Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

INTERVIEWER FALSIFICATION DETECTION USING DATA MINING

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

ABSTRACT

This paper describes an innovative use of data mining on response data and metadata to identify, characterize and prevent falsification by field interviewers on the National Survey on Drug Use and Health (NSDUH). Interviewer falsification is the deliberate creation of survey responses by the interviewer without input from the respondent. Undetected falsification can introduce bias into the population estimates if falsified responses do not match the values that would have been provided by respondents. Currently, the procedures required to detect interviewer fraud can be expensive and draw resources away from the study that could be applied to data quality procedures. Data mining can be used to program falsification propensity checks that may be run on a frequent basis, facilitating timely and relatively inexpensive detection and remediation, and serving as a possible deterrent to falsification.

KEY WORDS: Falsification; Data mining; Interviewers

1. INTRODUCTION

1.1 Interviewer Falsification

Interviewer falsification is the deliberate creation of survey responses by the survey interviewer without input from the persons sampled to be respondents. It can be detrimental to surveys in several ways. First, detected falsification can have a considerable impact on the survey budget if fraudulent cases are reinterviewed. Second, detection procedures can be expensive and can draw resources away from the study that could be applied to other data quality procedures. Finally, undetected interviewer falsification can introduce bias into the population estimates if falsified responses do not match the values that would have been provided by sample members if they were surveyed (Schraepler & Wagner, 2003). The greater the number of undetected falsified cases and the greater the difference between the true population values and the falsified values, the greater the bias introduced into the survey statistics.

Although it is believed that falsification of data by field interviewers (FIs) in the National Survey on Drug Use and Health (NSDUH) is a rare event, some damaging instances have been detected in recent years and have raised concerns. Furthermore, recent design changes to the study may have modified the propensity for interviewers to falsify. For example, the addition of monetary incentives (2002) may encourage falsification among unscrupulous interviewers and may even attract white-collar criminals to the interviewing profession.

This paper describes a special NSDUH study that explored the issue of falsification in order to develop a detection system for NSDUH that reduces monitoring costs and increases the rate of falsification detection. Data mining was selected as the analytical strategy for this study based on its ability to conduct automated searches of large multivariate data sets, as well as its rapidly emerging importance as the analytical strategy of choice in applied fraud detection. Data mining is an automated process that facilitates exhaustive searches of datasets and the identification of trends, patterns, and complex relationships that are not obvious or attainable through traditional analytical methodologies. One of the benefits of using data mining is that it provides the opportunity to effectively review extremely large datasets in a timely fashion and tailor specific fraud detection and reduction strategies. Survey

[1] RTI International, PO Box 12194, Research Triangle Park, North Carolina, USA (email contact: 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

response data and metadata are examined using the CRoss Industry Standard Process for Data Mining (CRISP-DM) model as a framework. More detail about the CRISP-DM model can be found at http://www.crisp-dm.org/.

1.2 Existing Falsification Detection Strategies on the NSDUH

The NSDUH is an annual cross-sectional face-to-face household survey that gathers data on substance use and abuse among the non-institutionalized civilian population of the United States aged 12 and older. Formerly called the National Household Survey on Drug Abuse (NHSDA), the NSDUH is conducted under contract by RTI International, Research Triangle Park, North Carolina. The Substance Abuse and Mental Health Services Administration (SAMHSA) sponsors the survey.

As part of NSDUH's Data Quality (DQ) Monitoring System, field interviewers (FIs) attempt to collect telephone numbers from all selected households. These telephone numbers are used to call individuals and check on the quality of the interviewers' work. This verification is conducted on the first two noninterview cases and the first two interview cases completed by each interviewer in each calendar quarter. In addition, at least 5 percent of each interviewer's completed household screenings and at least 15 percent of each interviewer's completed interviews are randomly selected for telephone verification. Trained professional telephone interviewers call to verify that the screening or interview occurred and was conducted in the correct manner. Interview cases selected for telephone verification that do not have a telephone number are verified by mail. If there is any suspicion about the performance of an interviewer, a greater proportion of his or her cases (up to 100 percent) can be "forced" into verification. Some data quality trends that may lead to increased verification include a high rate of missing or refused verification telephone numbers, an interviewer report of his or her own or another staff member's phone number for verification, unexpected duplicate use of any particular telephone number, and any problem that represents a serious protocol violation. In addition, time-stamp data on each interview are collected. These data are transmitted nightly from the interviewers' computers along with completed interview data. Any cases completed in less than 30 minutes or in more than 60 minutes are examined. When there are serious concerns about the validity of an interviewer's work, a mixed sample of the interviewer's completed cases is selected for in-person field verification. Those found guilty of falsification are terminated from employment immediately.

