Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs

par Sebastian Bredl, Peter Winker et Kerstin Kº tschau

Juin 2012

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 1-12
Statistique Canada, N° 12-001-X au catalogue

Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs

Sebastian Bredl, Peter Winker et Kerstin Kötschau [1]

Résumé

Les données d'enquête peuvent être falsifiées par les intervieweurs, la fabrication de données étant la forme de falsification la plus flagrante. Même un petit nombre d'interviews contrefaites peuvent fausser gravement les résultats d'analyses empiriques subséquentes. Outre l'exécution de réinterviews, certaines approches statistiques ont été proposées pour repérer ce genre de comportement frauduleux. À l'aide d'un petit ensemble de données, le présent article illustre comment la classification automatique, qui n'est ordinairement pas employée dans ce contexte, pourrait être utilisée pour repérer les intervieweurs qui falsifient les tâches qui leur sont assignées. Plusieurs indicateurs sont combinés pour classer les intervieweurs « à risque » en se fondant uniquement sur les données recueillies. Cette classification multivariée semble supérieure à l'utilisation d'un seul indicateur tel que la loi de Benford.

Mots clés : Fabrication de données ; falsificateur ; loi de Benford ; classification automatique.

1. Introduction

Lorsque la collecte des données se fait par interview, il convient de se préoccuper de la qualité des données. Cette dernière peut souffrir si le répondant fournit des réponses fausses ou imprécises ou que le questionnaire est mal conçu, ou encore si l'intervieweur s'écarte de la procédure d'interview établie. S'il le fait délibérément, on parle de « falsification des données par l'intervieweur » (Schreiner, Pennie et Newbrough 1988) ou de « tromperie » (Schräpler et Wagner 2003).

L'intervieweur peut falsifier les données de nombreuses façons (voir Guterbock 2008). Des formes assez subtiles de falsification consistent à interviewer le mauvais membre du ménage ou à mener l'enquête par téléphone quand l'interview devrait avoir lieu sur place. La forme de falsification la plus grave est la fabrication d'interviews complètes sans jamais prendre contact avec le ménage concerné. Dans notre analyse, nous traitons de ce dernier cas.

Les interviews contrefaites peuvent avoir de graves répercussions sur les statistiques produites d'après les données d'enquête. Schnell (1991), ainsi que Schräpler et Wagner (2003) donnent des preuves que l'effet sur les statistiques univariées pourrait être moins prononcé, à condition que la proportion de falsificateurs demeure suffisamment faible et que la « qualité » des données contrefaites soit élevée. En revanche, même une faible proportion d'interviews contrefaites peut suffire à introduire un biais important dans les statistiques multivariées. Schräpler et Wagner (2003) constatent que l'inclusion de données contrefaites provenant du panel socioéconomique de l'Allemagne (GSOEP) dans une régression multivariée réduit d'environ 80 % l'effet de la formation sur le logarithme de la rémunération brute, bien que la part d'interviews contrefaites soit inférieure à 2,5 %. Cet exemple montre qu'il est important de repérer ces interviews.

Le moyen le plus fréquent de déceler les intervieweurs qui falsifient les données consiste à procéder à une nouvelle interview (Biemer et Stokes 1989). Dans ce cas, un superviseur prend contact avec certains ménages qui auraient dû être interviewés afin de vérifier si l'intervieweur leur a effectivement rendu visite. Cependant, pour des raisons budgétaires, il est impossible de réinterviewer tous les ménages qui participent à une enquête (voir Forsman et Schreiner 1991). Par conséquent, il faut déterminer comment optimiser l'échantillon de nouvelles interviews de manière à déceler le mieux possible les falsificateurs. En général, il semble utile de sélectionner pour une nouvelle interview les ménages interrogés par un intervieweur qui, selon des caractéristiques associées aux réponses obtenues dans ces interviews, est plus susceptible que les autres de fabriquer des données. Dans ce contexte, Hood et Bushery (1997) utilisent le terme d'intervieweur « à risque ». Si la sélection des cas à réinterviewer se fait par échantillonnage en deux étapes, où les intervieweurs sont sélectionnés à la première étape et les enquêtés interrogés par ces intervieweurs, à la deuxième étape [comme le recommandent Forsman et Schreiner (1991)], les intervieweurs à risque pourraient être suréchantillonnés à la première étape.

Dans le présent article, nous montrons une approche purement statistique qui s'appuie sur les données contenues

1. Sebastian Bredl, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : sebastian.bredl@wirtschaft.uni-giessen.de ; Peter Winker, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : peter.winker@wirtschaft.uni-giessen.de ; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankeneser Landstrasse 7, Allemagne. Courriel : kkoetschau@hanse-parlament.eu.

dans les questionnaires pour définir un groupe d'intervieweurs à risque. L'idée n'est pas nouvelle ; on trouve dans la littérature spécialisée plusieurs exemples de ce genre d'approche (Hood et Bushery 1997 ; Diekmann 2002 ; Turner, Gribbe, Al-Tayyip et Chromy 2002 ; Schräpler et Wagner 2003 ; Swanson, Cho et Eltinge 2003 ; Murphy, Baxter, Eyerman, Cunningham et Kennet 2004 ; Porras et English 2004 ; Schäfer, Schräpler, Müller et Wagner 2005 ; Li, Brick, Tran et Singer 2009). Cependant, à l'exception des travaux de Li et coll. (2009), les tests effectués dans ces études reposent sur l'examen d'indicateurs uniques dérivés des données des intervieweurs pour déceler les falsificateurs. Certaines études comprennent le calcul de plusieurs indicateurs, mais en les considérant tous séparément. Nous combinons plusieurs indicateurs dans des analyses par classification automatique, ce qui permet d'obtenir une meilleure classification des falsificateurs éventuels comparativement aux approches antérieures. Autant que nous sachions, cette procédure est une innovation dans le contexte du dépistage des intervieweurs qui falsifient les données, mais elle a déjà été employée dans d'autres domaines afin de déceler des comportements frauduleux. L'idée fondamentale est que les caractéristiques des « cas » frauduleux (ce qui définit un cas dépend du contexte) présentent, comparativement aux cas honnêtes, des schémas frappants qu'il est possible de révéler si ces caractéristiques sont considérées simultanément dans une classification automatique. Murad et Pinkas (1999) essayent de déceler la fraude dans l'industrie des télécommunications par classification des profils d'appels des clients. Un appel est caractérisé par plusieurs indicateurs, dont l'heure de l'appel ou la destination de celui-ci. Thiprungsri (2010) regroupe les demandes d'indemnisation en exécution d'un contrat d'assurance-vie collective soumises par les clients des compagnies d'assurance-vie en s'appuyant sur plusieurs caractéristiques des demandes. Celles qui forment de très petits groupes sont considérées comme étant suspectes. Donoho (2004) utilise la classification automatique, entre autres, pour dégager les tendances des marchés des options susceptibles de révéler des opérations d'initié.

