Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

A statistical approach to detect interviewer falsification of survey data

by Sebastian Bredl, Peter Winker and Kerstin Kötschau

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 1-10
Statistics Canada, Catalogue No. 12-001-X

A statistical approach to detect interviewer falsification of survey data

Sebastian Bredl, Peter Winker and Kerstin Kötschau [1]

Abstract

Survey data are potentially affected by interviewer falsifications with data fabrication being the most blatant form. Even a small number of fabricated interviews might seriously impair the results of further empirical analysis. Besides reinterviews, some statistical approaches have been proposed for identifying this type of fraudulent behaviour. With the help of a small dataset, this paper demonstrates how cluster analysis, which is not commonly employed in this context, might be used to identify interviewers who falsify their work assignments. Several indicators are combined to classify 'at risk' interviewers based solely on the data collected. This multivariate classification seems superior to the application of a single indicator such as Benford's law.

Key Words: Data fabrication; Falsifier; Benford's law; Cluster analysis.

1. Introduction

Whenever data collection is based on interviews, one has to be concerned about data quality. Data quality can be affected by false or imprecise answers of the respondent or by a poorly designed questionnaire, as well as by the interviewer when he or she deviates from the prescribed interviewing procedure. If the interviewer does so consciously, this is referred to as 'interviewer falsification' (Schreiner, Pennie and Newbrough 1988) or 'cheating' (Schräpler and Wagner 2003).

Interviewer falsification can occur in many ways (*cf*. Guterbock 2008). Rather subtle forms consist of surveying a wrong household member or of conducting the survey by telephone when face-to-face interviews are required. The most severe form of falsifying is the fabrication of entire interviews without ever contacting the respective household. In our analysis, we deal with the latter case.

Fabricated interviews can have serious consequences for statistics based on the survey data. Schnell (1991) and Schräpler and Wagner (2003) provide evidence that the effect on univariate statistics might be less severe, provided the share of falsifiers remains sufficiently small and the 'quality' of the fabricated data is high. But even a small proportion of fabricated interviews can be sufficient to cause heavy biases in multivariate statistics. Schräpler and Wagner (2003) find that the inclusion of fabricated data from the German Socio Economic Panel (GSOEP) in a multivariate regression reduces the effect of training on log gross wages by approximately 80 percent, although the share of fabricated interviews was less than 2.5 percent. This indicates the importance of identifying these interviews.

The most common way to identify falsifying interviewers is the reinterview (Biemer and Stokes 1989). In this case, a supervisor contacts some of the households that should have been surveyed to check whether they were actually visited by the interviewer. However, for reasons of expense, it is impossible to reinterview all households participating in a survey (*cf*. Forsman and Schreiner 1991). Therefore, the question arises of how the reinterview sample can be optimized to best detect falsifiers. Generally, it seems useful to select households for reinterview if the interviews were done by an interviewer − identified by characteristics linked to the answers in his interviews − who is more likely than others to be fabricating data. In this context, Hood and Bushery (1997) uses the term 'at risk' interviewer. If reinterview participants are sampled in a two-stage setting, whereby interviewers are selected in the first stage and participants surveyed by those interviewers in the second stage (as recommended by Forsman and Schreiner (1991)) one might oversample the at risk interviewers in the first stage.

In this paper, we demonstrate a purley statistical approach that relies on the data contained in the questionnaries to define a group of at risk interviewers. This is not a new idea; literature provides several examples for this kind of approach (Hood and Bushery 1997; Diekmann 2002; Turner, Gribbe, Al-Tayyip and Chromy 2002; Schräpler and Wagner 2003; Swanson, Cho and Eltinge 2003; Murphy, Baxter, Eyerman, Cunningham and Kennet 2004; Porras and English 2004; Schäfer, Schräpler, Müller and Wagner 2005; Li, Brick, Tran and Singer 2009). However, with the exception of the work of Li *et al*. (2009), the tests conducted in these studies rely on the examination of single indicators derived from the interviewer's data to detect falsifiers. Some studies calculate several indicators but consider them all separately. We combine multiple indicators in cluster

1. Sebastian Bredl, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: sebastian.bredl@wirtschaft.uni-giessen.de; Peter Winker, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: peter.winker@wirtschaft.uni-giessen.de; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankeneser Landstrasse 7, Germany. E-mail: kkoetschau@hanse-parlament.eu.

analyses, allowing for a better classification of the potential falsifiers compared to previous approaches. To the best of our knowledge, this procedure is an innovation in the context of identifying interviewers who fabricate data, but has already been employed in other fields in order to detect fraudulent behaviour. The basic idea is that characteristics of fraudulent 'cases' (what a case is depends on the context) feature striking patterns compared to honest cases that can be revealed if those characteristics are jointly considered in a cluster analysis. Murad and Pinkas (1999) try to detect fraud in the telecommunication industry by means of clustering call profiles of clients. A call is characterized by several indicators like calling time or destination of the call. Thiprungsri (2010) clusters group life claims submitted from clients to life insurance companies based on several characteristics of the claims. Claims that form very small clusters are considered to be suspicious. Donoho (2004) uses cluster analysis, among others, to trace patterns in option markets that might indicate insider trading.

