les enquêteurs réalisent un bon travail et il n'y aurait pas lieu de distinguer les très bons (A+) des bons (A) enquêteurs car on n'observe pas de différence au niveau de la qualité des données qu'ils récoltent. La procédure d'évaluation pourrait ainsi

être allégée et raccourcie lorsque la passation ne pose pas de problème majeur (comme par exemple une écoute partielle de l'enquête).

Il semble par contre qu'une plus faible fluidité de la passation de l'enquête (B+) impacte la qualité de l'enquête. Il est donc préférable de veiller à ce que les enquêteurs ayant des problèmes à mener de manière fluide l'enquête, même s'il n'y a pas de problème d'encodage, ne fassent pas trop d'enquêtes, voire soient retirés du terrain.

En revanche, s'il y a des problèmes dans la fluidité de l'enquête et dans l'encodage (évaluation B), cela peut réduire la qualité des données des répondants qui ont déjà des prédispositions (faible intérêt) à fournir plus de non-réponses. Il est donc important de stopper ces enquêteurs ou à tout le moins de s'assurer qu'ils fassent le moins d'enquêtes possible.

Références

Beatty P. and D. Herrmann (2002), « To answer or not to answer: Decision processes related to survey item nonresponse », *Survey Nonresponse*, 71, p. 86.

Bethlehem J., F. Cobben, and B. Schouten (2011), *Handbook of nonresponse in household surveys*, John Wiley & Sons.

Beullens K., G. Loosveldt, and C. Vandenplas (2019), « Interviewer Effects Among Older Respondents in the European Social Survey », *International Journal of Public Opinion Research*, 31(4), pp. 609-625, <https://doi.org/10.1093/ijpor/edy031>.

Bishop G. F., A. J. Tuchfarber, and R. W. Oldendick (1986), « Opinions on fictitious issues: The pressure to answer survey questions », *Public Opinion Quarterly*, 50(2), pp. 240-250.

Bradburn N. M., S. Sudman, and B. Wansink (2004), *Asking questions: The definitive guide to questionnaire design—for market research, political polls, and social and health questionnaires*, John Wiley & Sons.

Colsher P. L. and R. B. Wallace (1989), « Data quality and age: Health and psychobehavioral correlates of item nonresponse and inconsistent responses », *Journal of Gerontology*, 44(2), pp. 45-52.

de Leeuw E., J. Hox, and A. Luiten (2018), « International nonresponse trends across countries and years: An analysis of 36 years of labour force survey data », *Survey Methods: Insights from the Field*, pp. 1-11, <https://surveyinsights.org/?p=10452>

Dykema J., N. C. Schaeffer, D. Garbarski, and M. Hout (2020), « The role of question characteristics in designing and evaluating survey questions », *Advances in Questionnaire Design, Development, Evaluation and Testing*, pp. 117-152.

Fowler Jr F. J. (1991), « Reducing interviewer-related error through interviewer training, supervision, and other means », in *Measurement errors in surveys*, Wiley Online Library, pp. 259-278.

Groves R. M. (2004), *Survey errors and survey costs*, John Wiley & Sons.

Groves R. M., F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009),

Survey Methodology, John Wiley & Sons.

Japac L. (2008), « Interviewer Error and Interviewer Burden », in *Advances in Telephone Survey Methodology*, John Wiley & Sons, pp. 185-211, <https://doi.org/10.1002/9780470173404.ch9>

Krosnick J. A. (1991), « Response strategies for coping with the cognitive demands of attitude measures in surveys », *Applied Cognitive Psychology*, 5(3), pp. 213-236.

Laurison D. (2015), « The Willingness to State an Opinion: Inequality, Don't Know Responses, and Political Participation », *Sociological Forum*, 30(4), pp. 925-948, <https://doi.org/10.1111/socf.12202>

Purdam K., J. Sakshaug, M. Bourne, and D. Bayliss (2020), « Understanding 'Don't know' answers to survey questions—An International comparative analysis using interview paradata », *Innovation*, <https://doi.org/10.1080/13511610.2020.1752631>

Schuman H. and S. Presser (1996), *Questions and answers in attitude surveys: Experiments on question form, wording, and context*, Sage.

Tourangeau R., F. G. Conrad, and M. P. Couper (2013), *The science of web surveys*, Oxford University Press.

Tourangeau R., L.J. Rips, and K. Rasinski (eds.) (2000), *The psychology of survey response*, Cambridge University Press, <https://doi.org/10.1017/CBO9780511819322>

Vercruyssen A., C. Wuyts, and G. Loosveldt (2017), « The effect of sociodemographic (mis)match between interviewers and respondents on unit and item nonresponse in Belgium », *Social Science Research*, 67, pp. 229-238, <https://doi.org/10.1016/j.ssresearch.2017.02.007>

West B. T. and A. G. Blom (2017), « Explaining interviewer effects: A research synthesis », *Journal of Survey Statistics and Methodology*, 5(2), pp. 175-211, <https://doi.org/10.1093/jssam/smw024>

West B. T., F. G. Conrad, F. Kreuter, and F. Mittereder (2018), « Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing », *Journal of Survey Statistics and Methodology*, 6(3), pp. 335-359, <https://doi.org/10.1093/JSSAM/SMX029>

West B. T., F. Kreuter, and U. Jaenichen (2013), « 'Interviewer' effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? », *Journal of Official Statistics*, 29(2), pp. 277-297.

Yan T. and R. Curtin (2010), « The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective », *International Journal of Public Opinion Research*, 22(4), pp. 535-551.
<https://doi.org/10.1093/ijpor/edq037>

Abandons dans une enquête sur internet : l'exemple de l'inclusion dans la cohorte Coset-MSA



Noémie SOULLIER¹

Santé publique France, Chargée de projet organisation et suivi des enquêtes

Hugo ROGIE²

Santé publique France, Stagiaire

Guilhem DESCHAMPS³

Santé publique France, Data scientist

Jean-Luc MARCHAND⁴

Santé publique France, Chef de projet COSET-Indépendants

Béatrice GEOFFROY-PEREZ⁵

Santé publique France, Cheffe de projet COSET-MSA

TITLE

Dropouts of Web surveys: lessons from the recruitment in the French Coset-MSA cohort study

RÉSUMÉ

Les enquêtes par internet sont de plus en plus utilisées y compris dans le domaine de la santé. Elles présentent l'avantage de pouvoir enquêter de nombreuses personnes à faible coût et de disposer de données enregistrées automatiquement qui peuvent éclairer sur les comportements de réponse. Nous nous proposons d'explorer les abandons en cours de questionnaire lors du recrutement de la cohorte française Coset-MSA, à la fois en fonction des profils des répondants mais également de leurs données de connexion. Le programme Coset a pour objectif de décrire et de surveiller l'état de santé de la population selon l'activité professionnelle en France. Entre fin 2017 et début 2018, 270 000 actifs âgés de 18 à 65 ans affiliés à la Mutualité Sociale Agricole (MSA) en 2016 ont été invités à remplir un questionnaire en ligne. À la fin de la collecte, 28 054 personnes avaient répondu au moins aux questions d'identité (sexe et âge) et parmi celles-ci, 2 004 personnes avaient abandonné le remplissage de leur questionnaire avant la fin. Le taux d'abandon était associé à l'âge, à l'état de santé perçu et au niveau d'études. L'abandon était également associé aux données de connexion : plus fréquent parmi les personnes qui s'étaient

1. noemie.soullier@santepubliquefrance.fr

2. hugo.rogie@hotmail.fr

3. guilhem.deschamps@santepubliquefrance.fr

4. Jean-Luc.MARCHAND@santepubliquefrance.fr

5. Beatrice.GEOFFROY-PEREZ@santepubliquefrance.fr

connectées une seule fois, celles qui ne s'étaient jamais connectées le week-end, celles qui s'étaient connectées pour la première fois la nuit et celles qui s'étaient connectées pour la première fois après la seconde relance. Ces parodonnées, disponibles à faible coût pour tous les individus, se révèlent pertinentes pour documenter les abandons lorsqu'on a peu de données auxiliaires. Ce travail offre un nouvel angle de vue sur l'analyse des parodonnées au regard des abandons, forme particulière de non-réponse partielle. Ces résultats pourront servir à orienter la correction de la non-réponse partielle, mais également à améliorer le questionnaire et le protocole de contact.

Mots-clés : *internet, enquête, santé, abandon, parodonnées.*

ABSTRACT

Internet surveys are developing, including in the health field. Internet makes it easy to survey many people at low cost and provides automatically recorded data, which can shed light on response behaviors. We intend to explore the questionnaire drop-outs during the recruitment of the French Coset-MSA cohort, both according to the respondents' characteristics but also according to their connection data. The Coset program aims to describe and follow the health status of the French population according to professional characteristics. Between late 2017 and early 2018, 270,000 working people aged 18 to 65 years and affiliated to the Mutualité Sociale Agricole (MSA) in 2016 were invited to complete an online questionnaire. At the end of the collection, 28,054 people had answered at least the identity questions (sex and age) and among these, 2,004 people had stopped their questionnaire before the end. The drop-out rate was associated with age, perceived health status and education. Drop-out was also associated with connection data: it was more frequent among those who had logged in only once, those who had never logged in on weekends, those who had logged in for the first time at night and those who had logged in for the first time after the second reminder. These paradata, available at low cost for all individuals, prove to be relevant for documenting questionnaire drop-outs when there is few auxiliary data. This work offers a new perspective on the analysis of paradata with regard to questionnaire drop-outs, a particular form of item non-response. These results could be used to guide the correction of item non-response, but also to improve the questionnaire and the contact protocol.

Keywords: *Web, survey, health, drop-out, paradata.*

1. Introduction

En lien avec l'augmentation de l'équipement et de l'utilisation d'internet ces dernières décennies (Legleye *et al.*, 2022), les enquêtes par internet sont de plus en plus utilisées y compris dans le domaine de la santé. Avec le développement de l'accès à internet, la couverture de ces enquêtes est désormais importante (Crouette & Muller, 2021), même si une part de la population reste éloignée d'internet en raison d'un accès lent ou par choix, soit en raison d'une crainte pour la protection de ses données, soit parce qu'elle tire peu de bénéfice d'internet (Felderer & Herzing, 2022). Les enquêtes par internet présentent par ailleurs l'avantage de pouvoir contacter de nombreuses personnes à faible coût (Couper, 2000).

La non-réponse aux enquêtes est classiquement divisée en non-réponse totale (la personne ne répond à aucune question) et non-réponse partielle (la personne ne répond pas à certaines questions), toutes deux étant des indicatrices de qualité d'une enquête (Groves, 1989). La collecte par internet souffre d'un taux de non-réponse totale plus important que pour les autres modes de collecte (Daikeler *et al.*, 2020), quand le taux de non-réponse partielle est quant à lui plutôt similaire entre les modes (Čehovin *et al.*, 2022). Ces deux types de non-réponses sont le résultat de deux mécanismes différents, puisque dans le premier cas la décision est prise avant de commencer le questionnaire, alors que dans le second cas elle est prise au cours du questionnaire. Cependant, il semble exister une relation positive entre les deux au niveau individuel : par exemple, les personnes les plus réticentes à participer à l'enquête sont également plus susceptibles de ne pas répondre à certaines questions (Couper, 1997). Ainsi, les personnes se placeraient sur un *continuum*, alliant leur propension à répondre à l'enquête et à chaque question de celle-ci (Yan & Curtin, 2010). Dans le cadre des enquêtes longitudinales, les individus qui ont émis une non-réponse partielle plus importante à une vague d'enquête sont moins susceptibles de répondre à la vague suivante (Loosveldt *et al.*, 2002).

Tout comme la non-réponse totale, la non-réponse partielle peut être source de biais si les non-répondants diffèrent des répondants et que seuls les répondants sont pris en compte dans l'estimation. Des méthodes d'imputation prenant en compte les caractéristiques des individus peuvent permettre de résoudre ce problème, sous l'hypothèse que les données sont manquantes au hasard conditionnellement aux caractéristiques prises en compte (de Leeuw *et al.*, 2003). Les enquêtes par internet permettent, grâce à des données supplémentaires enregistrées automatiquement telles que les données de connexion (paradonnées), de définir plus finement les comportements de non-réponse (Bosnjak & Tuten, 2001), mais également d'en éclairer la compréhension (Kreuter, 2013).

Nous nous proposons d'explorer les abandons en cours de questionnaire lors du recrutement de la cohorte française Coset-MSA. Ces abandons seront étudiés à la fois en fonction des profils des répondants, mais également de leurs données de connexion. Ces résultats permettront d'identifier les facteurs associés à l'abandon en cours de questionnaire, afin d'orienter la correction de la non-réponse partielle mais également de fournir des éléments pouvant servir à la compréhension de la non-réponse totale aux vagues suivantes, voire à la mise en place de protocoles ciblés.

2. Matériel et méthodes

Le programme Coset (Cohortes pour la Surveillance Epidémiologique en lien avec le Travail) est un dispositif longitudinal pour la surveillance épidémiologique en lien avec le travail (Geoffroy-Perez *et al.*, 2012), s'appuyant sur des données de cohortes concernant les actifs français. Il a pour objectif de surveiller l'état de santé de la population selon l'activité professionnelle en France. Dans ce cadre, deux cohortes ont été constituées par Santé publique France : Coset-Indépendants et Coset-MSA, dont les populations cibles sont les actifs âgés de 18 à 65 ans qui étaient affiliés respectivement au Régime Social des Indépendants et à la Mutualité Sociale Agricole (MSA) en 2016.

Ce travail porte sur le recrutement de la cohorte Coset-MSA par un questionnaire en ligne. Le recrutement de cette cohorte visait à inclure 30 000 cohortistes. Pour cela, 270 000 personnes ont été tirées au sort dans les bases de la MSA et invitées à participer. Les invitations, envoyées en plusieurs vagues, ont commencé en novembre 2017, sous la forme d'un courrier postal envoyé au domicile des personnes. Le courrier comprenait un identifiant et un mot de passe uniques, qui permettaient aux personnes de s'authentifier afin de répondre au questionnaire en ligne. Les personnes n'ayant pas répondu recevaient une première relance postale, puis une deuxième si elles n'avaient toujours pas répondu. Les dernières relances postales ont été envoyées fin mai 2018. L'accès au questionnaire en ligne a été ouvert le 20 novembre 2017 et fermé le 11 juillet 2018. À cette date, 28 054 personnes avaient rempli les questions obligatoires situées sur la première page du questionnaire (sexe et âge) et leurs réponses étaient cohérentes avec l'identité de la personne tirée au sort et invitée.

Le questionnaire se composait des grandes sections suivantes :

- Santé,
- Habitudes et cadre de vie,
- Activité professionnelle actuelle,
- Historique professionnel (autres activités et arrêts de travail).

Pour la partie concernant l'historique professionnel, il était demandé à la personne de renseigner et décrire chaque épisode professionnel (type d'activité, profession, période d'exercice, conditions de travail) et chaque arrêt de travail qu'elle a eu au cours de sa carrière professionnelle. Pour les personnes n'ayant pas eu d'autre emploi que leur emploi actuel ou

n'ayant jamais eu d'arrêt de travail au cours de leur carrière, ces parties n'étaient pas à remplir. L'étude de cette section montre qu'elle a été vraisemblablement mal remplie de manière générale, avec une sous-déclaration du nombre d'épisodes professionnels (1,5 en moyenne contre 2,4 déclarés par questionnaire papier lors de l'étude pilote menée sur un échantillon de 2 363 répondants).

Ainsi, afin de pouvoir bien distinguer les abandons en cours de questionnaire, on s'est intéressé dans cette analyse uniquement à la partie linéaire du questionnaire, c'est-à-dire sans prendre en compte la partie sur l'historique professionnel (voir tableau 1). Sur cette partie contenant 22 pages, le nombre minimal de questions posées à une personne était de 85 et le nombre maximal 267. Il est à noter que les pages 13 à 22 concernaient uniquement les personnes ayant déclaré être en activité au moment du remplissage du questionnaire ($n = 26\ 847$, 96%) et les pages 20 à 22 uniquement celles ayant déclaré une activité professionnelle actuelle qui n'était pas une activité de bureau ($n = 20\ 285$, 72%).

Dans ce travail, on considère qu'une personne a abandonné définitivement le remplissage de son questionnaire à la page p si elle n'a répondu à aucune des questions qui lui étaient posées à partir de la page p et jusqu'à la page 22.

Les variables étudiées pour l'association avec l'abandon étaient :

- des variables de l'auto-questionnaire rempli par la personne : sexe, âge, état de santé perçu, taille et poids (utilisés pour le calcul de l'indice de masse corporelle), les antécédents de cancer et le niveau d'études le plus élevé atteint par la personne ;
- des variables de la base de sondage : statut professionnel (salarié / non-salarié) et zone géographique (métropole / départements et régions d'outre-mer) au moment du tirage au sort (au dernier trimestre 2017) ;
- des variables de connexion au questionnaire (paradonnées) : une variable couplant le nombre de connexions au délai entre les connexions (une seule connexion / plusieurs connexions le même jour / plusieurs connexions en plusieurs jours), une variable indiquant s'il y avait eu au moins une connexion le week-end, une variable indiquant le créneau horaire de la première connexion (matin 6h-12h / midi 12h-14h / après-midi 14h-18h / soirée 18h-22h / nuit 22h-6h), et une variable indiquant après quel courrier avait eu lieu la première connexion (courrier d'invitation / première relance / seconde relance).

Les différences de taux d'abandon pour une caractéristique donnée ont été testées au moyen du test du khi-deux de Pearson. Des régressions logistiques multivariées ont été réalisées pour analyser l'abandon en fonction de plusieurs variables explicatives. Les analyses ont été effectuées avec le logiciel SAS® version 9.4.