Nous disposons d'un petit ensemble de données d'enquête (voir la sous-section 3.1 pour une description plus détaillée de cet ensemble) qui est constitué en partie de données falsifiées. Issu de 250 questionnaires administrés par 13 intervieweurs, l'ensemble de données est de taille assez limitée et la mesure dans laquelle nos résultats peuvent être généralisés à de plus grands ensembles de données n'est pas claire. Toutefois, cet ensemble de données nous permet d'illustrer notre approche. Le fait que nous sachions quelles données ont été recueillies honnêtement et lesquelles ont été contrefaites permet de procéder à une première évaluation de l'approche. Il convient toutefois de souligner que cette connaissance a priori n'est pas une condition préalable à l'emploi de la méthode.

La question de l'identification des intervieweurs à risque a été abordée durant les années 1980, mais la littérature traitant du sujet demeure peu abondante. En 1982, le U.S. Census Bureau a mis en œuvre l'Interviewer Falsification Study. En s'appuyant sur l'information recueillie dans le contexte de cette étude, Schreiner et coll. (1988) constatent que les intervieweurs ayant peu d'ancienneté sont plus susceptibles que les autres de fabriquer des données. Hood et Bushery (1997) utilisent plusieurs indicateurs pour découvrir les intervieweurs à risque dans la National Health Interview Survey (NHIS). Par exemple, ils calculent le taux par intervieweur de ménages désignés comme étant inadmissibles ou de ménages n'ayant pas de numéro de téléphone, et comparent ces taux aux données de recensement pour les régions pertinentes. En cas d'écart important, l'intervieweur est signalé et une nouvelle interview est effectuée. Les taux de détection parmi les intervieweurs signalés s'avèrent plus élevés que ceux observés pour des échantillons aléatoires de cas à réinterviewer. Turner et coll. (2002) constatent aussi, en examinant les données de la Baltimore STD and Behaviour Survey, que les intervieweurs qui fabriquent des données indiquent moins fréquemment des numéros de téléphone que les intervieweurs honnêtes. Dans le cas de l'interview assistée par ordinateur, Bushery, Reichert, Albright et Rossiter (1999), ainsi que Murphy et coll. (2004) proposent d'utiliser l'horodatage, c'est-à-dire l'enregistrement de l'heure et de la durée de l'interview par l'ordinateur, pour découvrir les intervieweurs suspects. Ceux qui ont besoin d'un temps anormalement long ou court pour administrer le questionnaire complet ou certains modules, ou ceux qui administrent un nombre étonnamment grand de questionnaires durant une période donnée pourraient alors être signalés comme des intervieweurs à risque. Schäfer et coll. (2005) supposent que les falsificateurs évitent les réponses extrêmes lorsqu'ils fabriquent les données. D'après des données du GSOEP, ces auteurs calculent la variance des réponses pour chaque question de tous les questionnaires administrés par un intervieweur et totalisent ces variances. Grâce à d'autres mécanismes de contrôle intégrés dans le GSOEP, les falsificateurs sont connus et il s'avère qu'on les retrouve parmi les intervieweurs ayant les variances globales les plus faibles. Porras et English (2004) adoptent une approche similaire et constatent aussi que les falsificateurs produisent des variances plus faibles que celles observées pour les questionnaires remplis honnêtement. Li et coll. (2009) combinent plusieurs indicateurs prédictifs dans un modèle de régression logistique dans lequel l'état de falsification connu d'un intervieweur sert de variable dépendante binaire. Les auteurs constatent que les échantillons de nouvelles

interviews dans lesquelles sont surpondérés les cas ayant une probabilité élevée d'être frauduleux selon le modèle de régression logistique donnent lieu à la détection d'un plus grand nombre de cas réels de fabrication de données que les échantillons sélectionnés purement au hasard. Cependant, il est évident qu'il faut disposer de données de réinterviews antérieures pour lesquelles l'état de falsification est connu pour exécuter la régression logistique.

Le nombre de combinaisons de réponses rares ou improbables dans les questionnaires soumis par un intervieweur (Murphy et coll. 2004 ; Porras et English 2004) et la comparaison de la composition ou des statistiques descriptives des ménages dans les questionnaires d'un intervieweur avec les données pour l'échantillon complet (Turner et coll. 2002 ; Murphy et coll. 2004) sont d'autres indicateurs discutés dans la littérature.

Un autre moyen de déceler les données contrefaites devenu populaire ces dernières années est l'utilisation de la loi de Benford (Schräpler et Wagner 2003 ; Swanson et coll. 2003 ; Porras et English 2004 ; Schäfer et coll. 2005), dont nous discuterons à la section 2, y compris le succès avec lequel elle a permis de détecter les interviews contrefaites dans le cadre d'études antérieures. La section 2 décrit aussi notre approche statistique en vue de dépister les falsificateurs. La section 3 présente les données sur lesquelles s'appuie notre analyse ainsi que nos résultats. Nous concluons l'article par une discussion de nos résultats.

2. Méthodes

2.1 Loi de Benford

Quand le physicien Frank Benford a remarqué que les pages des tables de logarithmes contenant les logarithmes des nombres faibles (1 et 2) étaient plus souvent utilisées que celles contenant les logarithmes de nombres plus élevés (8 et 9), il a commencé à étudier la distribution du premier chiffre d'une grande gamme de types de nombres, comme ceux figurant à la première page d'un journal, dans les adresses de voirie ou dans les poids moléculaires (Benford 1938). Benford a constaté que la distribution du premier chiffre non nul des nombres pouvait être décrite par la formule qui suit que l'on a appelée « loi de Benford » :

$$\text{Prob}(\text{premier chiffre} = d) = \log_{10}\left(1 + \frac{1}{d}\right). \qquad (1)$$

Toutefois, les séries de nombres étudiées par Benford (1938) ne semblaient pas se conformer toutes à cette loi. Par conséquent, la question qui se posait était de savoir quelle sorte de données on pouvait supposer produire des premiers chiffres de nombre obéissant à la loi. Des discussions de cette question peuvent être consultées dans Hill (1995), Nigrini (1996), Hill (1999), et Scott et Fasli (2001). La détection de la fraude financière est un domaine dans lequel l'application de la loi de Benford s'est beaucoup répandue durant la dernière décennie (Nigrini 1996 ; 1999 ; Saville 2006). Bien que les résultats de ces études ne soient pas pertinents dans notre contexte, il est intéressant de mentionner que le consensus dans la littérature semble être que l'on peut supposer que les valeurs monétaires suivent la loi de Benford. Swanson et coll. (2003) montrent que la distribution des premiers chiffres des nombres dans la Consumer Expenditure Survey des États-Unis est proche de la distribution de Benford.

La notion fondamentale qui sous-tend l'utilisation de la loi de Benford pour détecter les données contrefaites est que les falsificateurs ne connaissent vraisemblablement pas la loi ou qu'ils ne sont pas capables de fabriquer des données qui la suivent. Par conséquent, un écart important de la distribution des premiers chiffres par rapport à la distribution de Benford dans un ensemble de données indique que les données pourraient être contrefaites. Naturellement, il faut se demander si la nature des données permet de supposer qu'elles suivent la loi de Benford si elles sont authentiques. La loi de Benford ne peut pas être appliquée si les questionnaires ne contiennent que très peu de variables métriques, voire aucune.

