COMPROMIS ENTRE ERREUR DE NON-RÉPONSE ET ERREUR DE MESURE DANS COSET-MSA

Gaëlle Santin[1] (,**), Pauline Delézire (**), Béatrice Geoffroy (**), Laetitia Bénézet (**),Jean Bouyer (***), Alice Guéguen (*)*

(*) UMS 011 - Unité Cohortes épidémiologiques en population, Inserm, Uvsq
(**) InVS, Département Santé Travail
(***) UMRS 1018 Inserm, Ups, Uvsq - Equipe 4

Résumé

L'objectif était d'étudier l'erreur de non-réponse et l'erreur de mesure selon la difficulté à joindre les personnes dans le pilote de l'enquête Coset-MSA sans et avec correction de la non-réponse.
Dans cette enquête conduite en 2010, un questionnaire postal relatif à la santé et à l'emploi a été envoyé à 10000 personnes tirées au sort (enquête initiale) parmi les actifs du Régime agricole en 2008 (taux de réponse 24%). Une enquête complémentaire par interview avec un questionnaire restreint a été réalisée auprès d'un échantillon aléatoire de 500 non-répondants (taux de réponse 63%). La combinaison des deux enquêtes correspondait à une enquête en deux phases pour non-réponse. Pour l'ensemble des 10000 personnes de l'échantillon initial, les données individuelles de santé issues du SNIIRAM et des données professionnelles issues des systèmes d'informations de la MSA ont pu être exploitées. Quatre variables étaient à la fois recueillies par questionnaire et disponibles via ces bases médico-administratives (BMA) : statut d'emploi salarié, secteur d'activité primaire, surface agricole utile en ares pour les non-salariés, contrat de travail en CDI pour les salariés. En considérant les statistiques issues des BMA comme des statistiques de référence, l'erreur de mesure, l'erreur de non-réponse, et l'erreur totale, approchée par la somme des deux précédentes erreurs, ont été estimées pour chaque variable selon la difficulté à joindre les personnes sans et avec correction de la non-réponse. La difficulté à joindre les personnes était considérée comme faible pour les répondants à l'enquête initiale et élevée pour les répondants à l'enquête complémentaire.
Sans correction de la non-réponse, l'erreur de non-réponse était en général plus élevée pour l'enquête initiale que pour l'enquête en deux phases. Comparée à l'enquête initiale, l'erreur de mesure était légèrement supérieure à l'enquête en deux phases. Avec correction de la non-réponse, l'erreur de non-réponse était nettement réduite ; l'erreur totale était due avant tout à l'erreur de mesure et était en général équivalente pour l'enquête initiale et pour l'enquête en deux phases.
Ces résultats indiquent qu'avec correction de la non-réponse, l'enquête initiale seule permet d'obtenir des estimations de prévalence aussi satisfaisantes que celles obtenues grâce à l'enquête en deux phases.

Abstract

The aim was to study the nonresponse and the measurement errors related to the difficulty to reach persons in the Coset-MSA pilot study, with and without nonresponse adjustment.
The Coset-MSA sampling design was a two-phase sampling for nonresponse. The respondents to the first phase (initial survey) were categorized as easy-to-reach, whereas the respondents to the subsample of nonrespondents of the first phase (complementary survey) were categorized as hard-to-reach.
Without nonresponse adjustment, nonresponse error was generally greater for the initial survey than for the two-phase survey. Compared to the initial survey, measurement error was slightly greater for the two phase survey. With nonresponse adjustment, nonresponse error decreased significantly; the total survey error was mostly due to measurement error and was generally equivalent for the initial survey and for the two phase survey.

[1] gaelle.santin@inserm.fr

Mots-clés

Biais, erreur de non-réponse, erreur de mesure, bases médico-administratives, surveillance épidémiologique

1. Introduction

Dans le cadre de l'erreur totale dans les enquêtes [1], le taux de réponse est un paramètre important qui peut influer sur l'erreur de non-réponse et de l'erreur de mesure. L'idée qu'il faut maximiser autant que possible le taux de réponse aux enquêtes pour minimiser les biais de non-réponse est encore largement dominante malgré certains articles récents remettant en cause cette façon de procéder. D'un autre côté, maximiser le taux de réponse peut également augmenter les biais de mesure.

1.1. Revue de la littérature

Cette revue de la littérature a pour objet de faire une synthèse des travaux réalisés sur les liens entre probabilité de réponse et biais de non-réponse, probabilité de réponse et biais de mesure, probabilité de réponse et variance, et enfin, probabilité de réponse, biais de non-réponse et biais de mesure.

1.1.1. Probabilité de réponse et biais de non-réponse

De plus en plus d'articles récents suggèrent que la baisse des taux de réponse n'altère pas nécessairement les estimations issues d'une enquête [2] ; après avoir résumé les résultats de 30 articles ayant produit plus de 200 estimations, Groves conclut que la plupart des variations des biais de non-réponse surviennent à l'intérieur d'une même enquête mesurant des variables d'intérêt différentes plus qu'entre différentes enquêtes ayant des taux de réponse variant de 15% à 70%. Autrement dit, maximiser le taux de réponse global sans autre critère n'est pas nécessairement efficace en termes de biais de non-réponse. Par ailleurs, chercher à maximiser autant que possible le taux de réponse à une enquête peut s'avérer contre-productif. L'exemple tiré de l'article de Merkle et coll. [3] est significatif en ce sens. Il est issu d'une enquête pré-électorale qui avait pour objectif d'estimer les intentions de vote aux élections présidentielles américaines de 2008. Sans contrepartie, la participation des sympathisants démocrates et des sympathisants républicains à l'enquête était équivalente. En revanche, en offrant un stylo en cas de participation à l'enquête, les sympathisants démocrates participaient plus à l'enquête que les sympathisants républicains, ce qui conduisait finalement à une surestimation des intentions de vote envers le candidat démocrate. Le recours à un cadeau reposait sur l'hypothèse le taux de réponse pouvait augmenter sans que la covariance entre la variable d'intérêt et la propension à réponse ne soit modifiée, avec en corollaire, une diminution du biais. L'exemple ici montre que le recours au stylo a augmenté le taux de réponse, ce qui était attendu, mais a également augmenté la covariance entre la probabilité de réponse et la variable d'intérêt, et a ainsi augmenté le biais de non-réponse, malgré l'augmentation du taux de réponse.
Dans ce contexte, au lieu de viser un taux de réponse maximal global, de nouveaux protocoles d'enquête développés actuellement [4, 5] cherchent, pour un budget donné, à maximiser le taux de réponse de groupes de personnes particulières, potentiellement génératrices de biais de non-réponse. Ces protocoles d'enquête sont appelés « plan de collecte adaptative » (« responsive design » en anglais). Ils sont définis comme une « approche adaptative où l'information disponible est utilisée pour modifier la collecte des données pour les personnes restantes » [5]. Ils ont été formalisés par Groves et Heeringa [4]. Deux notions sont associées à ce schéma d'enquête : la phase et la capacité d'une phase. Une phase correspond à une période de la collecte de données pendant laquelle le même protocole est utilisé (par exemple, un questionnaire postal avec plusieurs relances). La capacité d'une phase est la condition de stabilité d'une estimation dans une phase spécifique ; autrement dit, une phase a atteint sa capacité lorsqu'il n'est plus utile de continuer la collecte des données selon le protocole choisi pour cette phase (par exemple lorsqu'on considère que le nombre maximal de répondants a été atteint sous un protocole d'enquête donné pour un budget donné).
Ainsi, même si ce formalisme est assez récent (2008), la pratique du « responsive design » est ancienne ; il est en effet courant depuis longtemps de changer de mode de recueil de données au cours de la collecte pour enquêter le plus de personnes possibles. De la même manière, les enquêtes en deux phases pour non-réponse [6] entrent également dans la définition du « responsive design ».
Depuis quelques années, des « responsive design » plus élaborés sont en pleine expansion. Ils nécessitent en général d'avoir précédemment réalisé une enquête similaire permettant d'accéder à des informations auxiliaires également disponibles dans l'enquête à réaliser.