Tableau 1 – Présentation des différentes pages de la partie linéaire du questionnaire

Page	Description de la page	Nombre de questions sur la page (min-max)
Page 1 - Santé générale 1/3	Sexe, âge, biométrie, état de santé générale perçu	7
Page 2 - Santé générale 2/3	Antécédents cardiovasculaires, métaboliques, psychiques	7-16
Page 3 - Santé générale 3/3	Antécédents de cancer	1-7
Page 4 - Santé respiratoire	Volet santé respiratoire de l'enquête <i>European Community Respiratory Health Survey</i> (ECRHS), autres antécédents de maladie respiratoire, allergie nasale	15-32

Page 5 - Santé musculo articulaire	Symptômes musculo-articulaires au cours des douze derniers mois	11-32
Page 6 - Santé - Moral	Symptomatologie dépressive (échelle <i>Center for Epidemiologic Studies- Depression (CES-D)</i>)	20
Page 7 - Santé - Autres	Maladies infectieuses au cours des 12 derniers mois, troubles de l'audition, eczéma de contact, autres problèmes de santé, recours aux soins, observance	13-15
Page 8 - Cadre de vie	Situation familiale et composition du foyer, situation professionnelle du conjoint	4-9
Page 9 - Rythme de vie	Chronobiologie, sommeil, rythme alimentaire	3
Page 10 - Habitudes vie - Tabac	Consommation de tabac actuelle et passée, utilisation de la cigarette électronique	1-11
Page 11 - Habitudes vie - Alcool	Consommation d'alcool : Audit C abrégé	1-4
Page 12 - Activité professionnelle - Situation actuelle	Niveau d'études, situation professionnelle actuelle	2-7
Page 13 - Activité professionnelle - Description 1/2	Volume horaire, polyactivité, statut et description de l'activité professionnelle actuelle principale (ou dernière activité professionnelle pour les inactifs)	0-18
Page 14 - Activité professionnelle - Description 2/2	Description des cultures produites et des animaux d'élevage concernés par cette activité	0-29
Page 15 - Activité professionnelle - Déplacements	Déplacements et horaires de travail (travail de nuit, rythme de travail, organisation du travail)	0-17
Page 16 - Activité professionnelle - Bien-être	Contact avec le public, déséquilibre efforts/ récompenses (<i>Effort-Reward Imbalance (ERI)</i>)	0-18
Page 17 - Activité professionnelle - Efforts 1	Efforts et contraintes physiques au travail : pénibilité physique (échelle de Borg), contraintes musculosquelettiques	0-16
Page 18 - Activité professionnelle - Efforts 2	Contraintes musculosquelettiques : postures	0-8
Page 19 - Activité professionnelle - Bruits	Exposition aux bruits	0-3
Page 20 - Activité professionnelle - Autres expositions 1	Entretien de machines, de bâtiments, utilisation de solvants, matériaux d'isolation, de construction, de peinture, de soudage en rapport avec cet entretien	0-20
Page 21 - Activité professionnelle - Autres expositions 2	Désinfection de bâtiments ou de matériel, activités de brûlage, exposition aux poussières	0-10
Page 22 - Activité professionnelle - Autres expositions 3	Utilisation et/ou application de produits phytopharmaceutiques sur les cultures ou les animaux	0-20

3. Résultats

Au total, 2 004 personnes (7%) ont abandonné le remplissage avant la fin du questionnaire. Parmi ces abandons, 43% ont eu lieu dans la partie santé, 9% dans la partie habitudes et cadre de vie, et 48% dans la partie activité professionnelle actuelle.

Les abandons les plus fréquents ont été observés aux pages 5 (Santé musculaire et articulaire)

($n = 163$, 8% des abandons), 6 (Santé-Moral) ($n = 349$, 17% des abandons), 13 (Description de l'activité professionnelle 1/2) ($n = 380$, 19% des abandons) et 14 (Description de l'activité professionnelle 2/2) ($n = 199$, 10% des abandons). Ainsi, les pages 5-6 de la partie santé totalisaient 25% des abandons et les pages 13-14 de la partie professionnelle totalisaient 29% des abandons. Les pages 5 et 6 comprenaient en moyenne une vingtaine de questions (respectivement 22 et 20) ; les pages 13 et 14 étaient les premières pages s'intéressant en détail à l'activité professionnelle actuelle et comprenaient en moyenne respectivement 14 et 8 questions.

L'âge et le sexe, renseignés sur la première page du questionnaire, sont des données nécessaires à la définition de la population d'étude (vérification de l'identité du participant). Par conséquent, il n'y a pas de données manquantes pour ces variables, ni d'abandon à la page 1. En revanche, pour les autres caractéristiques étudiées, des données manquantes existent : l'analyse se fait parmi les valeurs renseignées et exclut donc une partie des abandons (ceux qui ont eu lieu avant la page sur laquelle la caractéristique est renseignée et ceux qui ont eu lieu après mais pour lesquelles la réponse à la question est manquante).

Le taux d'abandon était identique selon le sexe (Tableau 2). Cependant, les hommes abandonnaient significativement plus tôt dans le questionnaire que les femmes : parmi les hommes qui ont abandonné, 47% se sont arrêtés à la partie « santé », 10% à la partie « habitudes » et 43% à la partie « activité », contre respectivement 37%, 8% et 55% des femmes qui abandonnent.

Le taux d'abandon diminuait avec l'âge, les 18-34 ans abandonnant 1,5 fois plus que les 50-65 ans.

Le taux d'abandon diminuait lorsque l'état de santé perçu était meilleur : ce taux est 1,5 fois moins important chez les personnes déclarant un bon état de santé perçu par rapport à celles se déclarant en mauvais état de santé.

On n'observe pas d'association significative avec l'indice de masse corporelle, ni avec les antécédents de cancer.

Tableau 2 – Taux d'abandon selon les variables de l'auto-questionnaire

	Effectif	% d'abandon ¹	Test de khi-deux (<i>p-value</i> ²)
Total	28 054	7,1%	
Sexe [page 1]		7,1%	NS
Homme	15 990	7,1%	
Femme	12 064	7,2%	
Age [page 1]		7,1%	< 0,001
18-34 ans	4 695	8,9%	
35-49 ans	10 197	7,0%	
50-65 ans	13 162	6,1%	
Etat de santé perçu (manquant=109³) [page 1]		7,0%	< 0,001
Bon (A-B)	13 756	6,5%	
Moyen (C-D-E)	12 753	7,3%	
Mauvais (F-G-H)	1 436	9,9%	

Indice de masse corporel (manquant=531⁴) [page 1]		6,8%	NS
Maigre (<18,5) 500	500	8,0%	
Normal (18,5-<25)	13 952	7,0%	
Surpoids (25-<30)	9 359	6,5%	
Obèse (>=30)	3 712	6,7%	
Antécédents de cancer (manquant=160⁵) [page 3]		6,8%	NS
Non	26 776	6,8%	
Oui	1 118	5,6%	
Niveau d'études (manquant=1227⁶) [page 12]		3,3%	< 0,001
Jamais scolarisé, Ecole primaire ou Collège	987	7,1%	
Enseignement technique court (CAP, BEP ou équivalent)	5 866	4,3%	
Lycée	4 042	3,3%	
1er cycle de l'enseignement supérieur (Bac + 1 à Bac + 3)	11 563	2,9%	
2 ^e et 3 ^e cycle de l'enseignement supérieur (Bac + 4 et plus)	4 369	2,4%	

1 calculé parmi les valeurs renseignées

2 niveau de significativité du test statistique ; NS = non-significatif (p -value > 0,05)

3 dont 43 abandons

4 dont 135 abandons

5 dont 117 abandons (75 avant la page 3, 24 à la page 3 et 18 après la page 3)

6 dont 1 110 abandons (1 033 avant la page 12, 50 à la page 12 et 27 après la page 12)

Enfin, concernant le niveau d'études, cette information était collectée tardivement dans le questionnaire (page 12) ; l'analyse porte donc uniquement sur les abandons tardifs, c'est-à-dire ceux qui ont eu lieu dans la section « activité professionnelle » du questionnaire (soit 48% des abandons). Parmi ces abandons tardifs, le taux d'abandon était fortement lié au niveau d'études, avec un gradient net : plus le niveau d'études déclaré était faible et plus le taux d'abandon était élevé.

Aussi, afin d'étudier tous les abandons dans un modèle multivarié, il est préférable d'utiliser des informations renseignées pour l'ensemble des répondants. Ces informations étaient de deux ordres : les informations de la base de sondage et les données de connexion.

Les données de connexion sont des marqueurs des conditions de remplissage du questionnaire ; ces conditions pouvant jouer sur la propension à abandonner, elles étaient donc intéressantes à étudier. Par ailleurs, les données de connexion étaient intéressantes car elles étaient associées aux caractéristiques des personnes. Par exemple, les personnes qui déclaraient un mauvais état de santé perçu effectuaient plus souvent plusieurs connexions en plusieurs jours, moins souvent leur première connexion le soir ou la nuit et plus souvent leur première connexion à la suite du premier courrier (invitation). Le niveau d'études était également associé aux données de connexion : les personnes déclarant un niveau d'études supérieur ou égal à Bac + 4 effectuaient plus souvent une seule connexion, plus souvent leur première connexion le matin et plus souvent cette connexion suite au premier courrier (invitation). Ainsi, les données de connexion représentent une alternative pour tenir compte des caractéristiques disponibles uniquement via l'auto-questionnaire mais qui sont entachées de données manquantes.

Les variables de connexion étaient toutes associées à l'abandon en univarié (Tableau 3). Les abandons étaient plus fréquents parmi les personnes qui se connectaient une seule fois, celles

qui ne se connectaient jamais le week-end, celles qui se connectaient pour la première fois la nuit et celles qui se connectaient pour la première fois après la seconde relance.

Tableau 3 – Taux d'abandon selon les variables de connexions

	Effectif	% d'abandon	Test de khi-deux (<i>p-value</i>)
Nombre de connexions			< 0,001
1 seule connexion	22 471	7,8%	
Plusieurs connexions le même jour	2 655	3,9%	
Plusieurs connexions en plusieurs jours	2 928	5,1%	
Au moins une connexion le week-end			< 0,05
Non	20 441	7,3%	
Oui	7 613	6,7%	
Créneau horaire de la première connexion			< 0,05
Matin (6h-12h)	6 972	6,7%	
Midi (12h-14h)	2 838	7,8%	
Après-midi (14h-18h)	8 754	7,2%	
Soirée (18h-22h)	8 076	6,9%	
Nuit (22h-6h)	1 414	9,1%	
Première connexion après le courrier			< 0,001
1 (invitation)	10 750	6,6%	
2 (relance 1)	12 080	7,2%	
3 (relance 2)	5 224	8,2%	

Concernant les données de la base de sondage, les abandons étaient plus fréquents parmi les non-salariés (7,6% vs 6,9% parmi les salariés, $p < 0,05$) et dans les départements et régions d'outre-mer (18,7% vs 7,0% en métropole, $p < 0,0001$).

Le modèle logistique multivarié incluait le sexe et l'âge issus de l'auto-questionnaire, le statut professionnel et la zone géographique issus de la base de sondage, et les données de connexion. Une sélection pas-à-pas (*stepwise*) a été appliquée. Toutes choses égales par ailleurs, l'abandon était associé à l'âge, à la zone géographique, au nombre de connexions et au courrier après lequel avait lieu la première connexion (Tableau 4).

Tableau 4 – Résultats de la régression logistique multivariée expliquant l'abandon

	Effectif	OR [IC 95%]	<i>p-value</i>
Sexe (auto-questionnaire)			NS
Âge (auto-questionnaire)			< 0,0001
18-34	4 695	1,37 [1,22 - 1,55]	
35-49	10 197	1,06 [0,96 - 1,18]	
50-65	13 162	1 Ref	
Statut professionnel (base de sondage)			NS

Zone géographique (base de sondage)			< 0,0001
Métropole	27 610	1 Ref	
Outre-mer	444	3,15 [2,46 - 4,02]	
Nombre de connexions			< 0,0001
1 seule connexion	22 471	1,55 [1,30 - 1,84]	
Plusieurs connexions le même jour	2 655	0,75 [0,58 - 0,97]	
Plusieurs connexions en plusieurs jours	2 928	1 Ref	
Au moins une connexion le week-end			NS
Créneau horaire de la première connexion			NS
Première connexion après le courrier			0,02
1 (invitation)	10 750	1 Ref	
2 (relance 1)	12 080	1,06 [0,96 - 1,18]	
3 (relance 2)	5 224	1,20 [1,06 - 1,36]	

Abréviations : OR = Odds-ratio, IC = Intervalle de confiance, Ref = Référence, NS = non-significatif (p -value > 0,05)

Lorsque l'on ajoutait la variable état de santé perçu au modèle (suppression de 109 individus dont 43 abandons), les mêmes variables restaient associées à l'abandon et la variable état de santé était également associée significativement ($p < 0,0001$) avec un risque d'abandon plus élevé pour les personnes déclarant un mauvais état de santé perçu (OR = 1,71 [1,42 - 2,07] pour un mauvais état de santé vs un bon état, et OR = 1,20 [1,09 - 1,32] pour un état de santé moyen vs un bon état).

4. Discussion

Ce travail propose une analyse des abandons en cours de questionnaire, dans le cadre d'un questionnaire en ligne destiné à l'inclusion dans une cohorte épidémiologique ciblant les risques professionnels. Cette analyse des abandons a pour intérêt d'éclairer la non-réponse partielle au questionnaire via une seule variable d'intérêt (l'abandon), permettant de tirer des conclusions globales pour l'ensemble du questionnaire. Ce travail offre un nouvel angle de vue en analysant également les paradonnées (ici les données de connexion) au regard des abandons en cours de questionnaire.

L'objectif du questionnaire Coset est de décrire l'état de santé en relation avec la situation professionnelle : ce sont là les deux grandes dimensions d'intérêt. Ainsi, les associations entre l'abandon et l'état de santé ou le niveau d'études interpellent sur le traitement des questions qui sont positionnées en fin de questionnaire : pour ces questions pour lesquelles il y aura eu de nombreux abandons antérieurs, il sera important de traiter rigoureusement la non-réponse partielle au risque que les résultats soient biaisés, par exemple si les personnes en meilleur état de santé remplissent plus souvent la partie professionnelle. Ces résultats sont également à prendre en compte pour la correction de la non-réponse totale à l'enquête. Les données de la cohorte Coset ont été appariées aux données de l'assurance maladie (Système National des Données de Santé, de la Caisse Nationale d'Assurance Maladie), pour les répondants et un échantillon aléatoire des non-répondants, lorsque ces personnes avaient été informées (courrier distribué) et qu'elles ne s'étaient pas opposées à l'appariement (Geoffroy-Perez *et al.*, 2019). Ces données appariées seront utilisées pour étudier les événements de santé rencontrés par les cohortistes, l'appariement permettant d'éviter un recueil par questionnaire qui aurait été difficile et souffrant généralement de biais de mémoire. Cet appariement a également été mis en place afin de disposer de données de remboursements de soins, utilisées comme proxys

de l'état de santé de la personne, pour la correction de la non-réponse totale. Lors du pilote de l'enquête Coset-MSA, il avait en effet pu être observé que la consultation d'un praticien de ville (médecin généraliste, dentiste ou spécialiste) dans l'année précédant l'enquête était positivement associée à la probabilité de répondre à l'enquête, quand l'hospitalisation de la personne l'année précédente était associée négativement à la probabilité de répondre (Soullier *et al.*, 2018). Les résultats observés ici sur les abandons en cours de questionnaire confortent ce choix d'utiliser cet appariement dans la correction de la non-réponse totale, en particulier compte tenu du faible taux de participation à l'enquête.

L'utilisation de parodonnées telles que les données de connexion d'une enquête par internet est intéressante car ces données sont peu coûteuses et ne souffrent pas de non-réponse partielle (Lynn & Nicolaas, 2010). En effet, ces données sont collectées de manière indirecte lors d'une enquête et ne pèsent donc pas sur le fardeau de réponse du répondant. De plus, elles sont renseignées pour tous les répondants, y compris ceux qui abandonnent le questionnaire en cours de remplissage. Lors du recrutement en ligne de la cohorte Coset-MSA, le nombre de connexions et le courrier après lequel avait lieu la première connexion étaient significativement associés à l'abandon, toutes choses égales par ailleurs. Les parodonnées peuvent donc jouer un rôle dans la correction de la non-réponse partielle, au même titre que les données de la base de sondage et que certaines données déclarées dans l'auto-questionnaire qui ne souffriraient pas de non-réponse partielle. En effet, il semble tout de même important d'inclure des données déclarées essentielles qui ne sauraient être totalement remplacées par les parodonnées, par exemple ici l'état de santé perçu qui contient peu de données manquantes.

L'intérêt d'étudier les abandons est aussi qu'on peut le faire en fonction de variables dont on ne dispose pas forcément pour les non-répondants, que ce soit les variables déclarées dans l'auto-questionnaire ou les parodonnées. Or, on pourrait voir la non-réponse totale comme un abandon qui a lieu avant le début du questionnaire. Ainsi, les facteurs associés à l'abandon pourraient aussi éclairer sur la non-réponse totale, observée à cette vague ou attendue aux vagues suivantes. Par exemple, l'association de l'abandon avec l'âge est identique aux associations connues avec la non-réponse totale : les plus jeunes abandonnent plus souvent, de même qu'ils répondent moins à l'enquête. Il pourrait être intéressant d'orienter des stratégies de contact ou de connexion afin de limiter les abandons et de regarder si ces stratégies permettent également de réduire la non-réponse totale. Coset étant une cohorte, ces adaptations de protocole pourraient servir à limiter l'attrition.

La non-réponse partielle, et donc l'abandon, est un indicateur de la qualité du questionnaire. Aussi le fait que les abandons augmentent avec les relances donne à réfléchir sur le compromis à faire entre non-réponse totale et non-réponse partielle : les relances permettent de diminuer la non-réponse totale mais on peut obtenir des questionnaires moins bien complétés. Aussi, il convient de relancer avec parcimonie le cas échéant.

