No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

DÉPISTAGE DE LA FALSIFICATION DES DONNÉES PAR L'INTERVIEWEUR GRÂCE À L'EXPLORATION DES DONNÉES

Joe Murphy[1], Joe Eyerman[1], Colleen McCue[1], Christy Hottinger[1] et Joel Kennet[2]

RÉSUMÉ

La communication décrit une application novatrice de l'exploration des données à des données de réponses et des métadonnées pour dépister, caractériser et prévenir la falsification sur le terrain par les intervieweurs de la National Survey on Drug Use and Health (NSDUH). La falsification des données par l'intervieweur est la création délibérée de réponses par l'intervieweur sans intervention du répondant. La falsification non décelée peut introduire un biais dans les estimations démographiques si les réponses falsifiées ne correspondent pas aux valeurs qui auraient été fournies par les répondants. À l'heure actuelle, les procédures requises pour dépister la fraude des intervieweurs peuvent être coûteuses et utiliser des ressources qui pourraient autrement être affectées aux procédures d'amélioration de la qualité des données. L'exploration des données peut être utilisée pour programmer des vérifications de la propension à la falsification exécutables fréquemment, qui facilitent l'exécution rapide et relativement peu coûteuse de la détection et des mesures de correction, et servent éventuellement de moyen dissuasif.

MOTS-CLÉS : falsification; exploration des données; intervieweurs

1. INTRODUCTION

1.1 Falsification des données par l'intervieweur

La falsification des données par l'intervieweur est la création délibérée de réponses par l'intervieweur sans intervention des personnes échantillonnées comme répondants. Elle peut nuire aux enquêtes de plusieurs façons. Tout d'abord, les cas de falsification dépistés peuvent avoir des répercussions considérables sur le budget d'une enquête lorsqu'ils doivent faire l'objet de nouvelles interviews. En deuxième lieu, les procédures requises pour dépister la fraude peuvent être coûteuses et utiliser des ressources qui pourraient autrement être affectées aux procédures d'amélioration de la qualité. Enfin, la falsification non décelée peut introduire un biais dans les estimations démographiques, si les réponses falsifiées ne correspondent pas aux valeurs qui auraient été fournies par les répondants de l'échantillon (Schraepler et Wagner, 2003). Plus le nombre de cas de falsification non dépistés est élevé, et la différence entre les valeurs réelles de population et les valeurs falsifiées est grande, plus le biais touchant les statistiques de l'enquête est important.

Même si on croit que la falsification des données par les intervieweurs sur le terrain, dans le cadre de la National Survey on Drug Use and Health (NSDUH), est rare, certains cas graves ont été dépistés ces dernières années et ont suscité des préoccupations. Par ailleurs, des changements récents apportés au plan de sondage de l'enquête pourraient avoir modifié la propension des intervieweurs à falsifier les données. Par exemple, l'adoption d'incitatifs monétaires (2002) pourrait encourager la falsification chez les intervieweurs sans scrupule et pourraient même inciter des criminels cols blancs à devenir intervieweurs.

La présente communication décrit une étude spéciale de la NSDUH portant sur la question de la falsification, en vue de mettre au point un système de dépistage pour la NSDUH qui réduit les coûts de la surveillance et augmente le taux de dépistage. L'exploration des données a été choisie comme stratégie d'analyse pour cette étude, étant donné qu'elle permet d'effectuer des recherches automatisées dans des ensembles importants de données

[1] RTI International, PO Box 12194, Research Triangle Park, North Carolina, USA (contact par courriel : jmurphy@rti.org).
[2] Substance Abuse and Mental Health Services Administration (SAMHSA), Office of Applied Studies (OAS), 1 Choke Cherry Road, Room 7-1044, Rockville, Maryland, USA.

multidimensionnelles et du fait de son importance émergente comme stratégie d'analyse de choix dans l'application du dépistage des fraudes. L'exploration des données est un processus automatisé qui facilite les recherches exhaustives dans des ensembles de données, ainsi que la détermination de tendances, de modèles et de relations complexes qui ne sont pas évidents ou même impossible à déterminer par les méthodes d'analyse traditionnelles. L'un des avantages de l'exploration des données est qu'elle fournit l'occasion de passer en revue efficacement et rapidement de très gros ensembles de données, ainsi que d'appliquer des stratégies adaptées de dépistage et de réduction des fraudes. Les données de réponses aux enquêtes et les métadonnées sont examinées en utilisant le modèle CRISP-DM (CRoss Industry Standard Process for Data Mining) comme cadre. Plus de détails concernant le modèle CRISP-DM se trouvent à l'adresse suivante : http://www.crisp-dm.org/.

