Bredl, Sebastian; Winker, Peter; Kötschau, Kerstin

Working Paper

A statistical approach to detect cheating interviewers

Discussion Paper, No. 39

Provided in Cooperation with:
Justus Liebig University Giessen, Center for international Development and Environmental Research (ZEU)

Suggested Citation: Bredl, Sebastian; Winker, Peter; Kötschau, Kerstin (2008) : A statistical approach to detect cheating interviewers, Discussion Paper, No. 39, Justus-Liebig-Universität Gießen, Zentrum für Internationale Entwicklungs- und Umweltforschung (ZEU), Giessen

This Version is available at:
https://hdl.handle.net/10419/39808

Zentrum für internationale Entwicklungs- und Umweltforschung der Justus-Liebig-Universität Gießen

A STATISTICAL APPROACH TO DETECT CHEATING INTERVIEWERS

by

Sebastian BREDL**
Peter WINKER*
Kerstin KÖTSCHAU**

Nr. 39

Gießen, December 2008

*)

FB Wirtschaftswissenschaften
Statistik u. Oekonometrie
Licher Strasse 64
D-35394 Giessen

**)

Zentrum f. internationale Entwicklungs- und Umweltforschung
Senckenbergstraße 3
D-35390 Giessen

Contents

Abstract

Survey data are potentially affected by cheating interviewers. 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 fabrication of interviews. As a novel tool in this context, cluster and discriminant analysis are used. Several indicators are combined to classify 'at risk' interviewers based solely on the collected data. An application to a dataset with known cases of cheating interviewers demonstrates that the methods are able to identify the cheating interviewers with a high probability. The multivariate classification is superior to the application of a single indicator such as Benford's law.

Keywords: Cheating interviewers; Benford's law; cluster analysis; data fabrication

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, but it can be affected as well by the interviewer, when he deviates from the prescribed interviewing procedure. If he does so consciously, this is referred to as interviewer falsification (Schreiner et al., 1988) or cheating (Schräpler and Wagner, 2003).

Interviewer cheating can occur in many ways. Rather subtle forms consist in surveying another household member than intended, or in conducting the survey per telephone when face-to-face interviews are required. The most severe form of cheating is the fabrication of entire interviews without ever contacting the respective household.[1] 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 cheaters 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 GSOEP data in a multivariate regression reduces the effect of training on log gross wages by approximately 80 per cent, although the share of fabricated interviews is less than 2.5 per cent. This indicates the importance of eliminating these interviews from the dataset.

The most common way to identify cheating interviewers is the reinterview (Biemer and Stokes, 1989). Here, a supervisor contacts some households which 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. So the question arises, how the reinterview sample can be optimized to best detect cheating interviewers. Generally, it seems useful to select households for the reinterview, which have been surveyed by an interviewer, who is - due to personal characteristics, or characteristics linked to the answers in his questionnaires - more likely than others to be cheating. In this context, Hood and Bushery (1997) use the term 'at risk' interviewer. In other cases like street surveys, in which the respondent's name is not recorded, the reinterview is not practicable at all. Hence, the identification of 'at risk' interviewers based on the above mentioned characteristics becomes even more important.

In our analysis we try to detect cheaters by a purely statistical approach relying on the data produced by the interviewers. This is not a new idea, literature provides several examples for this kind of approach (Hood and Bushery, 1997; Diekmann, 2002; Schräpler and Wagner, 2003; Swanson et al., 2003; Schäfer et al., 2005). However, the tests conducted in these studies rely on the examination of only one indicator derived from the interviewer's data to detect cheaters. We combine several of those indicators in cluster analyses, allowing

[1] The act of fabricating entire interviews is called 'curbstoning' by the US Bureau of the Census (Swanson et al., 2003)

for a better classification of the interviewers compared to previous approaches. To the best of our knowledge, this procedure is an innovation in the context of identifying cheating interviewers.

We have survey data available (see subsection 3.1 for a further description of our dataset) which was partly fabricated by cheaters. We know which data was collected by honest interviewers[2] and which data was fabricated. This knowledge allows us to evaluate our approach. However, this a priori knowledge is no prerequisite to employ the method.

The problem of identifying 'at risk' interviewers has already been addressed in the 1980s, however, literature on this issue is still scarce. In 1982 the US Bureau of the Census 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 cheat. 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 which have been labelled ineligible or the rate of households without telephone number[3] 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. For the case of computer assisted interviewing, Bushery et al. (1999) 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. Interviewers who complete an extremely high number of interviews during one day or spend very little time to complete the individual interviews are flagged as potential cheaters. Schäfer et al. (2005) assume that cheating interviewers avoid extreme answers when fabricating data. Using data of the German Socio Economic Panel (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, cheaters are known and it turns out that they could be found among the interviewers with the lowest overall variances.