Schräpler et Wagner (2003), ainsi que Schäfer et coll. (2005) utilisent la loi de Benford pour déceler la fabrication des données dans le GSOEP. Dans les deux études, tous les questionnaires administrés par chaque intervieweur sont combinés et vérifiés afin de déterminer si la distribution des premiers chiffres des nombres figurant dans les questionnaires s'écarte de manière significative de la loi de Benford. Cela peut se faire en calculant la statistique χ^2 :

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(h_{id} - h_{bd})^2}{h_{bd}} \qquad (2)$$

où n_i est le nombre de premiers chiffres dans l'ensemble des questionnaires provenant de l'intervieweur i, h_{id} est la proportion observée du premier chiffre d dans l'ensemble de premiers chiffres dans les questionnaires de l'intervieweur i et h_{bd} est la proportion du premier chiffre d dans l'ensemble des premiers chiffres sous la loi de Benford. Les valeurs élevées de χ^2 indiquent un écart par rapport à la distribution de Benford et signalent des intervieweurs à risque. Schräpler et Wagner (2003) utilisent différents types de variables continues dans leur analyse, tandis que Schäfer et coll. (2005) limitent la leur à des valeurs monétaires. Dans les deux études, les valeurs critiques de χ^2 sont supposées dépendre de la taille de l'échantillon n et sont par conséquent corrigées pour ce paramètre. Les résultats obtenus semblent prometteurs. L'ajustement de la distribution des premiers chiffres des nombres à la loi de Benford est en général nettement moins bon pour les questionnaires des

falsificateurs (qui étaient connus à l'avance) que pour ceux des intervieweurs honnêtes ; il semble donc approprié d'utiliser la loi de Benford comme moyen de dépistage des intervieweurs à risque.

Cependant, quand nous avons comparé les données des intervieweurs honnêtes figurant dans notre ensemble de données à la distribution de Benford, nous avons observé un écart important pour le chiffre 5. Cet écart pourrait être dû à l'arrondissement des nombres par les répondants. Le même problème est mentionné par Swanson et coll. (2003) et par Porras et English (2004), qui choisissent d'appliquer une autre approche « dans l'esprit de Benford » (Porras et English 2004, page 4224). Nous adoptons cette dernière approche qui consiste à comparer la distribution des premiers chiffres des nombres dans les questionnaires d'un intervieweur à la distribution des premiers chiffres des nombres dans l'ensemble des questionnaires sauf les siens. La valeur de χ^2 au niveau de l'intervieweur est calculée comme il est décrit plus haut, mais la proportion prévue d'un chiffre selon la loi de Benford h_{bd} est remplacée par la proportion du chiffre dans tous les autres questionnaires. Nous utilisons ensuite la valeur de χ^2 résultante comme indicateur dans la classification automatique.

Pour ce qui est de la sélection des variables dont nous examinons le premier chiffre des valeurs, nous suivons l'approche de Schäfer et coll. (2005) et n'incluons dans l'analyse que le premier chiffre des valeurs monétaires. L'enquête que nous utilisons pour les besoins de l'illustration contient des valeurs monétaires exprimées en devise locale pour les dépenses des ménages consacrées à divers articles, comme le loyer ou l'achat de terrains, de semences ou d'engrais, ou le paiement des impôts, et pour le revenu du ménage en provenance de diverses sources, comme le travail autonome agricole et non agricole et les transferts publics ou privés. Globalement, nous incluons le premier chiffre de 26 valeurs monétaires différentes par interview, en écartant les valeurs nulles déclarées. Ensuite, nous regroupons les premiers chiffres des nombres figurant dans tous les questionnaires livrés par un intervieweur et comparons leur distribution à celle obtenue pour toutes les autres interviews conformément à la méthode décrite plus haut. Le fait de se limiter aux valeurs monétaires constitue un critère catégorique durant le processus de sélection des données. En outre, comme nous l'avons mentionné plus haut, il est généralement reconnu que les données financières se prêtent à l'analyse fondée sur la loi de Benford. Il importe toutefois de mentionner que nous n'appuyons pas notre analyse sur la loi de Benford, mais sur une approche s'inspirant de cette loi.

2.2 Analyses multivariées

Notre idée consiste à combiner plusieurs indicateurs que nous dérivons directement des questionnaires remplis par chaque intervieweur et que nous supposons être différents pour les falsificateurs et les intervieweurs honnêtes. Nous recourons pour cela à la classification automatique et à l'analyse discriminante. Tous les indicateurs sont calculés au niveau de l'intervieweur. Cela signifie que nous regroupons pour l'analyse tous les questionnaires produits par un intervieweur, ce qui augmente la quantité de données sur laquelle est fondée la valeur de chaque indicateur unique. Les valeurs des indicateurs devraient donc être plus fiables et moins sensibles aux valeurs aberrantes. Par ailleurs, il est évident que le pouvoir discriminant des indicateurs au niveau des intervieweurs diminue lorsque ces derniers ne falsifient qu'une partie de leurs tâches. Par conséquent, l'examen des indicateurs au niveau du questionnaire semble préférable si la quantité de données par questionnaire est suffisamment grande.

La classification automatique constitue la méthode réelle de dépistage des intervieweurs à risque. Les intervieweurs sont répartis en deux groupes avec l'intention d'en obtenir un qui contient une part élevée de falsificateurs et l'autre, une part élevée d'intervieweurs honnêtes. La classification ne requiert pas d'information a priori sur les intervieweurs qui fabriquent des données et ceux qui ne le font pas. En fait, c'est ce que l'analyse est censée révéler. Puisque nous savons au départ à quel groupe appartient chaque intervieweur, nous pouvons découvrir si la classification automatique repère les « vrais falsificateurs » comme étant des intervieweurs à risque. Manifestement, l'hypothèse voulant que notre approche soit capable de faire parfaitement la distinction entre les deux groupes n'est pas réaliste. L'idée est plutôt d'obtenir un groupe d'intervieweurs à risque présentant une part élevée de falsificateurs comparativement à l'autre groupe. Si une nouvelle interview est possible, les efforts subséquents de réinterview seront axés sur les intervieweurs appartenant au groupe à risque.

Afin d'évaluer la performance de la classification automatique, nous considérons le nombre de falsificateurs non décelés, ainsi que le nombre de « fausses alarmes ». Les deux types d'« erreurs » ont un coût : les données des falsificateurs non décelés altèreront vraisemblablement les résultats des analyses statistiques subséquentes. Les fausses alarmes entraînent un coût en ce sens que l'on pourrait s'efforcer inutilement de réinterviewer les ménages concernés ou supprimer inutilement des données de l'échantillon. En outre, les intervieweurs honnêtes dont le travail fait l'objet de nouvelles interviews pourraient se démoraliser, particulièrement s'ils savent que c'est le travail des intervieweurs à risque qui est principalement visé. La façon de pondérer un falsificateur non détecté comparativement à une fausse alarme dans une fonction de perte est une question très subjective. En général, il paraît raisonnable d'accorder plus de poids au premier qu'à la seconde.

L'analyse discriminante requiert avant de l'exécuter que l'on sache pour chaque intervieweur s'il s'agit ou non d'un falsificateur. Par conséquent, ce n'est pas un instrument pour déceler les falsificateurs. Nous utilisons l'analyse discriminante pour vérifier nos hypothèses sur le comportement des falsificateurs, dont nous discuterons plus loin, et pour évaluer dans quelle mesure les indicateurs employés permettent de distinguer les deux groupes.