We have a small survey dataset available (see subsection 3.1 for a further description of our dataset), which partially consists of falsified data. With a total of 13 interviewers and 250 questionnaires, the size of the dataset is quite limited and it is not clear to what extent our findings can be generalized to larger datasets. However the dataset enables us to demonstrate our approach. The fact that we know which data was collected honestly and which data was fabricated allows for a first evaluation of our approach. It must be stated that this a priori knowledge is no prerequisite to employ the method.

The problem of identifying at risk interviewers was addressed in the 1980s, however, literature on this issue is still scarce. In 1982, the U.S. Census Bureau implemented the Interviewer Falsification Study. Based on the information collected in the context of this study, Schreiner *et al*. (1988) find that interviewers with a shorter length of service are more likely to fabricate data. Hood and Bushery (1997) use several indicators to find at risk interviewers in the National Health Interview Survey (NHIS). For example, they calculate the rate of households that have been labelled ineligible or the rate of households without telephone number per interviewer and compare the rates to census data from the respective area. When large differences occur, the interviewer is flagged and a reinterview is conducted. Detection rates among the flagged interviewers turn out to be higher than those in random reinterview samples. Turner *et al*. (2002) also find interviewers committing data fabrication to indicate telephone numbers less frequently than honest interviewers when examining the Baltimore STD and Behaviour Survey. For the case of computer assisted interviewing, Bushery, Reichert, Albright and Rossiter (1999) and Murphy *et al*. (2004) propose the use of date and time stamps – the recording of the time and the duration of the interview by the computer – to find suspect interviewers. Those who need a remarkably long or short time to complete the entire questionnaire or certain modules or complete remarkably many questionnaires within a given time period might be flagged as at risk interviewers. Schäfer *et al*. (2005) assume that falsifiers avoid extreme answers when fabricating data. Using data of the GSOEP, the authors calculate the variance of the answers for every question on all questionnaires of an interviewer and sum up all variances. Thanks to other control mechanisms in the GSOEP, falsifiers are known and it turns out that they could be found among the interviewers with the lowest overall variances. Porras and English (2004) use a similar approach and also find falsifiers to produce variances that are smaller to those found in honestly filled questionnaires. Li *et al*. (2009) combine several predictive indicators in a logistic regression model in which the known falsification status of an interview serves as a binary dependent variable. The authors find that reinterview samples that overweight cases with a high probability of being fraudulent according to the logistic regression model identify more cases of actual data fabrication than purely randomly drawn samples. However, it is evident that past reinterview data with known falsification status must be available to conduct the logistic regression.

Further indicators discussed in literature are the number of rare or unlikely response combinations in an interviewer's questionnaires (Murphy *et al*. 2004; Porras and English 2004) and the comparison of household compositions or descriptive statistics in interviewer's questionnaires with the entire sample (Turner *et al*. 2002; Murphy *et al*. 2004).

Another means of detecting fabricated data that has gained a lot of popularity in recent years is Benford's law (Schräpler and Wagner 2003; Swanson *et al*. 2003; Porras and English 2004; Schäfer *et al*. 2005), which will be discussed in section 2, along with its success in detecting fabricated interviews in previous studies. Furthermore, section 2 describes our statistical approach to identify falsifiers. Section 3 presents the data our analysis is based upon as well as our results. The paper concludes with a discussion of our findings.

2. Methods

2.1 Benford's law

When the physicist Frank Benford noticed that the pages in logarithmic tables containing the logarithms of low numbers (1 and 2) were more used than pages containing logarithms of higher numbers (8 and 9), he started to investigate the distribution of leading digits in a wide range

of different types of numbers like numbers on the first page of a newspaper, street addresses or molecular weights (Benford 1938). Benford found that the distribution of the leading non-zero digits could be described by the following formula which has become known as 'Benford's law:'

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

However, not all series of numbers Benford (1938) investigated seemed to conform to his law. Consequently, the question arose what kind of data can be supposed to produce first digits in line with the law. Discussions of this issue are provided by Hill (1995), Nigrini (1996), Hill (1999) and Scott and Fasli (2001). The detection of financial fraud is a field in which the application of Benford's law has gained much popularity during the recent decade (Nigrini 1996; 1999; Saville 2006). The results of those studies are not relevant in our context. However, it is interesting to note that there seems to be a consensus in literature that monetary values can be supposed to follow Benford's law. Swanson *et al*. (2003) show that the distribution of first digits in the American Consumer Expenditure Survey is close to Benford's distribution.