Il convient également de réfléchir à des adaptations spécifiques des questionnaires internet pour certains territoires, où la fracture numérique se fait sentir. Par exemple, le taux d'abandon est bien plus élevé dans les territoires ultra-marins, alors même que le taux de réponse y est également bien plus faible. Une analyse s'appuyant sur un zonage permettant d'appréhender les territoires urbains et ruraux pourrait apporter un éclairage important sur ce sujet. Cette catégorisation, non disponible au moment de cette analyse, pourra être construite à partir des adresses postales de contact des cohortistes. Pour ces territoires ou pour les groupes de population moins enclins à répondre sur internet, un protocole multimode pourrait également permettre de réduire la non-réponse totale et partielle (de Leeuw, 2005). Compte tenu du nombre important de personnes invitées et du budget contraint de l'enquête, cette possibilité n'a pas été envisagée pour le recrutement de la cohorte, mais pourrait l'être pour les vagues suivantes.

Le remplissage d'un questionnaire résulte d'un compromis entre l'intérêt porté par le répondant et le fardeau engendré par le questionnaire (Galesic, 2006). Un abandon peut donc être interprété comme un fardeau ressenti plus important que l'intérêt porté. Ainsi, même si on peut penser que répondre à un questionnaire en ligne est plus aisé pour les plus jeunes, ils abandonnent plus fréquemment en cours de questionnaire, soit parce que leur vie professionnelle est débutante et que le sujet les intéresse moins, soit parce que le questionnaire leur apparaît trop long et que subjectivement le fardeau de réponse leur paraît plus important. Il en est de même pour l'état de santé : l'état de santé perçu au moment du remplissage est très lié à l'abandon ce qui peut témoigner d'un fardeau de réponse ressenti plus lourd pour les personnes qui se déclarent en moins bon état de santé, même si elles sont intéressées pour évoquer les sujets de santé. On retrouve aussi cela dans les paradonnées : les personnes qui se connectent plusieurs fois abandonnent moins, ce qui peut être interprété comme un intérêt à répondre qui prévaut sur la difficulté à répondre. Une manière d'approcher le fardeau rencontré par le répondant est la durée (Yan & Tourangeau, 2008 ; Zhang & Conrad, 2014 ; Six *et al.*, 2016) ; cette donnée n'est malheureusement pas disponible de manière fiable et pour tous les répondants dans les paradonnées de Coset. Les déconnexions n'étaient en effet pas enregistrées avec certitude, et le temps de remplissage calculé en cas de connexions multiples est donc peu fiable. À titre d'illustration, la durée totale médiane de remplissage du questionnaire calculée parmi les personnes ayant rempli le questionnaire en une seule connexion était de 37 minutes. Le questionnaire était donc d'une durée relativement longue pour une enquête par internet (Revilla & Ochoa, 2017). On remarque par ailleurs que quatre pages du questionnaire concentrent plus de la moitié des abandons. Une analyse plus approfondie pourrait étudier le moment de l'abandon (page) en fonction du nombre de questions (variable dépendant de la page) et de la difficulté des questions (de binaire oui/non la plus simple à question ouverte la plus difficile) sur cette page mais également sur les précédentes. Cette analyse pourra permettre d'envisager des améliorations tenant à l'ergonomie du questionnaire, ce afin de réduire les abandons dans les enquêtes futures.

Dans la continuité de ce travail, on pourra également étudier l'impact de la prise en compte des paradonnées sur la correction de la non-réponse partielle. Cette analyse pourra être complétée en incluant d'autres données auxiliaires (non disponibles au moment de ce travail) appariées aux données d'enquête, telles que les données de l'assurance maladie ou les données professionnelles issues des bases de données de la MSA. L'analyse pourra également être pondérée par les poids de sondage corrigés pour la non-réponse totale (non disponibles au moment de ce travail) (Santin *et al.*, 2014).

En conclusion, l'étude des abandons permet de détecter des difficultés à répondre à l'enquête, liées aux caractéristiques du répondant (âge, état de santé, niveau d'études) mais aussi aux données de connexion (nombre de connexions, créneau horaire). Cette étude permet d'orienter le traitement de la non-réponse partielle et également de la non-réponse totale, mais aussi de proposer des pistes d'amélioration du questionnaire et du protocole de contact pour les vagues futures de la cohorte.

Références

- Bosnjak M. and T. Tuten (2001), « Classifying Response Behaviors in Web-Based Surveys », *Journal of Computer-Mediated Communication*, 6(3), <https://doi.org/10.1111/j.1083-6101.2001.tb00124.x>
- Čehovin G., M. Bosnjak, and K. Lozar Manfreda (2022), « Item Nonresponse in Web Versus Other Survey Modes: A Systematic Review and Meta-Analysis », *Social Science Computer Review*, 0(0), <https://doi.org/10.1177/08944393211056229>
- Couper M. P. (1997), « Survey Introductions and Data Quality », *Public opinion quarterly*, 61(2), pp. 317-338.
- Couper M. P. (2000), « Web Surveys: A Review of Issues and Approaches », *Public opinion quarterly*, 64, pp. 464-494.
- Croutte P. and J. Muller (2021), « Baromètre du numérique », Crédoc.
- Daikeler J., M. Bošnjak, and K. Lozar Manfreda (2020), « Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates », *Journal of Survey Statistics and Methodology*, 8(3), pp. 513-539.
- de Leeuw E., J. Hox, and M. Huisman (2003), « Prevention and treatment of item nonresponse », *Journal of Official Statistics*, 19, pp. 153-176.
- de Leeuw E. D. (2005), « To mix or not to mix data collection modes in surveys », *Journal of Official Statistics*, 21, pp. 233-255.
- Felderer B. and J. M. E. Herzing (2022), « What about the Less IT Literate? A Comparison of Different Postal Recruitment Strategies to an Online Panel of the General Population », *Field Methods*, 0(0), <https://doi.org/10.1177/1525822X221132940>
- Galesic M. (2006), « Dropouts on the web: Effects of interest and burden experienced during an online survey », *Journal of Official Statistics*, 22(2), pp. 313-328.
- Geoffroy-Perez B., J. Chatelot, G. Santin, L. Benezet, P. Delezire, and E. Imbernon (2012), « Coset : un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Numéro thématique. Surveillance épidémiologique des risques professionnels, quoi de neuf ? », *Bulletin Epidémiologique Hebdomadaire*, 22-23, pp. 276-277.
- Geoffroy-Perez B., N. Soullier, P. Delézire, L. Bénézet, G. Deschamps, E. Breuillard, J. Chesneau, and J.-L. Marchand (2019), « Cohortes pour la surveillance épidémiologique en lien avec le travail (Coset). Bilan de la phase d'inclusion de la cohorte Coset-MSA », *Santé publique France*, Saint-Maurice, 71 p.
- Groves R. M. (1989), *Survey Errors and Survey Costs*, New York, Wiley.
- Kreuter F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*, Hoboken, New Jersey, John Wiley & Sons, Inc.
- Legleye S., A. Nougaret, and L. Viard-Guillot (2022), « L'usage des technologies de l'information et de la communication par les ménages entre 2009 et 2021 », *Insee Focus*, 259, <https://www.insee.fr/fr/statistiques/6049348>

Loosveldt G., J. Pickery, and J. Billiet (2002), « Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey », *Journal of Official Statistics*, 18(4), pp. 545-557.

Lynn P. and G. Nicolaas (2010), « Making Good Use of Survey Paradata », *Survey Practice*, 3, pp. 1-5.

Revilla M. and C. Ochoa (2017), « Ideal and Maximum Length for a Web Survey », *International Journal of Market Research*, 59(5), pp. 557-565.

Santin G., B. Geoffroy, L. Benezet, P. Delezire, J. Chatelot, R. Sitta, J. Bouyer, and A. Gueguen (2014), « In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse », *J. Clin. Epidemiol.*, 67(6), pp. 722-730.

Six M., A. Kowarik, and M. Plate (2016), « Using Paradata to Assess the Quality of an Online Questionnaire », *Statistics Austria*.

Soullier N., B. Geoffroy-Perez, A. Gueguen, L. Bénézet, J. Chatelot, M. Zins, and G. Santin (2018), « Correction de la non-réponse et estimation de prévalences : résultats issus de trois cohortes épidémiologiques ciblant les risques professionnels », « 10^{ème} Colloque francophone sur les sondages » (Lyon, France).

Yan T. and R. Curtin (2010), « The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective », *International Journal of Public Opinion Research*, 22(4), pp. 535-551.

Yan T. and R. Tourangeau (2008), « Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times », *Applied Cognitive Psychology*, 22(1), pp. 51-68.

Zhang C. and F. Conrad (2014), « Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining », *Survey Research Methods*, 8(2), pp. 127-135, <https://doi.org/10.18148/srm/2014.v8i2>.

L'appariement statistique des bases de données SILC, HBS et HFCS : aspects méthodologiques et applications à l'étude de la pauvreté et des inégalités



Huyen TRAN¹

STATEC Luxembourg et Luxembourg Income Study (LIS)



Guillaume OSIER²

STATEC Luxembourg et Luxembourg Income Study (LIS)

TITLE

Statistical matching between the SILC, HBS and HFCS data: methodological aspects and application to the analysis of poverty and social inequalities

RÉSUMÉ

Les techniques d'appariement statistique connaissent depuis plusieurs années un regain d'intérêt parmi les producteurs de statistiques sociales. Cet article présente les résultats d'un exercice d'appariement réalisé à partir des bases de données sur les revenus (Statistics on Income and Living Conditions - SILC), les dépenses (Household Budget Survey - HBS) et le patrimoine des ménages (Household Finance and Consumption Survey - HFCS) au Luxembourg. Les aspects méthodologiques sont abordés et trois exemples d'indicateurs obtenus à partir des données appariées sont présentés : le taux de pauvreté multidimensionnelle basé sur le revenu, la consommation et le patrimoine, le taux d'épargne et le taux de précarité énergétique. Si l'appariement statistique offre une solution à bas coût pour produire des indicateurs sophistiqués à partir d'informations disponibles dans plusieurs sources de données, ces méthodes restent cependant basées sur de la modélisation et doivent donc être employées avec discernement, en tenant compte des hypothèses sous-jacentes.

Mots-clés : appariement statistique, revenus, consommation, patrimoine, pauvreté, épargne, énergie.

ABSTRACT

Statistical matching techniques have been gaining interest for several years among the producers of social statistics. This article presents the outcome from a statistical matching exercise between household income (Statistics on Income and Living Conditions - SILC), household expenditure (Household Budget Survey - HBS) and household wealth (Household Finance and Consumption Survey - HFCS) data in Luxembourg. Methodological aspects are addressed, and three examples of indicators based on the fused dataset are presented: the multidimensional poverty rate based on income, consumption and wealth, the saving rate, and the energy poverty rate. Although statistical matching offers a cheap and convenient solution to yield sophisticated indicators using information available from several databases, these methods are model-based and must yet be used with caution, taking into account the main underlying assumptions.

Keywords: statistical matching, income, consumption, wealth, poverty, saving, energy.

1. thi.tran@eib.org
2. Guillaume.Osier@statec.etat.lu

1. Introduction

L'appariement statistique (*Statistical matching* en anglais) est une technique connue depuis longtemps, mais qui connaît un regain d'intérêt depuis une vingtaine d'années dans le contexte de la modernisation des statistiques sociales au niveau européen. Des initiatives internationales, comme le rapport Stiglitz-Sen-Fitoussi (2009) sur la mesure de la performance économique et du progrès social, ont notamment mis en avant l'importance d'exploiter davantage des micro-données issues des enquêtes sur les ménages et les personnes en complément des statistiques macroéconomiques plus classiques comme le Produit Intérieur Brut. La technique de l'appariement statistique offre une solution à bas coût pour l'intégration des différentes sources de données disponibles et la construction d'indicateurs complexes prenant en compte la multi-dimensionnalité des phénomènes sociaux.

Malgré ses avantages, l'appariement reste néanmoins une technique qui repose essentiellement sur la modélisation des relations entre les variables dans les différentes bases de données. La validité de la méthode repose sur des hypothèses sous-jacentes qu'il est souvent compliqué de valider, en particulier l'hypothèse d'indépendance conditionnelle (*Conditional Independence Assumption* – CIA). Cet aspect ne doit surtout pas être oublié lorsque l'on travaille sur une base de données appariées, qui ne sont pas des données réelles collectées sur le terrain mais plutôt des données reconstruites *ex post*.

Cet article présente un cas concret d'appariement de données statistiques sur la population résidente du Luxembourg, dans lequel les micro-données des enquêtes SILC (*Statistics on Income and Living Conditions*) sur les revenus des ménages, HBS (*Household Budget Survey*) sur la consommation et HFCS (*Household Finance and Consumption Survey*) sur le patrimoine sont fusionnées en une unique source de données. Des indicateurs nouveaux, qu'il n'aurait pas été possible de construire auparavant, peuvent alors être calculés. Cet article présente trois exemples d'indicateurs ainsi obtenus à partir de la distribution jointe du revenu, de la consommation et du patrimoine :

- le taux de pauvreté multidimensionnelle,
- le taux d'épargne des ménages et
- le taux de pauvreté énergétique.

Des estimations sont présentées pour l'ensemble de la population ainsi que pour certaines sous-populations d'intérêt définies d'après des caractéristiques comme l'âge, le genre, le niveau de vie ou encore le statut d'occupation du logement. La validité de l'appariement sera également évoquée.

2. Rappels sur l'appariement statistique

L'appariement est une technique statistique plutôt ancienne, dont on trouve déjà des traces dans les années 70 (Ruggles, 1974 ; Kadane, 1978). Depuis les années 2000, une littérature abondante a de nouveau émergé autour de l'utilisation de cette technique dans le contexte des statistiques sociales au niveau européen (D'Orazio, 2010). Un cadre théorique rigoureux a également été décrit dans d'Orazio (2006).

Dans ses fondements, l'appariement statistique est une approche assez intuitive qui permet de fusionner deux bases de données à partir de critères communs à ces deux bases. Si l'on considère deux bases de données A et B , A étant la « donneuse » et B la « receveuse », et un ensemble de variables X de nature qualitative ou quantitative et communes à A et B , l'appariement aboutit à la construction d'une base synthétique qui correspond en fait à la base « receveuse » B enrichie de l'information disponible dans la base « donneuse » A . Comme A et B ne couvrent généralement pas les mêmes unités, les variables Y qui se trouvent dans la base A mais pas dans la base B se

retrouvent dans la base finale synthétique sous la forme d'une imputation \hat{Y} de la valeur réelle que l'on obtiendrait sur les unités de la base B si l'information était réellement collectée. Comme pour les variables communes X , le cadre des méthodes d'appariement est suffisamment large pour permettre de traiter aussi bien des variables quantitatives que qualitatives.

Tableau 1 – Principe de l'appariement de deux bases de données A et B

Base A « donneuse »	Base B « receveuse »	Base synthétique
Y		
X	X	X
	Z	Z, \hat{Y}

La validité de l'approche précédente repose sur l'hypothèse dite d'indépendance conditionnelle (*Conditional Independence Assumption* – CIA), qui postule que la corrélation entre les variables Y et Z s'explique uniquement au travers d'un ensemble de variables X observables et communes aux deux bases de données (D'Orazio, 2010). Autrement dit, il n'y a pas de lien causal direct entre les variables Y et Z .

Cette hypothèse CIA est cependant rarement vérifiée dans la pratique et, de toute façon, ne peut pas être testée à partir des seules bases de données qui entrent en considération. Néanmoins, des techniques ont été développées (Rässler, 2004) pour tester la sensibilité des résultats de l'appariement au relâchement de cette hypothèse. En outre, des approches alternatives (Paass, 1986) intègrent la possibilité d'utiliser des informations auxiliaires lors de l'appariement afin de contourner l'hypothèse d'indépendance conditionnelle.

La mise en œuvre d'un appariement statistique requiert un travail préalable de réconciliation des bases de données. Cela signifie qu'elles doivent couvrir les mêmes populations d'intérêt et se référer aux mêmes périodes de temps. Si besoin, des ajustements sont réalisés pour permettre la mise en conformité des données, par exemple *via* la suppression d'unités « hors-champ » ou encore l'ajustement de certaines valeurs. Pour une description plus détaillée des étapes préalables de réconciliation des bases de données à partir d'exemples concrets, on pourra consulter Eurostat (2013).

3. Détermination des variables de l'appariement statistique

L'hypothèse fondamentale d'indépendance conditionnelle montre que le choix des variables partagées X est déterminant pour garantir un bon niveau de précision aux résultats de l'appariement. Ces variables doivent être disponibles dans les deux bases de données A et B , et collectées d'après les mêmes concepts et définitions. Par exemple, si l'on utilise l'âge de la personne, celui-ci doit être calculé au même point³ dans chacune des deux bases. De la même façon, on ne pourra pas utiliser le statut d'activité d'une personne si d'un côté celui-ci a été obtenu à partir de réponses auto-déclarées tandis que de l'autre il a été déterminé à partir de la nomenclature utilisée par l'Organisation Internationale du Travail (OIT) et qui est reprise dans l'enquête européenne sur les forces de travail. Il faudra aussi veiller à ce que les variables impliquées dans l'appariement ne souffrent pas d'erreurs de mesure ou encore de la présence d'un nombre élevé de valeurs manquantes ou de valeurs imputées.

3. Généralement l'âge peut être calculé en nombre d'années révolues (c'est-à-dire au 31 décembre de l'année dernière), au moment de l'enquête ou à la fin de l'année de l'enquête.