1.2 Stratégies existantes de dépistage de la falsification dans la NSDUH

La NSDUH est une enquête annuelle transversale auprès des ménages, qui est réalisée en personne et conçue pour recueillir des données sur l'usage et l'abus de substances psychoactives chez les membres de la population civile des États-Unis âgés de 12 ans et plus, non placés en établissement. Anciennement appelée la National Household Survey on Drug Abuse (NHSDA), la NSDUH est effectuée à contrat par RTI International, Research Triangle Park, Caroline du Nord. C'est la Substance Abuse and Mental Health Services Administration (SAMHSA) qui parraine l'enquête.

Dans le cadre du Système de surveillance de la qualité des données de la NSDUH, les intervieweurs sur le terrain tentent de recueillir des numéros de téléphone auprès de certains ménages. Ces numéros servent à appeler des personnes et à vérifier la qualité du travail des intervieweurs. Ces vérifications sont effectuées à l'égard des deux premiers cas de non-interview et des deux premières interviews menées par chaque intervieweur pour chaque trimestre. Par ailleurs, au moins 5 % des présélections effectuées par l'intervieweur et au moins 15 % des interviews menées par l'intervieweur font l'objet d'une sélection aléatoire pour une vérification par téléphone. Des intervieweurs spécialement formés pour les interviews téléphoniques appellent pour vérifier si la présélection ou l'interview a eu lieu et a été effectué correctement. Les cas d'interviews sélectionnés pour la vérification par téléphone et pour lesquels on n'a pas de numéro de téléphone sont vérifiés par courrier. S'il existe des doutes quant au rendement d'un intervieweur, une proportion plus élevée de ses cas (jusqu'à 100 %) font l'objet d'une vérification « forcée ». Parmi les problèmes en matière de qualité des données qui pourraient entraîner un niveau plus élevé de vérification figure un taux élevé de numéros de téléphone manquants ou dont la vérification a été refusée, la déclaration par l'intervieweur de son propre numéro de téléphone ou du numéro de téléphone d'un autre membre du personnel pour la vérification, l'utilisation en double inattendue de tout numéro de téléphone, et tout problème constituant une violation grave du protocole. Par ailleurs, des données sur le moment de l'interview sont recueillies pour chaque interview. Elles sont transmises tous les soirs à partir des ordinateurs des intervieweurs, en même temps que les données d'interview complètes. Tous les cas terminés en moins de 30 minutes ou en plus de 60 minutes sont examinés. Lorsqu'il existe des préoccupations graves concernant la validité du travail d'un intervieweur, un échantillon mixte des cas terminés par l'intervieweur est sélectionné pour une vérification sur place. Les personnes trouvées coupables de falsification sont congédiées immédiatement.

Parmi les ajouts récents aux stratégies de dépistage de la falsification de la NSDUH figurent l'examen des réponses, des modèles de réponse et des durées d'interview inhabituels (Murphy et coll., 2004). D'autres procédures sont en place pour déterminer les modèles de réponses incompatibles à l'intérieur de l'enquête. Ces stratégies, même si elles sont efficaces, prennent beaucoup de temps et nécessitent beaucoup de main-d'œuvre. Le nombre important de cas devant faire l'objet d'un dépistage, d'un examen et d'une validation additionnels a donné lieu à la détermination de nombreux cas de fraude de la part des intervieweurs, qui n'auraient pas été décelés si ces procédures n'avaient pas été en place. Toutefois, le nombre important de cas devant faire l'objet d'un examen additionnel, et le nombre élevé de faux positifs liés à cette approche fait ressortir la nécessité d'élaborer des approches mieux ciblées pour le dépistage des fraudes. Par ailleurs, il est probable qu'il existe d'autres modèles de fraude qui n'ont pas encore été déterminés. La capacité de cibler les stratégies d'identification et de dissuasion en fonction de modèles particuliers de fraudes permet aux chercheurs de déployer plus efficacement des ressources pour le dépistage des fraudes et d'élaborer des stratégies utiles qui portent directement sur les modèles particuliers de fraude et d'inconduite de la part des intervieweurs.