L'un de nos indicateurs est la valeur de χ^2, calculée en comparant la distribution des premiers chiffres des nombres figurant dans les questionnaires fournis par chaque intervieweur à la distribution correspondante dans tous les autres questionnaires, comme nous l'avons décrit à la sous-section précédente. En outre, nous calculons trois autres indicateurs en partant d'hypothèses concernant le comportement des falsificateurs de données. Schäfer et coll. (2005) supposent que les falsificateurs ont tendance à répondre à chaque question, donc à produire moins de valeurs manquantes. En outre, comme Porras et English (2004), ils s'attendent à ce que les falsificateurs choisissent des réponses moins extrêmes pour les questions ordinales. Hood et Bushery (1997) émettent l'hypothèse que les falsificateurs vont « essayer de ne pas compliquer les choses et de fabriquer un minimum de données falsifiées » (Hood et Bushery 1997, page 820).

Partant de ces hypothèses, nous calculons trois proportions, qui serviront de variables indicatrices dans les analyses multivariées, de même que la valeur de χ^2. Les trois variables indicatrices sont calculées comme il suit :

- Le « ratio de non-réponse partielle » est la proportion de questions qui demeurent sans réponse dans l'ensemble des questions. Nous nous attendons à ce que ce ratio soit plus faible pour les falsificateurs que pour les intervieweurs honnêtes.
- Le « ratio de réponses extrêmes » concerne les réponses mesurées sur une échelle ordinale. Le ratio indique la part de réponses extrêmes (la catégorie la plus faible ou la plus élevée sur l'échelle) dans toutes les réponses ordinales. Selon les hypothèses susmentionnées, ce ratio devrait être plus faible pour les falsificateurs.
- Le « ratio de réponses 'Autre' » concerne les questions qui, outre plusieurs réponses définies, offrent l'option « Autre » comme réponse possible. Le choix de cette option requiert la déclaration explicite d'une autre réponse. Si les falsificateurs ont tendance à ne pas compliquer les choses, nous pouvons nous attendre à ce qu'ils préfèrent les réponses définies à la déclaration d'une autre réponse. Donc, ce ratio également (calculé comme étant la proportion de réponses « Autre » dans l'ensemble des réponses pour lesquelles l'option « Autre » peut être choisie) devrait être plus faible pour les falsificateurs.

Naturellement, la liste des variables indicatrices susceptibles d'être incluses dans la classification automatique peut être allongée. En général, il est possible de dériver de nombreuses autres variables de ce genre des hypothèses sur le comportement des intervieweurs qui fabriquent des données ou d'utiliser celles qui ont déjà été proposées dans la littérature, quoique dans d'autres contextes que celui de la classification automatique. Par exemple, en se fondant sur l'hypothèse que les falsificateurs essayent de fabriquer un minimum de données falsifiées, Hood et Bushery (1997) s'attendent à ce qu'ils sélectionnent souvent de manière disproportionnée la réponse « Non » aux questions, qui selon la réponse, mènent à un ensemble de nouvelles questions ou les évitent (en supposant que « Non » est généralement la réponse qui évite les questions supplémentaires). On pourrait donc calculer le ratio de réponses « Non » à ce genre de questions et l'utiliser comme variable dans l'analyse par classification automatique. Nous ne nous servons pas de ce ratio, car deux versions légèrement différentes du questionnaire ont été utilisées dans notre exemple empirique. Il n'y a qu'un faible nombre de questions menant ou non à de nouvelles questions selon la réponse, le même dans les deux versions du questionnaire.

En outre, quand l'interview assistée par ordinateur permet d'utiliser les données d'horodatage comme il est discuté dans Bushery et coll. (1999), le temps moyen nécessaire pour mener une interview ou le nombre d'interviews réalisées en une journée peuvent servir d'indicateur. Les enquêtes par panel fournissent certains renseignements supplémentaires pour construire des indicateurs. Stokes et Jones (1989) proposent de comparer le taux réel de membres d'un ménage non appariés dans les questionnaires d'un intervieweur particulier au taux prévu de non-appariement calculé conditionnellement à plusieurs caractéristiques du ménage. Les auteurs emploient cette procédure dans l'enquête postcensitaire qui est menée à titre d'enquête de suivi du recensement des États-Unis. Si le taux réel de non-appariement dépasse fortement le taux prévu, les auteurs considèrent qu'il s'agit d'un indicateur de données contrefaites. Habituellement, cette approche peut être appliquée aussitôt que l'on dispose des données d'au moins deux vagues d'une enquête par panel.

Il devient évident que les premières étapes de notre approche consistent à examiner la structure du questionnaire et d'autres types de données, comme les données d'horodatage recueillies durant les opérations d'enquête. Ensuite, il convient de déterminer les indicateurs susceptibles d'être différents pour les falsificateurs et pour les intervieweurs honnêtes que l'on peut dériver de ces sources. Une autre

approche consiste à recourir à des techniques d'exploration de données (*data mining*) pour dégager les tendances qui sont fréquentes dans les données contrefaites ou celles qui ne sont pas les mêmes dans les données contrefaites que dans les données recueillies honnêtement (Murphy, Eyerman, McCue, Hottinger et Kennet 2005). S'il est possible de les déceler, de telles tendances pourraient servir d'indicateurs à la place de ceux dérivés d'après des hypothèses quant au comportement des falsificateurs. Toutefois, cette approche nécessite un énorme ensemble de données comprenant des cas connus de falsification afin de procéder à l'exploration des données. Ce genre d'ensemble des données n'est pas toujours disponible.

3. Résultats

3.1 Sources des données

Les données sur lesquelles porte la présente étude proviennent d'enquêtes-ménages menées en novembre 2007 et en février 2008 dans un pays de la Communauté des États indépendants (CEI) (ancienne Union soviétique). L'enquête a été réalisée dans le cadre d'un projet de recherche international sur les réformes agraires et la pauvreté rurale. Nous avions l'intention d'interviewer 200 ménages dans quatre villages en 2007. Après avoir déterminé que toutes les interviews avaient été contrefaites dans le premier village étudié, nous avons interrompu l'enquête et lancé un nouveau cycle avec de nouveaux intervieweurs dans d'autres villages en février 2008. Tous les villages ont été sélectionnés en s'appuyant sur des critères qualitatifs tels que la structure de la production agricole et la mise en œuvre de réformes agraires. Dans chaque village, les ménages ont été échantillonnés aléatoirement en se servant de listes de ménages fournies par les maires des villages. Non seulement cette procédure assurait que tous les ménages soient sélectionnés au hasard, mais elle fournissait aussi le fondement pour des réinterviews, puisque tous les ménages étaient définis exactement. Cependant, ces nouvelles interviews n'avaient pas été planifiées au tout début de l'enquête. Comme les ménages possédaient rarement un téléphone, les appels de vérification n'étaient pas possibles et les nouvelles interviews de ces ménages ont nécessité des déplacements dans le village pour procéder à la réinterview sur place, ce qui a entraîné d'importantes dépenses d'argent et de temps. Cinq intervieweurs ont été recrutés au moment de la première enquête en 2007. Deux d'entre eux étaient les partenaires locaux du projet de recherche. Ils avaient participé à l'élaboration du questionnaire et étaient responsables de la coordination des enquêtes dans leur pays. Les trois autres intervieweurs étaient des étudiants engagés par les partenaires. Le questionnaire comprenait diverses sections concernant les caractéristiques du ménage, la richesse en ressources, ainsi que le revenu et les dépenses. La plupart des questions étaient de type fermé. Quelques-unes seulement comprenaient une échelle. Des données sur des variables métriques ont été recueillies pour les dépenses des ménages, comme le loyer ou l'achat de terrains, de semences ou d'engrais, ou les impôts, ainsi que pour le revenu du ménage provenant de diverses sources, comme le travail autonome agricole et non agricole, et les transferts publics et privés.