Peytchev [7] propose d'utiliser l'enquête précédente pour modéliser la probabilité de réponse en fonction des informations auxiliaires, d'affecter ce modèle de prédiction à l'échantillon auprès duquel l'enquête doit être réalisée afin de disposer d'une estimation de la probabilité de réponse avant la collecte et de faire plus d'efforts de collecte auprès des personnes avec une probabilité de réponse *a priori* faible. C'est un point important car il considère qu'une personne avec une propension de réponse initialement faible peut avoir une probabilité de réponse plus élevée avec un processus de collecte de données différent. Il conclut cependant que cette stratégie n'est pas forcément efficace en terme de biais et qu'il vaudrait mieux identifier les groupes de personnes induisant potentiellement le plus de biais dans les estimations d'un paramètre pour une variable d'intérêt donnée plutôt que les groupes de personnes avec une probabilité de réponse basse.
L'approche adaptative proposée par Lundquist et Sarndal [5] se base sur une distance à minimiser entre la moyenne estimée chez les répondants et la moyenne estimée pour l'échantillon tiré au sort pour des variables auxiliaires expliquant la variable d'intérêt.
Schouten [8] propose l'indicateur R construit à partir d'une probabilité de réponse prédite par un modèle ; l'indicateur R est dit de bonne qualité si les probabilités de réponse prédites ont une petite variabilité. La collecte des données peut être conduite en fonction de cet indicateur afin de réduire au maximum la variabilité des réponses prédites par ce modèle en utilisant, comme dans le cas de Peytchev, les probabilités de réponse prédites dans une enquête similaire. Cette approche est critiquée par Beaumont et col. [9] pour deux raisons : si le modèle de non-réponse inclut uniquement des variables peu associées à la non-réponse alors qu'il existe des variables associées à la probabilité de réponse et aux variables d'intérêt, les probabilités de réponse prédites par le modèle vont être peu dispersées et l'indicateur de représentativité va conduire à un indicateur R de bonne qualité alors que les « vraies » probabilités de réponse sont en fait dispersées et n'ont pas été correctement estimées. Une autre limite vient du fait que des probabilités de réponse dispersées ne signifient pas nécessairement qu'il y a un biais de non-réponse ; en effet, si les variables associées à la non-réponse ne sont pas associées à la variable d'intérêt, il n'y aura pas de biais de non-réponse.

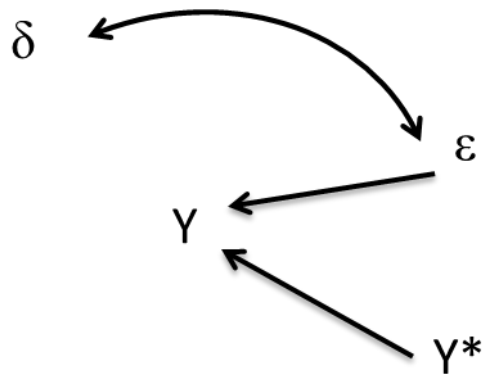
Quelle que soit l'approche choisie, elle permet de diminuer le biais lié à la non-réponse si les informations auxiliaires utilisées expliquent le lien entre la probabilité de réponse et la variable d'intérêt. Mais, comme le soulignent Beaumont et col. [9], cette approche, qui nécessite des informations auxiliaires disponibles chez les répondants et les non-répondants, n'apporte pas plus en terme de correction du biais de non-réponse qu'un ajustement pour la non-réponse par repondération qui utiliserait ces mêmes variables. Beaumont et col [9], plutôt que d'utiliser des « responsive design » pour diminuer les biais de non-réponse, proposent de les utiliser plutôt pour diminuer la variance liée à la non-réponse.

La stratégie optimale pour recueillir les données afin de limiter les effets du lien entre la probabilité de réponse et le biais de non-réponse est donc assez complexe. Pour Beaumont et col [9], pour diminuer le biais de non-réponse, le plan de sondage adaptatif le mieux indiqué est l'enquête en deux phases pour non-réponse ; dans ce cas, il suffit de tirer au sort des personnes parmi les non-répondants et que celles-ci répondent pour obtenir une estimation sans biais. Groves suggère pour sa part en conclusion de son article relatif à cette question [10] que, dans un contexte où les taux de réponse aux enquêtes diminuent de plus en plus, il faut avant tout rechercher autant que possible des informations auxiliaires qui peuvent être exploitées pour réduire les effets de la covariance entre la probabilité de réponse et les variables d'intérêt d'une enquête.

1.1.2. Probabilité de réponse et biais de mesure

Le lien entre la probabilité de réponse et le biais de mesure a été évoqué dès 1963 [11] et peut s'expliquer ainsi : les répondants les moins motivés pour participer à une enquête, et pour lesquels on suppose que la probabilité de réponse est faible, ont tendance à répondre de manière plus imprécise voire erronée que les répondants spontanés [12]. Si on suppose que ce manque motivation est corrélé avec la difficulté à joindre les personnes, alors les personnes qui n'auraient pas répondu spontanément mais qui répondent après plusieurs relances seraient susceptibles de générer des biais de mesure plus grands que les personnes ayant répondu spontanément [13, 14].
Ce lien potentiel a été formalisé par Groves [2] et représenté par la Figure 1.

Figure 1 : Représentation graphique du lien entre probabilité de réponse et biais de mesure (d'après [2])



Soit *Y** la vraie valeur de la variable d'intérêt, ε l'erreur de mesure et *Y* la mesure de la variable d'intérêt ; on suppose qu'il n'existe pas de lien entre *Y** et ε. Supposons que la probabilité de réponse δ soit liée à l'erreur de mesure ε (par exemple par le manque d'intérêt pour l'enquête de la personne interrogée). Ce lien va engendrer une liaison entre la probabilité de réponse et la variable d'intérêt mesurée *Y*, même si en réalité il n'existe pas de lien entre la probabilité de réponse et la variable d'intérêt *Y**.
Il est difficile de quantifier ce biais de mesure car il faudrait pour cela disposer, pour chaque répondant, à la fois de la vraie valeur de la variable d'intérêt et de sa valeur mesurée. Comme c'est rarement le cas, des études par simulation ont été réalisées pour évaluer quantitativement ce problème [12].
En pratique, on peut mesurer la proportion de données manquantes partielles selon la probabilité de réponse pour avoir une idée des biais de mesure potentiels ; Fricker et Tourangeau [14] qui étudient des variables socioéconomiques et Dahlhamer [13] qui étudie des variables relatives au recours au soin ou à l'emploi, ont trouvé que la proportion de données manquantes partielles tendait à augmenter chez les personnes avec une probabilité de réponse faible en comparaison aux personnes avec une probabilité de réponse élevée.
Peu d'études ont mesuré le lien entre probabilité de réponse et biais de mesure. Les rares qui ont pu le faire disposaient pour les répondants à leurs enquêtes de variables identiques disponibles à la fois dans le questionnaire et dans des systèmes d'information existants, ces dernières pouvant servir de gold standard [15, 16]. Deux études ont montré que généralement le biais de mesure augmentait légèrement ou restait stable lorsqu'on le comparait chez les personnes avec une propension à être contactée élevée ou dans l'échantillon de répondants complet ; les variables étudiées étaient relatives au statut matrimonial (durée du mariage, nombre de mois depuis le divorce, nombre de mariages dans l'étude de Olson) et sociodémographiques ou de santé (bénéficiaire de l'aide sociale, statut vis-à-vis de l'emploi, âge, nationalité étrangère). Olson a trouvé des résultats similaires en étudiant le biais de mesure selon la propension à coopérer (probabilité de réponse des personnes ayant réussi à être contactées par un enquêteur). Ces deux études ne montrent pas d'augmentation significative des biais de mesure selon la probabilité de réponse.

1.1.3. Probabilité de réponse et variance

La non-réponse génère une variance supplémentaire assimilable à celle observée dans une enquête en deux phases. Ainsi, dans une enquête simple, plus le taux de réponse est élevé, plus la variance liée à la non-réponse est petite. Néanmoins, si, pour obtenir un taux de réponse maximal, on a recours à une enquête en deux phases pour non-réponse, la variance a plutôt tendance à augmenter.
L'erreur de mesure génère une variance supplémentaire ; si on suppose que le biais de mesure augmente lorsque la probabilité de réponse diminue, on peut également supposer qu'il engendre une plus grande variabilité dans les estimations donc une augmentation de la variance.

1.1.4. Relation entre probabilité de réponse, biais de mesure, et biais de non-réponse

Les points précédents montrent que les relations entre probabilité de réponse et biais de non-réponse, biais de mesure ou variance ne sont pas simples à comprendre même lorsqu'ils sont appréhendés indépendamment les uns des autres. Quelques auteurs ont tenté de les étudier conjointement quand il est possible de disposer de données gold standard.