Les variables communes \mathbf{X} doivent en outre avoir des distributions proches dans les deux bases de données (Eurostat, 2013). Cette condition doit permettre d'assurer une meilleure robustesse aux résultats de l'appariement. La comparaison des distributions entre A et B peut se faire en utilisant des fonctions de distance comme la distance de Hellinger :

$$HD(V, V') = \sqrt{\frac{1}{2} \sum_{i=1}^K (\sqrt{P(V=i)} - \sqrt{P(V'=i)})^2},$$

où V et V' désignent deux vecteurs de covariables. Une formule analogue existe dans le cas d'une variable continue X avec deux fonctions de densité f_X et g_X :

On peut également recourir à des tests statistiques pour évaluer la similarité entre les distributions, comme les tests du Chi-deux ou de Kolmogorov-Smirnov. Dans le cas de variables

$$HD(f_X, g_X) = \sqrt{\frac{1}{2} \int (\sqrt{f_X(x)} - \sqrt{g_X(x)})^2 dx.}$$

continues, on peut aussi réaliser des tests de Student.

Parmi l'ensemble des variables communes aux deux bases de données et collectées suivant les mêmes définitions, il faut ensuite déterminer celles qui sont les plus corrélées avec les variables d'intérêt de l'appariement, c'est-à-dire les variables \mathbf{Y} et \mathbf{Z} présentes dans les bases A et B et que l'on cherche à réunir dans une seule et même base. Pour cela, on peut recourir à des méthodes de sélection automatique du type régression pénalisée (par exemple, Ridge, LASSO ou Elastic Net) ou à des algorithmes basés sur des arbres de décision (par exemple, Random Forest). À ce sujet on pourra consulter Schork (2018). Préalablement, il faudra veiller à ne retenir pour l'appariement que des variables avec un faible pourcentage de valeurs manquantes (Van Buuren, 2018).

Le package R *StatMatch*⁴ permet de mettre en œuvre de façon concrète l'appariement statistique en utilisant pour l'imputation des données dans la base « receveuse » des méthodes de type hot-deck. La méthode du hot-deck est l'archétype des méthodes d'imputation par donneur. Elle consiste à imputer les valeurs dans la base « receveuse » en prenant les valeurs observées dans la base « donneuse » sur des individus jugés suffisamment « proches ». Cette proximité entre observations se mesure à partir d'une fonction de distance, par exemple la distance euclidienne, définie sur la base des valeurs prises par les variables communes aux deux bases de données. Pour une observation donnée, le « donneur » est en fait choisi de façon aléatoire parmi un ensemble de candidats possibles. Cela permet d'obtenir plus de variabilité dans la distribution des valeurs imputées. Pour une revue des méthodes de hot-deck on pourra consulter (Andridge & Little, 2010). La grande force de ces méthodes est qu'elles conduisent à imputer des valeurs à partir de données réellement observées, ce qui permet de préserver les relations entre les variables et d'éviter de possibles incohérences qui pourraient survenir entre elles lorsqu'on utilise des méthodes d'imputation basées sur des modèles.

4. <https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf>

4. Les limites de la mesure de la pauvreté

Chaque année, le STATEC calcule des indicateurs clés sur la pauvreté et les inégalités au Luxembourg à partir de son enquête SILC sur les revenus et les conditions de vie des ménages. Cette enquête est conduite chaque année par le STATEC auprès d'un échantillon de 4 000 ménages et 10 000 individus représentatif de la population résidente au Grand-Duché⁵. Des enquêteurs réalisent les entretiens en face-à-face au domicile des ménages. À partir de 2022, un protocole multimodal associant le face-à-face, le téléphone et l'internet a remplacé le protocole précédemment en vigueur.

Les méthodes actuelles de calcul du taux de pauvreté au Luxembourg mais aussi en Europe se basent principalement sur le revenu disponible des ménages, c'est-à-dire leur revenu calculé après la prise en compte des cotisations sociales et des impôts directs. Cependant, le revenu n'est pas le seul facteur affectant le bien-être d'un ménage. La mesure de la pauvreté basée uniquement sur le revenu présente en fait trois faiblesses principales. Premièrement, le revenu peut fluctuer dans le temps. Ceci est particulièrement vrai pour les travailleurs indépendants ou les chômeurs de courte durée. Meyer & Sullivan (2012) et Brewer & O'Dea (2012) constatent qu'il est préférable d'utiliser la consommation des ménages plutôt que leur revenu si l'on souhaite mieux appréhender le niveau de vie des personnes défavorisées aux États-Unis et au Royaume-Uni. Par ailleurs, la théorie du « revenu permanent » de Friedman suggère que les comportements décisionnels des ménages sont fondés sur les attentes de revenu à long terme plutôt que sur le niveau de revenu actuel. Deuxièmement, le revenu reflète mal la consommation de biens durables comme les voitures ou les logements. Troisièmement, les ménages peuvent compter sur leur épargne et leur patrimoine financier pour lisser leur consommation au cours d'une année difficile et ainsi conserver le même niveau de vie qu'auparavant. Ceux des ménages qui disposent d'un niveau élevé de patrimoine et d'épargne accumulés sont donc moins susceptibles d'être exposés au risque de pauvreté que les ménages disposant de peu de ressources. Ces limites soulignent l'importance d'un cadre multidimensionnel tenant compte à la fois du revenu, de la consommation et du patrimoine afin d'obtenir une meilleure évaluation du bien-être économique des ménages.

C'est pourquoi il est souhaitable de disposer d'un taux de pauvreté multidimensionnelle combinant à la fois le revenu, la consommation et le patrimoine financier des ménages (STATEC, 2022). Cependant, une telle analyse nécessiterait d'avoir une source de données unique fournissant des informations conjointes sur toutes ces dimensions. Actuellement, il n'existe aucune source de données de ce type au Luxembourg. Par conséquent, nous utilisons des techniques d'appariement statistique pour faire correspondre les informations sur la consommation tirée de l'enquête sur le budget des ménages (*Household Budget Survey* – HBS) avec le niveau de patrimoine obtenu dans l'enquête de la Banque Centrale du Luxembourg sur la consommation et le patrimoine financier des ménages (*Household Finance and Consumption Survey* – HFCS) et avec le revenu tel que collecté dans l'enquête SILC sur les revenus et les conditions de vie. Bien que ces trois sources concernent des périodes différentes (2021 pour HBS, 2020 pour les revenus de l'enquête SILC et 2018 pour le HFCS), elles ont néanmoins été rapprochées pour les besoins de cet exercice d'appariement.

Dans cet exercice, SILC a été considérée comme la base « receveuse », HBS et HFCS étant des bases « donneuses ». La méthodologie présentée dans le chapitre précédent a été déroulée afin de déterminer la liste des variables retenues pour l'appariement, avec des listes différentes selon que l'appariement concernait SILC et HBS ou SILC et HFCS :

- **Pour l'appariement SILC/HBS** : déciles du revenu du ménage (disponible et par équivalent-adulte) ; taille du ménage ; nombre d'adultes et d'enfants dans le ménage ;

5. Voir à ce sujet : <https://statistiques.public.lu/fr/enquetes/enquetes-particuliers/silc-conditions-vie.html>

nombre de pièces du logement ; nombre de membres du ménage en emploi ; nombre d'étudiants dans le ménage ; nombre d'étrangers dans le ménage ; âge, niveau d'éducation, statut marital et pays de naissance du chef de ménage ; région de résidence et statut d'occupation du logement.

- **Pour l'appariement SILC/HFCS** : déciles du revenu du ménage (disponible et par équivalent-adulte) ; taille du ménage ; nombre d'adultes et d'enfants dans le ménage ; nombre de voitures que possède le ménage ; nombre de membres du ménage en emploi ; nombre d'étrangers dans le ménage ; nombre d'individus de 65 ans ou plus dans le ménage ; âge, niveau d'éducation, statut d'activité, statut dans l'emploi et nombre d'heures travaillées par le chef de ménage ; statut d'occupation du logement.

5. Résultats : taux de pauvreté multidimensionnelle

Le taux de pauvreté multidimensionnelle est défini au croisement de la pauvreté de revenu, la pauvreté de consommation et la pauvreté de patrimoine :

- **Pauvreté de revenu** : le revenu de l'individu⁶ est inférieur à 60% du revenu médian.
- **Pauvreté de consommation** : la consommation de l'individu⁷ est inférieure à 60% de la consommation médiane.
- **Pauvreté de patrimoine** : le patrimoine financier de l'individu⁸ est inférieur à 3 mois de ressources monétaires⁹. Ici, nous utilisons uniquement le patrimoine financier pour mesurer la pauvreté car celui-ci peut être facilement liquidé afin de lisser la consommation du ménage.

Les résultats sont présentés dans la Figure 1, où les calculs ont été réalisés au niveau individuel et au niveau ménage. Au niveau individuel, nous avons constaté que le taux de pauvreté selon le revenu était de 18,1% en 2021, tandis que les taux de pauvreté de consommation et de patrimoine étaient respectivement de 19,8% et 32,7%. Calculés au niveau des ménages, ces taux étaient respectivement de 15,6%, 18,5% et 29,9% pour les dimensions du revenu, de la consommation et du patrimoine.

6. Le revenu d'un individu est défini en partant du revenu disponible de son ménage, c'est-à-dire le revenu après impôts directs et cotisations sociales, et en le divisant par le nombre d'équivalents-adulte calculé selon l'échelle d'équivalence dite « OCDE modifiée ». L'ajustement par les échelles d'équivalence permet de comparer le niveau de vie de ménages avec des tailles et des compositions différentes.

7. Même approche que pour la pauvreté de revenu.

8. Là aussi, le patrimoine financier d'un individu correspond au patrimoine financier de son ménage ajusté par le nombre d'équivalents-adulte dans le ménage.

9. En 2021, le seuil de pauvreté monétaire par personne au Luxembourg est de 2 124 euros par mois (source : SILC), ce qui représente 60% du revenu médian des ménages ajustés par le nombre d'unités de consommation. Ainsi, trois mois de ressources correspondent à 6 372 euros par nombre d'unités de consommation. Il est à noter qu'en 2020, le seuil de pauvreté est de 1 892 euros par mois et par unité de consommation, donc le seuil en 2021 augmente de 12%.

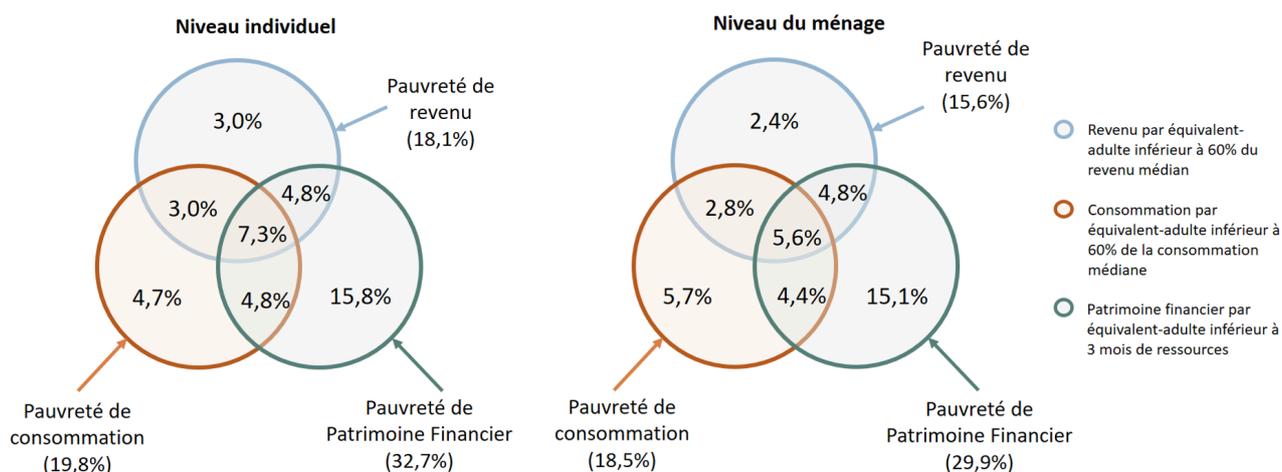


Figure 1 – Taux de pauvreté multidimensionnelle basée à la fois sur le revenu, la consommation et le patrimoine financier des ménages

[Source : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021 et HFCS 3^{ème} Vague (2018). Il convient de noter que le revenu collecté dans EU-SILC 2021 provient en fait de l'année 2020.]

Finalement, un résultat important est que la pauvreté au Luxembourg, qui dépasse 18% lorsque sa mesure se base uniquement sur le revenu, chute à 7,3% si l'on tient compte à la fois du revenu, de la consommation et du patrimoine financier.

La proportion de personnes pauvres en patrimoine est nettement plus élevée que pour les deux autres dimensions. Cela reflète le constat selon lequel le patrimoine est beaucoup plus inégalement réparti entre les ménages que ne le sont le revenu et la consommation. Le chevauchement entre ces trois dimensions marque le taux de pauvreté multidimensionnelle. Il est de 7,3% au niveau individuel et 5,6% au niveau ménage. Ces deux chiffres représentent une situation de pauvreté « extrême », dans laquelle les ménages touchés ne peuvent s'appuyer ni sur des niveaux de revenus suffisants ni sur des réserves monétaires ou de l'aide financière extérieure pour maintenir un niveau de vie adéquat.

6. Autres applications de l'appariement des bases de données SILC, HBS et HFCS : le taux d'épargne des ménages et le taux de pauvreté énergétique

Sur la base des données appariées, nous arrivons également à produire d'autres indicateurs synthétiques, comme le taux d'épargne des ménages, qui est un indicateur important sur l'accumulation de richesse, et le taux de pauvreté énergétique. Cette approche est conforme aux lignes directrices recommandées dans le rapport Stiglitz, Sen et Fitoussi (2009) et par l'OCDE (2013).

Le taux d'épargne d'un ménage est obtenu en retranchant de son revenu disponible le montant de ses dépenses de consommation finale, et en exprimant cette différence en pourcentage du revenu disponible. Le taux d'épargne moyen par ménage a ainsi été estimé à 28%. Sur la Figure 2, on voit cependant des différences marquées dans les taux moyens d'épargne selon le quintile du revenu : plus le revenu du ménage est élevé, plus le taux d'épargne est grand. On observe même une désépargne, c'est-à-dire une épargne négative, parmi les ménages du premier quintile.

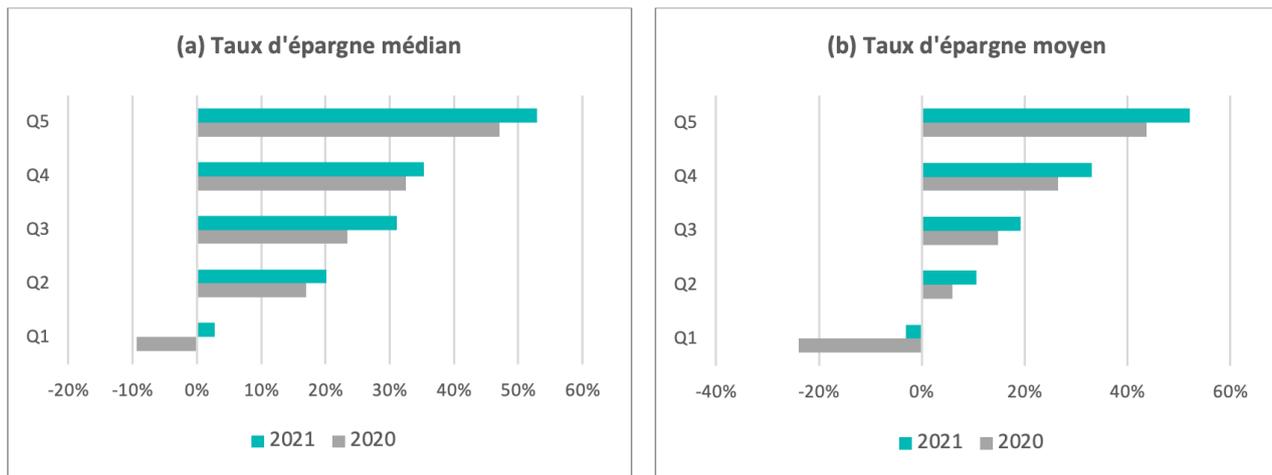


Figure 2 – Variation du taux d'épargne moyen et médian par quintile entre 2020 et 2021

[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus à partir de l'appariement d'EU-SILC 2021, d'EBM 2021 et de HFCS 3^e vague (2018) pour les taux d'épargne en 2021 ; et d'EU-SILC 2020, d'EBM 2020 et de HFCS 3^e vague (2018) pour les taux d'épargne en 2020. Il convient de noter que les revenus collectés dans EU-SILC 2021 se rapportent à l'année 2020 et que ceux collectés dans EU-SILC 2020 se rapportent à l'année 2019.]

Un autre sujet d'importance de cette étude est celui de la pauvreté énergétique. Dans le contexte actuel de crise énergétique, la pauvreté énergétique est devenue un sujet central pour les statistiques sociales. Celle-ci est cependant définie différemment entre les pays en développement et les pays développés. Dans le premier cas, elle est généralement comprise comme un manque d'accès aux services énergétiques, alors que dans le second cas elle est attribuée au poids excessif des dépenses énergétiques par rapport aux revenus des ménages. C'est cette seconde approche que nous avons retenue dans le cas du Luxembourg (Di Falco *et al.*, 2021). Pour mesurer quantitativement la pauvreté énergétique, nous avons utilisé les deux indicateurs suivants :

- i. Taux d'effort énergétique (TEE) élevé :

$$\text{TEE} = \frac{\text{Dépenses énergétiques du ménage}}{\text{Revenu du ménage}} > 2 \times \text{Valeur médiane nationale en 2012} ;$$

- ii. Bas Revenus, Dépenses Élevées (BRDE) :

$$\text{BRDE} = \begin{cases} \frac{\text{Dépenses énergétiques du ménage}}{\text{par unité de consommation}} > \text{Valeur médiane nationale en 2012} \\ (\text{Revenu net du ménage} - \text{Charges du logement}) < 60\% \times \\ \text{Médiane du (Revenu net du ménage} - \text{Charges du logement) en 2012.} \end{cases}$$

Les mesures TEE sont fréquentes dans la littérature (Di Falco *et al.*, 2021) car elles sont faciles à calculer et à expliquer. Cependant, elles ne tiennent pas compte des niveaux de revenu et

pourraient alors inclure des ménages ayant un niveau de revenu élevé mais faisant un possible gaspillage d'énergie, ce qui entrainerait des dépenses énergétiques importantes. L'indicateur BRDE, quant à lui, est plus compliqué car il inclut deux conditions simultanées : pour être considéré en précarité énergétique, un ménage doit cumuler à la fois un faible revenu et des dépenses énergétiques élevées. Les ménages qui se situent en dessous du seuil conventionnel de 60% du revenu médian et dont les dépenses d'énergie sont supérieures au niveau médian sont considérés comme étant en situation de pauvreté énergétique. Suivant Di Falco *et al.* (2021), nous avons choisi l'année 2012 comme point de référence pour notre mesure de la pauvreté énergétique. Pour l'indicateur TEE, la valeur médiane nationale en 2012 était de 7,2%, donc tous les ménages dont la part des dépenses énergétiques par rapport au revenu est supérieure au double de ce seuil (soit 14,4%) sont considérés en situation de pauvreté énergétique. Concernant l'indicateur BRDE, la dépense énergétique médiane ajustée par le nombre d'unités de consommation dans le ménage s'élevait à 1 194 euros par an en 2012. Quant au seuil défini par 60% du revenu net médian après la prise en compte des charges de logement, sa valeur était de 17 263 euros par an et par ménage en 2012, soit 1 438 euros par mois.