The basic idea of using Benford's law to detect fabricated data is that falsifiers are unlikely to know the law or to be able to fabricate data in line with it. Therefore a strong deviation of the leading digits from Benford's distribution in a dataset indicates that the data might be faked. Of course, one has to be concerned if the nature of the data is such that it can be supposed to follow Benford's law if it is authentic. Benford's law cannot be applied if the questionnaires do not contain any or contain only very few metric variables.

Schräpler and Wagner (2003) and Schäfer *et al*. (2005) use Benford's law to detect data fabrication in the GSOEP. In both studies, all questionnaires delivered by every single interviewer are combined and checked for whether the distribution of the first digits in the respective questionnaires deviates significantly from Benford's law. This can be done by calculating the χ^2-statistic:

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

where n_i is the number of leading digits in all questionnaires from interviewer i, h_{id} is the observed proportion of leading digit d in all leading digits in interviewer i's questionnaires and h_{bd} is the proportion of leading digit d in all leading digits under Benford's distribution. High χ^2-values indicate a deviation from Benford's distribution and indicate at risk interviewers. Schräpler and Wagner (2003) use different kinds of continuous variables in their analysis, whereas Schäfer *et al*. (2005) restrict theirs to monetary values. In both studies, the critical χ^2-values are assumed to be dependent on the sample size n and are consequently adjusted for this parameter. The results obtained look promising. The fit of the leading distribution of first digits to Benford's distribution in the questionnaires of falsifiers (which were already known in advance) is, in general, much worse than for honest interviewers. Thus it seems appropriate to use Benford's law as a means to identify at risk interviewers.

However, when we compared the data of the honest interviewers in our dataset to Benford's distribution, we observed a large deviation for the digit 5. This might be due to rounding of numbers by the respondents. The same problem is mentioned by Swanson *et al*. (2003) and Porras and English (2004) who opt for applying an alternative approach "in the spirit of Benford" (Porras and English 2004, page 4224). We adopt this approach which consists of comparing the distribution of leading digits in the questionnaires of an interviewer to the distribution of first digits in all questionnaires except their own. The χ^2-value on the interviewer level is calculated as described above but the expected proportion of a digit according to Benford's law h_{bd} is replaced by the proportion of the digit in all other questionnaires. We then use the resulting χ^2-value as one indicator in the cluster analysis.

With regard to the selection of variables whose first digits are examined, we stick to the approach of Schäfer *et al*. (2005) and include only the first digits of monetary values in the analysis. The survey we are using for demonstration purposes contains monetary values expressed in local currency referring to household expenditures for different items like leasing or buying land, seeds, fertilizer, taxes, and to household income from different sources like agricultural or non agricultural self employment and public or private transfers. Overall we include first digits of 26 different monetary values per interview, ignoring values that were reported to be zero. We then pool first digits of all questionnaires delivered by one interviewer and compare the distribution of first digits to the one for all other interviews according to the method described above. The restriction to monetary values constitutes a clear criterion during the process of selecting data. Furthermore, as mentioned above, financial data is broadly agreed upon to be apt for the analysis with Benford's law. This is important, although we do not ground our analysis on Benford's distribution but on an approach based on it.

2.2 Multivariate analyses

Our idea is to combine several indicators, which we derive directly from the questionnaires of each interviewer and which we suppose to be different for falsifiers and honest interviewers. We do this by means of cluster and discriminant analysis. All indicators are derived on the

interviewer level. This implies that we pool all questionnaires of one interviewer for the analysis, which increases the amount of data on which every single indicator value is based. This should make the indicator values more reliable and less sensitive to outliers. On the other hand, it is obvious that the discriminatory power of interviewer-level indicators decreases as soon as interviewers only fake parts of their assignments. Looking at indicators on the questionnaire level, therefore, seems to be preferable if the amount of data per questionnaire is sufficiently high.