1.3 Historique récent de la falsification dans le cadre de la NSDUH

De 2000 à 2001, la prévalence des falsifications dépistées relativement à la NSDUH était relativement faible (environ un dixième d'un pour cent des interviews terminées étaient considérées comme falsifiées). Toutefois, à la fin de 2002, on a identifié quatre intervieweurs qui avaient inscrit leur propre numéro de téléphone ou le numéro de téléphone d'un autre intervieweur aux fins de la vérification, ce qui contrevient au protocole du projet. Par suite d'une vérification sur place de l'ensemble de l'affectation de travail de ces intervieweurs, il a été déterminé que les quatre intervieweurs s'étaient adonnés fréquemment à la falsification. Parmi les 760 présélections et 464 interviews menées par les 4 intervieweurs, 287 présélections et 134 interviews ont été considérées comme valides. Au total, 473 présélections et 330 interviewes ont été considérées comme falsifiées et ont par conséquent fait l'objet d'autres travaux. À partir de 2003, le niveau de dépistage était le même qu'en 2000 et en 2001.

Le taux de réponse et les données temporelles au niveau de la question pour les cas falsifiés en 2002 ont fait l'objet d'un examen et d'une comparaison avec des données provenant d'interviews valides dans les mêmes États. Des différences significatives entre les cas falsifiés et les cas valides ont été décelées, ce qui a mené à l'élaboration et à l'intégration de modèles d'évaluation par score des données sur les réponses et des modèles de métadonnées, en vue de prédire la propension future à la falsification (Murphy et coll., 2004). Les travaux d'exploration des données dont il est question dans le présent document reposent sur cet examen initial et visent à améliorer le processus de dépistage.

2. ANALYSE AU MOYEN DU MODÈLE CRISP-DM

2.1 Objectifs et défis

L'application des techniques d'exploration des données à la falsification, dans le contexte de la NSDUH, suscite deux défis dans l'immédiat. Tout d'abord, les techniques d'exploration des données ne sont pas beaucoup utilisées dans le domaine des sciences sociales[3], ni dans les ouvrages spécialisés sur les méthodes d'enquête. Par conséquent, il n'existe pas de vocabulaire commun aux deux domaines, de normes ou de méthodes largement acceptées, ni de conventions de rapport communes. En deuxième lieu, l'exploration des données est plus fréquemment utilisée dans des applications commerciales (marketing, dépistage de la fraude, etc.). Les normes de rapport sont généralement adaptées aux besoins particuliers des entreprises et ne peuvent pas servir de guide pour les rapports en sciences sociales. Par ailleurs, les analyses sont souvent confidentielles sur le plan des affaires et ne sont pas disponibles dans le public, ce qui rend difficile de trouver des exemples dans les ouvrages publiés. Enfin, les données des ouvrages publiés sont moins exhaustives que celles utilisées en sciences sociales et dans les méthodes d'enquête. Le modèle CRISP-DM (CRoss Industry Standard Process for Data Mining) a été choisi afin de relever ces deux défis, du fait qu'il fournit un format uniformisé d'analyse et de rapport, qui a été approuvé par des experts de l'exploration des données. Cet effort visait principalement à améliorer le système de dépistage utilisé dans le cadre de la NSDUH, de façon à réduire les coûts de la surveillance et à augmenter le niveau de précision du dépistage de la falsification.