La Figure 3 montre un TEE de 4,9% et un BRDE de 3,9% pour l'année 2021, tandis que 1,8% des ménages au Luxembourg sont en situation de précarité énergétique d'après les deux définitions. Ces chiffres appellent cependant une mise à jour pour tenir compte des poussées inflationnistes récentes que l'on enregistre sur les prix de l'énergie. Pour plus d'éléments sur cette question, on pourra consulter STATEC (2022).

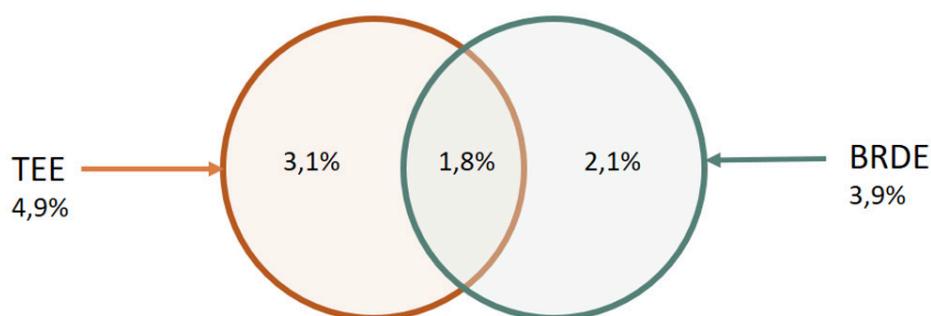


Figure 3 - Indicateurs de précarité énergétique au Luxembourg (TEE et BRDE)

[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021. Il convient de noter que le revenu collecté dans EU-SILC 2021 provient de l'année 2020.]

Si les indicateurs TEE et BRDE sont aujourd'hui couramment utilisés, ils doivent être complétés par des indicateurs de nature plus subjective sur le ressenti des ménages (Charlier *et al.*, 2015). C'est ce qui est illustré dans le Tableau 2 où l'on voit que les ménages en situation de précarité énergétique sont les plus modestes en termes de revenu et de mesures subjectives. Ces mesures subjectives proviennent des réponses données par les ménages aux questions de savoir si : (i) le ménage ne peut pas se chauffer suffisamment en hiver, (ii) le ménage rencontre des difficultés pour les paiements des factures énergétiques. On voit qu'il y a une corrélation entre le fait pour un ménage d'être en situation de précarité énergétique selon les indicateurs TEE et BRDE, et les difficultés à payer son loyer, rembourser ses emprunts ou encore payer ses factures courantes en lien avec son logement. De ce point de vue, on peut conclure que les mesures subjectives « valident » les indicateurs objectifs de pauvreté énergétique (TEE et BRDE).

Tableau 2 – Indicateurs objectifs et subjectifs de précarité énergétique

	TEE		BRDE		TEE & BRDE	
	Oui	Non	Oui	Non	Oui	Non
Statistiques de base						
- Revenu net moyen par unité de consommation (EUR/an)	20 934	51 566	20 015	51 435	15 545	50 884
- Dépenses énergétiques moyennes (EUR/an)	2 957	1 623	2 258	1 657	2 668	1 662
Mesures subjectives						
- Ne pas pouvoir chauffer suffisamment son logement	0,8%	2,4%	4,6%	2,3%	2,2%	2,3%
- Arriérés sur le paiement des factures énergétiques	7,6%	2,6%	9,6%	2,6%	15,9%	2,6%
- Arriérés sur le paiement des prêts hypothécaires ou des loyers	4,1%	2,0%	6,2%	1,9%	9,9%	1,9%
- Arriérés sur le paiement des autres emprunts	13,8%	4,2%	13,6%	4,4%	22,6%	4,4%

[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021. Il convient de noter que les revenus collectés dans EU-SILC 2021 proviennent de l'année 2020.]

7. Conclusion

L'appariement statistique est une technique puissante et rentable qui permet d'exploiter tout le potentiel des bases de données disponibles afin de construire des indicateurs beaucoup plus riches et bien plus intéressants que ceux établis à partir d'une unique source de données. Le taux de pauvreté multidimensionnelle en est un parfait exemple, dans la mesure où il fournit une image de la pauvreté qui est beaucoup plus fine que celle définie seulement à partir du niveau de revenu des ménages. Les deux autres exemples présentés dans cet article sur l'épargne et la précarité énergétique des ménages sont tout aussi pertinents à l'aune des crises que traversent actuellement les pays européens dans le domaine économique et dans celui de l'énergie.

Toutefois, nous conseillons de réaliser une analyse de sensibilité pour toutes les variables appariées. De façon générale, il convient d'être prudent lors de l'utilisation de mesures synthétiques basées sur des données appariées, car celles-ci reposent sur des estimations qui sont de nature expérimentale et doivent donc faire l'objet de tests et de validations supplémentaires. Des analyses plus poussées sur la qualité et la robustesse de l'appariement sont en cours. Celles-ci consistent notamment à comparer les distributions des principales variables socio-économiques observées sur les bases de données originelles et sur la base de données synthétiques issue de l'appariement. Pour plus de détails on pourra consulter Eurostat (2013).

L'hypothèse d'indépendance conditionnelle constitue une autre limite. Dans cet exercice d'appariement des bases de données SILC, HBS et HFCS, nous avons supposé que l'hypothèse d'indépendance conditionnelle était vérifiée, ce qui est probablement une simplification de la réalité. On peut toutefois considérer une telle simplification comme raisonnable dans la mesure

où les variables utilisées pour l'appariement incluent le revenu du ménage, qui est disponible dans les trois bases de données sous revue et qui est lié de manière directe aux indicateurs qui nous intéressent. De façon générale, on trouvera dans Rässler (2004) et Lamarche (2017) des éléments de réponse sur la validité ou non de l'hypothèse d'indépendance conditionnelle et de quelle manière la procédure d'appariement statistique doit être adaptée lorsque cette hypothèse ne peut pas être vérifiée.

Il faut enfin mentionner que les années de référence des trois bases de données avec lesquelles nous avons travaillé n'étaient pas les mêmes : 2021 pour HBS, 2020 pour SILC et 2018 pour HFCS. Cette situation est malheureuse et contraire aux lignes de bonnes pratiques, mais il s'agissait là d'une contrainte liée au fait que SILC collecte les revenus détaillés pour l'année précédant l'enquête et que l'enquête HFCS n'est conduite que tous les 3 ou 4 ans.

Références

Andridge R. R. and R. J. A. Little (2010), « A Review of Hot Deck Imputation for Survey Non-Response », *International Statistical Review*, 78(1), pp. 40-64.

Brewer M. and C. O'Dea (2012), « Measuring living standards with income and consumption: evidence from the UK », *ISER Working Paper Series*, N° 2012-05, Institute for Social and Economic Research (ISER), Essex.

Charlier D., A. Risch et C. Salmon (2015), « Les indicateurs de la précarité énergétique en France », *Revue d'Économie Française*, 4, pp. 187-230.

D'Orazio M., M. Di Zio and M. Scanu (2006), *Statistical Matching: Theory and Practice*, Chichester, Wiley.

D'Orazio M. (2010), « Evaluation of the accuracy of statistical matching, Report WP1 ESS-net », Statistical Methodology Project on Integration of Surveys and Administrative Data.

Di Falco E., O. Thunus et G. Zardet (2021), « Analyse sur la précarité énergétique au Luxembourg », Document de travail du STATEC, <https://statistiques.public.lu>

Eurostat (2013), « Statistical matching: a model based approach for data integration », *Methodologies and Working Papers*, <https://ec.europa.eu/eurostat>

Kadane J. B. (1978), « Some statistical problems in merging data files », Department of Treasury, Compendium of Tax Research, Washington, DC: US Government Printing Office, pp. 159-179.

Lamarche P. (2017), « Estimating consumption in the HFCS - Experimental results on the first wave of the HFCS », European Central Bank, Statistics Paper Series, 22, <https://www.ecb.europa.eu/>

Meyer B. D. and J. Sullivan (2012), « Identifying the Disadvantaged: Official Poverty, Consumption Poverty, and the New Supplemental Poverty Measure », *Journal of Economic Perspectives*, 26(3), pp. 111-136.

OCDE (2013), « OECD Framework for Statistics on the distribution of income, consumption and wealth », OECD Publishing, <https://www.oecd.org/statistics/framework-for-statistics-on-the-distribution-of-household-income-consumption-and-wealth-9789264194830-en.htm>

Paass G. (1986), « Statistical match: evaluation of existing procedures and improvements by using additional information », in G. H Orcutt, J. Merz, and H Quinke (eds.), *Microanalytic Simulation Models to Support Social and Financial Policy*, Elsevier, pp. 401-422.

Rässler S. (2002), *Statistical Matching, a Frequentist Theory, Practical Applications and Alternative Bayesian Approach*, Springer.

Rässler S. (2004), « Data fusion: identification problems, validity, and multiple imputation », *Austrian Journal of Statistics*, 33(1-2), pp. 153-171.

Ruggles N. and R. Ruggles (1974), « A strategy for merging and matching microdata sets », *Annals of Economic and Social Measurement*, 1(3), pp. 353-371.

Schork J. (2018), « Automatic Variable Selection for Imputation Models: Common Methods Applied to EU-SILC », *STATEC Working Paper*, N° 98/2018, <https://statistiques.public.lu>

STATEC (2022), « D'une crise à l'autre : la cohésion sociale sous pression », Rapport travail et cohésion sociale, Analyses, <https://statistiques.public.lu/>

Stiglitz J., A. Sen et J.-P. Fitoussi (2009), *Rapport de la Commission sur la mesure des performances économiques et du progrès social*, Éditions Odile Jacob.

Van Buuren S. (2018), *Flexible imputation of missing data*, Chapman & Hall/CRC, <https://stefvanbuuren.name/publication/vanbuuren-2018/>

Un entretien avec Marc Hallin



Marc HALLIN¹
Université libre de Bruxelles

Jean-Jacques DROESBEKE²
Université libre de Bruxelles

TITLE

An interview with Marc Hallin

RÉSUMÉ

Marc Hallin est professeur émérite à l'Université libre de Bruxelles, où il a fait toute sa carrière. Il est l'un des statisticiens-mathématiciens parmi les plus brillants de notre époque. Il est reconnu internationalement pour ses travaux en inférence statistique fondée sur les rangs pour des observations multivariées sous des densités non spécifiées, ainsi que dans le domaine des séries chronologiques en grande dimension et des modèles à facteurs dynamiques. Il a aussi joué un rôle important dans la Société royale belge de Statistique et la Société Française de Statistique.

Au cours de cet entretien réalisé en décembre 2022, Marc Hallin nous a parlé de ses années de formation, de ses activités à l'Université libre de Bruxelles, de ses travaux de recherche, de sa vision du futur de sa discipline et de ses goûts personnels.

Mots-clés : *statistique mathématique, inférence fondée sur les rangs, séries chronologiques en grande dimension.*

ABSTRACT

Marc Hallin is professor emeritus at the Université libre de Bruxelles, where he spent his entire career. One of the most brilliant mathematical statisticians of our time, he is internationally recognized for his contributions to rank-based statistical inference for multivariate observations under unspecified densities and to high-dimensional time series analysis based on dynamic factor models. He also played an important role in the Royal Belgian Statistical Society and the French Statistical Society.

In this interview conducted in December 2022, Marc Hallin tells us about his training in statistics, his activities at the Université libre de Bruxelles, his research, his vision of his discipline for the future, and his personal tastes.

Keywords: *mathematical statistics, rank-based inference, high-dimensional time series.*

1. Marc.Hallin@ulb.be / désigné par les initiales MH dans cet entretien.
2. Jean-Jacques.Droesbeke@ulb.be / désigné par les initiales JJD dans cet entretien.

JJD : Bonjour, Marc. Permits-moi de te présenter en quelques mots pour celles et ceux qui ne te connaissent pas. Tu es professeur émérite à l'Université libre de Bruxelles, où tu as fait toute ta carrière. Tu es l'un des statisticiens-mathématiciens parmi les plus brillants de notre époque, récompensé notamment par le Prix Pierre Simon de Laplace que t'a décerné en juin 2022 la Société Française de Statistique et, deux mois plus tard, par le Gottfried Noether Distinguished Scholar Award de l'American Statistical Association. J'ai suivi toute ta carrière comme collègue et surtout comme ami, ce qui me permet de ne pas me contenter des aspects purement professionnels de ta vie. Peux-tu me parler en quelques mots de ta jeunesse et de tes années de formation ?

MH : Merci, Jean-Jacques. Ma jeunesse n'a rien de très extraordinaire ni d'exceptionnel, caractérisée cependant par une certaine mobilité européenne, à une époque où l'Union Européenne n'existait pas encore. Je suis né à Gand, d'un père gantois et d'une mère brugoise. Très rapidement après ma naissance, nous nous sommes établis, mes parents et moi, à Bruxelles avant de partir pour les Pays-Bas, revenir en Belgique, et nous installer finalement en France. J'ai passé la plus grande partie de mon adolescence, jusqu'à l'âge de 18 ans, dans les environs de Paris, à Marly-Le-Roi précisément. Mon père était ingénieur militaire spécialisé dans les techniques du radar, ce qui, dans l'immédiat après-guerre, explique cette mobilité européenne. J'ai fait toutes mes études secondaires au lycée Marcel Roby de Saint-Germain-en-Laye et ai passé le baccalauréat section « Math Elem » en 1967. Mon père, ensuite, a décidé que, pour mes études supérieures, il était préférable que je rentre en Belgique, et c'est ainsi que, mon bac en poche, je me suis retrouvé, à l'âge de dix-huit ans, seul à Bruxelles, une ville qui m'était complètement étrangère, où je ne connaissais personne, et où j'ai entamé, à l'Université libre de Bruxelles, des études de mathématiques. Près de soixante ans plus tard, j'y suis toujours, tout en restant profondément marqué par mon « adolescence parisienne ».

Le choix des mathématiques fut le résultat d'un compromis laborieux entre mon père et moi – comme beaucoup d'ingénieurs, il pensait que son métier était le plus beau du monde et aurait voulu me voir emprunter la même voie. Je me sentais davantage attiré vers des branches littéraires ou historiques, et nous avons transigé sur les mathématiques. Je dois dire que je ne l'ai jamais regretté depuis. Cependant, tout en étant un bon étudiant, je n'avais pas d'enthousiasme particulier pour les mathématiques, et ce n'est que plus tard, quand j'ai commencé à en faire réellement, à créer des choses, que j'y ai vraiment pris goût ; cela n'a plus changé depuis.

JJD : Le compromis est une histoire bien belge dont tu as pu profiter. Est-ce que tu peux nous parler un peu de ton entrée en statistique ? Y es-tu entré comme on entre en religion ?

MH : Non, pas du tout. En fait, je suis entré en statistique par le plus grand des hasards, j'oserais même dire par accident, au sens propre du terme. En effet, c'est à la suite d'un mauvais accident de voiture que je me suis tourné vers la statistique. Au terme de mes études de mathématiques, je ne connaissais strictement rien dans ce domaine. Le cursus des mathématiciens de l'époque comportait bien un cours de statistique mathématique, enseigné par un analyste qui avait pris pour référence le *Mathematical Methods of Statistics* (1946) de Harald Cramér, ouvrage d'une importance historique indéniable mais qui, un quart de siècle après sa parution, commençait à dater quelque peu. J'ai donc suivi ce cours au terme duquel je ne savais pas grand-chose – je n'avais pas même entendu parler du Lemme de Neyman-Pearson ! Par ailleurs, mes projets n'étaient pas du tout orientés vers une carrière académique : j'avais, à l'issue de mes quatre années de licence en mathématiques, déjà largement entamé une licence en sciences actuarielles (programme d'une durée théorique de trois années). Il me restait une année à faire avant d'obtenir le diplôme d'actuaire et, pour arrondir mes fins de mois, j'avais accepté un poste d'assistant auprès de Simone Huyberegts dans le Service de Statistique de la Faculté des Sciences sociales, politiques et économiques, où tu occupais toi-même un poste analogue depuis quelques années.