Peytchev [17], dans une étude cherchant à quantifier la prévalence d'avortements à partir des données d'une enquête nationale américaine (taux de réponse 68%), a trouvé que les personnes avec une probabilité de réponse faible étaient plus susceptibles de sous-déclarer une expérience d'avortement que les personnes avec une probabilité de réponse élevée. Il suppose que biais de mesure et biais de non-réponse proviennent d'une cause commune : la désirabilité sociale qui incite les personnes ayant connu un avortement à ne pas répondre spontanément et qui les pousse à répondre de manière erronée lorsqu'on fait des efforts particuliers pour qu'elles répondent. Même si la donnée « gold standard » pour cette étude provient d'une information déclarée au cours de l'enquête, l'étude permet des discussions intéressantes. Après approximation de l'erreur quadratique moyenne (EQM), Peytchev conclut que l'EQM est supérieure lorsqu'il inclut le groupe de personnes avec une probabilité de réponse faible pour estimer la prévalence d'avortement que lorsqu'il ne l'inclut pas. En conclusion, l'auteur met en garde contre des maximisations naïves des taux de réponse qui peuvent ne pas améliorer la qualité générale des prévalences estimées d'avortement du fait de liens entre biais de non-réponse et biais de mesure.

Olson [16] est, à notre connaissance, la première à étudier et à mesurer la balance entre biais de non-réponse et biais de mesure selon la probabilité de réponse à partir de variables gold standard disponibles via des systèmes d'information existants pour l'ensemble des personnes tirées au sort et de variables identiques mesurées par questionnaire, donc uniquement pour les répondants à l'enquête. Elle se base sur une étude relative au statut matrimonial de la personne (taux de contact 80% et taux de coopération, égal au nombre de répondants divisé par le nombre de personnes contactées, 88%) et étudie trois variables : la durée de mariage, la durée écoulée depuis le divorce, le nombre total de mariages. Elle considère par ailleurs deux composantes de la probabilité de réponse : la propension à être contacté et la propension à coopérer après avoir été contacté. Le biais de mesure diminue avec la propension à être contacté pour l'estimation de la moyenne de deux variables ; il augmente avec la propension à coopérer pour l'estimation de la moyenne d'une seule variable. Quelle que soit la variable et la composante utilisée pour la probabilité de réponse, le biais de non-réponse tend à diminuer après inclusion des personnes ayant la probabilité de réponse la plus faible. Par ailleurs, quelle que soit la variable d'intérêt, le biais total (estimé ici par la somme du biais de mesure et du biais de non-réponse) diminue lorsque sont rajoutées dans l'analyse les personnes avec la propension d'être contacté la plus faible ; en revanche, le biais total ne diminue que pour une seule variable lorsque les personnes ayant la propension à coopérer la plus faible sont incluses dans l'analyse. Olson conclut que le lien entre probabilité de réponse, biais de non-réponse et biais de mesure dépend de la variable étudiée, de la statistique estimée et du type de non-réponse étudié.

Kreuter [15] étudie elle aussi les liens entre biais de mesure et biais de non-réponse grâce à l'utilisation de variables gold-standard issues de systèmes d'information existants. A partir d'une enquête relative à l'emploi (taux de réponse 27%), elle étudie plus exactement l'évolution du biais de non-réponse, du biais de mesure, de la variance et de l'erreur quadratique moyenne selon la propension à être contacté. Pour les trois variables relatives à l'emploi, elle trouve d'une part que les biais de non-réponse diminuent avec l'inclusion de personnes ayant une faible propension à être contacté et d'autre part que les biais de mesure ont tendance à augmenter avec l'inclusion de personnes ayant une faible propension de contact ; au final, le biais total est inchangé. Pour la variable relative au fait d'être bénéficiaire de l'aide sociale, elle observe un résultat à première vue surprenant : alors que le biais de non-réponse diminue avec la propension à être contacté et que le biais de mesure et la variance restent stables, l'erreur quadratique moyenne augmente. Cela se produit parce que les biais de non-réponse et les biais de mesure qui affectent cette variable sont contraires : si les personnes ayant une faible propension à être contactées sont des personnes ayant une probabilité de bénéficier de l'aide sociale élevée et si tous les répondants, quelle que soit leur propension à être contactés, surestiment leur probabilité de bénéficier de l'aide sociale, alors le biais total va être supérieur si on inclut les personnes avec une propension à être contactées faible.

Ces trois études montrent donc que pris conjointement, les liens entre probabilité de réponse, biais de non-réponse et biais de mesure sont complexes et que leur impact sur le biais total varie. Aucune d'entre elles cependant n'étudie le lien entre probabilité de réponse et biais de mesure après correction de la non-réponse.

1.2. Résumé de la problématique

Nous considérons ici qu'une personne difficile à joindre est une personne pour laquelle un niveau important d'efforts dans la collecte doit être mis en place pour obtenir sa réponse.
L'hypothèse que l'inclusion de personnes difficiles à joindre diminue l'erreur de non-réponse est de plus en plus remise en question [2, 18]. De plus, chercher à joindre autant que possible des personnes difficiles à joindre peut augmenter l'erreur de mesure car ces dernières peuvent répondre avec moins de précision, voire de manière sciemment erronée aux questions posées. On peut donc se demander si en termes d'erreur totale, il est vraiment utile de chercher à augmenter autant que possible le taux de réponse aux enquêtes si le gain en termes d'erreur de non-réponse est discutable et s'il peut entrainer une augmentation des erreurs de mesure.
Les quelques enquêtes ayant discuté ce compromis entre erreur de non-réponse et erreur de mesure selon la difficulté à joindre les personnes montrent des résultats différents. Par ailleurs, à notre connaissance, aucune de ces études n'a étudié cette balance après correction de la non-réponse et on peut se demander quel est l'intérêt de maximiser le taux de réponse dans le cas où il est possible de corriger le biais de non réponse grâce à des données auxiliaires.

1.3. Objectifs

L'objectif de ce travail est d'étudier le compromis entre erreur de non-réponse et erreur de mesure selon la difficulté à joindre une personne dans Coset-MSA avant et après correction de la non-réponse.
L'enquête pilote Coset-MSA étant une enquête en deux phases pour non-réponse, nous avons considéré que la difficulté à joindre les personnes était faible pour les répondants à l'enquête de première phase (appelée par la suite enquête initiale) et élevée pour les répondants à l'enquête réalisée auprès d'un sous-échantillon des non-répondants de l'enquête de première phase (appelée par la suite enquête complémentaire).

2. Population et méthodes

2.1. La cohorte pilote Coset-MSA

2.1.1. Présentation générale

La présente étude s'appuie sur les données de la phase pilote de la cohorte Coset-MSA (Cohorte pour la surveillance épidémiologique en lien avec le travail chez les travailleurs affiliés à la Mutualité Sociale Agricole) à l'inclusion [19]. Cette dernière a pour objectifs principaux de décrire à un instant « t » la morbidité des actifs affiliés à la Mutualité Sociale Agricole (MSA) selon l'activité professionnelle et son évolution, et de décrire l'évolution des liens entre la morbidité des actifs et les expositions professionnelles. La MSA est un régime d'assurance obligatoire qui couvre en maladie et en retraite des actifs du monde agricole en France : ils peuvent être non-salariés (agriculteurs ou chefs d'entreprise agricole) ou salariés (ouvriers agricoles, travailleurs de certaines banques et assurances, employés de coopératives agricoles, enseignants en lycée agricole etc.).

Avant de mettre en place la cohorte Coset-MSA à l'échelle nationale, les modalités de recrutement d'actifs du régime agricole ont été testées en 2010 sur cinq départements lors d'une phase pilote. Compte tenu du faible taux de réponse attendu et de la durée du suivi envisagée (au moins 20 ans), l'objectif de la phase pilote était par ailleurs d'étudier le mieux possible les biais potentiels de non-réponse à l'inclusion. C'est pour cela qu'en plus du recueil par questionnaire postal et de la consultation des bases de données issues de systèmes d'information existants, une enquête complémentaire auprès d'un échantillon de non-répondants à l'enquête postale a été réalisée spécifiquement pour l'enquête pilote. Les accords de la Cnil nécessaires ont été obtenus préalablement à la mise en œuvre des différentes étapes (dossiers Cnil n°909091, Cnil n°910191 et Cnil n°910176).