2.2 Données

Les données disponibles de la NSDUH ont été réparties entre les données du processus d'interview et les résultats de l'enquête, comme le montre le tableau 1. Les métadonnées comprenaient le registre des visites (RV) et la piste de vérification, tandis que les données d'enquête comprenaient les réponses proprement dites fournies par les répondants aux questions de l'enquête. On a utilisé les données de l'enquête pour 2004 dans le cadre de ce projet, étant donné qu'il s'agissait de l'année complète la plus récente disponible au moment de l'analyse. Afin de réduire les délais liés à la préparation des données, un seul trimestre de données a été utilisé pour certaines analyses.

[3] Cette question a été abordée par suite d'une subvention versée par la National Science Foundation, à Richard Berk et à ses collègues. Des exemples de travaux non publiés dans ce domaine figurent dans la bibliographie (Berk, sous presse; Lennert, Cody et Berk, à l'étude; Berk, Kriegler et Baek, à l'étude).

Tableau 1 : Données de la NSDUH disponibles pour l'analyse

Source des données	Nombre d'enregistrements	Description de l'enregistrement
RV	1,169,054	Chaque enregistrement comprend une visite à l'unité de logement ou une tâche de disposition dirigée par un superviseur des opérations sur le terrain.
Piste de vérification	70,356	Chaque enregistrement représente une entrée dans l'instrument d'IAO*.
Données brutes de l'IAO	67,913	Chaque enregistrement comprend les données d'enquête brutes d'une interview.

IAO = interview assistée par ordinateur; RV = registre des visites.
*Une seule interview comportant plusieurs ruptures peut être liée à plusieurs enregistrements de piste de vérification.

2.3 Processus d'exploration des données

Le modèle CRISP-DM comprend six étapes : compréhension des activités, compréhension des données, préparation des données, modélisation, évaluation et déploiement. Il se peut que l'étape la plus importante du processus d'exploration des données soit la compréhension des pratiques en vigueur et des objectifs globaux du projet. Par conséquent, au cours de l'étape de la compréhension des activités du processus d'exploration des données, les procédures actuelles et les rapports antérieurs de dépistage de la falsification dans le cadre de la NSDUH/NHSDA ont été passés en revue, en vue de servir de guide pour l'exploration initiale des données, ainsi que pour le choix d'approches et d'algorithmes statistiques particuliers au cours du processus de modélisation.

L'identification et la caractérisation des données falsifiées ont été effectuées au moyen de méthodes de dépistage des fraudes similaires à celles utilisées par d'autres organismes, y compris l'Internal Revenue Service (IRS) et le secteur des soins de santé (p. ex., dépistage des fraudes de Medicaid) (Mena, 2003). Ces méthodes générales peuvent être réparties en deux catégories : modèles d'évaluation par score et techniques de détection des anomalies. Ces méthodes peuvent être utilisées à la fois avec les données du processus d'interview et les réponses à l'enquête.

Les **modèles d'évaluation par score** servent à examiner les profils existants de données et activités de collecte frauduleuses, afin que les profils similaires dans des cas futurs puissent faire l'objet d'une évaluation additionnelle. La détermination de ces profils peut laisser supposer une falsification de la part de l'intervieweur ou une déception du répondant, ces deux situations nécessitant une analyse additionnelle. Ces profils tirent parti du fait que les personnes n'affichent pas souvent un comportement créatif, ni unique, lorsqu'elles falsifient les données, c'est-à-dire qu'elles ont tendance à falsifier les données de façon similaire. Jusqu'à maintenant, les modèles d'évaluation par score ont permis d'identifier de façon précoce au moins trois falsificateurs dans le cadre de la NSDUH.

Une autre approche, appelée **détection des anomalies**, caractérise les profils « normaux » de réponses et d'activités de collecte des données. Les données qui s'écartent de ces profils normaux font l'objet d'une évaluation plus poussée. Parmi les exemples figurent les délais de réponse à l'enquête particulièrement courts, les taux de réponse extrêmement élevés et les quantités démesurées de données recueillies au cours d'une seule séance ou journée. Ces résultats extrêmes ne correspondent généralement pas aux autres ensembles calculés et nécessitent un examen et un suivi additionnels. De même, les profils de réponse inhabituellement uniformes ou les profils de réponse qui sont incompatibles avec les modèles connus de comportement ou qui s'en écartent justifient aussi un suivi et un examen. Même si ces données anormales ou valeur aberrantes ne constituent pas nécessairement une indication d'activités frauduleuses, elles méritent qu'on fasse un suivi, afin d'assurer un niveau élevé de qualité et de précision dans les données. À partir d'une combinaison de modèles d'évaluation par score et de détection des anomalies, nous pouvons nous pencher sur des profils connus de comportements frauduleux, ainsi que déceler les profils qui pourraient indiquer des comportements nouveaux ou suspects et la violation des protocoles.