J'ai donc commencé ma carrière d'enseignant par des séances de travaux dirigés pour les étudiants de Sciences économiques et de l'École de Commerce. Or, à cette époque-là, je ne rêvais que de voyages lointains. Avec ma petite coccinelle Volkswagen, j'avais déjà sillonné les routes d'Espagne et du Portugal, de l'Anatolie, de la Syrie, d'Irak et du Liban, du Sud algérien, ... Et, au terme de ma première année d'assistant, mon projet de voyage était l'Iran et l'Afghanistan. J'y suis parti avec un ami, et, en arrivant à Kaboul, nous avons eu, sur une route de montagne, un très mauvais accident. J'ai eu la cheville droite écrasée et le visage fort abîmé. C'est sur une civière que je suis rentré à Bruxelles après une semaine éprouvante à l'hôpital de Kaboul (cela se passe avant l'invasion soviétique du pays) où mes blessures se sont vilainement gangrenées. J'ai subi plusieurs opérations et frôlé l'amputation. Finalement, grâce à l'acharnement d'un chirurgien orthopédiste et après bien des mois passés dans un lit d'hôpital, mon pied a été sauvé. Je m'ennuyais ferme, et un de mes amis qui terminait une thèse me proposa de relire son manuscrit, qui me parut tout à fait intéressant, ce qui me décida à l'imiter. C'est ainsi que j'entamai la rédaction d'une thèse de doctorat... en Théorie des Jeux. En sortant de l'hôpital, j'ai conservé le poste d'assistant que, dans mes projets initiaux, je pensais abandonner pour une carrière dans les assurances et ai pris goût à la vie à l'université, où j'ai fait de nombreuses rencontres. En particulier, c'est à cette époque-là que nous avons fait connaissance, Jean-Jacques, puisque tu étais l'assistant en chef de ce gros service de Statistique et de Recherche opérationnelle de la Faculté des Sciences sociales, politiques et économiques où j'exerçais mes activités.

En 1976, j'ai soutenu ma thèse (sur les Jeux à Information incomplète qu'avaient récemment introduits John Harsanyi et Reinhard Selten – ce qui leur valut, en 1994, de partager avec John Nash le Prix Nobel en Économie). À cette époque – tu t'en souviens certainement – une thèse consistait en une « thèse principale » doublée d'une « thèse annexe ». Cette dernière devait prendre la forme d'un énoncé, qui ne devait pas être accompagné d'une preuve écrite mais faisait l'objet de questions lors de la soutenance. L'élégance voulait que cet énoncé se rattache à un sujet distinct de celui de la thèse principale. Puisque j'étais dans une équipe de statistique, je me suis naturellement tourné, à la recherche d'une idée, vers mes collègues et amis statisticiens. L'un d'eux, Guy Mélard, terminait la rédaction d'une thèse sur les séries temporelles et m'a suggéré l'idée d'un énoncé de thèse annexe sur les modèles ARMA à coefficients dépendant du temps. Et c'est ainsi, par une porte dérobée en quelque sorte, que j'ai fait une très discrète « entrée en statistique ».

Au bout de quelque temps, toutefois, j'ai compris que la Théorie des Jeux, si je désirais rester dans le milieu académique, n'était pas le choix idéal. La Théorie des Jeux ne m'ouvrirait pas les portes du Département de Mathématiques et, n'étant pas économiste, j'avais peu de chances de me faire recruter dans un Département d'Économie. La Statistique, en revanche, offrait des perspectives en mathématiques autant qu'en économie et en économétrie. J'ai donc décidé de me reconverter et me suis tourné tout naturellement vers ce qui avait été le sujet de ma thèse annexe. Encore fallait-il se construire une crédibilité scientifique dans un domaine auquel, il faut bien le dire, je ne connaissais vraiment pas grand-chose et dans lequel je n'avais d'autre publication que celle de ma thèse annexe, parue dans le *Journal of Multivariate Analysis*. J'ai donc essayé de trouver un problème pouvant passer pour statistique, et ai rédigé un manuscrit sur la factorisation spectrale des processus q -dépendants non stationnaires... un sujet plus proche de la géométrie des espaces L^2 que de la statistique inférentielle. Ne sachant à quelle revue envoyer mon texte, j'ai parcouru les rayons de la bibliothèque et avisé une revue à l'élégante couverture bleu pastel, imprimée avec soin sur du beau papier. J'ai ainsi envoyé mon manuscrit, en trois exemplaires, aux *Annals of Statistics* dont – autre preuve de mon ignorance abyssale – j'ignorais jusqu'au nom. Huit mois plus tard m'est revenue, sur ce papier bleu léger qui servait alors au courrier transatlantique, une lettre de l'éditeur en chef m'annonçant que mon article était accepté, qu'on avait constaté une petite erreur à la page 8 (j'avais écrit *subtract* au lieu de *subtract*) qu'il me faudrait corriger sur les épreuves qui me seraient envoyées très bientôt.

Je me suis dit : « Voilà, c'est comme ça que ça se passe... On envoie son article et puis huit mois plus tard, il est accepté moyennant la correction de quelques coquilles... » Inutile de te dire que cela ne m'est plus jamais arrivé ! Mon penchant pour le bleu pastel m'a toutefois bien servi peu après, quand un poste vacant s'est présenté en statistique au Département de Mathématiques où personne avant moi n'avait jamais publié dans les *Annals of Statistics* : j'ai hérité, avec beaucoup de chance – pour des raisons budgétaires, il s'agissait d'une vacance interne – des cours de statistique mathématique des années terminales du diplôme de licence en sciences mathématiques. Et, comme on dit en anglais, « if you don't know something, teach it ». Cela se passait en 1978. J'ai donc empoigné ce qui était considéré comme de bons ouvrages en statistique mathématique, principalement les livres d'Erich Lehmann, et j'ai appris tout en enseignant. C'est ainsi que je suis devenu *statisticien par accident*.

JJD : Voilà une aventure intéressante. Es-tu d'accord de parler, à présent, des rencontres qui t'ont permis de progresser, d'établir des contacts avec l'extérieur ?

MH : Oui. Je pense — et je ne suis sûrement pas le seul dans ce cas — que nos carrières sont grandement construites sur les rencontres que nous avons (ou pas) la chance de faire. Ces rencontres sont autant de bifurcations aléatoires qui déterminent notre évolution intellectuelle. J'ai déjà parlé de ma rencontre avec Guy Mélard, qui a été le point de départ de mon activité dans le domaine des séries chronologiques (un domaine dans lequel je travaille toujours). Et il y en a eu d'autres, bien entendu. Toujours à la même époque, je me suis lié d'amitié avec Jean-François Ingenbleek, un autre doctorant, qui travaillait sur les tests de rangs. À cette époque-là, nous avions un séminaire, mais les ressources ne permettaient pas, comme on le fait aujourd'hui, d'inviter des intervenants extérieurs à qui il aurait fallu payer un déplacement, une nuit d'hôtel, etc. Le séminaire consistait essentiellement à choisir un livre que nous lisions collectivement, chacun d'entre nous en présentant, semaine après semaine, un chapitre. Cette année-là, le livre que nous étudions était celui de Puri et Sen intitulé *Nonparametric Methods in Multivariate Analysis* (1971). Et c'est ainsi, et grâce au fait que Jean-François Ingenbleek rédigeait sa thèse sur des problèmes de tests de rangs, que j'ai rencontré le deuxième sujet qui a été un des intérêts principaux de ma carrière, et sur lequel je travaille encore aujourd'hui : l'inférence fondée sur les rangs. La thèse d'Ingenbleek était un travail tout à fait remarquable dans lequel il introduisait des statistiques de rangs sérielles, permettant l'inférence dans les modèles de séries chronologiques. Nous avons naturellement commencé à collaborer sur ce sujet et avons fait la connaissance de Madan Puri, qui nous a fortement encouragés. Ce fut mon premier contact avec le monde académique américain et, dans les années qui suivirent, je fis de nombreux séjours au Département de Mathématiques de l'Indiana University à Bloomington.

D'autres rencontres ont ponctué ma carrière et ont joué un rôle important dans ma vision de la statistique. Certaines, et qui ne sont pas des moindres, se sont faites à travers mes lectures. Parmi les ouvrages qui m'ont marqué figure le livre de Hajek et Sidak, *Theory of Rank Tests* (1967), et tout particulièrement le chapitre consacré à l'asymptotique où apparaissaient pour la première fois les fameux trois « lemmes de Le Cam », qui auraient aussi bien pu être ceux de Hajek lui-même. Cette « asymptotique à la Le Cam » était à l'époque difficile à cerner car les quelques articles publiés par Lucien Le Cam étaient assez difficiles à lire, parfois contradictoires, hésitants sur la terminologie et les définitions, et surtout très parcellaires. Il était bien évident que les « trois Lemmes » du Hajek et Sidak n'étaient que la pointe émergée d'une théorie infiniment plus puissante et générale. Et quand enfin, près de vingt ans plus tard, a été annoncée, en 1986, la parution chez Springer de *Asymptotic Methods in Statistical Decision Theory*, cet ouvrage que tout le monde attendait, nous avons tous pensé que, finalement, nous allions tout comprendre. Malheureusement, la lecture de ce livre monumental (près de huit cents pages) n'est pas des plus aisées, et il me fallut personnellement plusieurs années pour en absorber les éléments essentiels. C'est un livre tout à fait remarquable, et Lucien Le Cam à mes yeux figure au premier rang des statisticiens les plus importants du XX^e siècle. Un de mes sujets de fierté est de lui

avoir obtenu le seul doctorat honoris causa qui lui fut jamais attribué, et qui lui fut décerné par l'Université libre de Bruxelles en 1997. Je suis également à l'origine de la création de la *Conférence Lucien Le Cam* organisée chaque année depuis 2001 en son honneur dans le cadre des *Journées de Statistique*, conférence annuelle de la Société Française de Statistique.

C'est vers la même époque, en 1996, que j'ai fait la connaissance de Bas Werker, qui venait de soutenir à Tilburg une thèse sur l'inférence dans les modèles semi-paramétriques de séries chronologiques. Bas m'a si fortement impressionné lors de notre rencontre que, lorsque peu après un poste a été déclaré vacant à Bruxelles en statistique, je l'ai convaincu de poser sa candidature. Il a été sélectionné et a passé chez nous quatre années pendant lesquelles nous avons beaucoup travaillé ensemble, à la fois sur l'asymptotique à la Le Cam et sur ses conséquences en inférence semi-paramétrique. Cette rencontre aussi a beaucoup compté dans mon parcours.

Une autre rencontre que je dois absolument mentionner est celle de Marco Lippi, à la fin des années quatre-vingt-dix. Professeur à la Sapienza à Rome, Marco est un mathématicien qui s'est tourné vers l'économie et l'économétrie, et qui m'a énormément appris. Cette rencontre a été le point de départ de mon activité dans le domaine des séries chronologiques en grande dimension et des modèles à facteurs dynamiques. Et c'est aussi un sujet sur lequel je continue à travailler.

Pour terminer, j'ai eu la chance d'avoir également à la même époque (début des années 2000) un étudiant particulièrement brillant, Davy Paindaveine, avec qui j'ai fait les premiers pas en direction d'une théorie de l'inférence fondée sur les rangs pour des observations multivariées. Et ce que nous avons fait ensemble constitue une théorie assez complète de l'inférence fondée sur les rangs et les signes que nous avons appelés *rangs* et *signes elliptiques* ou *de Mahalanobis*. Ces rangs jouissent, dans le contexte multivarié, des mêmes propriétés que les rangs ordinaires dans le cas univarié.

Malheureusement, l'utilisation des rangs et signes de Mahalanobis est limitée à des observations de loi elliptique. Cette hypothèse d'ellipticité, qui affaiblit considérablement les hypothèses gaussiennes de l'analyse multivariée classique, reste cependant fort restrictive, l'idéal étant d'arriver à une théorie d'inférence fondée sur les rangs pour des densités entièrement quelconques. Ce n'est pas un problème simple parce que, bien évidemment, l'espace réel, à partir de la dimension 2, n'est plus totalement ordonné comme l'est la droite réelle. La définition des concepts de rangs et de signes, ainsi que de celui de quantile, devient donc problématique. Ce problème est resté ouvert pendant un demi-siècle, et a été à l'origine, notamment, du développement des diverses notions de *profondeur statistique*. Malgré mes efforts dans cette direction, je tournais un peu en rond sans trouver de solutions, jusqu'au jour où j'ai fait, à nouveau, une rencontre aussi décisive qu'improbable. Me trouvant, suite à une invitation, à l'Université de Kyoto et me promenant dans les couloirs, je découvre l'annonce d'un séminaire au Département d'Économie, sur un sujet qui ne m'attirait pas particulièrement. Mais l'orateur, Marc Henry, était de mes amis, et je ne l'avais plus vu depuis un certain temps. Par amitié et pour le plaisir de le revoir, j'ai décidé d'aller l'entendre. Marc Henry, en fait, parlait d'une théorie, les transports de mesures, qui m'était alors inconnue. Et ce fut l'illumination ! J'ai tout de suite compris que les transports de mesures étaient l'outil dont j'avais besoin pour définir des rangs et des quantiles multivariés jouissant des propriétés désirées. Ce séminaire à Kyoto a ainsi été le point de départ de mes résultats récents, qui étendent au cas général les méthodes d'inférence précédemment développées dans le cadre elliptique. Une leçon peut être tirée de cette anecdote : toujours participer au séminaire, même quand le sujet annoncé n'est pas le centre de nos préoccupations !

JJD : Je crois que tu nous as très clairement présenté les étapes essentielles de ton parcours scientifique. Je te propose à présent de revenir un petit peu en arrière, à l'époque de tes travaux avec Puri et Ingenbleek, quand nous nous sommes tournés vers la communauté des statisticiens français regroupés, à l'époque, au sein de l'ASU, l'*Association des Statisticiens Universitaires*, ancêtre de la *Société Française de Statistique* (SFdS).

MH : Oui, c'était une période de pionniers. Nous étions alors un petit groupe à l'Université de Bruxelles à travailler sur les séries chronologiques. Il y avait toi, il y avait également Guy Mélard et moi, auxquels il faut ajouter Jean-François Ingenbleek. Nous nous sentions un peu isolés dans une Belgique où, de surcroît, les tensions communautaires étaient encore vives et où les communautés francophone et néerlandophone se regardaient un peu en chiens de faïence après les remous consécutifs à la scission de l'Université de Louvain. Dieu merci, les choses ont bien changé depuis, les réticences ont été aplanies, et nous sommes désormais les meilleurs amis du monde. Ce n'était cependant pas encore le cas en 1978, et notre soif de contacts scientifiques nous a amenés tout naturellement à regarder vers la France. Je crois que c'est toi, Jean-Jacques, qui nous a appris qu'une société de statistique française, l'Association des Statisticiens Universitaires (ASU), organisait une rencontre annuelle qui, cette année-là, se tenait à Nice. Nous avons de commun accord, décidé d'acheter chacun un aller-retour Bruxelles-Nice en train-couchettes et avons débarqué au petit matin sur les bords de la Méditerranée. Ce fut le début d'une participation importante de la Belgique dans les activités de cette société qui allait devenir au fil du temps la Société Française de Statistique. L'ASU, à cette époque, était une société relativement restreinte. Ce colloque de Nice était sans sessions parallèles ni grande foule, et la communauté des statisticiens français se trouvait encore marquée de la querelle entre analyse des données et statistique inférentielle. Les années passant, cette société a grandi et sa réunion annuelle est devenue la principale rencontre des statisticiens de langue française. De façon quasi permanente, le Bureau élu de l'ASU, puis celui de la SFdS, a compté un membre belge ; nous en avons été tous les deux, et avons chacun tenu, à un certain moment, le rôle de vice-président de la société.

La tradition s'est instaurée, par ailleurs, tous les 10 ans, d'organiser la conférence annuelle – les *Journées de Statistique* – à Bruxelles. Nous avons commencé, toi (Président du Comité d'Organisation) et moi (Président du Comité Scientifique) en 1982. Puis il y a eu 1992, 2002, 2012, année où Catherine Vermandele nous a rejoints, et il aurait dû y avoir 2022 si la pandémie n'avait tout remis à 2023.

JJD : Quittons l'international, si tu le veux bien, pour revenir en Belgique et à la manière dont tu as vécu l'évolution de la statistique dans ce pays, et à l'Université libre de Bruxelles en particulier.

MH : En fait, ce qui manquait à la statistique belge, au moment où nous avons intégré ce domaine, c'est une structure qui favorise les rencontres et une certaine identité statistique nationale. Il existait, je devrais dire il avait existé, une Société belge de Statistique, mais cette société était tombée, comme la belle au bois dormant, dans une catalepsie profonde. Et depuis plusieurs décennies, elle n'avait plus aucun membre faute de paiement de cotisations, de réunions, et faute d'activités. Cette société était ainsi devenue une société fantôme. Et nous avons décidé, un groupe de statisticiens de notre génération – tu en étais, Jean Jacques, mais également nos collègues de l'Université catholique de Louvain, Léopold Simar, Pierre Dagnelie des Facultés agronomiques de Gembloux, Jan Beirlant de la KU Leuven, Noël Veraverbeke de l'Université de Hasselt, et quelques autres encore – qu'il était temps de posséder au niveau belge une structure qui nous permette de nous connaître, de nous rassembler, et nous rendre plus visibles. Et il était naturel que cette structure s'appelât en français *Société belge de Statistique* et, en néerlandais, *Belgische Vereniging voor Statistiek*. Nous nous sommes rapidement aperçus que nous ne pouvions utiliser légalement ce nom parce qu'il existait déjà. Et c'est toi, Jean-Jacques, qui a découvert à la lecture des statuts de cette société morte, que même en l'absence

de paiement de cotisations, l'ensemble des membres belges de l'*International Statistical Institute* (ISI) en étaient membres *ex-officio*. L'idée a donc été de ressusciter la société au départ de ce noyau de membres de l'ISI. Une des difficultés était le grand déséquilibre communautaire parmi les membres belges de l'ISI – sept francophones et un ou deux néerlandophones seulement, ce qui à l'époque menaçait de poser des difficultés. Néanmoins, nous avons réuni ces membres, qui se sont constitués en assemblée en vue d'une réforme des statuts, et avons pu faire passer toutes les procédures légales en vue d'un redémarrage et du recrutement de nouveaux membres. L'opération fut un grand succès. Les problèmes linguistiques furent aplanis par un recours systématique à l'anglais. Jef Teugels fut (1991-1992) le premier Président élu de la société ressuscitée, qui tint en 1993 son premier colloque annuel au bord de la mer du Nord, au Coq (De Haan en néerlandais). Je succédai (1993-1995) comme Président à Jef Teugels. Tout s'est fort bien déroulé depuis, la société a grandi. Désormais, tous les statisticiens, certainement tous les statisticiens académiques du pays, en sont membres et le colloque annuel, organisé par alternance dans nos deux communautés linguistiques, est l'occasion d'amicales retrouvailles. Un épisode piquant de la saga de cette résurrection est l'apparition, lors d'une des premières assemblées générales, d'un inconnu d'un certain âge, porteur d'un gros sac de cuir, qui dans le silence général s'est avancé vers la tribune du président de séance (moi, en l'occurrence) en annonçant d'une voix grave : « Je suis le trésorier de la Société belge de Statistique ». Et il s'agissait effectivement du dernier trésorier de la société qui seul, pendant toutes ces années de dormance, avait soigneusement payé sa cotisation de membre, et avait géré les avoirs de la société. Sa sacoche de cuir contenait l'ensemble des documents comptables et il venait demander décharge de sa fonction de trésorier. Nous avons ainsi pu récupérer les quelques biens de la société. Nous avons également bénéficié de l'enthousiaste soutien de Claude Chéruiy, qui était à l'époque directeur de l'*Institut National de Statistique belge*. Ce fut un combat que nous avons mené avec succès, tous ensemble, au sein de la communauté statistique belge.