Recent additions to the NSDUH falsification detection strategies include the review of unusual responses, response patterns, and interview length (Murphy, et al., 2004). Additional procedures are in place to identify incompatible patterns of responses within the survey. These current strategies, while effective, are time consuming and labor intensive. The large number of cases identified for additional screening, review, and validation has resulted in the identification of numerous cases of interviewer fraud that would not have been detected had these procedures not been in place. However, the large number of cases flagged for additional review and the high number of false positives associated with this approach underscore the need for developing more targeted approaches to fraud detection. Moreover, it is likely that there are other patterns of interview fraud that have yet to be identified. The ability to target identification and deterrence strategies to specific patterns of interviewer fraud offers researchers the ability to deploy fraud detection resources more effectively and to develop meaningful strategies that directly address the specific pattern of fraud and interviewer misconduct.

1.3 Recent Historical Record of NSDUH Falsification

From 2000 to 2001, detected falsification prevalence on the NSDUH was relatively low (about one-tenth of a percent of completed interviews were found to be falsified). However, in late 2002, four interviewers were identified who had recorded their own or another interviewer's telephone number for verification, a violation of project protocol. Through in-person verification of the entire work assignment for these interviewers, it was determined that all four interviewers had committed frequent falsification. Of the total 760 screenings and 464 interviews completed by the four interviewers, 287 screening cases and 134 interview cases were deemed valid. A total of 473 screenings and 330 interviews were determined to have been falsified and were subsequently reworked. Beginning in 2003, the level of detection has been similar to 2000 and 2001.

The interview response and question-level timing data from the 2002 falsified cases were examined and compared with data from valid interviews completed in those States. Some significant differences between the falsified and valid cases were detected leading to the development and incorporation of response data and metadata-based scoring models for prediction of future falsification propensity (Murphy, et al., 2004). The data mining work in this paper builds upon this initial investigation to improve the detection process.

2. ANALYSIS USING THE CRISP-DM MODEL

2.1 Goals and Challenges

The application of data mining techniques to falsification on the NSDUH faced two immediate challenges. First, data mining techniques are not widely used in the social sciences[3] or the survey methods literature. As a result, the two fields lack a common vocabulary, a standard or widely accepted method, and common reporting conventions. Second, data mining is most frequently used in business applications (marketing, fraud detection, etc.). Reporting standards are generally tailored to specific business needs and cannot be used as a guide for social science reporting. Furthermore, the analyses are often business confidential and are not available for public review, making it difficult to identify and reference examples from public literature. Finally, the published literature is less comprehensive than what is found in the social sciences or survey methods. The CRoss Industry Standard Process for Data Mining (CRISP-DM) model was selected to address these two challenges by providing a standardized analysis and reporting format that has been vetted with data mining experts. The primary goal of this effort was to enhance the detection system used on NSDUH in a way that reduces monitoring costs and increases the accuracy of falsification detection.

2.2 Data

The available NSDUH data were categorized into interview process data and survey results as shown in Table 1. The metadata included the record of calls (ROC) and the audit trail, while the survey data included the actual responses to the survey questions provided by the survey respondent. Survey Year 2004 data were used for this project since this was the most recent, complete year available at the time of analysis. To reduce the time associated with data preparation, single quarters of data were used for some analyses.

Table 1. Available NSDUH Data for Analysis

Data Source	Number of Records	Description of Record
ROC	1,169,054	Each record includes one visit to the dwelling unit or disposition assignment directed by a field supervisor.
Audit Trail	70,356	Each record represents one entry into the CAI instrument.*
Raw CAI	67,913	Each record includes the raw survey data for one interview.

CAI = computer-assisted interviewing; ROC = record of calls.
*A single interview with multiple breakoffs can be associated with multiple audit trail records.

2.3 Mining Process

The CRISP-DM model incorporates six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Perhaps the most important phase of the data mining process includes gaining an understanding of the current practices and overall objectives of the project. Therefore, during the business understanding phase of the data mining process, current procedures and previous NSDUH/NHSDA falsification detection reports were reviewed to provide guidance during initial exploration of the data, as well as the selection of specific statistical approaches and algorithms during the modeling process.

The identification and characterization of falsified data was conducted using fraud detection methods similar to those employed by other agencies, including the Internal Revenue Service (IRS) and the health care industry (e.g.,

[3] This issue is being addressed through a National Science Foundation grant to Richard Berk and colleagues. Some examples of unpublished work in this area are included in the references (Berk, in press; Lennert-Cody & Berk, under review; Berk, Kriegler, & Baek, under review).

detection of Medicaid fraud) (Mena, 2003). These general methods can be divided into two areas: scoring models and anomaly detection techniques. These methods can be employed with both the interview process data as well as the actual survey responses.