2.1.2. Echantillonnage

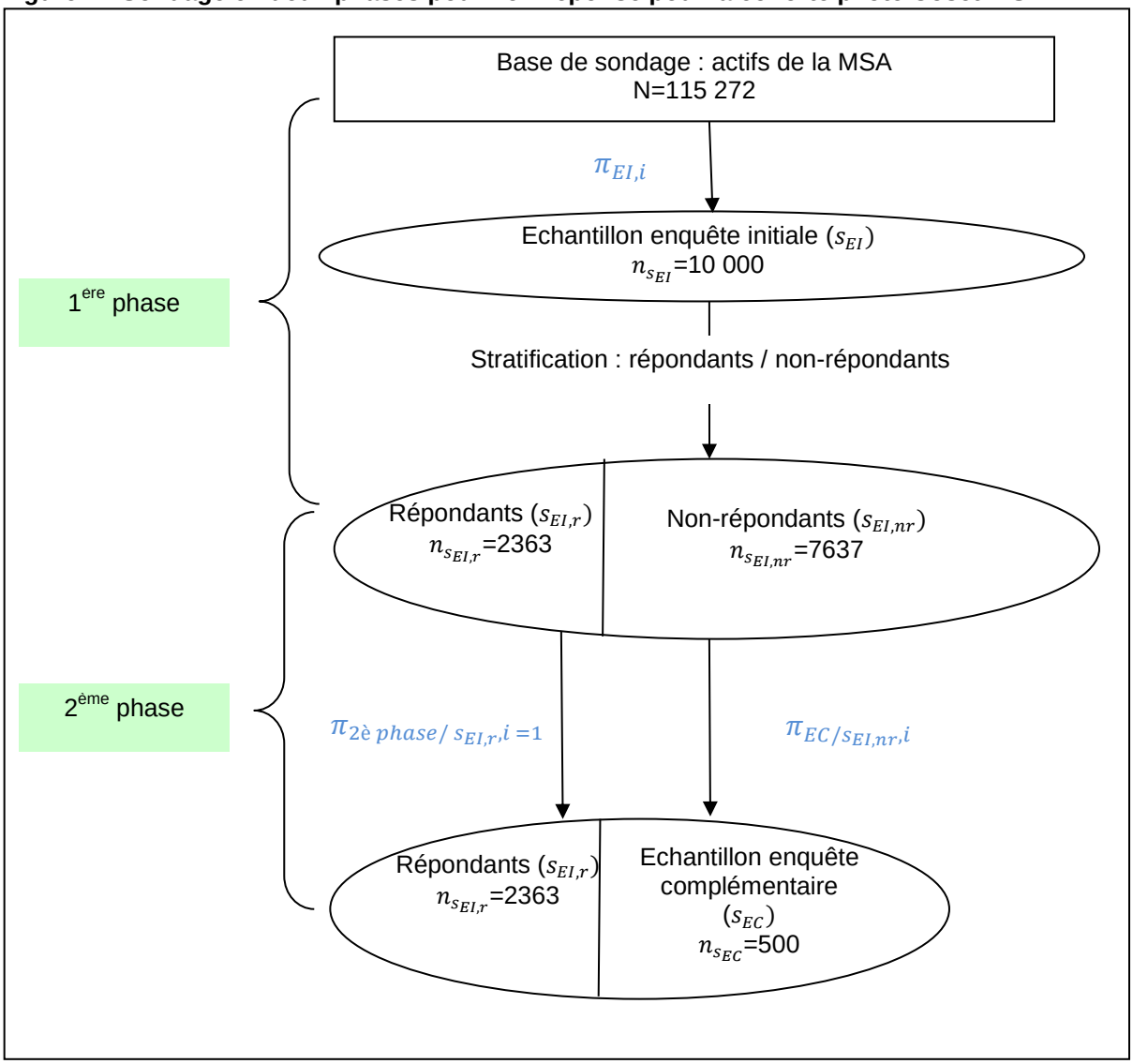
2.1.2.1. Population source

La population source pour la phase pilote correspond aux non-salariés agricoles et aux salariés affiliés à une des cinq caisses départementales pilotes (Bouches-du-Rhône, Finistère, Pas-de-Calais, Pyrénées Atlantiques, Saône-et-Loire), âgés de 18 à 65 ans et ayant travaillé au moins 90 jours en 2008 en tant qu'affiliés au Régime agricole, quel que soit leur type d'activité. Elle comprend environ 100 000 personnes.

2.1.2.2. Plan de sondage

Pour réaliser le tirage au sort, la base de sondage utilisée était la base d'assurance retraite, qui contient toutes les personnes ayant travaillé au moins une fois en tant qu'affiliées à la MSA. Le plan de sondage mis en place pour la phase pilote était un plan de sondage en deux-phases pour non-réponse (Figure 2).
Pour le pilote de Coset-MSA, la première phase a consisté à tirer au sort 10 000 personnes, soit 2 000 personnes dans chacun des cinq départements participant à la phase pilote. Dans chaque département, le tirage au sort, par sondage aléatoire simple sans remise, était stratifié selon le sexe, l'âge et le statut d'emploi (salarié ou non salarié) et était proportionnel à la taille des strates. Cette première phase correspondait au plan de sondage envisagé pour l'extension nationale, France entière. L'enquête de deuxième phase a été réalisée auprès d'un échantillon de 500 non-répondants ; 100 non-répondants par département ont été tirés au sort par sondage aléatoire simple.
L'enquête de première phase sera appelée par la suite enquête initiale, et l'enquête additionnelle menée lors de la deuxième phase, l'enquête complémentaire.
Pour simplifier la lecture du texte, l'enquête en deux-phases pour non-réponse sera appelée par la suite « enquête en deux phases ».

Figure 2 : Sondage en deux phases pour non-réponse pour la cohorte pilote Coset-MSA



N : taille de la population
EI : enquête initiale
EC : enquête complémentaire
s_{EI} *: échantillon issu de EI et* $n_{s_{EI}}$ *sa taille*
$s_{EI,r}$ *: sous-échantillon de répondants à EI et* $n_{s_{EI,r}}$ *sa taille*
$s_{EI,nr}$ *: sous-échantillon de non-répondants à EI et* $n_{s_{EI,nr}}$ *sa taille*
s_{EC} *: échantillon issu de EC et* $n_{s_{EC}}$ *sa taille*
$\pi_{EI,i}$ *: probabilité d'inclusion d'un individu i à l'enquête initiale*
$\pi_{2\text{è}\,phase/s_{EI,r},i}$ *: probabilité d'inclusion d'un individu i dans l'échantillon de deuxième phase sachant qu'il appartient à* $s_{EI,r}$
$\pi_{EC/s_{EI,nr},i}$ *: probabilité d'inclusion d'un individu i dans l'échantillon de deuxième phase sachant qu'il appartient à* $s_{EI,nr}$

2.1.3. Données recueillies

Deux types de données ont été recueillies : des données de questionnaire et des données issues de systèmes d'information existants, appelées données auxiliaires (Tableau 1).

Les données de questionnaires étaient disponibles pour les répondants à l'enquête initiale et à l'enquête complémentaire. Pour l'enquête initiale, le questionnaire, envoyé par courrier comportait 40 pages. Pour l'enquête complémentaire, le questionnaire était administré par un enquêteur (téléphone

ou face-à-face) ; il était issu du questionnaire de l'enquête initiale et l'interview durait environ 10 minutes.
Les questions posées étaient relatives à la santé, aux comportements à risque pour la santé, à l'emploi et aux conditions de travail.

Via les systèmes d'information existants, trois sources de données étaient disponibles théoriquement pour les 10 000 personnes tirées au sort : les variables de stratification issues de la base de sondage relatives à des variables sociodémographiques, les données extraites de l'Assurance Maladie (SNIIRAM) relatives à la santé et les données extraites de la MSA relatives au dernier emploi principal en date en tant qu'affilié à la MSA, et aux accidents du travail et maladies professionnelles.

- **Variables de stratification** *(en 2008)* : Sexe, classe d'âge (18-34 ans, 35-49 ans, 50-65 ans), département, statut (SA : salarié, NSA : non salarié).
- **Données de santé** (Assurance maladie, SNIIRAM) *(de 2008 à 2010)* :
  - *Remboursement de soins :* Nombre de recours à un professionnel de santé et type de professionnel de santé, nombre et type de boites de médicaments, montant total des prestations ;
  - *Absentéisme pour raison de santé - Indemnités journalières :* Montant total des versements, durée totale, type d'arrêts (pour maladie, accident du travail ou maternité) ;
  - *Hospitalisations* : Nombre d'hospitalisations, type d'établissement (public ou privé), motif (chirurgical ou non chirurgical), pathologie (grands chapitres de la 10ème révision de la classification internationale des maladies CIM10).
- **Données professionnelles** (Cotisations professionnelles, MSA) :
  - *Dernier emploi principal en date en tant qu'affilié à la MSA (2008, 2009 ou 2010*) : Statut (SA, NSA), secteur d'activité (selon la classification « code risque » interne à la MSA), durée d'emploi, nombre d'emplois salariés dans l'année ;
  - *Accidents du Travail et Maladies Professionnelles (ATMP) reconnus (2003 à 2008*) : Nombre d'ATMP, par gravité, par année.