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; Schäfer et al., 2005) which will be discussed in Section 2 along with its success to detect faked interviews in previous studies. Furthermore, Section 2 describes our statistical approach to identify cheating interviewers. 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] Of course one can never be absolutely sure if the assumed honest interviewers were really honest. However, given the circumstances in which these interviewers collected the data, makes cheating from their side extremely improbable.

[3] As reinterviews are often conducted by telephone, Hood and Bushery assume cheaters to be less likely to provide the telephone numbers in order to remain undetected.

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':[4]

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

Benford also provided distributions for the second, third and higher digits but this paper deals exclusively with the leading digit distribution for reasons discussed below.

However, not all series of numbers Benford 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. Hill (1995, 1999) postulates that random sampling from randomly selected distributions would create such data. The idea is that drawing from many different distributions leads to scale and base neutral numbers, which in turn implicates the applicability of Benford's law. Scott and Fasli (2001) find out that producing data by multiplying several random numbers results in leading digits distributed according to Benford's law. This procedure is the same as adding logarithms of random numbers. According to the central limit theorem this sum tends to a normal distribution so the data itself will be lognormally distributed. The authors conclude that data from a sufficiently positively skewed distribution whose modal value is not zero and whose values are positive can be expected to conform to the law. According to Nigrini (1996) Benford's law applies to numbers that describe similar phenomena like market values of listed enterprises or populations but have not been assigned like social security numbers.

The basic idea of using Benford's law to detect fabricated data is that cheaters are unlikely to know the law or to be able to fabricate data in line with it.[5] So a strong deviation of the leading digits in a dataset from Benford's distribution indicates that the data might be faked. Of course one has to be

[4] In fact, Benford was not the first one to describe the law: Simon Newcomb already mentioned it in 1881 (Newcomb, 1881).

[5] When Diekmann (2002) asked sociology students to invent regression coefficients in order to make them 'fit' to a given hypothesis, he found that the students were able to produce coefficients in line with the leading digit distribution (although the students were not familiar with it) but that the second digit distribution significantly deviated from the distribution derived by Benford. So Diekmann proposes not to focus on first digits to detect fraudulent data. However, other studies discussed below find that leading digits seem to be a good indicator to uncover such data in other contexts.

concerned if the nature of the data is such that it can be supposed to follow Benford's law if it is authentic.

The detection of financial fraud is a field in which the application of Benford's law has gained much popularity in recent years (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 are appropriate to be analyzed with 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.

Schräpler and Wagner (2003) and Schäfer et al. (2005) use Benford's law to detect cheating interviewers in the German Socio-Economic Panel Study (SOEP). In both studies all questionnaires delivered by every single interviewer are combined and it is checked 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:

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(h_{d_i} - h_{b_d})^2}{h_{b_d}} \tag{2}$$

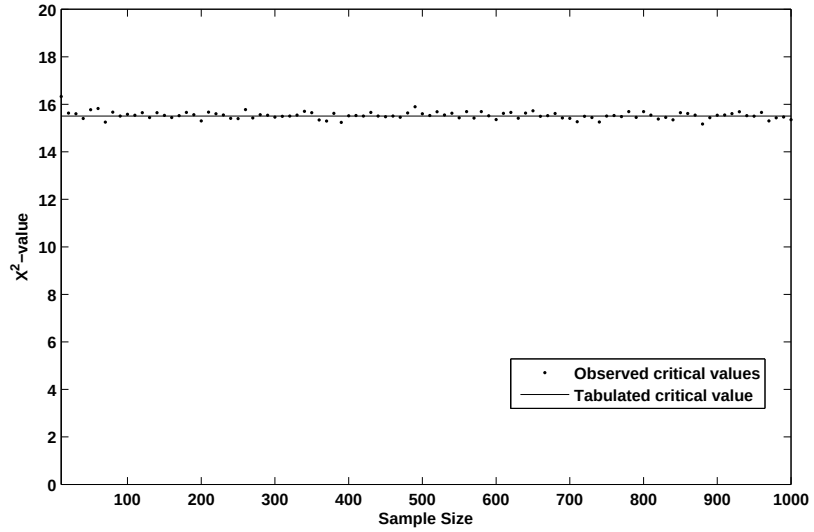
where n_i is the number of leading digits in all questionnaires from interviewer i, h_{d_i} is the observed proportion of leading digit d in all leading digits in interviewer i's questionnaires and h_{b_d} is the proportion of leading digit d in all leading digits under Benford's distribution. High χ^2-values indicate a deviation from Benford's distribution and thus indicate 'at risk' interviewers. Schräpler and Wagner (2003) use different kinds of continuous variables, Schäfer et al. (2005) restrict their analysis to monetary values. Both studies assume the critical χ^2-values to be dependent on the sample size n. Instead of using the χ^2-value to detect cheaters, Schräpler and Wagner (2003) construct a goodness of fit measure, which relates the observed χ^2-value to the highest possible χ^2-value, given the sample size. But this measure seems even more dependent on the sample size than the χ^2-value itself. Schäfer et al. (2005) use a bootstrap method to calculate the plausibility of obtaining a larger χ^2-value than the one observed, given a certain sample size.