The cluster analysis constitutes the real method of identifying at risk interviewers. The interviewers are clustered in two groups with the intention of obtaining one that contains a high share of falsifiers and another one that contains a high share of honest interviewers. Clustering does not require a priori information on who is fabricating data and who is not. In fact, this is what it is supposed to reveal. Since we know from the outset which interviewer belongs to which group, we can discover whether the cluster analysis identifies the 'true falsifiers' to be at risk. Clearly, the assumption that our approach is able to separate both groups perfectly is not realistic. The idea is rather that we obtain an at risk interviewer cluster exhibiting a higher share of falsifiers compared to the other cluster. If a reinterview is feasible, subsequent reinterview efforts might be focused on interviewers in the at risk cluster.

To judge the performance of the cluster analysis, we consider the number of undetected falsifiers as well as the number of 'false alarms.' Both types of 'errors' inccur costs: data of undetected falsifiers is likely to impair the results of further statistical analysis. False alarms inccur costs in the sense that an unnecessary effort to reinterview the respective households might be taken or data is unnecessarily removed from the sample. Furthermore, it might be demoralizing for honest interviewers if they see their work being subject to a reinterview, particuliarly if they are aware of the fact that predominantly the work of at-risk interviewers is picked. How to weight an undetected falsifier compared to a false alarm in a loss function is a highly subjective issue. Generally, it seems reasonable to assign more weight to the former than to the latter.

The discriminant analysis requires knowledge on the falsifiers versus non-falsifiers status of each interviewer before it can be conducted. Therefore, it is not an instrument to detect falsifiers. We use the discriminant analysis to verify our hypotheses on the behaviour of falsifiers, which will be discussed below, and to evaluate how well the employed indicators can separate the two groups.

One of the indicators we use is the χ^2-value, calculated by comparing the distribution of first digits in the questionnaires of each interviewer with the respective distribution in all other questionnaires as described in the previous subsection. Furthermore, we derive three other indicators from hypotheses concerning the behaviour of falsifiers fabricating data. Schäfer *et al.* (2005) assume that falsifiers have a tendency to answer every question, thus producing less missing values. Furthermore, in line with Porras and English (2004), they expect falsifiers to choose less extreme answers to ordinal questions. Hood and Bushery (1997) hypothesize that falsifiers will "try to keep it simple and fabricate a minimum of falsified data" (Hood and Bushery 1997, page 820).

Based on these assumptions, we calculate three proportions, which serve as indicator variables in the multivariate analyses along with the χ^2-value. The three indicator variables are calculated as follows:

- The 'item-non-response-ratio' is the proportion of questions which remain unanswered in all questions. We expect this ratio to be lower for falsifiers than for honest interviewers.
- The 'extreme-answers-ratio' refers to answers which are measured in ordinal scales. The ratio indicates the share of extreme answers (the lowest or highest category on the scale) in all ordinal answers. According to the above-mentioned assumptions, this ratio should also be lower for falsifiers.
- The 'others-ratio' refers to questions which, besides several framed responses offer the item 'others' as a possible answer. The choice of this item requires the explicit declaration of an alternative. If falsifiers tend to keep it simple, we can expect them to prefer the framed responses to the declaration of an alternative. Thus, this ratio too (calculated as the proportion of 'others' answers in all answers where the others item is selectable) should be lower for falsifiers.

Of course, the list of indicator variables, which might be included in the cluster analysis, can be extended. Generally, it is possible to derive many more of those variables from hypotheses on the behaviour of interviewers who fabricate data or to use those which have already been proposed in the literature, albeit not in the context of cluster analysis. For example, based on the assumption that falsifiers try to fabricate a minimum of falsified data, Hood and Bushery (1997) expect them to disproportionately often select the answer 'No' to questions, which either lead to a set of new questions or avoid it (assuming that 'No' is generally the answer that avoids further questions). So one could calculate the ratio of 'No' answers to such questions and use this ratio as a variable in the cluster analysis. We do not use this ratio, as two slightly different versions of the questionnaire were used in our empirical sample. There are only a small number of questions that lead to new questions or avoid

them depending on the answers, which are identical in both versions of the questionnaire.

Furthermore, when computer assisted interviewing allows the use of date and time stamps as discussed by Bushery *et al*. (1999), the average time needed to conduct an interview or the number of interviews conducted in one day might serve as indicators. Panel surveys offer some additional information to construct indicators. Stokes and Jones (1989) propose to compare the actual rate of nonmatched household members in an interviewer's questionnaires to expected nonmatch rates that are calculated conditional on several household characteristics. The authors employ this procedure in the post-enumeration survey that is conducted as follow-up survey for the U.S. Census. If the actual rate of nonmatches strongly exceeds the expected rate, the authors consider this to be an indicator for fabricated data. Generally, this approach is applicable as soon as one has two or more waves of a panel survey available.