Les données de l'Assurance maladie et des cotisations professionnelles n'étaient pas disponibles pour 693 personnes pour une des raisons suivantes :

- leur adresse postale était invalide (donc ces personnes n'ont pas été informées de l'étude et n'ont pas pu exprimer d'éventuel refus de participer à l'enquête) ;
- elles ont exprimé un refus d'accès aux bases SNIIRAM ou MSA ;
- leurs données SNIIRAM ou MSA n'ont pas pu être appariées.

Tableau 1 : Données recueillies et modes de recueil des données à l'enquête initiale et à l'enquête complémentaire

	Enquête initiale (EI)	Enquête complémentaire (EC)
Période d'enquête	Février-mars 2010	Novembre 2010-février 2011
Mode de recueil des données	Autoquestionnaire postal	Téléphone ou face-à-face avec enquêteur
Taille de l'échantillon :		
Tiré au sort	10 000	500
Avec données issues des SI (*)	9 307	454
Répondants au questionnaire	2 363	313
Données recueillies pour :		
Personnes tirées au sort	Variables de stratification	
Personnes tirées au sort n'ayant pas exprimé de refus d'accès aux bases médico-administratives	Données du SNIIRAM (de 2008 à 2010) : remboursements de soins, indemnités journalières, hospitalisations Données de la MSA : dernier emploi principal en date (2008, 2009 ou 2010), accidents du travail et maladies professionnelles reconnus (de 2003 à 2008),	
Répondants au questionnaire	Santé perçue, échelles de santé (TMS, symptômes dépressifs, asthme), consommation de tabac et d'alcool, historique d'emplois et d'expositions professionnelles	Santé perçue, consommation de tabac et d'alcool, emploi actuel, expositions professionnelles actuelles *(Questionnaire issu de l'EI)*

(*) SI : systèmes d'information existants= Assurance maladie et cotisations professionnelles

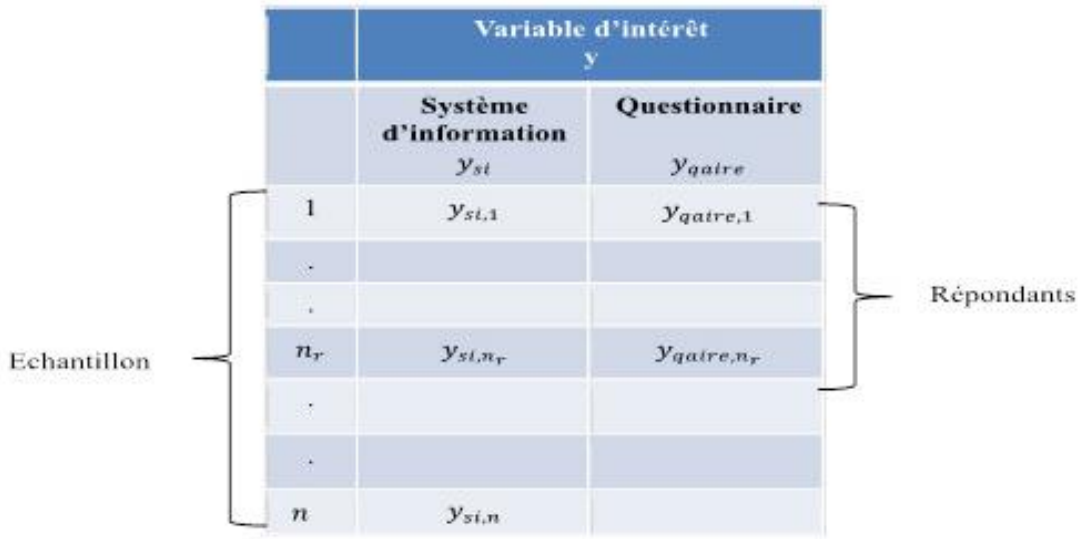
2.2. Analyses statistiques : erreur de mesure, erreur de non réponse et erreur totale selon difficulté à joindre les personnes

2.2.1. Présentation générale de la méthode

Cette méthode a été proposée par Olson [16]. L'objectif est de quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes. Dans la littérature, la difficulté à joindre les personnes a été définie de deux manières selon les auteurs : par l'estimation de la probabilité de réponse d'une personne (une personne avec une probabilité de réponse petite étant considérée comme difficile à joindre), ou par des informations relevées pendant la collecte des données (par exemple, un répondant après une relance ou après un changement de protocole d'enquête est considéré comme plus difficile à joindre).Pour quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes, deux variables mesurant la même variable d'intérêt, notée y, sont nécessaires (Figure 3) :

- une variable dont la mesure est considérée comme la vraie valeur de la variable d'intérêt, disponible pour l'ensemble de l'échantillon. Elle ne présente donc pas d'erreur de mesure pour y ; on l'appellera variable gold-standard, ou permettant d'obtenir des estimations gold-standard. En général elle provient de systèmes d'information existants. On la note y_{si}. On a donc $y = y_{si}$;
- une variable qui mesure la variable d'intérêt par questionnaire, et qui n'est donc disponible que pour les répondants. On la note y_{qaire}. C'est cette variable qui présente potentiellement des erreurs de mesure pour y.

Figure 3 : Représentation d'un fichier de données nécessaire pour l'étude des erreurs de non-réponse et de mesure



Soit :

- $\hat{\bar{y}}_{s,si}$ la moyenne de y_{si} estimée à partir de l'échantillon tiré au sort *s* ; $\hat{\bar{y}}_{s,si}$ est une estimation sans biais de la moyenne de *y* ;
- $\hat{\bar{y}}_{s_r,qaire}$ la moyenne de y_{qaire} estimée à partir de l'échantillon de répondants (s_r) ;
- $\hat{\bar{y}}_{s_r,si}$ la moyenne de y_{si} estimée à partir de l'échantillon de répondants (s_r).

L'erreur de non-réponse notée $\hat{E}_{NR}$, est estimée par $\hat{E}_{NR} = \hat{\bar{y}}_{s,si} - \hat{\bar{y}}_{s_r,si}$.
L'erreur de mesure, notée $\hat{E}_M$, est estimée par $\hat{E}_M = \hat{\bar{y}}_{s_r,si} - \hat{\bar{y}}_{s_r,qaire}$.
L'erreur totale, notée $\hat{E}_{tot}$, est estimée par $\hat{E}_{tot} = \hat{E}_{NR} + \hat{E}_M$.
Supposons que la difficulté à joindre les personnes soit constituée de k catégories, la première catégorie correspondant aux personnes les plus faciles à joindre, la dernière catégorie aux plus difficiles à joindre (l'ensemble des catégories correspond à l'échantillon complet de répondants).
Selon Olson, quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes revient à estimer $\hat{E}_{NR}$ et $\hat{E}_M$ une première fois chez les personnes faciles à joindre (de la catégorie 1), puis de recalculer ces estimations en incorporant à la catégorie précédente les données de la catégorie suivante (de la catégorie 2) et de recommencer les calculs *k-1* fois jusqu'à ce que l'échantillon total des répondants soit intégré.

2.2.2. Application à Coset-MSA

2.2.2.1. Données étudiées

Afin d'appliquer le principe général de la méthode aux données de Coset-MSA, il faut, d'une part disposer de variables gold standard et d'autre part de variables de questionnaire mesurant la même variable d'intérêt que les variables gold standard. Les variables issues des systèmes d'information existants, qui sont disponibles chez les répondants et les non-répondants, et qui sont mesurées de façon indépendante des personnes tirées au sort sont de bonnes candidates pour être des variables gold standard.

Il n'existe aucune variable mesurée à la fois par questionnaire et par le SNIIRAM.
On dispose en revanche de 4 variables disponibles dans les bases de la MSA et mesurées par questionnaire:

- le statut d'emploi salarié (variable binaire) ;
- le secteur d'activité primaire (variable binaire) ;
- la surface agricole utile en ares pour les non-salariés (variable quantitative) ;
- le contrat de travail en CDI pour les salariés (variable binaire).

Même si les variables issues des systèmes d'information de la MSA sont disponibles pour 93,7% des répondants et des non-répondants, elles ne sont pas directement acceptables comme variables gold standard. En effet, d'une part, les systèmes d'information de la MSA n'incluent pas les emplois affiliés aux autres régimes d'assurance vieillesse (Cnav, RSI) et d'autre part il est difficile de faire le lien entre un emploi décrit dans les systèmes d'information de la MSA et un emploi décrit dans le questionnaire quand la personne a eu plusieurs emplois.
Pour ces raisons, l'étude a été restreinte aux personnes ayant eu un seul emploi affilié à la MSA en 2010 dans les bases de la MSA (*n* = 7 484), soit environ 80% de l'échantillon initial ; parmi elles, 1 896 ont participé à l'enquête initiale et 237 à l'enquête complémentaire.