JJD : Peux-tu aussi nous parler de la statistique au sein de l'Université libre de Bruxelles ?

MH : Il s'agit là d'un autre combat dans lequel, une fois de plus, nous avons été partenaires, mais qui se solda, lui, par une cuisante défaite. L'Université de Bruxelles avait été, dans l'immédiat après-guerre, la première en Belgique à créer un *Institut de Statistique et de Recherche opérationnelle*. Cet institut n'avait pas tout à fait le statut de Faculté ; en particulier, il ne contrôlait pas le cadre et les postes de statistique dans l'ensemble de l'université. Néanmoins, il regroupait en son sein l'ensemble des statisticiens de l'université, qu'ils soient membres de la Faculté des Sciences ou de la Faculté des Sciences économiques et politiques, de la Faculté de Médecine ou de l'École des ingénieurs ; il organisait des enseignements et délivrait des diplômes en statistique et en recherche opérationnelle. Cette structure avait été créée, dans un mouvement vraiment visionnaire, par mon prédécesseur Paul Gillis, analyste de renom, au retour, dans les années d'après-guerre, d'un voyage aux États-Unis. Il avait entrevu là-bas la révolution dans les mathématiques appliquées que constituait l'émergence de la statistique et de la recherche opérationnelle, et conclu qu'il fallait absolument faire quelque chose en Europe pour leur emboîter le pas. Ainsi fut créé cet *Institut de Statistique et de Recherche opérationnelle* qui, au fil du temps, par manque de moyens et de cadre, s'était peu à peu assoupi. Au début des années 1990, il nous est apparu (toi, moi, Guy Mélard, Philippe Vincke et quelques autres) éminemment regrettable que cet instrument de recherche et d'enseignement reste tristement sous-utilisé et avons entrepris d'y remédier. Ensemble, nous avons ainsi revu les statuts de cet Institut et avons essayé de lui donner une visibilité et des activités plus importantes. En particulier, nous nous sommes tous chargés volontairement, sans rémunération, bien entendu, de cours supplémentaires afin de développer l'offre d'enseignement dans le domaine de la statistique. Cette entreprise a connu un grand succès. Nous avons des étudiants, des étudiants en thèse, un séminaire régulier. L'économétrie qui, au sein de la Faculté des Sciences économiques et de l'École de Commerce, connaissait quelques difficultés, s'est rapprochée de nous qui avons le grand privilège légal de pouvoir décerner des diplômes. Et nous avons opéré une forme de fusion

entre la statistique et la partie quantitative de la recherche en économie à Bruxelles. Un centre de recherche de pointe de la faculté des sciences économiques qui s'appelait *ECARE – European Center for Advanced Research in Economics* – a rejoint en partie l'Institut de Statistique comme, disons, faculté de tutelle, se transformant en *ECARES – European Center for Advanced Research in Economics and Statistics*. Ensemble, nous avons créé un DEA en économie et statistique, en fait un DEA en économétrie à forte composante statistique. Grande innovation à l'époque, ce DEA était entièrement enseigné en anglais de façon à attirer des étudiants étrangers, ce que nous avons remarquablement réussi à faire, grâce notamment au recrutement de collègues étrangers. Quelqu'un qui a joué un rôle très important dans cette évolution est Lucrezia Reichlin, maintenant professeur à la London Business School après avoir dirigé le département recherche de la Banque centrale européenne à Francfort. Nous avons ainsi pu attirer de très nombreux étudiants étrangers, allant jusqu'à diplômer 25 étudiants par an, dont beaucoup ont poursuivi par des thèses de doctorat. Ce succès, en interne, je pense, n'a pas plu vraiment, ni à la Faculté des Sciences, ni à la Faculté des Sciences économiques. Au moment où toutes les autres universités belges et européennes, voire mondiales, développaient les enseignements de statistique et de science des données, l'Université de Bruxelles, qui au niveau belge avait fait la course en tête pendant tant d'années, a dissout son Institut de Statistique et réduit à néant des années d'efforts.

Cette décision m'a profondément affecté, et a considérablement assombri la fin de ma carrière à Bruxelles, conduisant à ma demande de mise à la retraite anticipée à l'âge de 60 ans. Mais de cette décision regrettable au plan collectif est né un bénéfice personnel. À 60 ans, je me trouvais bien entendu beaucoup trop jeune pour arrêter mes activités, et j'ai écrit à un certain nombre d'universités américaines de la côte est, en proposant mes services. La première qui a répondu à mes messages est Princeton, ce qui m'a valu de passer là, pendant les six années qui ont suivi, un semestre par an, en charge du cours de statistique mathématique. Princeton est un lieu très inspirant, très confortable sur le plan professionnel, où j'ai eu d'excellents étudiants et d'excellents collègues. Ces six années ont été, d'une certaine façon, la partie la plus heureuse de ma carrière académique.

JJD : Je crois que c'est peut-être le moment de passer à un autre sujet intéressant, à savoir ta vision de l'évolution de la statistique pendant ta carrière et la façon dont tu vois cette discipline dans un avenir plus ou moins immédiat.

MH : Je crois que peu de disciplines, en un demi-siècle, ont connu des transformations, voire des révolutions, aussi fondamentales et profondes que la statistique. La statistique d'aujourd'hui n'a plus grand-chose à voir avec celle que nous pratiquions il y a cinquante ans. Jamais nous n'aurions pu prédire les bouleversements qu'a connus notre discipline. Et il y a là-dedans du bon et du moins bon. Je veux dire qu'il y a des progrès évidemment nécessaires et même salutaires, qu'il faut saluer. Il y en a d'autres qui peut-être sont regrettables.

Nous avons vu arriver successivement des techniques et des idées comme la naissance de la statistique non paramétrique, les estimateurs à noyau, les techniques d'apprentissage statistique et d'intelligence artificielle. On s'est mis à estimer des densités et des fonctions, à travailler sur des variétés et dans des espaces de grande dimension... Toutes ces choses étaient inimaginables il y a 50 ans, où on était bien content alors de se débrouiller avec les familles exponentielles où tout se passait bien. Après la statistique non paramétrique, il y a eu l'asymptotique à la Le Cam et, dans la foulée, toute l'inférence semi-paramétrique, sans oublier les techniques de rééchantillonnage et le bootstrap, et les méthodes comme la régression quantile, qui sont absolument fantastiques comme outil, comme quantité d'informations que l'on peut extraire de données.

Et puis, il y a d'autres choses qu'on a vues apparaître puis s'éclipser rapidement, et qui sont

peut-être des modes plus passagères. À un certain moment, tout le monde faisait des modèles ARCH et GARCH, des modèles de cointégration, des processus à mémoire longue, des lois à queues lourdes, des estimateurs de James-Stein... Beaucoup de ces sujets ont été à une certaine époque l'objet d'une très grande activité et le sont peut-être un peu moins aujourd'hui.

Et puis aujourd'hui, ce à quoi on fait face est lié également à un besoin très réel dans les applications, ce sont les méthodes d'inférence en grande dimension. Cela, je pense que c'est une vague de fond et pas simplement une mode passagère. Il y a également tout ce qui nous vient de disciplines annexes, principalement l'informatique, qui sont les réseaux neuronaux dont on ne peut que saluer l'efficacité, sans savoir toujours très bien à quoi elle est due, tout ce qui est méthode d'apprentissage et d'intelligence artificielle. Et peut-être que cette évolution-là est la plus significative, mais aussi la plus dangereuse pour notre discipline dans la mesure où elle nous vient pour une large partie de l'extérieur. Les départements d'informatique sont partout bien plus gros et les chercheurs y sont bien plus nombreux que les départements de statistique. Et la recherche qui a trouvé son origine dans ces départements d'informatique ne connaît pas toujours ou ne fait pas toujours suffisamment attention à l'accumulation de connaissances et de la sagesse due à un demi-siècle de réflexion statistique. Les problèmes qu'ils traitent sont des problèmes statistiques même si on les habille désormais d'une terminologie différente. On va nous parler de science des données, on va nous parler de Data Analysis, mais tout ça, c'est de la statistique. Et il faudrait prendre garde à ce que la sagesse accumulée par plus d'un demi-siècle de réflexion statistique ne soit pas perdue dans l'enthousiasme de ces nouvelles méthodes et de ces nouveaux développements. Des résultats tels que les théorèmes de Rao-Blackwell, de Lehmann-Scheffé, de Basu, ... ne doivent pas être oubliés au prétexte que les réseaux neuronaux sont une panacée d'une redoutable efficacité !

Je pense que l'avenir réside dans une meilleure et plus étroite collaboration entre les statisticiens et ceux qui, à l'extérieur de la statistique, dans les équipes d'informatique essentiellement, font de l'analyse des données, de l'apprentissage et de l'intelligence artificielle. Je pense que nous avons, de façon réciproque, beaucoup de choses à nous apporter. Le danger réside dans le déséquilibre des effectifs de chercheurs et des moyens financiers. Nous sommes relativement petits, eux numériquement très importants. Je pense que c'est dans un meilleur dialogue et une coopération renforcée que se situent les défis et les enjeux contemporains de notre discipline.

JJD : On vient de faire le tour du Marc Hallin statisticien. Je crois qu'il faudrait que tu nous parles un peu de l'homme que tu es, de ses goûts, de ses centres d'intérêt. Que peux-tu nous dire aussi de ta famille, de ce qui fait que tu es « un homme comblé », si tu l'es ?

MH : Je suis l'heureux père de trois filles, toutes trois remarquables dans des styles différents, et le grand-père de huit petits-enfants qui tous habitent à distance de marche de chez nous. Et donc, mon épouse Dominique et moi les voyons avec bonheur de façon assez régulière. Maintenant que je suis retraité, je passe pas mal de temps avec mon épouse dans une maison que nous avons la chance de posséder en Provence. En particulier, chaque année, je procède à la récolte de mes olives, que je porte religieusement au moulin coopératif pour en retirer une huile de qualité exceptionnelle et couronnée de plusieurs médailles d'or et d'argent au Salon de l'Agriculture de Paris...

JJD : Je voudrais aussi, pour les avoir partagés, que tu nous parles un petit peu de tes goûts pour la lecture, pour l'art, pour la musique.

MH : Une chose qui a joué dans ma vie un rôle très important tout au long de ces années au cours desquelles j'ai été très absorbé par les devoirs familiaux et les activités académiques, c'est la pratique de la musique. En particulier, j'ai chanté pendant de très nombreuses années dans le pupitre des ténors de la Chorale royale protestante de Bruxelles. Ce fut pour moi une expérience

très importante. Musicalement, cela m'a permis non seulement d'écouter de la musique, mais d'en faire ; la position des ténors dans le chœur est particulièrement intéressante. Le chant est principalement aux mains des sopranos ; les basses, comme leur nom l'indique, s'occupent des parties basses de l'harmonie. Les ténors et les altos se trouvent entre les deux tessitures, et c'est eux qui tissent le lien entre le chant et l'accompagnement. J'ai ainsi eu la chance de chanter les œuvres majeures du répertoire choral, les messes de Mozart et les passions de Bach, les Requiem de Brahms, Mozart, Cherubini, Verdi, les Stabat Mater de Rossini et Dvorak... Toutes ces œuvres extraordinaires, chacun peut les écouter, mais les chanter, c'est toucher du doigt leur beauté de façon bien plus intense. Et en même temps, ces années de chorale m'ont permis de rencontrer une variété de personnes qui sont devenues souvent des amis et que je n'aurais jamais rencontrées dans mon milieu professionnel ni dans mon milieu social. L'amour de la musique rassemble en effet des gens venus d'horizons extrêmement divers. Hélas, j'ai dû interrompre cette activité lorsque j'ai commencé à enseigner à Princeton et n'ai plus pu participer de façon régulière aux répétitions. À Princeton, j'ai fait partie également de la chorale de l'université. Mais cette chorale-là était davantage dirigée vers le chant à l'office. Et donc, j'ai chanté très régulièrement dans la prestigieuse chapelle de Princeton, le dimanche matin, en tige noire et orange. Mais c'était moins gratifiant que le travail que nous faisons ici à la Chorale royale protestante où vraiment on travaillait en profondeur des œuvres plus exigeantes. À mon retour, j'ai été gagné par la limite d'âge – la voix ne s'améliore pas avec le temps, et il y a un âge-limite pour le chant comme pour l'enseignement académique.

JJD : Est-ce que tu peux nous parler de ton goût pour les vieux livres, pour la peinture, pour l'art ?

MH : J'ai toujours été un grand amateur de voyage. Comme je l'ai dit, mes expériences de jeunesse se sont terminées, malheureusement, de façon un peu abrupte à Kaboul. Mais professionnellement, ce qui est magnifique dans notre métier, c'est que nous avons l'occasion de voyager dans des conditions qui ne sont pas celles d'un touriste normal, parce que nous sommes invités dans divers endroits où nous rencontrons des gens remarquables, collègues ou étudiants, avec qui nous partageons un langage et des intérêts communs. Mes pérégrinations académiques m'ont permis de découvrir des pays et des lieux absolument fascinants, auxquels je suis très attaché. Je pense au Japon, par exemple, où je suis allé de très nombreuses fois grâce à mon ami Masanobu Taniguchi. Je pense aussi à Rome, où j'ai passé pendant six années de suite, le mois de janvier à l'invitation de mon ami Marco Lippi. Je pense bien entendu à Princeton, à Madrid ou à Bloomington, où j'ai passé beaucoup de temps et dont je conserve un souvenir ému. Pour le reste, je suis plutôt un voyageur des villes que des campagnes. J'adore déambuler dans les villes. Peut-être est-ce lié au fait que j'ai longtemps été empêché de marcher à cause de mon pied accidenté ? À côté des promenades urbaines, j'ai une prédilection toute particulière pour les musées, les musées de peinture surtout. Quand je vais quelque part, ma première visite est toujours pour le musée des beaux-arts local.

JJD : Merci Marc, de nous avoir parlé de toi, du scientifique que tout le monde connaît, mais aussi de l'homme que j'ai la chance d'avoir pour ami.

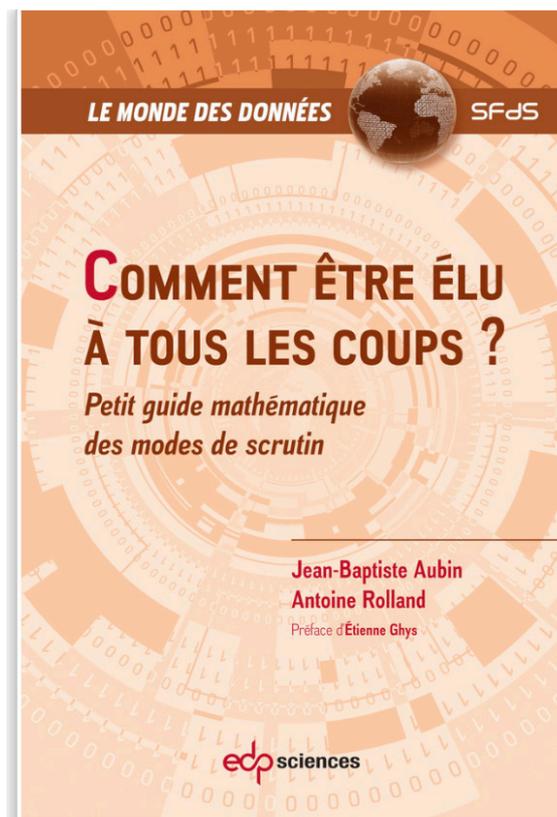
Comment être élu à tous les coups ? Petit guide mathématique des modes de scrutin

de
Jean-Baptiste AUBIN et Antoine ROLLAND
(2022)



Antoine HOULOU-GARCIA¹

Chercheur associé au CESPRA (Centre d'études sociologiques et politiques Raymond-Aron)



Livre (185 pages)

Direction : Jean-Baptiste AUBIN et Antoine ROLLAND

Édition : EDP Sciences (Collection : Le monde des données) – 2022

ISBN : 978-2759826841

1. antoine.houlou@ehess.fr

La première chose à dire concernant cet ouvrage concerne son titre : *Comment être élu à tous les coups ?* n'est clairement pas la question à laquelle il répond. Il semble que ce soit un choix de l'éditeur plus que des auteurs car, dans sa préface, Etienne Ghys se réfère au sous-titre qui devait probablement être le titre originel : *Petit guide mathématique des modes de scrutin*. Ce sous-titre correspond en effet à l'intention affichée des auteurs.