Scoring models examine existing patterns of fraudulent data and collection activity so similar patterns in future cases can be flagged for additional evaluation. Identification of these patterns may suggest interviewer falsification or respondent deception, either of which warrants additional analysis. These models capitalize on the fact that people frequently are not creative or unique when they falsify data, that is, they tend to falsify data in similar ways. To date, scoring models have been successful in leading to the early identification of at least three falsifiers on the NSDUH.

Another approach, known as **anomaly detection**, characterizes "normal" patterns of responses and data collection activity. Data deviating from these normal patterns were identified for further evaluation. Examples include unusually quick survey response times, extremely high response rates, and inordinate amounts of data collected during a single session or day. These extreme data generally would not be consistent with the other sets derived and would warrant additional review and follow-up. Similarly, unusually consistent response patterns, or response patterns that are incompatible or that deviate from known patterns of behavior, are also cause for follow-up and review. While these anomalous data or outliers do not necessarily indicate fraudulent activity specifically, they are worthy of follow-up to ensure a high level of data quality and accuracy. Using a combination of scoring models and anomaly detection allowed us to address the known patterns of fraudulent behavior, as well as flag patterns that may indicate new or suspicious behaviors and protocol violations.

It should be noted that predicting low-frequency events like fraud can be particularly challenging. Overall accuracy of the model can be somewhat misleading with low frequency events. For example, a model would be correct 97 percent of the time if it always predicted "no" for an event with an expected frequency of 3 percent. Clearly, overall accuracy would be an unacceptable measure for the predictive value of this type of model. In these cases, the nature and direction of errors can provide a better estimate of the overall value of the model. By adjusting the "costs" associated with false positives or misses, the model can be refined to better predict low-frequency events. These costs can be balanced to create a model that accurately identifies cases of interest ("true positives") while limiting the number of false alarms.

Significant domain expertise or extensive knowledge of the overall project requirements, data resources, procedures, and goals is essential to creating predictive models that are operationally reasonable in their criteria. The investigators included project staff with significant experience with the NSDUH project, as well as the ongoing fraud detection protocols. The report team worked closely with the DQ Team to ensure that the models not only reflect current knowledge regarding interviewer fraud but also are valuable and actionable in their setting.

3. FINDINGS

The CRISP-DM model revealed several patterns of protocol violations and patterns of interviewing that were associated with suspected patterns of fraudulent behavior. Though space in this proceedings volume does not allow for full presentation of the patterns detected through the mining process, a summary follows.

Interviewers with a large number of "breakoffs" or disruptions were associated with known falsification. Interviews completed late in the field period had different timing characteristics, were generally "difficult" cases, and were less likely to be verified due to end of quarter close-out procedures.

Among falsified interviews, the occurrence of a "Code 01" (no one home at screening call) was higher (28.21 percent) than among valid interviews by falsifiers (21.01 percent) or valid interviews among nonfalsifiers (20.77 percent). If all FI work associated with the falsified cases was fraudulent, then the number of 01s might be higher because falsifiers were attempting to show a significant amount of valid work when there was none. A more plausible explanation is that cases with many 01 codes were more difficult to complete, and the decision to falsify the interview was made after experiencing the difficulty to contact. Because the number of detected falsified interviews was sufficiently low, tests for significance did not register significant differences, but that does not necessarily mean that there was no relationship between distribution of call disposition and falsification.

An initially surprising finding was that the rate of "Code 32s" (two dwelling unit members selected for interview) was much lower (2.56 percent) among call dispositions for falsified cases than the two other groups (5.84 percent for valid cases by falsifiers and 6.57 percent for nonfalsifiers). The selection process is automatically completed by the FI's electronic screening device and cannot be directly altered by the FI. If the falsifying FIs were working in areas where Code 32s were uncommon (such as areas where most householders live alone), the lower rate of 32s would also be seen in their valid cases, but this was not so. What might have happened is that falsifiers decided to pursue valid interviews with Code 32 DUs since these represented opportunities to complete two interviews in one visit. Response rates are generally higher when two DU members are selected rather than one, so this also might have factored into falsifying FIs' decisions on which cases to validly complete.