- Indicateur de difficulté à joindre les personnes

L'indicateur considéré pour la difficulté à joindre les personnes comprend deux niveaux : réponse à l'enquête initiale (facile à joindre) *vs* réponse à l'enquête complémentaire (difficile à joindre).
Cet indicateur correspond bien à un niveau d'efforts consentis pour obtenir la réponse d'une personne.

2.2.2.2. Analyses statistiques

L'analyse consiste à estimer les erreurs de non-réponse et de mesure ainsi que l'erreur totale via les données de l'enquête initiale et de l'enquête en deux phases sans et avec correction de la non-réponse, puis à comparer ces estimations selon les différentes configurations.

2.2.2.2.1. Estimation des probabilités de réponse à l'enquête initiale et à l'enquête complémentaire

Sans correction de la non-réponse, l'hypothèse implicite était que les processus de non-réponse à l'enquête initiale et à l'enquête complémentaire étaient MCAR (Missing Completely At Random). Ainsi, les probabilités de réponse correspondaient aux taux de réponse moyen observés dans chacune des enquêtes.

Avec correction de la non-réponse, l'hypothèse implicite était que les processus de non-réponse à l'enquête initiale et à l'enquête complémentaire étaient MAR (Missing At Random) conditionnellement aux variables auxiliaires disponibles. L'enquête Coset-MSA étant une enquête relative à la santé et au travail, nous avons fait l'hypothèse que le processus de non-réponse était MAR conditionnellement à des variables sociodémographiques, relatives à la santé et à l'emploi. Comme nous disposions pour l'ensemble de l'échantillon tiré au sort des variables de stratification, du SNIIRAM et de la MSA, nous avons utilisé ces variables comme variables auxiliaires [20, 21].
Ainsi, pour l'enquête initiale et l'enquête complémentaire, les probabilités de réponse ont été estimées par régression logistique en fonction des variables de stratification, du SNIIRAM et de la MSA. Puis, à partir des probabilités de réponse prédites par le modèle, des groupes homogènes de réponse (GHR) ont été estimés par la méthode des scores [22-24]. Les taux de réponse estimés dans chacun des GHR (10 GHR pour l'enquête initiale, 5 GHR pour l'enquête complémentaire) ont été utilisés comme estimations des probabilités de réponse à l'enquête initiale et à l'enquête complémentaire.

2.2.2.2.2. Estimation des prévalences

Sur l'échantillon complet, $\hat{\bar{y}}_{s,si}$ a été estimé par :

$$\hat{\bar{y}}_{s,si} = \frac{\sum_{i \in s_{EI}} w_i y_i}{\sum_{i \in s_{EI}} w_i}$$

où $w_i = \frac{1}{\pi_{EI,i}}$

Sans correction de la non-réponse, $\hat{\bar{y}}_{s_r,qaire}$ et $\hat{\bar{y}}_{s_r,si}$ ont été estimées par :

- ***Pour l'enquête initiale :***

$$\hat{\bar{y}}_{s_{EI,r},MCAR_{EI}} = \frac{\sum_{i \in s_{EI,r}} w_i y_i}{\sum_{i \in s_{EI,r}} w_i}$$

où $w_i = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MCAR_{EI},i}}$; $\hat{\delta}_{MCAR_{EI},i}$ représente le taux de réponse moyen observé dans l'Enquête Initiale

et y_i représente soit la variable issue du questionnaire soit la variable issue des systèmes d'information.

- ***Pour l'enquête en deux phases :***

$$\hat{\bar{y}}_{s_{EI,r} \cup s_{EC,r};MCAR_{EC}} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_i y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_i}$$

où $w_i = \begin{cases} \frac{1}{\pi_{EI,i}} \text{ si } i \in s_{EI,r} \\ \frac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MCAR_{EC},i}} \text{ si } i \in s_{EC,r} \end{cases}$; $\hat{\delta}_{MCAR_{EC},i}$ représente le taux de réponse moyen observé dans l'enquête complémentaire

et y_i différent selon qu'il était issu du questionnaire ou des systèmes d'information.

Avec correction de la non-réponse, $\hat{\bar{y}}_{s_r,qaire}$ et $\hat{\bar{y}}_{s_r,si}$ ont été estimées par :

- ***Pour l'enquête initiale :***

$$\hat{\bar{y}}_{s_{EI,r};MAR_{EI}(X,V)} = \frac{\sum_{i \in s_{EI,r}} w_i y_i}{\sum_{i \in s_{EI,r}} w_i}$$

où $w_i = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MAR_{EI},i}}$; $\hat{\delta}_{MAR_{EI},i}$ représente le taux de réponse dans le GHR (pour l'enquête initiale) auquel appartient le sujet i

et y_i différent selon qu'il était issu du questionnaire ou des systèmes d'information.

- ***Pour l'enquête en deux phases :***

$$\hat{\bar{y}}_{s_{EI,r} \cup s_{EC,r};MAR_{EC}(X,V)} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_i y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_i}$$

où $w_i = \begin{cases} \frac{1}{\pi_{EI,i}} \text{ si } i \in s_{EI,r} \\ \frac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MAR_{EC},i}} \text{ si } i \in s_{EC,r} \end{cases}$; $\hat{\delta}_{MAR_{EC},i}$ représente le taux de réponse dans le GHR (pour l'enquête complémentaire) auquel appartient le sujet i

et y_i différent selon qu'il était issu du questionnaire ou des systèmes d'information.

3. Résultats

3.1. Probabilités de réponse à l'enquête initiale et à l'enquête complémentaire

Les estimations des probabilités de réponse à l'enquête initiale et à l'enquête complémentaire ont déjà fait l'objet de publications [20, 21]. C'est pourquoi les résultats sont simplement résumés ci-après.

Sous l'hypothèse MCAR, donc sans correction de la non-réponse, les probabilités de réponse estimées à l'enquête initiale et à l'enquête complémentaire étaient respectivement $\hat{\delta}_{MCAR_{EI}} = 24\%$ et $\hat{\delta}_{MCAR_{EC}} = 63\%$.
Sous l'hypothèse MAR conditionnellement aux variables sociodémographiques, de santé et relatives à l'emploi, les probabilités de réponse estimées $\hat{\delta}_{MAR_{EI},i}$ et $\hat{\delta}_{MAR_{EC},i}$ variaient respectivement de 10% à 45% et de 33% à 83%.

3.2. Erreur de non-réponse et erreur de mesure selon la difficulté à joindre les personnes

Sans correction de la non-réponse (partie gauche de la Figure 4, Tableau 2), l'erreur de non-réponse est représentée par l'écart entre les courbes bleu clair (correspondant à la prévalence « gold standard » dans l'échantillon complet) et bleu foncé (correspondant à la prévalence « gold standard » chez les répondants) alors que l'erreur de mesure est représentée par l'écart entre les courbes bleu foncé et rouge (correspondant à la prévalence mesurée chez les répondants).
Avec correction de la non-réponse (partie droite de la Figure 4, Tableau 2), l'erreur de non-réponse est représentée par l'écart entre la courbe bleu clair et la courbe verte (correspondant à la prévalence « gold standard » chez les répondants) alors que l'erreur de mesure est représentée par l'écart entre la courbe verte et la courbe violette (correspondant à la prévalence mesurée chez les répondants).

- Secteur d'activité primaire

Avant correction de la non-réponse, l'erreur de non-réponse est de 7,6% pour les personnes faciles à joindre (enquête initiale) et de 0,1% pour après inclusion des personnes difficiles à joindre ('enquête en deux phases) alors que l'erreur de mesure est de 1% pour les personnes faciles à joindre et de 0,4% après inclusion des personnes difficiles à joindre. L'erreur de non-réponse est donc prépondérante par rapport à l'erreur de mesure pour les personnes faciles à joindre.
Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite pour l'enquête initiale puisqu'elle est estimée à 1,2% et elle reste stable pour l'enquête en deux phases. L'erreur de mesure reste également stable pour les deux enquêtes. L'erreur totale est estimée à environ 1,8% pour l'enquête initiale versus 0,6% pour l'enquête en deux phases.

- Statut salarié

Sans correction de la non-réponse, l'erreur de non-réponse est de 5,7% à l'enquête initiale et de 4,6% à l'enquête en deux phases alors que l'erreur de mesure est de 7,2% à l'enquête initiale et de 4,4% à l'enquête en deux phases. Quelle que soit l'enquête, la part de l'erreur de non-réponse est équivalente à celle de l'erreur de mesure.
Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale et à l'enquête en deux phases puisqu'elle est estimée respectivement à 1,6% et 0,6%. L'erreur de mesure reste également stable dans les deux enquêtes. L'erreur totale est estimée à environ 8% à l'enquête initiale versus 4,5% à l'enquête en deux phases.