Au moment où les interviews de l'enquête de 2007 ont été menées, aucun des chercheurs allemands n'était présent dans les villages. Les questionnaires ont été recueillis immédiatement après la réalisation de l'enquête dans le premier village. Un premier examen de ces questionnaires a suscité des soupçons, parce que le papier des questionnaires paraissait très propre et très blanc. Le papier ne présentait pas de salissures ni de coins cornés. En comparant les réponses figurant sur divers questionnaires soumis par un intervieweur, nous avons découvert deux questionnaires contenant des réponses identiques. Étant donné que nous avions demandé d'indiquer le montant du revenu en provenance de diverses sources en valeur métrique, il était fort peu probable que les réponses sur les deux questionnaires soient identiques. Comme nous n'obtenions aucune explication de la part des partenaires du projet, nous avons réinterviewé sur place un sous-échantillon de 10 % de l'échantillon original. Aucun des ménages réinterviewés n'a déclaré qu'il avait déjà été interviewé. Après que nous ayons détecté la fabrication des interviews, les partenaires ont reconnu que toutes les interviews avaient été contrefaites. Nous avons naturellement cessé de travailler avec tous les intervieweurs et partenaires, et créé un nouveau groupe de recherche local.

En février 2008, l'enquête a été répétée dans le même pays. Comme nous l'avons mentionné plus haut, nous avons sélectionné de nouveaux villages et ménages conformément aux critères susmentionnés. Nous avons recruté neuf étudiants pour les interviews et organisé une supervision sur place durant l'enquête. Dans la plupart des cas, les interviews ont eu lieu dans une école ou à l'hôtel de ville afin que nous puissions surveiller tous les intervieweurs. Quand les interviews ont eu lieu au domicile des familles participant à l'enquête, nous avons assisté à certaines d'entre elles. Étant donné cette procédure, nous présumons que les réponses au questionnaire de l'enquête de 2008 n'ont pas été contrefaites.

Le présent article porte sur un total de 250 interviews de ménages réalisées par 13 intervieweurs, dont quatre étaient des falsificateurs de l'enquête de 2007 (les interviews soumises par l'un d'eux ont été exclues, car il n'a remis que trois questionnaires) qui ont définitivement contrefait les

résultats et qui sont désignés par F1 à F4, et neuf intervieweurs censés être honnêtes, désignés par H1 à H9. Le tableau 1 donne un aperçu du nombre de questionnaires par intervieweur qui ont été inclus dans l'analyse.

Tableau 1
Nombre de questionnaires par intervieweur

Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Nombre de questionnaires	10	12	10	10	22	23	23	24	23	23	23	23	24

3.2 Classification automatique

À la présente sous-section, nous présentons les résultats de notre analyse par classification automatique. D'après ces résultats, nous évaluons dans quelle mesure notre procédure permet de détecter les intervieweurs qui fabriquent des données. Comme nous l'avons déjà mentionné, nous utilisons quatre variables indicatrices dans la classification automatique, à savoir le ratio de non-réponse partielle, la proportion de réponses extrêmes sur échelle ordinale parmi l'ensemble des réponses sur échelle ordinale que nous appelons ratio de réponses extrêmes, la proportion de réponses pour lesquelles l'option « Autre » nécessitant une autre réponse a été choisie parmi l'ensemble des réponses offrant cette option (appelée ratio de réponses « Autre ») et la valeur de χ^2 découlant de la comparaison de la distribution du premier chiffre dans les nombres figurant dans les questionnaires d'un intervieweur à la distribution correspondante pour l'ensemble des autres questionnaires.

Le tableau 2 donne les valeurs des quatre variables indicatrices incluses dans la classification automatique pour les 13 intervieweurs. Il montre que le ratio de non-réponse partielle et le ratio de réponses « Autre » sont manifestement plus faibles pour les quatre falsificateurs que pour les intervieweurs honnêtes. F1 et F4 n'ont pas choisi l'option « Autre » du tout. Pour le ratio de réponses extrêmes, les résultats paraissent moins clairs. Toutes les valeurs sont comprises entre 40 % et 70 %, sauf celles pour l'intervieweur F1, qui est clairement plus faible. Les valeurs de χ^2 sont assez élevées pour les falsificateurs F2 et F4. Les valeurs pour les deux autres diffèrent peu de celles observées pour les intervieweurs honnêtes.

L'idée générale de la classification automatique est de déterminer des sous-groupes d'éléments dans un espace d'éléments qui sont tous caractérisés par des mesures multivariées (voir Härdle et Simar (2007) pour une introduction à la classification automatique). À la première étape, il faut choisir une mesure pour évaluer la distance ou la ressemblance entre les éléments. À la deuxième étape, les éléments sont affectés à divers sous-groupes ou classes. Les éléments d'une classe particulière doivent se ressembler du point de vue de la mesure choisie, tandis que les éléments appartenant à des classes différentes doivent être distincts. Une grande variété de méthodes sont applicables pour affecter les éléments aux classes, le nombre de classes pouvant être fixe ou déterminé par la méthode de classification.

Tableau 2
Valeurs des variables incluses dans la classification automatique pour chaque intervieweur (toutes les valeurs sauf celles de χ^2 sont exprimées en pourcentage)

Intervieweur	Non-réponse partielle	Réponses « Autre »	Réponses extrêmes	Valeur de χ^2
F1	1,36	0,00	28,33	19,63
F2	0,71	0,65	40,85	29,70
F3	0,68	2,33	56,90	11,34
F4	0,51	0,00	58,62	27,33
H1	3,85	18,01	65,12	14,48
H2	1,99	2,40	59,42	6,91
H3	3,10	9,47	70,07	15,49
H4	4,52	13,04	56,43	16,61
H5	1,18	4,48	70,07	12,16
H6	3,46	1,37	50,75	15,42
H7	2,51	12,72	45,65	9,11
H8	1,77	10,95	69,85	3,63
H9	0,14	1,61	69,44	19,14

Nous avons choisi comme mesure de proximité le carré de la distance euclidienne et employé plusieurs méthodes de classification pour vérifier la robustesse des résultats. Dans tous les cas, les intervieweurs ont été classés en deux groupes avec l'intention d'obtenir un « groupe de falsificateurs » et un « groupe d'intervieweurs honnêtes ». L'avantage de cette approche est que l'on obtient une classification nette. Par contre, lorsque l'on examine séparément les variables indicatrices, il n'est pas évident de savoir où tirer la ligne qui sépare les falsificateurs des interviews honnêtes. Avant de procéder à la classification automatique, nous avons transformé toutes les variables en variables centrées réduites de moyenne nulle et de variance égale à l'unité. Les effets d'échelle sont ainsi éliminés, car les distances sont mesurées en écarts-types et non en différentes unités.