It becomes obvious that the first steps of our approach consist of examining the structure of the questionnaire and other types of data like date or time stamps collected during the survey process. Then one might consider which indicators could be derived from those sources that are likely to differ between falsifiers and honest interviewers. Another approach is the use of data mining techniques to identify patterns that are common in fabricated data or patterns in which fabricated data differs from honestly collected data (Murphy, Eyerman, McCue, Hottinger and Kennet 2005). If those patterns are detected, they might be used as indicators instead of deriving indicators from hypothesis on falsifier behaviour. However, this approach requires a huge dataset with known cases of falsification in order to conduct the data mining process. Such a dataset is not always available.

3. Results

3.1 Data sources

The data used in this study are derived from household surveys conducted in November 2007 and February 2008 in a Commonwealth of Independent States (CIS) (*i.e.*, former Soviet Union) country. The survey was part of an international research project on land reforms and rural poverty. We intended to interview 200 households in four villages in 2007. After identifying that all interviews had been fabricated in the first surveyed village we broke the survey off and started a new round with new interviewers in other villages in February 2008. All villages had been selected by qualitative criteria like the agricultural production structure and the implementation of land reforms. The households within one village had been selected by random sample based on household lists, which were provided by the mayors of the villages. This procedure not only assured that all households had been selected at random, but also provided the basis for reinterviews as all households were exactly defined. However, these reinterviews were not planned in the very beginning. Because the households rarely owned telephones, check-calls were not possible and reinterviews in these households were associated with high costs and expenditure of time for traveling to the village for a face-to-face reinterview. Five interviewers were engaged in the first 2007 survey. Two of them had been the local partners of the research project. They had been involved in the development of the questionnaire and were responsible for the coordination of the surveys in their country. The other three interviewers were students hired by the partners. The questionnaire was composed of different sections with regard to household characteristics, resource endowment as well as income and expenditures. Most of the questions were closed questions. Only a few questions included a scale. Metric variables were collected for household expenditures like leasing or buying land, seeds, fertilizer or taxes and household income from different sources like agricultural or non-agricultural self employment and public or private transfers.

When the interviews of the 2007 survey were conducted, none of the German researchers were present in the villages. The questionnaires were collected right after the survey of the first village. In a first review of the questionnaires, we became suspicious because the paper of the questionnaires looked very clean and white. There was no dirt or dog-ears on the paper. Comparing the answers of different questionnaires of one interviewer we found two questionnaires with identical answers. Considering the fact that we asked for the amount of income from different sources in metric numbers it was very unlikely that the answers of two questionnaires would have been identical. Not getting any explanations from the project partners, we reinterviewed a sub-sample of 10% of the original sample face-to-face. None of the reinterviewed households reported having been surveyed. After detecting the fabrication of the interviews, the partners acknowledged that all interviews had been fabricated. As a matter of course, we stopped working with all interviewers and partners and implemented a new local research group.

In February 2008, the survey was repeated in the same country. As mentioned before, we selected new villages and households according to the above-mentioned criteria. We hired nine students for the interviews and arranged the survey with on-site supervision. In most cases, the interviews took place in a school or the city hall so that we could monitor all interviewers. When the interviews took place in the houses of the surveyed families we attended some of them. Due to this procedure, we presume that the questionnaires from the 2008 survey are not fabricated.

In this paper, we use a total of 250 household interviews by 13 interviewers, of which four were falsifiers from the 2007 survey (the interviews submitted by one falsifier were excluded as he filled in only three questionnaires) who definitely faked the results, referred to as F1-F4, and nine interviewers who are supposed to be honest, labelled H1-H9. Table 1 provides an overview of the number of questionnaires per interviewer, which were included in the analysis.

Table 1
Number of questionnaires per interviewer

Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Number of questionnaires	10	12	10	10	22	23	23	24	23	23	23	23	24

3.2 Cluster analysis

In this subsection, we present the results of the cluster analysis. Based on the results, we evaluate the success of our procedure in identifying interviewers who fabricate data. As already mentioned, we use four indicator variables in the cluster analysis: the item-non-response ratio, the proportion of extreme ordinally scaled answers in all ordinally scaled answers referred to as extreme ratio, the proportion of answers where the others item including an alternative was selected in all answers which offered this item (referred to as others ratio) and the χ^2-value stemming from the comparison of the leading digit distribution in the questionnaires of an interviewer with the respective distribution in all other questionnaires.