- Surface agricole utile

Sans correction de la non-réponse, l'erreur de non-réponse est estimée à 514 ares à l'enquête initiale et à 435 ares à l'enquête en deux phases alors que l'erreur de mesure est estimée à 1033 ares à l'enquête initiale et à 1143 ares à l'enquête en deux phases. Quelle que soit l'enquête, la part de l'erreur de mesure est prépondérante par rapport à l'erreur de non-réponse.

Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale et à l'enquête en deux phases puisqu'elle est estimée respectivement à 285 et 346 ares. L'erreur de mesure reste également stable dans les deux enquêtes. L'erreur totale est estimée à environ 1500 ares quelle que soit l'enquête considérée.

- Contrat de travail en CDI

Sans correction de la non-réponse, l'erreur de non-réponse est estimée à 6% à l'enquête initiale et à 5% à l'enquête en deux phases alors que l'erreur de mesure est estimée à 6,9% à l'enquête initiale et à l'enquête en deux phases. Quelle que soit l'enquête, la part de l'erreur de mesure est supérieure à celle de l'erreur de non-réponse.
Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale puisqu'elle est estimée 0,5% et reste stable pour l'enquête en deux phases. L'erreur de mesure reste également stable pour l'enquête en deux phases et augmente de 2% pour l'enquête initiale. L'erreur totale est estimée à environ 10% quelle que soit l'enquête considérée.

Figure 4 : Moyenne ou prévalence à l'enquête initiale (EI) ou à l'enquête en deux phases (EDPNR) sans correction de la non-réponse (partie gauche) et avec correction de la non-réponse (partie droite)

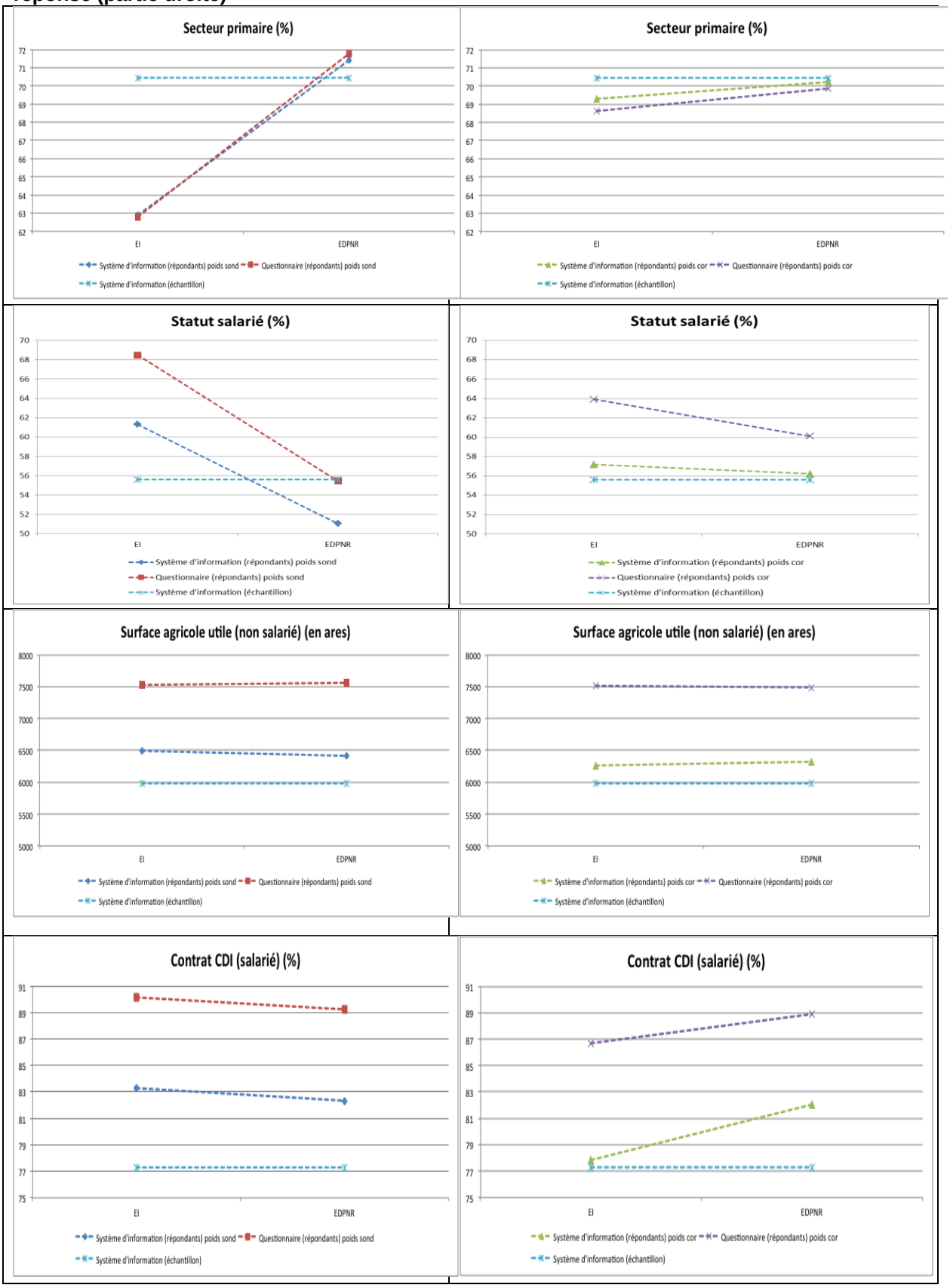


Tableau 2 : Erreur de non-réponse, erreur de mesure et erreur totale à l'enquête initiale et à l'enquête en deux phases pour non-réponse sans et avec correction de la non-réponse

	Enquête initiale			Enquête en deux phases pour non-réponse		
	Erreur de non-réponse	**Erreur de mesure**	**Erreur totale**	**Erreur de non-réponse**	**Erreur de mesure**	**Erreur totale**
Secteur d'activité primaire (%)						
Sans correction de la non-réponse	7,6	0,1	7,7	-1,0	-0,4	-1,3
Avec correction de la non-réponse	1,1	0,7	1,8	0,2	0,3	0,5
Statut salarié (%)						
Sans correction de la non-réponse	-5,7	-7,2	-12,9	4,6	-4,4	0,1
Avec correction de la non-réponse	-1,6	-6,7	-8,3	-0,6	-3,9	-4,5
Surface agricole utile pour les non salariés (ares)						
Sans correction de la non-réponse	-513,9	-1033,7	-1547,6	-435,2	-1143,3	-1578,4
Avec correction de la non-réponse	-285,2	-1244,9	-1530,1	-346,2	-1164,6	-1510,8
Contrat CDI pour les salariés (%)						
Sans correction de la non-réponse	-6,0	-6,9	-12,9	-5,0	-6,9	-11,9
Avec correction de la non-réponse	-0,5	-8,9	-9,4	-4,7	-6,9	-11,6

4. Discussion générale

Avant correction de la non réponse, quelle que soit la variable considérée, l'erreur de non-réponse est soit équivalente, soit plus élevée pour l'enquête initiale que pour l'enquête en deux phases ; hormis pour le statut salarié, l'erreur de mesure est soit équivalente, soit légèrement supérieure pour l'enquête en deux phases comparativement à l'enquête initiale. L'erreur totale est plus élevée pour l'enquête initiale que pour l'enquête en deux phases, sauf pour la surface agricole utile où elle est équivalente dans les deux enquêtes.
Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite, que ce soit à l'enquête initiale ou à l'enquête en deux phases. Ce résultat était attendu, plusieurs études précédentes sur ces données ayant montré l'apport des informations auxiliaires utilisées pour corriger la non-réponse [20, 21].
Avec correction de la non-réponse, les erreurs de mesure sont du même ordre de grandeur à l'enquête initiale et à l'enquête en deux phases. Elles sont à peu près équivalentes pour deux variables (secteur primaire et surface agricole utile), légèrement plus importante à l'enquête initiale pour le contrat CDI et légèrement moins importante à l'enquête initiale pour le statut salarié.
A priori, on aurait pu s'attendre à peu d'erreur de mesure pour les variables « secteur d'activité primaire », « statut salarié » et « emploi en CDI » pour les salariés, car ces variables sont des variables factuelles, binaires et *a priori* faciles à renseigner. Cependant, il n'y a que pour le « secteur d'activité primaire » qu'on observe une faible erreur de mesure. Ce résultat à première vue étonnant pour les variables « statut salarié » et « emploi en CDI » montre que des questions qui nous semblent simples ne le sont pas nécessairement pour les personnes enquêtées. Néanmoins, avant de lancer l'étude pilote, des tests de questionnaire ont été réalisés en face-à-face auprès de personnes travaillant en tant qu'affiliées à la MSA et nous avons pu constater dans ce contexte que ce n'était pas toujours aisé pour les personnes interrogées de situer leur position vis-à-vis de leur emploi (en tant que salarié, ou bien de leur temps de travail ou de leur type de contrat pour les personnes salariées).
Pour la variable « surface agricole utile », on observe une surestimation de la taille de l'exploitation par les répondants. On peut supposer que la question a mal été comprise et que les exploitants ont renseigné, non pas la surface agricole utile qui correspond à une surface agricole excluant les bois et les forêts, mais la surface agricole totale.