La première méthode de classification que nous avons utilisée est la classification hiérarchique. Il s'agit d'une procédure classique qui peut aussi être appliquée à de plus grands ensembles de données et qui est implémentée dans les progiciels statistiques classiques. La classification hiérarchique consiste à fusionner les classes pas à pas, en combinant les deux classes les plus proches. Au début, chaque élément est considéré comme constituant une classe distincte. Nous mesurons la distance entre deux classes comme étant le carré de la distance euclidienne entre toutes les paires d'éléments possibles, où le premier élément de la paire provient d'une classe et le deuxième, de l'autre classe. Nous avons utilisé le progiciel STATA avec l'option « average linkage » pour effectuer la classification hiérarchique.

Dans la classification hiérarchique, deux éléments demeurent dans la même classe une fois qu'ils ont été appariés. Donc, la procédure n'aboutit pas forcément à un optimum global par rapport à une mesure de distance

donnée. Dans notre cas, le nombre relativement faible d'intervieweurs nous permet de réaliser une autre analyse en examinant simplement toutes les compositions de classes possibles et de choisir la meilleure en fonction d'une certaine fonction cible. (L'analyse a été exécutée en MATLAB et le code du programme peut être obtenu sur demande.) Cette procédure est clairement supérieure à la classification hiérarchique, car elle fait en sorte que soit déterminée la composition globalement optimale des classes. Cependant, nous donnons aussi les résultats de la classification hiérarchique, car elle est assez pratique comparativement à l'approche consistant à essayer toutes les compositions possibles qui requiert d'importants calculs quand le nombre d'intervieweurs augmente. À la place, on pourrait recourir à des techniques d'optimisation heuristiques.

Pour examiner toutes les compositions de classes possibles, nous employons deux fonctions cibles. La première combine deux notions, à savoir qu'une grande distance entre deux centres de classe est admissible, de même qu'une petite distance entre les éléments d'une classe et le centre de la classe. Nous recherchons la composition des classes qui maximise l'expression suivante :

$$\frac{\sum_{i=1}^{4}(\bar{d}_{1i} - \bar{d}_{2i})^2}{\sum_{j=1}^{n_1}\sum_{i=1}^{4}(d_{ij} - \bar{d}_{1i})^2 + \sum_{j=n_1+1}^{13}\sum_{i=1}^{4}(d_{ij} - \bar{d}_{2i})^2}. \qquad (3)$$

L'indice i représente les quatre variables indicatrices différentes $\bar{d}_{ai}$ où $a = 1, 2$ est la moyenne de la variable i dans la classe a, j symbolise les différents éléments (intervieweurs) dans la classe 1 et dans la classe 2, d_{ij} est la valeur de la variable i pour l'élément j, et n_1 est le nombre d'éléments dans la classe 1. Donc, le numérateur mesure la distance entre les deux classes, et le dénominateur, la distance à l'intérieur des classes, et la distance est mesurée par le carré de la distance euclidienne.

Il serait également intéressant de voir quelle composition optimale des classes on obtiendrait si, au lieu de maximiser l'équation (3), on minimisait la moyenne du carré de la distance euclidienne entre toutes les paires possibles dans une classe. En fait, cette idée ressemble fort à la fonction cible pertinente dans les procédures de classification hiérarchique présentées plus haut. Notre deuxième mesure de distance, qui cette fois doit être minimisée, est calculée comme il suit :

$$\frac{\sum_{j=1}^{n_1-1}\sum_{k=j+1}^{n_1}\mathrm{CDE}_{jk} + \sum_{j=n_1+1}^{13-1}\sum_{k=j+1}^{13-1}\mathrm{CDE}_{jk}}{(n_1(n_1-1))/2 + ((13-n_1)(13-n_1-1))/2} \qquad (4)$$

où CDE_{jk} est le carré de la distance euclidienne entre les éléments j et k, calculée comme $\mathrm{CDE}_{jk} = \sum_{i=1}^{4}(d_{ij} - d_{ik})^2$. Le numérateur est égal à la somme des distances entre toutes les paires possibles d'éléments dans la même classe. En divisant cette somme par le nombre de paires possibles, on obtient la distance intra-classe moyenne.

Le tableau 3 donne les résultats des trois méthodes de classification. Dans l'analyse hiérarchique avec lien entre groupes, les trois falsificateurs F1, F2 et F4 forment la classe 1, et le falsificateur F3 ainsi que tous les intervieweurs honnêtes, la classe 2. Donc, nous arrivons à séparer les deux groupes d'intervieweurs, à l'exception d'un falsificateur. Cependant, si nous ne savions pas au départ quels intervieweurs ont fabriqué des données et lesquels sont honnêtes, nous devrions décider laquelle des deux classes contient les intervieweurs à risque. Cela peut se faire en comparant les moyennes des variables indicatrices dans chaque classe présentées au tableau 4. Pour la classification hiérarchique, les moyennes pour le ratio de non-réponse partielle et pour le ratio de réponses « Autre » sont manifestement plus faibles dans la classe 1. Il en est de même de la moyenne pour le ratio de réponses extrêmes, quoique la différence entre les deux classes soit moins frappante. Enfin, une moyenne plus élevée de la valeur de χ^2 peut être observée pour la classe 1. Étant donné ces résultats, en nous en tenant aux hypothèses susmentionnées concernant le comportement des falsificateurs, nous déterminerions correctement que la classe 1 est celle qui contient tous les intervieweurs à risque. Nous avons également essayé d'améliorer les résultats de la classification hiérarchique en utilisant les moyennes de classe présentées au tableau 4 comme point de départ pour la classification par la méthode des K-moyennes. Cependant, l'application de l'algorithme des K-moyennes n'a donné lieu à aucune modification de la composition des classes.

La composition des classes qui maximise l'équation (3) est identique à celle obtenue en utilisant la classification hiérarchique. Par conséquent, comme le montre le tableau 4, les moyennes des indicateurs dans les deux classes sont également identiques.

La composition des classes qui minimise l'équation (4) est légèrement différente. La classe 1 contient maintenant tous les falsificateurs et un intervieweur honnête. Les moyennes des variables indicatrices indiquent de nouveau clairement que la classe 1 est celle qui contient les intervieweurs à risque. Ce résultat est très satisfaisant. Tous les falsificateurs sont repérés et une seule fausse alarme est produite. Il ne faut toutefois pas perdre de vue que cela ne signifie pas que cette méthode de classification particulière est celle qui donne les meilleurs résultats lorsqu'elle est appliquée à un autre ensemble de données.