Generally, the results obtained look promising. The goodness of fit measure in the study of Schräpler and Wagner is rather low for cheaters (which were already known in advance) compared to those of honest interviewers. Schäfer et al. find the plausibilities derived from the bootstrap for cheaters to be among the lowest of all interviewers. Thus it seems appropriate to use Benford's law as a means to identify 'at risk' interviewers.

To investigate whether there is indeed a relation between the χ^2-value and the sample size, we run a Monte Carlo simulation. In this simulation we draw samples from a Benford distribution with the sizes $n = 10, 20, 30, \ldots, 1\,000$[6] and calculate the resulting χ^2-values. For each sample size we conduct 10 000 repetitions. Thus we obtain for each sample size the critical χ^2-value for any

[6]This covers the range of first digit sample sizes for all our interviewers, which goes from 91 to 175.

Figure 1: Relation between sample size and simulated 95%-critical values



significance level by simply ordering the 10 000 values. Figure 1 plots the 95%-critical values against the sample size. Obviously there seems to be no relation. In fact, all the plotted values are close to the tabulated value for the 95%-percentile of a χ^2-distribution with $9 - 1 = 8$ degrees of freedom. Based on this finding, we consider the fit of the leading digit distribution to Benford's distribution to be independent from the sample size. Thus we assume that there is no need to further modify the χ^2-value.

We use the χ^2-value on a per interviewer basis as one element in our multivariate analyses. Concerning the selection of variables, we stick to the approach of Schäfer et al. (2005) and include only monetary values into the examination of leading digits. These values refer to household expenditures for different items like leasing or buying land, seeds, fertilizer or taxes and to household income from different sources like agricultural or non agricultural self employment and public or private transfers. Overall we include 26 different monetary values. The restriction to monetary values constitutes a clear criterion during the process of selecting data for the examination that ensures that all the data describe similar phenomena (in our case household income and expenditures) as proposed by Nigrini (1996). Furthermore, as mentioned above, financial data is broadly agreed upon to be apt for the analysis with Benford's law. The fact that we use different types of expenditure and income raises confidence that overall the monetary data might be scale and base neutral, a fact which, according to Hill (1995, 1999), implicates the applicability of Benford's law. Finally, a rough graphical appraisal of the different monetary variables reveals that most of them seem to follow a lognormal-like distribution, with the majority of values

being small compared to some outliers upwards. This is not at all surprising as income is typically assumed to be lognormally distributed. Following Scott and Fasli (2001), one can be confident that such data conforms to Benford's law.

Our examination is restricted to the leading digit. We have also experimented with the second digit, but the results were unusable. The reason is that many of the monetary values are rounded, which has led to an extremely high portion of zeros and fives among second digits. At this point it must be stated that it is not clear in which way the rounding of the second digits influences the leading digit distribution.

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 cheaters and non-cheaters. We do this by means of cluster and discriminant analysis.

The cluster analysis constitutes the real 'at risk interviewer identifying' approach. The interviewers are clustered in two groups with the intention of obtaining one group that contains the cheaters and another that contains the honest interviewers. This approach requires no a priori information on who is cheating and who is not. In fact this is what it is supposed to reveal. As we know from the outset which interviewer belongs to which group, we can discover whether the cluster analysis identifies the 'true' cheaters to be 'at risk'. In contrast, the discriminant analysis requires knowledge on the cheater respectively non cheater status of each interviewer before it can be conducted. We use the discriminant analysis to verify our hypotheses on the behaviour of cheaters, 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 Benford distribution as described in the previous subsection. Furthermore, we derive three other indicators from hypotheses concerning the behaviour of cheaters when fabricating data. Schäfer et al. (2005) assume that cheaters have a tendency to answer every question, thus producing less missing values. Furthermore, they expect cheaters to choose less extreme answers to ordinal questions. Hood and Bushery (1997) hypothesize that cheaters will "try to keep it simple and fabricate a minimum of falsified data" (Hood and Bushery, 1997, p. 820).