Le point de départ de Jean-Baptiste Aubin et Antoine Rolland est concret et nous ramène aux fameuses « soirées électorales » où plusieurs candidats revendiquent la victoire, où l'on tente de décrypter la volonté du peuple à partir des bulletins qu'il a glissés dans l'urne. Dès lors, la question qu'ils se posent est d'investiguer « la manière de rendre compte des préférences collectives des votants » (p. 13).

Ainsi, les auteurs se proposent d'établir un catalogue raisonné des méthodes de vote les plus usitées dans la pratique et les plus connues et étudiées par la théorie mathématique du vote. Il s'agit d'un vrai guide présentant les procédures électorales avec de brefs éléments historiques contextuels (qui les a imaginées, à quelle époque). Ce guide a l'avantage, grâce à la distinction visuelle entre texte et encadrés, d'être clair pour le lecteur rétif aux mathématiques et assez précis pour un lecteur habitué aux formules.

On balaye ainsi les grands résultats (théorème d'Arrow, théorème de Gibbard-Satterthwaite, etc.), les grandes notions (vainqueur et perdant de Condorcet, anonymat et neutralité, etc.), ainsi, bien sûr, que de nombreuses procédures électorales (Borda, Copeland, Minimax, etc.). On notera la volonté pédagogique, omniprésente au fil des pages, de rendre pratique le discours théorique sur les procédures : en invitant par exemple le lecteur à sélectionner les conditions du théorème d'Arrow qu'on trouve les plus importantes (p. 84) ; ou encore en illustrant le caractère non trivial de la propriété de monotonie par un astucieux développement sur l'élection présidentielle française de 2002 (p. 44) ; en précisant enfin la notion de perdant de Condorcet avec le cas de l'élection présidentielle sénégalaise de 2012 (p. 40). Le chapitre 7 est dédié à l'application des différentes procédures à des cas théoriques mais assez parlants.

Le choix d'une procédure de vote ou d'une autre peut impacter directement le résultat d'une élection. Prenons un exemple, qui n'est pas dans le livre, mais qui a été proposé par Borda en 1770. On demande aux 21 électeurs de classer par ordre de préférence les trois candidats A, B et C. Le tableau suivant donne les votes :

N° de l'électeur	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Premier choix	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	C	C	C	C	C	C
Deuxième choix	B	C	C	C	C	C	C	C	C	C	C	C	C	C	C	B	B	B	B	B	B
Troisième choix	C	B	B	B	B	B	B	B	A	A	A	A	A	A	A	A	A	A	A	A	A

Si on s'en tient à la première préférence (scrutin uninominal, comme on le fait généralement en France), alors A doit gagner avec 8 voix contre B qui en a 7 et C qui en a 6. Mais on voit également que A est très souvent classé dernier. Changeons la procédure et attribuons, à peu près comme à l'Eurovision ou pour le Ballon d'or, 3 points à un candidat classé en premier choix par un électeur, 2 pour en deuxième et 1 en troisième ; on peut calculer le score de chaque candidat :

- A est classé 8 fois premier et 13 fois troisième, il obtient donc $8 \times 3 + 13 \times 1 = 37$ points ;
- B est classé 7 fois premier, 7 fois deuxième et 7 fois troisième, il obtient donc $7 \times 3 + 7 \times 2 + 7 \times 1 = 42$ points ;
- C est classé 6 fois premier, 14 fois deuxième et 1 fois troisième, il obtient donc $6 \times 3 + 14 \times 2 + 1 \times 1 = 47$ points.

La comparaison des résultats donne ainsi C vainqueur. De plus, A, qui avait gagné, se retrouve bon dernier. La sensibilité du résultat à la procédure est donc un sujet particulièrement

important.

Pire : de façon contre-intuitive, l'abstention peut faire gagner votre camp. Voici un extrait du livre d'Aubin et Rolland qui l'explique :

« Revenons dans le passé en 1988 pour l'élection présidentielle française, qui voyait s'affronter François Mitterrand (gauche), Raymond Barre (centre) et Jacques Chirac (droite). D'après les sondages de l'époque, Raymond Barre était certainement le candidat de Condorcet, c'est-à-dire qu'il aurait certainement été élu au deuxième tour contre tout autre candidat. Mais les deux candidats en tête du premier tour ont été F. Mitterrand et J. Chirac, et le deuxième tour a vu la réélection de F. Mitterrand. La majorité des électeurs de droite, partisans de J. Chirac, auraient probablement préféré l'élection de R. Barre à celle de F. Mitterrand. Et si certains de ces électeurs n'étaient pas allés voter au premier tour, J. Chirac, privé de ses voix, serait arrivé troisième du premier tour et R. Barre deuxième. Le second tour aurait donc vraisemblablement vu la victoire de R. Barre contre F. Mitterrand. Nos électeurs de droite auraient donc eu un résultat plus satisfaisant en s'abstenant qu'en participant au premier tour ! »

Dans cette veine, on trouve divers cas passionnants de résultats ou stratégies très rusées qui ont permis à des élections d'être gagnées alors qu'elles devaient être perdues dans un ouvrage du politologue américain William Riker, *The Art of Political Manipulation* (New Haven, Yale University Press, 1986), qu'Aubin et Rolland ne semblent pas connaître : ils ne le citent pas en bibliographie et ne citent pas les exemples astucieux de l'élection de Lincoln, du jugement raconté par Pline le Jeune et bien d'autres qui auraient pu leur servir d'illustration réelle de stratégies de vote et des effets de changement de procédure.

On soulignera l'humour très présent dans l'ouvrage, qui en rend la lecture d'autant plus agréable, ainsi que les nombreuses citations, mises en épigraphes de façon très régulière, qui « sortent » l'analyse mathématique de son aspect nécessairement un peu rugueux. On notera néanmoins qu'aucune n'est précisément référencée et que de nombreuses sont dues à des auteurs que le lecteur ne connaîtra pas forcément, même de nom, ce qui rend leur pertinence parfois délicate².

Afin d'éviter tout malentendu, précisons enfin que ce livre, malgré quelques intentions apparentes, ne permet pas d'apprécier en toute généralité le rôle du vote dans la démocratie, ni l'histoire de l'analyse mathématique du vote³. Ce n'est pas davantage un ouvrage proposant une réflexion épistémologique : on part du principe qu'il existe une « volonté du peuple » à mesurer par le biais de procédures électorales, or rien n'assure que la volonté du peuple puisse être mesurée ni que l'expression ait un sens ; de vifs débats existent à ce sujet. Et le regard posé sur le sens de certaines propriétés mathématiques n'apparaît pas toujours assez critique : comme p. 37, où il est indiqué que « dans une démocratie digne de ce nom, il semble impossible d'imaginer un système de vote où certains votants verraient leurs préférences invalidées avant même de voter ». Or la question de savoir si une personne antidémocratique peut s'exprimer dans une démocratie est une vraie question de fond, à laquelle, par exemple, l'Athènes du V^e siècle avait répondu par la négative.

Ces quelques remarques n'enlèvent toutefois rien à l'intérêt de l'ouvrage, qui se présente en définitive comme un petit manuel très utile au débat, dans la perspective de donner des outils de comparaison sur les procédures électorales et permet de vivifier un débat qui a besoin de clarifications dans la sphère citoyenne.

2. En guise d'exemple à ce propos, p. 111 « A. Pierce » est en réalité le facétieux Ambrose Bierce, peut-être ici confondu avec Charles Sanders Peirce, dans son *Dictionnaire du Diable*.

3. Une erreur a d'ailleurs été relevée à ce sujet p. 65, où est présentée la méthode de Borda, qui en a proposé en réalité deux, et citée une phrase qui n'est pas présente dans son discours de 1770.

Adolphe Quetelet

Passeur d'idées

de
Jean-Jacques DROESBEKE
(2021)



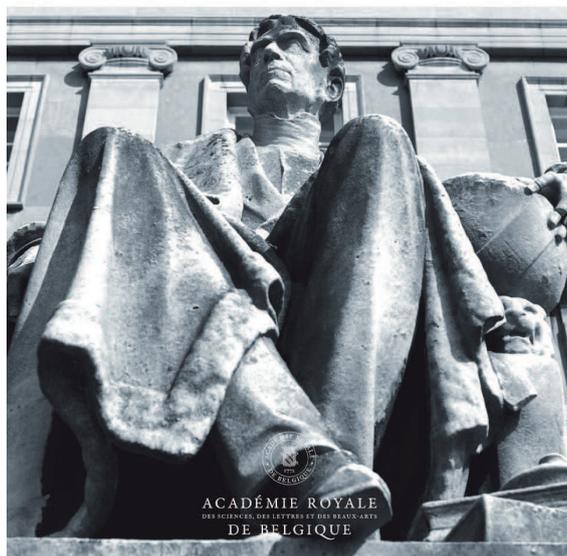
Emmanuel DIDIER¹

Directeur de recherche au CNRS, Membre du Centre Maurice-Halbwachs
(ENS et EHESS)

JEAN-JACQUES DROESBEKE

ADOLPHE QUETELET
PASSEUR D'IDÉES

NOUVELLE BIOGRAPHIE NATIONALE. SUPPLÉMENT 2



Livre (422 pages)

Direction : Jean-Jacques DROESBEKE

Édition : Académie royale des sciences, des lettres et des beaux-arts
de Belgique (Collection : Nouvelle biographie nationale. Supplément
2) – 2021

ISBN : 978-2-8031-0806-0

1. emmanuel.didier@ens.fr

Jean-Jacques Droesbeke nous livre ici une nouvelle biographie en 422 pages de ce grand scientifique belge que fut Adolphe Quetelet. L'auteur, qui est lui aussi Belge, membre de l'Académie Royale de Belgique et statisticien, comme son héros, avoue dans l'avant-propos qu'il y travaille depuis 40 ans ! Il s'agit donc de l'œuvre d'une vie, à savourer comme telle.

On entre dans le texte en contemplant d'abord la photo de couverture, qui demande cependant une explication. En effet, celle-ci montre la statue hiératique du grand homme, en contre-plongée, regardant bien au-dessus du lecteur. Une photo de la grandeur. Mais le livre, et c'est son grand intérêt, ne prend pas cette grandeur pour acquise. Au contraire, il raconte, en restant tout du long rigoureusement à hauteur d'homme, comment Quetelet a lui-même gravi les échelons de la société pour produire cette position dominante. En fait, la photo dévoile la fin de l'histoire, mais ne rend pas compte de sa progression.

D'autres avant Jean-Jacques Droesbeke, et pas des moindres comme Steve Stigler ou Ted Porter, s'étaient déjà attelés, au moins partiellement, à la biographie de cet homme. Mais l'ouvrage dont il est question ici est exceptionnel d'abord pour la quantité d'informations qu'il nous livre sur la vie de son protagoniste, vie privée comme professionnelle. Des documents nouveaux sont produits – on a même un bulletin de notes du jeune Adolphe (qui était excellent ; p. 63) ou encore le portrait fait par son beau-fils du personnage vieillissant jouant avec sa petite-fille (p. 254). En observant ce tableau, on voit que Quetelet portait volontiers un bonnet, ce qui nous permet d'insister sur le fait que, souvenons-nous-en une fois pour toutes, son nom, lui, ne porte pas d'accent. On écrit bel et bien Quetelet (même si on prononce « Quètelet »).

Insistons aussi sur la beauté de ce livre comme objet : le papier est splendide, les reproductions impeccables. Seule une Académie peut encore nous offrir des volumes si précieux. Regrettons simplement qu'il soit si mal référencé : un an après sa sortie, nulle trace dans le catalogue de la Bibliothèque nationale de France, pas davantage dans celui de la librairie de Harvard ou du Congrès américain. Il est regrettable que l'Académie Royale de Belgique ne fasse pas les efforts supplémentaires qui permettraient de mieux diffuser et faire connaître ses belles productions.

Entrons maintenant dans le texte. Le style est très vivant, très agréable à lire. L'argument n'est pas « internaliste » pour reprendre un concept classique de l'histoire des sciences. Il ne s'agit pas d'expliquer la pensée scientifique de Quetelet. Cet aspect du personnage est traité ailleurs, dans un autre livre en deux volumes, écrits par Michel Armatte et (encore une fois) par Jean-Jacques Droesbeke, intitulé *Quetelet et les probabilités*, qui sortira dans les mois prochains aux Éditions de l'INED. Droesbeke, le grand connaisseur de Quetelet devant l'Éternel, a donc préféré séparer la vie de l'œuvre de son personnage principal.

Ce livre-ci est une histoire sociale centrée sur un personnage. Le contexte est très finement dessiné. L'auteur précise avec une méticulosité impressionnante les évolutions institutionnelles de la Belgique dans laquelle a vécu Quetelet, laquelle a changé plusieurs fois de régime et traversé une révolution pendant la période. Les transformations de ce que nous appellerions aujourd'hui l'enseignement supérieur et la recherche dans lequel navigue Quetelet sont pointées avec une étonnante acuité ; la Révolution belge de 1830 explicitée avec une grande clarté.

Dans ce contexte, certains aspects de la vie de Quetelet sont particulièrement mis en avant. D'abord, insistons sur son ambition. Jeune, il est polymathe, s'intéresse à tout, publie des poèmes, soutient une thèse de doctorat en mathématiques. Il est d'une curiosité inextinguible. En même temps, il cherche à faire carrière et à se faire remarquer. Il construit un réseau de personnes qui l'appuieront ou qui lui serviront de ressources. Les hommes qui ont exercé sur lui le plus d'influence sont d'ailleurs présentés en fin d'ouvrage dans une série de brèves biographies, annexes d'une grande utilité. Quetelet conquiert aussi des institutions de plus en

plus prestigieuses : d'abord à Gand, puis à Bruxelles où il est élu membre de l'Académie Royale en 1820. Il voyage aussi beaucoup dans toute l'Europe, d'abord à Paris qui bouleversera sa vie par les rencontres qu'il parviendra à y faire avec Arago, Fourier, Laplace et bien d'autres, puis aussi, plus tard, en Allemagne, en Italie et en Grande-Bretagne.

Cette ascension sociale ne se fait pas sans vie familiale, qui est elle aussi précisément rapportée : celle de sa famille ascendante comme celle qu'il construit avec sa femme et leurs enfants, la famille de sa femme, avec qui il s'entendait bien et qui par ailleurs a pu aussi être utile pour sa carrière.

Un des points sur lequel insiste Droesbeke est que, contre toute attente, l'investissement intellectuel de Quetelet dans les statistiques semble en fait fortuit. La grande et véritable affaire de sa vie – outre l'ambition personnelle – était l'astronomie. La part la plus importante de son énergie était d'abord de parvenir à faire construire un observatoire astronomique à Bruxelles, puis d'y habiter et de le diriger.

Son engagement dans la statistique n'apparaît ainsi que comme un simple passe-temps, alors que ses ambitions astronomiques n'avancent pas ! Voici comment Droesbeke narre cet épisode : nous sommes en 1830, une grande conférence internationale réunie à Londres décidera du sort politique de la Belgique et indirectement de l'observatoire désiré par Quetelet puisque celui-ci ne pourra pas être construit sans volonté ni financement gouvernemental. « En attendant [que l'avenir de la Belgique et donc de son observatoire soit scellé], pourquoi ne pas développer d'autres centres d'intérêt ? L'intensité magnétique reste bien sûr un sujet de prédilection, mais cela ne suffit pas à remplir ses journées. Il se lance dans une série de réflexions issues de ses travaux statistiques auxquels il s'est récemment intéressé » (p. 190). L'intuition qui a fait la célébrité mondiale de Quetelet, celle de l'homme moyen, ne lui est donc venue que par désœuvrement, en attendant de pouvoir faire progresser sa véritable passion, l'astronomie ! Beau paradoxe. Insistons cependant sur le fait que le succès rencontré dès le premier ouvrage que Quetelet a écrit sur les statistiques fut immense et qu'il a su, ensuite, capitaliser dessus.

Il ressort de ce livre une image de Quetelet moins comme inventeur de concepts que d'abord comme créateur d'institutions : une revue, un observatoire, la Commission centrale de statistique – ancêtre d'une institution de la statistique publique fondée en 1840 –, l'Institut International de la Statistique (IIS) dont il a organisé la réunion inaugurale en 1853.

On garde aussi une image fort aimable de Quetelet. L'auteur n'a pas rapporté les critiques qui n'ont pas manqué d'être exprimées contre un homme qui, gravissant les échelons sociaux, a nécessairement contredit les ambitions d'autres personnes. Une exception, une citation de Mailly, son ami, qui tance sa hâte perpétuelle : « Dès qu'il avait obtenu un résultat, il s'empressait de le faire connaître, souvent même avant que l'idée ne fût parvenue à maturité, et avant que le résultat ne fût parfaitement sûr » (p. 212). Droesbeke a préféré faire de son héros un modèle plutôt que d'en détailler les facettes moins dignes d'être imitées. Il a suivi le parcours qui va vers la statue.

Pour finir, permettons-nous de regretter le choix initial de séparer la vie de l'œuvre. Nous ne lirions probablement pas la biographie de Quetelet s'il n'avait inventé l'homme moyen, or l'homme moyen est presque absent de ce livre. Nous le lisons donc comme à l'ombre de cette grande idée. Pourtant, l'historiographie contemporaine a souvent montré l'intérêt qu'il y a à conjoindre les préoccupations personnelles et théoriques. Les *Sciences and Technology Studies*, qui ne sont presque pas citées par l'auteur, ont remis en cause la frontière entre la science et son contexte de façon souvent très productive. Entrelacer les arguments contenus dans ce livre et ceux de l'autre livre à venir aurait sans doute permis de présenter une image de Quetelet et de son temps encore plus complète et riche.

Mais ne boudons pas notre plaisir ! Ce livre-ci restera et sera lu avec le plus grand profit, plaisir et intérêt par tous les passionnés des personnalités engagées dans la statistique, de Belgique, du XIX^e siècle et d'astronomie.