Afin de déterminer dans quelle mesure un plus grand nombre d'indicateurs produit de meilleurs résultats, nous avons répété notre approche de classification fondée sur les équations 3 et 4 en nous servant de toutes les combinaisons

possibles d'indicateurs, y compris les cas ne s'appuyant que sur un seul indicateur. Les résultats (voir le tableau 7 en annexe) indiquent généralement que l'augmentation du nombre d'indicateurs améliore les résultats. Toutefois, il existe aussi des combinaisons comptant un plus petit nombre d'indicateurs qui produisent des résultats semblables à ceux fondés sur les quatre indicateurs pris ensemble. Pour déterminer quelle combinaison d'indicateurs est la meilleure, il faudrait s'appuyer sur l'établissement hautement subjectif du coût relatif de la non-détection d'un falsificateur comparativement à celui d'une fausse alarme. Toutefois, on peut déterminer quelles combinaisons d'indicateurs sont non Pareto-dominées en ce sens qu'il n'existe aucune autre combinaison présentant moins de falsificateurs non décelés (fausses alarmes) et en même temps ne présentant pas plus de fausses alarmes (falsificateurs non décelés). La combinaison d'indicateurs constituée des quatre indicateurs est la seule qui est non Pareto-dominée, quelle que soit l'équation utilisée. Par contre, les combinaisons ne comprenant qu'un seul indicateur sont Pareto-dominées dans six cas sur huit.

Tableau 3
Résultats des trois méthodes de classification employées

Classification hiérarchique													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance entre les classes divisée par la distance intra-classe													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance entre les éléments d'une classe													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	1	1	2	2	2	2	2	2	2	2	1

Tableau 4
Moyennes des variables indicatrices par classe pour les trois compositions des classes

	Non-réponse partielle		Réponses « Autre »		Réponses extrêmes		Valeur de χ^2	
Classification hiérarchique								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,86	2,32	0,22	7,64	42,60	61,37	25,55	12,43
Distance entre les classes divisée par la distance intra-classe								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,86	2,32	0,22	7,64	42,60	61,37	25,55	12,43
Distance entre les éléments d'une classe								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,68	2,80	0,92	9,06	50,83	60,92	21,43	11,73

3.3 Analyse discriminante

Enfin, nous nous tournons vers l'analyse discriminante pour vérifier si les hypothèses concernant le comportement des falsificateurs sur lesquelles est fondée notre classification automatique sont valides. L'analyse discriminante peut être appliquée si les classes sont connues afin d'évaluer dans quelle mesure les indicateurs employés dans l'analyse permettent de bien séparer les divers groupes et si l'appartenance à un groupe peut être prédite correctement [voir Härdle et Simar (2007) pour une introduction à l'analyse discriminante]. Dans une analyse discriminante linéaire, les coefficients b_0 et b_i de la fonction discriminante $D = b_0 + \sum_{i=1}^{n} b_i x_i$ sont déterminés de façon telle qu'ils maximisent une fonction qui augmente avec l'écart entre les valeurs D moyennes des deux groupes distincts et simultanément diminue avec les écarts entre les valeurs D des éléments à l'intérieur des groupes. Dans notre cas, les x_i sont nos quatre variables indicatrices et nous obtenons deux groupes en séparant les falsificateurs et les intervieweurs honnêtes.

Nous utilisons les probabilités a priori correspondant à la taille de groupe relative (4/13 et 9/13) afin de prédire l'appartenance à un groupe. Le tableau 5 donne les résultats. Manifestement, les quatre variables permettent d'obtenir une bonne séparation des falsificateurs et des intervieweurs honnêtes, car l'appartenance à un groupe est prédite correctement dans tous les cas sauf un.

Comme le montre le tableau 5, les valeurs négatives de la fonction discriminante sont associées au groupe de falsificateurs. Par conséquent, le tableau 6 indique que les signes de trois des quatre coefficients sont en harmonie avec le comportement prévu des falsificateurs. Des ratios de non-réponse partielle et de réponses extrêmes plus élevés mènent à une plus forte probabilité d'observer un intervieweur honnête, de même qu'une valeur de χ^2 plus faible. Le coefficient estimé pour les autres ratios est négatif. Donc, une augmentation des autres ratios, toutes choses étant égales par ailleurs, augmente la probabilité qu'un intervieweur ait fabriqué des données. Cela pourrait sembler contredire nos hypothèses susmentionnées. Une explication possible serait que l'effet des autres ratios est déjà reflété par le ratio de non-réponse partielle. En fait, le coefficient de corrélation entre les deux variables est assez élevé, sa valeur étant de 0,71. Le lambda de Wilks de l'analyse discriminante est statistiquement significatif au seuil de signification de 5 %.

Tableau 5
Résultats de l'analyse discriminante par intervieweur

Intervieweur	Groupe prédit	Groupe réel	Fonction discriminante
F1	1	1	-2,878
F2	1	1	-3,376
F3	2	1	-0,541
F4	1	1	-1,955
H1	2	2	1,828
H2	2	2	1,060
H3	2	2	1,747
H4	2	2	1,616
H5	2	2	0,706
H6	2	2	0,777
H7	2	2	-0,041
H8	2	2	1,765
H9	2	2	-0,710

Tableau 6
Coefficients estimés standardisés et non standardisés (analyse discriminante)

Variable	Coefficient (non standardisé)	Coefficient (standardisé)
Non-réponse partielle	0,767	0,917
Réponses « Autre »	-0,025	-0,129
Réponses extrêmes	0,075	0,821
Valeur de χ^2	-0,092	-0,562
Constante	-4,250	–
Lambda de Wilks (Prob > F)	0.0254	

4. Conclusion

Les données d'enquête peuvent être affectées par les intervieweurs qui fabriquent des données. La fabrication de données est un problème qu'il ne faut pas négliger, car il peut causer des biais importants. Même une petite quantité de données contrefaites peut altérer gravement les résultats des analyses empiriques ultérieures. Nous étendons les approches antérieures en vue de repérer les intervieweurs à risque en combinant plusieurs indicateurs dérivés directement des données d'enquête par classification automatique. Afin de démontrer notre approche, nous l'appliquons à un petit ensemble de données qui a été fabriqué en partie par des falsificateurs. Le fait que nous sachions dès le départ quels sont les falsificateurs nous permet d'évaluer les résultats de la classification automatique et de procéder par après à une analyse discriminante pour révéler dans quelle mesure les deux groupes d'intervieweurs peuvent être bien séparés au moyen des variables indicatrices. Des classifications automatiques de divers types sont effectuées. Toutes donnent lieu à la détermination d'une classe d'intervieweurs à risque, le ratio de non-réponse partielle et le ratio de réponses « Autre » étant les deux indicateurs les plus clairs. Nous n'arrivons pas à identifier parfaitement les falsificateurs. Cependant, dans tous les cas, la classe des intervieweurs à risque contient une part nettement plus élevée de falsificateurs que la deuxième classe. Les avantages de la classification automatique tient au fait que l'on obtient une classification nette des intervieweurs qui sont à risque et des autres intervieweurs, ce qui n'est pas le cas quand des indicateurs tels que la valeur de χ^2 sont examinés individuellement. En outre, elle nous permet de combiner l'information provenant de plusieurs indicateurs. En étudiant la performance de tous les sous-ensembles possibles d'indicateurs, nous constatons qu'en général, un grand nombre d'indicateurs permet de mieux repérer les falsificateurs. Le fait que diverses méthodes de classification produisent des résultats différents ne devrait pas être nécessairement considéré comme un défaut de notre approche. Selon la pondération choisie du coût d'un falsificateur non détecté comparativement à celui d'une fausse alarme, on pourrait en dernière analyse n'affecter au groupe des falsificateurs éventuels que les intervieweurs qui se retrouvent systématiquement dans la classe des intervieweurs à risque, quelle que soit la méthode de classification appliquée (ce qui impliquerait un coût élevé des fausses alarmes), affecter au groupe des falsificateurs éventuels tous les intervieweurs qui se retrouvent dans la classe des intervieweurs à risque au moins une fois (ce qui impliquerait un coût élevé des falsificateurs non détectés) ou choisir une solution intermédiaire.