Based on these assumptions, we calculate three rations, which along with the χ^2-value, serve as indicator variables in the multivariate analyses. All ratios are like the χ^2-value on the interviewer-level. This means that we pool all questionnaires for every single interviewer.

- The 'non-response-ratio' is the proportion of questions which remain unanswered in all questions. We expect this ratio to be lower for cheaters 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 cheaters.

- The 'others-ratio' refers to questions which, besides several framed responses offer the item 'others' as a possible answer. The choice of this 'others'-item requires the explicit declaration of an alternative. If cheaters 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 item 'others' is selectable) should be lower for cheaters.

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 cheating interviewers, 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 cheaters try to fabricate a minimum of falsified data, Hood and Bushery (1997) expect cheaters 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.[7] 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 highest number of interviews conducted in one day might serve as variables.

3 Results

3.1 Data Sources

The data used in this study are derived from household surveys from November 2007 and February 2008 in one non-OECD country. The target group of the surveys were rural households in different villages which were selected by random sample. 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 income and expenditure categories.

The first survey in November 2007 resulted in faked interviews. Five interviewers, most of them well-known, filled in 50 questionnaires within one village.

[7] 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 which lead to new questions or avoid them depending on the answer, which are identical in both versions of the questionnaire.

Table 1: Number of questionnaires per interviewer.

Interviewer	C1	C2	C3	C4	H1	H2	H3
Number of questionnaires	10	12	10	10	22	23	23
Interviewer	H4	H5	H6	H7	H8	H9	
Number of questionnaires	24	23	23	23	23	24	

The results of one interviewer have been removed from this study as he filled in only 3 questionnaires, which is too little for statistical analysis. Two of the interviewers had been involved in developing the questionnaire as well as in selecting the villages and getting in contact with the administration before surveying. After we selected the target households, the interviews were conducted by the five interviewers without supervision. As most surveyed households do not own a telephone, no check calls were possible and scheduled. After receiving the filled-in questionnaires we became suspicious due to the neat and seemingly unused papers as well as due to some given answers. As no check calls were possible the targeted households were re-interviewed face to face and the fabrication of all interviews could be detected. As mentioned earlier, this method of re-interviewing is very reliable. However, generally it is not practicable due to high costs (Biemer and Stokes, 1989). The larger the sample group, the higher are the time and effort which are necessary to implement this method.

A second survey in different villages with different interviewers was conducted in February 2008. Besides some minor changes, the questionnaire remained the same. As before, the target households were selected by random sample based on household lists provided by the local administration. This time the survey was arranged with supervision on the spot. Nine interviewers conducted the interviews. No interview faking could be observed this time, so we presume that none of these questionnaires were faked.

In total we use 250 household interviews by 13 interviewers, thereof 4 cheaters who definitely faked the results, referred to as C1-C4, as well as 9 presumed honest interviewers who probably did not fake questionnaires, labeled H1-H9. Table 1 provides an overview of the number of questionnaires per interviewer, which were included in the analysis.

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 cheating interviewers. As already mentioned, we use four indicator variables in the cluster analysis: the 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 item 'others' including an alternative was selected in

Table 2: Values of the variables included in the cluster analysis for each interviewer (all values in per cent)

Interviewer	Non-response	Others	Extreme	χ^2-value
C1	1.36	0.00	28.33	15.40
C2	0.71	0.65	40.85	38.62
C3	0.68	2.33	56.90	15.49
C4	0.51	0.00	58.62	37.26
H1	3.85	18.01	65.12	28.86
H2	1.99	2.40	59.42	10.38
H3	3.10	9.47	70.07	17.89
H4	4.52	13.04	56.43	21.85
H5	1.18	4.48	70.07	19.04
H6	3.46	1.37	50.75	24.94
H7	2.51	12.72	45.65	14.15
H8	1.77	10.95	69.85	11.06
H9	0.14	1.61	69.44	6.69

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 all questionnaires of an interviewer with Benford's distribution.