Table 2 provides the values of the four indicator variables included in the cluster analysis for all 13 interviewers. It shows that the item-non-response ratio and the others ratio are clearly lower for the four falsifiers than for the honest interviewers. F1 and F4 have not chosen the others item at all. For the extreme ratio, things seem to be less clear. All the values range between 40% and 70% except the value of interviewer F1, which is clearly lower. The χ^2-values are quite high for falsifiers F2 and F4. The values of the other two falsifiers do not differ much from the ones observed for honest interviewers.

The general idea of cluster analysis is to identify subgroups of elements in a space of elements that are all characterized by multivariate measurements (see Härdle and Simar (2007) for an introduction to cluster analysis). In the first step, a measure to determine either distance or similarity between elements has to be chosen. In the second step, elements are assigned to different subgroups or clusters. Elements within one cluster should be similar according to the selected measure whereas elements in different clusters should be distant. There is a large variety of methods according to which elements can be assigned to clusters whereby the number of clusters might either be fixed or determined by the cluster method.

Table 2
Values of the variables included in the cluster analysis for each interviewer (all values except χ^2-value in percent)

Interviewer	Item-Non-Response	Others	Extreme	χ^2-value
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

We measured distance as squared Euclidian distance and employed several cluster procedures in order to check the robustness of the results. In all cases, the interviewers have been clustered in two groups with the intention to obtain one 'falsifier group' and one 'honest interviewer group.' The advantage of this approach is that a clear classification is obtained. In contrast, when one of the indicator variables is examined separately, it is not clear where to draw the line separating falsifiers and honest interviewers. Before conducting the cluster analysis, we standardized all variables on a mean of zero and on a variance of unity. This eliminates the scale effect as distances are measured in standard deviations and not in different units.

The first clustering method we use is hierarchical clustering. This is a standard procedure that can also be applied to larger datasets and is implemented in standard statistical software packages. Hierarchical clustering merges clusters step by step, combining the two closest clusters. At the beginning, every element is considered as a separate cluster. We measure distance between two clusters as the average squared Euclidian distance between all possible pairs of elements with the first element of the pair coming from one cluster and the second element from the other cluster. We used the software package STATA with the option 'average linkage' to conduct the hierarchical cluster analysis.

In hierarchical cluster analysis, two elements will stay in the same cluster once they are merged together. Thus, the procedure does not necessarily lead to a global optimum with regard to a given distance measure. In our case the relatively low number of interviewers allows us to conduct an alternative analysis by simply examining all possible cluster compositions and select the best one with regard to a certain target function. (The analysis was carried out in MATLAB, the programm code is available upon request.)

This procedure is clearly superior to hierarchical clustering as it ensures that the globally optimal cluster composition is identified. However, we also provide the results of hierarchical clustering as it is rather feasible compared to the computationally intensive approach of trying all possible compositions when the number of interviewers rises. Alternatively, one might resort to heuristic optimization techniques.

When examining all possible cluster compositions we use two target functions. The first one combines the ideas that a large distance between the two cluster centers is eligible as well as a small distance between the elements of a cluster and the cluster center. We look for the cluster composition, which maximizes the following expression:

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

The index i represents the four different indicator variables, $\bar{d}_{ai}$ with $a = 1, 2$ is the mean of variable i in cluster a, j symbolizes the different elements (interviewers) in cluster 1 and cluster 2, d_{ij} is the value of variable i for element j, and n_1 is the number of elements in cluster 1. Thus the numerator measures the distance between the two clusters, the denominator the distance within clusters and distance is measured in squared Euclidian form.

Alternatively, it could be interesting to see what optimal cluster composition results if instead of maximizing Equation (3) the average squared Euclidian distance between all possible pairs of elements within one cluster is minimized. In fact, this idea is very similar to the relevant target function in the hierarchical cluster procedures we presented before. Our second distance measure, which this time is to be minimized, is calculated as follows:

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

SED_{jk} is the squared Euclidian distance between elements j and k, calculated as $\mathrm{SED}_{jk} = \sum_{i=1}^{4}(d_{ij}-d_{ik})^2$. The numerator is the sum of distances between all possible pairs of elements in the same cluster. By dividing this sum by the number of possible pairs, one obtains the average within cluster distance.

Table 3 reveals the results of the three cluster procedures. In the hierarchical analysis with linkage between groups, the three falsifiers F1, F2 and F4 form cluster 1, falsifier F3 and all honest interviewers form cluster 2. Thus, we are able to separate both groups of interviewers, except one falsifier. However, without knowing from the outset which interviewers fabricated data and which were honest, one would have to decide which of the two clusters contains the at risk interviewers. This can be done by comparing the means of the indicator variables for each cluster displayed in Table 4. For the hierarchical procedure, means of the item-non-response ratio and the others ratio are clearly lower in cluster 1. The same is true for the mean of the extreme ratio, albeit the difference between the two clusters is less striking. Finally, a higher mean of the χ^2-value can be observed for cluster 1. Given these results, one would – according to the above mentioned hypotheses on the behaviour of falsifiers – correctly identify cluster 1 to be the cluster containing the at risk interviewers. We also tried to improve the results of the hierarchical clustering procedure using the cluster means displayed in Table 4 as starting point for the K-means analysis. However, the application of K-means clustering did not lead to any changes in the cluster composition.