Case-level data showed that the mean calls per case were higher (6.88) for the falsified interviews than for valid interviews by falsifiers (5.47) or valid interviews by nonfalsifiers (6.17). The percent of interviews that were assigned a noncontact disposition at some point prior to completion was much higher for falsified interviews (82.35 percent) than for the other two categories (53.19 percent and 57.53 percent). Subtracting the median call date from the final call date, we find that the falsified interviews were more likely to have been completed later than the middle of the time spent working on the case (6.24 days) than valid interviews by falsifiers (3.32 days) or valid interviews by nonfalsifiers (4.76 days).

Finally, as an extension of existing scoring models, questionnaire responses were analyzed and it was found that falsified interviews may be more likely to show response patterns associated with no reported drug use. Interviewers may favor what they believe to be modal responses to avoid detection.

The current results support the finding that suspicious or fraudulent behavior can assume many forms that can change over time. Analytical tools and reports regularly employing anomaly detection and scoring model algorithms would enhance significantly the ability to identify and characterize possible suspicious or fraudulent patterns of behavior, particularly those that were previously unidentified. These recommended reports include screening for possible protocol violations, unusual interview yields, late producers, high-refusal converters, and contact efficacy, as well as existing screens that address the CAI metadata.

Table 2. Recommended Falsification Screening Reports

Name of the Report	Purpose
Protocol Violation Report	Screen audit trail data for possible protocol violations, including unusual time of interview (midnight to 6:00 a.m.), number of interviews per day, interview interval (e.g., time between interviews), and number and frequency of breakoffs.
Interview Yield Report	Interviewers reporting unusually high yields and high refusal converters will be flagged for follow-up.
Late Producers	Interviewers consistently back-loading interviews, holding interviews, or completing a large percentage of their interviews late in the collection cycle will be flagged for follow-up.
ROC Contact Efficacy Report	Screen number of contacts per completed interview in the ROC data.
Metadata Screen (Existing Report)	Maintain current metadata reports that determine reported lifetime use patterns at the State level, unusual patterns of substance use, and comparisons of individual item timing data indicating possible shortcutting.

CAI = computer-assisted interviewing; ROC = record of calls.

4. DISCUSSION

This project yielded mixed results. It clearly demonstrated that there is potential for data mining to be used as a falsification detection tool on the National Survey on Drug Use and Health (NSDUH). In particular, the complementary findings noted in different data resources underscore the value that can be obtained by using a combination of automated search strategies and expert review on the various NSDUH databases. For example, a putative association between difficult to complete interviews and falsification was supported by the ROC) data, which suggested that the decision to falsify was made after the interviewer had experienced difficulty in contacting the subject, and the finding that an increased number of breakoffs was associated with interviewers who had been placed under increased scrutiny by the DQ Team.

There is a great deal of interest from the project analytic staff to expand these techniques to include additional data sources (e.g., consistency checks) and to integrate data mining with the current automated reports. This interest has increased even more since the study findings have validated some of the existing suspicions and concerns of the DQ Team, particularly in the area of protocol violations. Furthermore, mining software, such as SPSS Clementine or SAS Enterprise Miner, is approachable to novice users, making it an effective tool for the DQ Team.

On the other hand, falsification detection is an especially challenging application of the data mining process, making it difficult to rapidly identify and apply the standard techniques. This is not to suggest that these techniques cannot be used for falsification detection on NSDUH, but that more effort will need to be expended to identify the best methods that fit the data and benefit the project. The current results support the finding that suspicious or fraudulent behavior can assume many forms that also can change over time. Therefore, additional training on data mining software and other analytical tools would enhance significantly the ability to identify and characterize possible suspicious or fraudulent behaviors, particularly those that are previously unidentified and other emerging or changing patterns of inappropriate interviewer behavior.

REFERENCES

Berk, R. A., "An introduction to ensemble methods for data analysis", *Sociological Methods and Research*. UCLA Statistics Preprint #417, currently in press. Retrieved August 11, 2005, from http://preprints.stat.ucla.edu/

Berk, R. A., Kriegler, B. and Baek, J., "Forecasting dangerous inmate misconduct: An applications of ensemble statistical procedures", UCLA Statistics Preprint #424, currently under review. Retrieved August 11, 2005, from http://preprints.stat.ucla.edu/

Lennert-Cody, C. E. and Berk, R.A. (under review), "Statistical learning procedures for monitoring regulatory compliance: An application to fisheries data", UCLA Statistics Preprint #426, currently under review. Retrieved August 11, 2005, from 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. and Barker, P. (2004), "A system for detecting interviewer falsification", Paper presented at the 59th annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Schraepler, J. P. and 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 (forthcoming). IZA Discussion Paper No. 969. Available at http://ssrn.com/abstract=487402