Il convient de souligner que la prédiction des événements peu fréquents comme les fraudes peut être particulièrement difficile. Dans l'ensemble, l'exactitude du profil peut être trompeuse dans une certaine mesure pour

les événements peu fréquents. Par exemple, un modèle pourrait être correct 97 % du temps s'il prédisait toujours une réponse de « non » pour un événement ayant une fréquence prévue de 3 %. De toute évidence, la précision globale constituerait une mesure inacceptable de la valeur prédictive de ce type de profil. Dans ces cas, la nature et le sens des erreurs peut fournir une meilleure estimation de la valeur globale du profil. En rajustant les « coûts » liés aux faux positifs ou aux manques, le modèle peut être précisé, afin de mieux prédire les événements peu fréquents. Ces coûts peuvent être équilibrés, afin de créer un modèle qui permet de déterminer avec précision les cas intéressants (les « vrais positifs »), en limitant le nombre de fausses alertes.

Il est essentiel de détenir une expertise importante du domaine ou une connaissance exhaustive des exigences globales du projet, des ressources de données, des procédures et des objectifs, afin de créer des modèles de prédiction dont les critères sont opérationnellement raisonnables. Les responsables de la recherche comprenaient des responsables de projets ayant une expérience significative de la NSDUH, ainsi que des protocoles permanents de dépistage des fraudes. L'équipe chargée du rapport a collaboré étroitement avec l'équipe chargée de la qualité des données, afin de s'assurer que les profils rendaient non seulement compte des connaissances actuelles en matière de fraude par des intervieweurs, mais qu'ils étaient aussi valables et pratiques dans ce contexte.

3. CONCLUSIONS

Le modèle CRISP-DM a révélé plusieurs profils de violation de protocole et profils d'interview laissant supposer des profils de comportement frauduleux. Étant donné que le contexte actuel ne permet pas la présentation complète des profils établis dans le cadre du processus d'exploration des données, un sommaire suit.

Les intervieweurs comptant un grand nombre de « ruptures » ou perturbations ont été associés à des cas de falsification connus. Les interviews terminées tard dans le cycle avaient des caractéristiques temporelles différentes et correspondaient généralement à des cas « difficiles », et elles étaient moins susceptibles d'être vérifiés à cause des procédures de fermeture.

Parmi les interviews falsifiées, l'occurrence d'un « code 01 » (personne à la maison au moment de la visite de présélection) était plus élevée (28,21 %) que parmi les interviews valides menées par des falsificateurs (21,01 %) ou les interviews valides menées par des non-falsificateurs (20,77 %). Si toutes les interviews sur place associées aux cas falsifiés étaient frauduleuses, le nombre de codes 01 aurait été plus élevé, les falsificateurs ayant tenté de démontrer qu'une part importante de leurs travaux étaient valides, alors qu'ils ne l'étaient pas. Une explication plus plausible est que les cas comportant de nombreux codes 01 ont été plus difficiles à compléter et que la décision de falsifier est venue après la difficulté d'entrer en rapport avec les répondants. Étant donné que le nombre d'interviews falsifiées dépistées était suffisamment faible, les tests d'hypothèse n'ont pas fait ressortir de différences significatives, mais cela ne signifie pas nécessairement qu'il n'existait pas de rapport entre la répartition des dispositions de visites et la falsification.