Table 3
Results of the three employed clustering procedures

Hierarchical clustering													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between clusters divided by distance within clusters													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between elements in one cluster													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	1	1	2	2	2	2	2	2	2	2	1

Table 4
Indicator variable means by cluster for the three cluster compositions

	Item-Non-Response		Others		Extreme		χ^2-value	
Hierarchical clustering								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between clusters divided by distance within clusters								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between elements in one cluster								
Cluster	1	2	1	2	1	2	1	2
Mean	0.68	2.80	0.92	9.06	50.83	60.92	21.43	11.73

The cluster composition that maximizes Equation (3) is identical to the one obtained using hierarchical clustering. Consequently, as can be seen from Table 4, the indicator means within the two clusters are identical as well.

The cluster composition minimizing Equation (4) is slightly different. Cluster 1 now contains all falsifiers and one honest interviewer. The means of the indicator variables again clearly indicate cluster 1 to be the cluster containing the at risk interviewers. This is a very satisfying result. All falsifiers are identified and only one false alarm is produced.

However, it should be kept in mind that this does not mean that this particular cluster method works best when applied to another dataset.

To evaluate to what extent a higher number of indicators leads to better results, we repeated our cluster approach based on Equations 3 and 4 with all possible combinations of indicators, including cases that only rely on one indicator. The results (see Table 7 in the appendix) generally indicate that an increasing number of indicators improves the results. However, there are also combinations with a smaller number of indicators that lead to similar results compared to those based on all four indicators. Determining which indicator composition is the best would require the highly subjective fixation of the relative cost caused by non-identified falsifiers compared to the cost caused by a false alarm. But one can determine which indicator compositions are not Pareto dominated in the sense that there is no other composition that exhibits less non-identified falsifiers (false alarms) and at the same time not more false alarms (non-identified falsifiers). The indicator composition including all four indicators is the only one that is not Pareto dominated no matter which equation is used. In contrast, compositions including only one indicator are Pareto dominated in six out of eight cases.

3.3 Discriminant analysis

Finally, we turn to the discriminant analysis to check whether the hypotheses on falsifiers' behaviour our cluster analysis is based upon are valid. Discriminant analysis can be used if the clusters are known in order to assess how well the indicators in the analysis can separate the different groups and whether group membership can be predicted correctly (see Härdle and Simar (2007) for an introduction to discriminant analysis). In a linear discriminant analysis, the coefficients b_0 and b_i of the discriminant function $D = b_0 + \sum_{i=1}^{n} b_i x_i$ are determined in such a way that they maximize a function that increases with the difference of the mean D-values of the two different groups and at the same time decreases with the differences of the D-values of elements within the groups. In our case, the x_i are our four indicator variables and we obtain two groups by separating falsifiers and honest interviewers.

We use prior probabilities corresponding to the relative group size (4/13 and 9/13) in order to predict group membership. Table 5 shows the results. Obviously the four variables allow a good separation of the falsifiers and the honest interviewers, as the group membership is correctly predicted in all cases but one.

As can be seen from Table 5 negative values of the discriminant function are associated with the falsifier group. Consequently, Table 6 indicates that three of the four coefficients' signs are in line with the expected falsifier behaviour. Higher item-non-response and extreme ratios lead to a higher probability to observe an honest interviewer as does a lower χ^2-value. The estimated coefficient for the others ratio is negative. Thus an increase in the others ratio ceteris paribus raises the probability that an interviewer has fabricated data. This might appear as a contradiction to our above-mentioned hypotheses. One possible explanation might be that the effect of the others ratio is already captured by the item-non-response ratio. In fact, the correlation coefficient between the two variables is quite high with a value of 0.71. The Wilks' lambda of the discriminant analysis is statistically significant on the 5%-level.