L'application à un petit ensemble de données démontre un autre mérite de notre approche : elle a été testée et a donné de bons résultats dans une situation dans laquelle le nombre de questionnaires par intervieweur était assez limité (trois des falsificateurs n'ont soumis que dix questionnaires). Si un petit nombre de questionnaires par intervieweur est suffisant pour effectuer l'analyse, on pourrait aussi imaginer de la mettre en œuvre durant la période principale de travail sur le terrain, quand les intervieweurs n'ont soumis qu'un certain pourcentage de leurs questionnaires. Les falsificateurs pourraient alors être remplacés par d'autres intervieweurs qui mèneront l'enquête auprès des unités qui auraient dû être interviewées par les falsificateurs.

Évidemment, en examinant nos résultats, nous ne devons pas perdre de vue que nous avons appliqué notre méthode à un ensemble de données dans lequel une forme très grave de fabrication des données a eu lieu : d'une part, nous avons des falsificateurs qui ont contrefait les données de tous leurs questionnaires (presque) complètement et d'autre part, nous avons des intervieweurs qui (on le présume) ont effectué tout leur travail honnêtement, ce qui facilite la discrimination entre les intervieweurs honnêtes et les intervieweurs malhonnêtes. En outre, la taille de notre échantillon, qui ne comprend que 13 intervieweurs, est assez limitée. Il serait intéressant d'explorer l'utilité de notre approche lorsqu'on l'applique à de plus grands ensembles de données, étant donné que la part d'interviews falsifiées dans les grandes enquêtes s'avère plus faible que dans notre cas. Qui plus est, les grands ensembles de données pourraient permettre de construire des indicateurs additionnels pour la classification automatique. Si l'enquête comprend un programme de réinterview, il serait possible d'évaluer l'utilité de notre approche en comparant le « succès » d'une réinterview aléatoire avec celui d'une réinterview axée sur les intervieweurs considérés comme étant à risque. Nous avons également l'intention de poursuivre l'analyse dans des conditions expérimentales. Des conditions appropriées permettent de s'assurer que l'on obtient un ensemble de données qui a été recueilli partiellement en menant des interviews réelles et partiellement fabriquées en disant à certains participants à l'expérience de remplir le questionnaire eux-mêmes.

Annexe

Tableau 7
Résultats des classifications automatiques fondées sur les équations 3 et 4 pour toutes les combinaisons possibles de classes

Indicateurs				Équation 3		Équation 4	
Non-réponse partielle	Réponses « Autre »	Réponses extrêmes	Valeur de χ^2	Falsificateurs non décelés	Fausses alarmes	Falsificateurs non décelés	Fausses alarmes
			X	2	0	1	1
		X		2	1	2	2
		X	X	2	0	1[1]	0
	X			0[1]	4	0	4
	X		X	2	0	0	2
	X	X		3	0	0	3
	X	X	X	1[1]	0	1	1
X				0[1]	4	0	4
X			X	2	1	0	2
X		X		3	0	-[2]	-
X		X	X	1[1]	0	1	1
X	X			0[1]	4	0	4
X	X		X	1	1	0	2
X	X	X		0[1]	4	0	4
X	X	X	X	1[1]	0	0[1]	1

[1] Combinaison d'indicateurs non Pareto-dominée.
[2] Les valeurs moyennes de classe n'ont pas permis de déterminer la classe « à risque ».

Remerciements

Nous tenons à souligner l'appui financier de la Deutsche Forschungsgemeinschaft par la voie du projet « PP 1292: *Survey Methodology* ». Nous remercions en outre John Bushery et quatre examinateurs anonymes de leurs commentaires constructifs concernant notre article.

Bibliographie

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(1), 551-572.

Biemer, P., et Stokes, S. (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics*, 5(1), 23-29.

Bushery, J., Reichert, J., Albright, K. et Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 316-320.

Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Rapport technique manuscript 06/2002, Institut für Technikfolgenabschätzung (ITA), Wien.

Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429.

Eyerman, J., Murphy, J., McCue, C., Hottinger, C. et Kennet, J. (2005). Dépistage de la falsification des données par l'intervieweur par l'exploration de données. Dans le *Recueil : Symposium 2005, Défis méthodologiques reliés aux besoins futurs d'information*. Statistique Canada.

Forsman, G., et Schreiner, I. (1991). The design and analysis of reinterview: An overview. Dans *Measurement Errors in Surveys*, (Éds., P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz et S. Sudman), New York : John Wiley & Sons, Inc, 279-301.

Guterbock, T.M. (2008). Falsification. Dans *Encyclopedia of Survey Research Methods*, (Éd., P.J. Lavrakas), Sage Publications, Thousand Oaks, 1, 267-270.

Härdle, W., et Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2e Édition. Springer, Berlin.

Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science*, 10(4), 354-363.

Hill, T. (1999). The difficulty of faking data. *Chance*, 26, 8-13.

Hood, C., et Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 820-824.

Li, J., Brick, J., Tran, B. et Singer, P. (2009). Using statistical models for sample design of a reinterview program. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4681-4695.

Murad, U., et Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. Lecture Notes in Computer Science, 1704, 251-261.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. et Kennet, J. (2004). A system for detecting interviewer falsification. Article présenté à l'American Association for Public Opinion Research 59th Annual Conference.

Nigrini, M. (1996). A taxpayers compliance application of Benford's law. *Journal of the American Taxation Association*, 18, 72-91.

Nigrini, M. (1999). I've got your Number. *Journal of Accountancy*, 187(5), 79-83.

Porras, J., et English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4223-4228.

Saville, A. (2006). Using Benford's law to predict data error and fraud - An examination of companies listed on the JSE Securities Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354.

Schäfer, C., Schräpler, J., Müller, K. et Wagner, G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, 125, 183-193.

Schnell, R. (1991). Der einfluss gefälschter Interviews auf survey ergebnisse. *Zeitschrift für Soziologie*, 20(1), 25-35.

Schräpler, J., et Wagner, G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys - An analysis by means of genuine fakes in the raw data of SOEP. Document de discussion IZA séries, 969.

Schreiner, I., Pennie, K. et Newbrough, J. (1988). Interviewer falsification in census bureau surveys. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 491-496.

Scott, P., et Fasli, M. (2001). Benford's law: An empirical investigation and a novel explanation. Rapport technique de la CSM, Department of Computer Science, University Essex.

Stokes, L., et Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 696-698.

Swanson, D., Cho, M. et Eltinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4172-4177.

Thiprungsri, S. (2010). Cluster Analysis for Anomaly Detection in Accounting Data. Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop San Francisco, California.

Turner, C., Gribbe, J., Al-Tayyip, A. et Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Ébauche de prépublication). Papier technique sur l'Health and Behavior Measurement. Washington DC : Research Triangle Institute. No. 53.