Parmi les conclusions surprenantes au départ figurait le fait que le taux de « codes 32 » (deux membres de l'unité de logement sélectionnés pour l'interview) était beaucoup plus faible (2,56 %) parmi les dispositions de visites des cas falsifiés que dans les deux autres groupes (5,84 % pour les cas valides de falsificateurs, et 6,57 % pour les non-falsificateurs). Le processus de sélection est effectué automatiquement par l'outil de présélection électronique des interviews sur place et ne peut être modifié directement par les intervieweurs sur place. Si les intervieweurs sur place à l'origine de falsifications travaillaient dans des régions où les codes 32 n'étaient pas répandus (par exemple, dans des régions où la plupart des ménages étaient des ménages à une personne), on aurait noté un taux plus faible de codes 32 dans les cas valides, ce qui ne s'est pas produit. Il se peut que les falsificateurs aient décidé de poursuivre des interviews valides auprès d'UL de code 32, étant donné que cela leur fournissait l'occasion de mener deux interviews en une visite. Les taux de réponse sont généralement plus élevés lorsque deux membres de l'UL sont sélectionnés, plutôt qu'un seul, ce qui peut avoir contribué aux décisions des intervieweurs sur place à l'origine de falsifications relativement aux cas à compléter.

Les données au niveau du cas montrent que le nombre moyen de visites par cas était plus élevé (6,98) pour les interviews falsifiées que pour les interviews menées par des falsificateurs (5,47) ou les interviews valides menées par des non-falsificateurs (6,17). Le pourcentage d'interviews attribuées à une disposition sans contact, à un moment donné avant la fin de l'interview, était beaucoup plus élevé pour les interviews falsifiées (82,35 %) que pour les

deux autres catégories (53,19 % et 57,53 %). Lorsque l'on soustrait la date de visite médiane de la date de visite finale, nous voyons que les interviews falsifiées étaient plus susceptibles d'avoir été menées après le milieu de la période consacrée aux cas (6,24 jours) que les interviews valides par des falsificateurs (3,32 jours) ou les interviews valides par des non-falsificateurs (4,76 jours).

Enfin, en plus des modèles existants d'évaluation par score, les réponses au questionnaire ont été analysées, et on a déterminé que les interviews falsifiées étaient plus susceptibles de montrer des profils de réponses caractérisés par l'absence de déclaration de consommation de drogue . Il se peut que les intervieweurs favorisent ce qu'ils pensent être des réponses modales pour éviter la détection.

Les résultats actuels appuient la conclusion selon laquelle un comportement suspect ou frauduleux peut prendre de nombreuses formes, qui peuvent changer au fil du temps. Les outils et les rapports d'analyse utilisant régulièrement des algorithmes de détection des anomalies et des modèles d'évaluation par score augmenteraient de façon significative la capacité d'identifier et de caractériser les profils possibles de comportement suspect ou frauduleux, et particulièrement ceux qui n'ont pas été identifiés au préalable. Parmi les rapports recommandés figurent le dépistage des violations de protocole possibles, les résultats d'interview inhabituels, les données produites en retard, les taux élevés de conversion de refus, et l'efficacité des contacts, ainsi que les écrans existants portant sur les métadonnées d'IAO.

Tableau 2 : Rapports recommandés de dépistage de la falsification

Titre du rapport	Objectif
Rapport de violation de protocole	Vérification des données de piste de vérification, en vue de déceler les violations possibles de protocole, y compris les heures inhabituelles d'interview (entre minuit et 6 h du matin), le nombre d'interviews par jour, les intervalles des interviews (le délai écoulé entre les interviews), ainsi que le nombre et la fréquence des ruptures.
Rapport du résultat des interviews	Les intervieweurs déclarant des résultats inhabituellement élevés et des taux élevés de conversion des refus feront l'objet d'un suivi.
Présentation tardive des données	Les intervieweurs qui reportent constamment les interviews, qui ont des interviews en suspens ou qui terminent un pourcentage important de leurs interviews tard dans le cycle de collecte feront l'objet d'un suivi.
Rapport d'efficacité des contacts au moyen du RV	Vérification du nombre de contacts par interview terminée dans les données du RV.
Écran de métadonnées (rapport existant)	Rapports de métadonnées courants faisant état des profils déclarés d'utilisation toute la vie durant au niveau de l'État et des profils inhabituels de consommation de drogue et comportant des comparaisons de données temporelles individuelles correspondant à des raccourcis possibles.