Table 5
Results of the discriminant analysis by interviewer

Interviewer	Predicted group	Actual group	Discriminant function
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

Table 6
Standardized and non-standardized estimated coefficients (discriminant analysis)

Variable	Coefficient (non-standardized)	Coefficient (standardized)
Item-Non-Response	0.767	0.917
Others	-0.025	-0.129
Extreme	0.075	0.821
χ^2-value	-0.092	-0.562
Constant	-4.250	–
Wilks' lambda (Prob > F)	0.0254	

4. Conclusion

Survey data are potentially affected by interviewers who fabricate data. Data fabrication is a non-negligible problem as it can cause severe biases. Even a small amount of fraudulent data might seriously impair the results of further empirical analysis. We extend previous approaches to identify at risk interviewers by combining several indicators derived directly from the survey data by means of cluster analysis. To demonstrate our approach, we apply it to a small dataset which was partialy fabricated by falsifiers. The fact that we know the falsifiers from the outset allows us to evaluate the results of the cluster analysis and to furthermore conduct a discriminant analysis to reveal how well the two

groups of interviewers can be separated by the indicator variables. Different types of cluster analyses are conducted. All of them lead to the identification of an at risk interviewer cluster, with the item-non-response ratio and the others ratio being the clearest indicators. We are not able to identify falsifiers perfectly. However, in all cases the at risk interviewer contains a much higher share of falsifiers than the second cluster. The advantage of clustering is that one obtains a clear classification of interviewers who are at risk and the other interviewers, something that is not the case when indicators like the χ^2-value are examined separately. Furthermore, it allows us to combine the information of several indicators. By investigating the performance of all possible subsets of indicators we find that generally a larger number of indicators is more apt to identify falsifiers. The fact that different clustering methods lead to different results should not necessarily be considered a shortcoming of our approach. Depending on how one weights the costs of an undetected falsifier relative to a false alarm, one might finally assign only those interviewers to the potential falsifier group that always fall into the at risk cluster, no matter what clustering method is applied (which would imply high costs of false alarms), one might assign all interviewers to the potential falsifier group that fall into the at risk cluster at least once (which would imply high costs of undetected falsifiers) or choose a solution in between.

The application to a small dataset demonstrates another merit of our approach: it was tested and worked well in a situation in which the number of questionnaires per interviewer was quite limited (three of the falsifiers only submitted 10 questionnaires). If a small number of questionnaires per interviewer is sufficient to perform the analysis, one might also think about implementing it during the main field period when interviewers have only submitted a certain percentage of their questionnaires. Falsifiers could then be replaced by other interviewers who survey the units that should have been surveyed by the falsifiers.

Of course, when examining our results one has to keep in mind that we applied our method to a dataset in which a very severe form of data fabrication occurred: on the one hand we have falsifiers that faked all of their questionnaires (nearly) completely, on the other hand we have interviewers that (presumably) did all of their work honestly, which eases the discrimination between honest interviewers dishonest interviewers. Furthermore, with 13 interviewers, the size of our sample is quite limited. It would be interesting to explore the usefulness of our approach applied to larger datasets, given that the share of falsified interviews in large surveys has been found to be smaller than in our case. Additionally, larger datasets might allow the construction of additional indicators for the cluster analysis. If the survey has a reinterview program it would be possible to evaluate the usefulness of our approach by comparing the 'success' of a random reinterview with the success of a reinterview focusing on interviewers that were labeled as being at risk. We also intend to pursue the analysis in an experimental setting. An appropriate setting can ensure that one obtains a dataset which was partly collected by conducting real interviews and partly fabricated by telling some participants in the experiment to fill their questionnaires themselves.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft through the project 'SPP 1292: Survey Methodology' is gratefully acknowledged.

We furthermore thank John Bushery and four anonymous referees for providing useful comments on our paper.

Appendix

Table 7
Results of the cluster analyses based on Equations 3 and 4 for all possible cluster combinations

Indicators				Equation 3		Equation 4	
Item-Non-Response	Others	Extreme	χ^2-value	Undetected falsifiers	False Alarms	Undetected falsifiers	False Alarms
			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] Indicator composition not Pareto dominated.
[2] Mean cluster values did not allow for an identification of the 'at risk' cluster.

References

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

Biemer, P., and 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. and Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. In *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. Technical Report Manuskript 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. and Kennet, J. (2005). Interviewer falsification detection using data mining. In *Proceedings*: *Symposium 2005, Methodological Challenges for Future Information Needs*. Statistics Canada.

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

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

Härdle, W., and Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd Edition. 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., and Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 820-824.

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

Murad, U., and 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. and Kennet, J. (2004). A system for detecting interviewer falsification. Paper Presented at the 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., and English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. In *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. and 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., and 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. IZA Discussion Paper Series, 969.

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

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

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

Swanson, D., Cho, M. and Eltinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. In *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. and Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Prepublication Draft). Technical Papers on Health and Behavior Measurement. Washington DC: Research Triangle Institute. No. 53.