Table 2 provides the values of the four indicator variables included in the cluster analysis for all 13 interviewers. The ratios are expressed in per cent. The table shows that the non-response ratio and the others ratio are clearly lower for the four cheaters than for the honest interviewers. C1 and C4 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 C1, which is clearly lower. The χ^2-values seem to be quite high for the group of cheaters but the same is true for the group of non-cheaters. The tabulated 95%-critical value of the χ^2-distribution with 8 degrees of freedom is 15.507 and is surpassed by two of the four cheaters (the other two stay only slightly below this threshold) but also by five out of the nine honest interviewers. This implies that an analysis restricted to the examination of the leading digit distribution would not have been very effective in our case.

Several different clustering procedures have been employed 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 'cheater' and one 'non-cheater 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 cheaters and non-cheaters. Furthermore, we have standardized all variables on a mean of zero and on a variance of unity.

The procedures employed include hierarchical clustering and the K-means

Table 3: Results of the hierarchical clustering with linkage between groups and squared Euclidian distances

Interviewer	C1	C2	C3	C4	H1	H2	H3
Cluster	1	1	2	1	2	2	2
Interviewer	H4	H5	H6	H7	H8	H9	
Cluster	2	2	2	2	2	2	

Table 4: Indicator variable means by cluster for cluster composition displayed in Table 3

Variable	Non-Response		Others		Extreme		χ^2-value	
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	30.42	17.04

approach. Hierarchical clustering merges clusters step by step, combining the two closest clusters. Consequently two elements will definitely stay in the same cluster once they are merged together. K-means clustering is an iterative process aiming at reducing the distance between the elements and the respective cluster center. Starting from arbitrarily defined cluster centers, each element is assigned to the cluster to whose center the distance is the shortest. Subsequently new centers are calculated and the process is repeated until the cluster composition does not change any more. Both procedures use - in the case of hierarchical clustering depending on the exact specification - a certain distance measure to determine the next step. However, the procedures do not necessarily lead to a global optimum which minimizes or maximizes this measure. In our case the relatively low number of interviewers allows us to try all possible cluster compositions and select the best one with regard to a certain distance measure. This approach is the most computationally intensive one and delivers exact results.

In our hierarchical cluster analyses the distance between two clusters is measured 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. Alternatively, it is measured as the average squared Euclidian distance between all possible pairs of elements in the two clusters, including pairs with both elements from the same cluster. The first procedure is referred to as linkage between groups, the latter as linkage within groups.

Table 3 reveals the result of the hierarchical analysis with linkage between groups. The three cheaters C1, C2 and C4 form cluster 1, cheater C3 and all honest interviewers form cluster 2. Thus, we are able to identify both groups of interviewers, except one cheater. However, without knowing from the outset which interviewers cheated and which were honest, one would have to decide which of the two clusters contains the 'at risk' interviewers. This can be done

Table 5: Results of the K-Means clustering

Interviewer	C1	C2	C3	C4	H1	H2	H3
Cluster	1	1	1	1	2	1	2
Interviewer	H4	H5	H6	H7	H8	H9	
Cluster	2	1	1	2	2	1	

Table 6: Indicator variable means by cluster for cluster composition displayed in Table 5

Variable	Non-Response		Others		Extreme		χ^2-value	
Cluster	1	2	1	2	1	2	1	2
Mean	1.25	3.15	1.61	12.84	54.30	61.42	20.98	18.76

by comparing the means of the indicator variables for each cluster displayed in Table 4. Means of the 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 cheaters - correctly identify cluster 1 to be the cluster containing the 'at risk' interviewers.

The use of linkage within groups leads to a slightly different result. Interviewer C1 changes the cluster, so one cluster contains the two cheaters C2 and C4, the other cluster contains the rest of the interviewers. This can be interpreted as a worsening of the result, as the separation between cheaters and non-cheaters becomes less clear-cut.

K-means clustering leads to the cluster composition shown in table 5. All cheating interviewers can be found in cluster 1, but four honest interviewers are assigned to this cluster as well. Table 6 indicates that one would correctly identify cluster 1 to be the 'at risk' interviewer cluster given the means of the variables: the means of all three ratios are lower, the mean for the χ^2-value is slightly higher compared to cluster 2. Thus, in our case K-means clustering leads to a relatively high number of 'false alarms'.

In our case, the fact that the number of interviewers is not large allows us to determine the best cluster composition given a certain distance measure by simply comparing all possible cluster compositions.[8] The advantage of this procedure is that it definitely leads to the composition minimizing or maximizing the measure and it is not necessary to select a specific clustering method. However, as the number of interviewers increases, this procedure becomes computationally too intensive. In this case heuristic optimization methods could be

[8] With 13 interviewers the number of possible compositions of cluster 1 and cluster 2 is simply $2^{13} = 8192$. As it does not matter if all elements from cluster 1 are put in cluster 2 and at the same time all elements from cluster 2 in cluster 1, there are in fact only $2^{13}/2 = 4096$ different compositions.