Cette étude comporte certaines limites. En effet, elle est restreinte à l'étude de variables relatives à l'emploi. Il aurait bien entendu été intéressant d'étudier les variables relatives à la santé, mais ce travail a été envisagé une fois les données recueillies ; la correspondance entre les variables recueillies par questionnaire et issues des systèmes d'information a été réalisée *a posteriori* et aucune variable disponible dans le SNIIRAM n'avait d'équivalent recueilli par questionnaire. C'est aussi pour cette raison que l'étude a porté sur un nombre limité de variables.
Une autre limite vient du fait que l'enquête complémentaire a été construite pour étudier l'erreur de non-réponse sans que les données soient collectées de la même manière qu'à l'enquête initiale. Il est donc difficile de différencier les erreurs de mesure liées à la difficulté à joindre les personnes ou à un changement de mode de collecte des données ; pour ce faire, il aurait fallu intégrer un troisième groupe « enquête par questionnaire postal » à l'enquête complémentaire. Il nous semble néanmoins plus probable que les différences entre les erreurs de mesure nettement plus faibles à l'enquête complémentaire qu'à l'enquête initiale sont dues aux protocoles d'enquête différents, les enquêteurs pouvant expliciter une question mal comprise, ce qui n'était pas le cas pour les variables recueillies par questionnaire postal (malgré le numéro vert via lequel les personnes tirées au sort pouvaient contacter l'équipe Coset).
Cette hypothèse est corroborée par une étude précédente qui comparait les estimations obtenues à l'enquête initiale et à l'enquête en deux phases sur plusieurs variables recueillies par questionnaire mais non disponibles dans les bases médico-administratives [20] où un écart de 7% avait été estimé pour la différence entre la prévalence estimée à l'enquête initiale et à l'enquête en deux phases pour la question « exposition à des bruits intenses ». Cet écart est probablement lié à un biais de mesure lié au questionnaire de l'enquête initiale. La variable relative à l'exposition à des bruits intenses était recueillie à la 33ème page du questionnaire (sur 40) et la première modalité proposée pour répondre à la question sur les bruits intenses était « non ». Il est donc possible que les répondants à l'enquête initiale aient répondu à la fin du questionnaire avec moins d'attention que les répondants à l'enquête complémentaire qui avaient un questionnaire court. Ceci expliquerait que la proportion de personnes exposée à des bruits intenses via l'enquête initiale est inférieure à celle estimée via l'enquête en deux phases.

Ainsi, pour l'extension de Coset-MSA à l'échelle nationale, qui ne comportera pas d'enquête complémentaire, des efforts devront être mis en œuvre pour minimiser les erreurs de mesure, qui semblent liées à la longueur et à la complexité du questionnaire. Une première recommandation est de pouvoir les estimer tout au long du questionnaire : pour cela, il sera pertinent de rajouter des questions relatives à la santé et au travail également disponibles dans les systèmes d'information existants.
Bien que cette étude ait été réalisée sur un nombre limité de variables, elle présente certains atouts. Elle montre, comme dans les études d'Olson [16] et de Kreuter [15], un nouvel intérêt d'exploiter des variables issues des systèmes d'information, qui n'avait pas été anticipé initialement. Elles peuvent permettre d'étudier la qualité des informations collectées, en termes d'erreur de mesure et d'erreur de non-réponse. Si cette possibilité avait été anticipée initialement, le questionnaire aurait pu être construit en prenant en compte cet objectif ; il aurait en effet été possible d'ajouter, tout au long du questionnaire, des questions disponibles également dans les systèmes d'information pour suivre l'évolution de l'erreur de mesure et de l'erreur de non-réponse en fonction de la taille et de la complexité du questionnaire.
L'apport de notre étude par rapport aux études d'Olson et Kreuter, est que l'évolution des erreurs de mesure et de non-réponse en fonction de la difficulté à joindre les personnes a été estimée sans et avec correction de la non-réponse. Autant pour comprendre le processus, il est intéressant de comparer l'erreur totale de l'enquête initiale et de l'enquête en deux phases sans correction de la non-réponse autant pour les aspects appliqués (choisir de faire une enquête complémentaire ou non), il est plus judicieux de comparer l'erreur totale de l'enquête initiale et de l'enquête en deux phases avec correction de la non-réponse, car ce sont ces estimations qui sont présentées en pratique.

Par ailleurs, cette étude nous a amenés à nous interroger sur la notion de difficulté à joindre les personnes et de probabilité de réponse. Certains auteurs considèrent que la probabilité de réponse et que les efforts consentis pour obtenir une réponse sont des notions proches [14, 25]. Un travail complémentaire pour discuter ce point a été réalisé mais n'est pas présenté ici.

Bibliographie

[1] Groves RM, Lyberg L. Total survey error: past, present, future. Public Opin Q 2010;74(5):849-79.

[2] Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 2006;70(5):646-75.

[3] Merkle D, Edelman M, Dykeman K, Brogan C. An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error. 1998.

[4] Groves RM, Heeringa SG. Responsive design for household surveys: tools for actively controlling survey errors ans costs. J R Statist Soc A 2006;169:439-57.

[5] Lundquist P, Särndal CE. Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. Journal of Official Statistics 2013;29(4):557-82.

[6] Hansen MH, Hurwitz WN. The problem of nonresponse in sample surveys. JASA 1946;41:517-29.

[7] Peytchev A, Rosen J, Riley S, Murphy J, Lindbad M. Reduction of nonresponse bias in surveys through case prioritization. Survey research methodology paper 2010;4(1):21-9.

[8] Schouten B, Cobben F, Bethlehem J. Indicators for the representativeness of survey response. Survey Methodol 2009;35(1):101-13.

[9] Beaumont JF, Haziza D, Bocci C. An adaptive data collection procedure for call prioritization. Journal of Official Statistics 2014;to appear.

[10] Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 2006;70(5):646-75.

[11] Cannell C, Fowler F. A Study of the Reporting Visits to Doctors in the National Health Survey. Ann Arbor; 1963.

[12] Stang A, Jockel KH. Studies with low response proportions may be less biased than studies with high response proportions. Am J Epidemiol 2004 Jan 15;159(2):204-10.

[13] Dahlhamer JM. The Intersection of Response Propensity and Data Quality in the National Health Interview Survey (NHIS). 2012 p. 4509-20.

[14] Fricker S, Tourangeau R. Examining the relationship between nonresponse propensity and data quality in two national household surveys. Public Opin Q 2010;74(5):934-55.

[15] Kreuter F, Muller G, Trappmann M. Nonresponse and Measurement Error in Employment Research. Making use of Administrative Data. Public Opin Q 2010;74(5):880-906.

[16] Olson K. Survey partipication, nonresponse bias, measurement error bias, and total bias. Public Opin Q 2006;70(5):737-58.

[17] Peytchev A, Peytcheva E, Groves RM. Measurement error, unit nonresponse, and self-reports of abortion experiences. Public Opin Q 2010;74(2):319-27.

[18] Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opin Q 2008;72(2):167-89.

[19] Geoffroy-Perez B, Chatelot J, SG, Bénézet L, Delézire P, Imbernon E. Coset : un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Bull Epidemiol Hebd 2012;22-23:276-7.

[20] Santin G, Geoffroy B, Bénézet L, Delézire P, Bouyer J, Guéguen A. Une enquête complémentaire auprès de non-répondants est-elle nécessaire si on dispose de données auxiliaires nombreuses et variées ? Résultats de l'étude Coset-MSA. 2013.

[21] Santin G, Geoffroy B, Benezet L, Delezire P, Chatelot J, Sitta R, et al. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. J Clin Epidemiol 2014 Jun;67(6):722-30.

[22] Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol 1997;23:33-40.

[23] Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007;75(1):25-43.

[24] Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986;54:139-57.

[25] Tourangeau R, Groves RM, Redline CD. Sensitive topics and reluctant respondents : demonstrating a link between nonresponse bias ans measurement error. Public Opin Q 2010;74(3):413-32.