IAO = interview assistée par ordinateur; RV = registre des visites.

4. DISCUSSION

Le projet a donné des résultats mixtes. Il a clairement permis de démontrer qu'il existe des possibilités d'utilisation de l'exploration des données comme outil de dépistage de la falsification dans le cadre de la National Survey on Drug Use and Health (NSDUH). De façon plus particulière, les conclusions complémentaires figurant dans différentes ressources de données font ressortir l'utilité d'utiliser une combinaison de stratégies de recherche automatisée et d'examen par les experts des diverses bases de données de la NSDUH. Par exemple, une association putative entre la difficulté de mener les interviews et la falsification est ressortie des données du RV, ce qui laisse supposer que la décision de falsifier a été prise par l'intervieweur après qu'il ait eu de la difficulté à communiquer

avec le sujet, de même qu'avec la conclusion selon laquelle le nombre accru de ruptures est lié à des intervieweurs qui ont fait l'objet d'un examen plus étroit par l'équipe chargée d'assurer la qualité des données.

Les responsables de l'analyse du projet sont très intéressés à élargir ces techniques, en vue d'inclure des sources de données additionnelles (p. ex., des vérifications de la cohérence) et d'intégrer l'exploration des données dans les rapports automatisés courants. Cet intérêt a augmenté encore davantage depuis que les conclusions de l'étude ont validé certains des doutes et des préoccupations existants de l'équipe chargée de la qualité des données, particulièrement dans le domaine de la violation des protocoles. Par ailleurs, les logiciels d'exploration des données comme SPSS Clementine ou SAS Enterprise Miner sont faciles à utiliser par les utilisateurs novices, ce qui en fait des outils efficaces pour l'équipe chargée de la qualité des données.

Par ailleurs, le dépistage de la falsification constitue une application particulièrement difficile du processus d'exploration des données, ce qui complique le choix rapide et l'application des techniques courantes. Cela ne signifie toutefois pas que ces techniques ne peuvent être utilisées pour dépister la falsification dans le cadre de la NSDUH, mais cela veut dire que davantage d'efforts devront être déployés pour déterminer les meilleures méthodes, en fonction des données et des avantages pour le projet. Les résultats actuels appuient la conclusion selon laquelle un comportement suspect ou frauduleux peut prendre diverses formes, qui peuvent aussi changer au fil du temps. Par conséquent, de la formation additionnelle concernant le logiciel d'exploration des données ou d'autres outils d'analyse améliorerait de façon significative la capacité de déterminer et de caractériser les comportements suspects ou frauduleux, particulièrement ceux qui n'ont jamais été décelés, et d'autres profils émergents ou changeants de comportements inappropriés de la part des intervieweurs.

RÉFÉRENCES

Berk, R. A., "An introduction to ensemble methods for data analysis", *Sociological Methods and Research*. UCLA Statistics Preprint #417, présentement sous presse. Obtenu le 11 août 2005 du site http://preprints.stat.ucla.edu/

Berk, R. A., Kriegler, B. et Baek, J., "Forecasting dangerous inmate misconduct: An applications of ensemble statistical procedures", UCLA Statistics Preprint #424, présentement sous révision. Obtenu le 11 août 2005 du site http://preprints.stat.ucla.edu/

Lennert-Cody, C. E. et Berk, R.A., "Statistical learning procedures for monitoring regulatory compliance: An application to fisheries data", UCLA Statistics Preprint #426, présentement sous révision. Obtenu le 11 août 2005 du site http://preprints.stat.ucla.edu/

Mena, J. (2003), *Investigative Data Mining for Security and Criminal Detection*. Boston: Butterworth-Heinemann.

Murphy, J., Baxter, R.K., Eyerman, J., Cunningham, D. et Barker, P. (2004), "A system for detecting interviewer falsification", Article présenté à la 59e rencontre annuel de l'*American Association for Public Opinion Research*, Phoenix, AZ.

Schraepler, J. P. et Wagner, G.G. (2003), "Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP", DIW Research Note, Berlin (à paraître bientôt). IZA Discussion Paper No. 969. Disponible sur le site http://ssrn.com/abstract=487402