Table 7: Cluster composition that maximizes distance between clusters divided by distance within clusters

Interviewer	C1	C2	C3	C4	H1	H2	H3
Cluster	2	1	2	1	2	2	2
Interviewer	H4	H5	H6	H7	H8	H9	
Cluster	2	2	2	2	2	2	

Table 8: Indicator variable means by cluster for cluster composition displayed in Table 7

Variable	Non-Response		Others		Extreme		χ^2-value	
Cluster	1	2	1	2	1	2	1	2
Mean	0.61	2.23	0.33	6.94	49.74	58.37	37.94	16.89

an alternative.

The first distance measure we use in this context 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:

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

The index i represents the four different indicator variables, $\overline{d}_{ai}$ 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. The values of the four variables are again standardized. Thus the enumerator measures the distance between the two clusters, the denominator the distance within clusters and distance is measured in squared Euclidian form. The cluster composition that maximizes Equation (3) is shown in Table 7. The result is identical to the one obtained using linkage within groups as clustering method. The mean variable values correctly indicate that cluster 1 contains the 'at risk' interviewers, as can be seen from Table 8

Equation (3) takes into account the distances between the elements and the cluster centers, as well as the distance between the cluster centers. 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:

Table 9: Cluster composition that minimizes distance within clusters

Interviewer	C1	C2	C3	C4	H1	H2	H3
Cluster	1	1	1	1	2	2	2
Interviewer	H4	H5	H6	H7	H8	H9	
Cluster	2	2	2	2	2	1	

Table 10: Indicator variable means by cluster for cluster composition displayed in Table 9

Variable	Non-Response		Others		Extreme		χ^2-value	
Cluster	1	2	1	2	1	2	1	2
Mean	0.68	2.80	0.92	9.06	50.83	60.92	22.96	18.52

$$\frac{\sum_{j=1}^{n_1-1}\sum_{k=j+1}^{n_1} SED_{jk} + \sum_{j=n_1+1}^{13-1}\sum_{k=j+1}^{13-1} SED_{jk}}{(n_1(n_1-1))/2 + ((13-n_1)((13-n_1-1))/2} \tag{4}$$

SED_{jk} is the squared Euclidian distance between elements j and k, calculated as $SED_{jk} = \sum_{i=1}^{4}(d_{ij} - d_{ik})^2$. The enumerator is the sum of distances between all possible pairs of elements which are in the same cluster. By dividing this sum by the number of possible pairs one obtains the average within cluster distance. The cluster composition minimizing this function is displayed in Table 9. Cluster 1 contains all cheaters and one non-cheater and the means of the indicator variables again clearly indicate cluster 1 to be the cluster containing the 'at risk' interviewers as shown by Table 10. This is a very satisfying result, all cheaters are identified and only one 'false alarm' is produced.

3.3 Discriminant Analysis

Finally, we turn to the discriminant analysis to check whether the hypotheses on cheater behaviour our cluster analysis is based upon are valid. In a 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 cheaters and honest interviewers. In this case we have to rely on a priori information of the category.

Table 11 shows the results. Obviously the four variables allow a clear separation of the cheaters and the honest interviewers, as the group membership is correctly predicted in all cases. Given the fact that negative values of the discriminant function are associated with the cheater group, Table 12 indicates that three of the four coefficients' signs are in line with the expected cheater

Table 11: Results of the discriminant analysis by interviewer

Interviewer	Predicted group	Actual group	Discriminant function
C1	1	1	-2.501
C2	1	1	-4.135
C3	1	1	-0.824
C4	1	1	-2.744
H1	2	2	0.468
H2	2	2	1.604
H3	2	2	1.079
H4	2	2	2.309
H5	2	2	2.173
H6	2	2	0.409
H7	2	2	0.494
H8	2	2	0.052
H9	2	2	1.616

Table 12: Standardized and non-standardized estimated coefficients (discriminant analysis)

Variable	Coefficient (non-standardized)	Coefficient (standardized)
Non-Response	0.952	1.138
Others	-0.006	-0.030
Extreme	0.082	0.894
χ^2-value	-0.088	-0.809
Constant	-4.789	-

behaviour. Higher non-response and extreme rations 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 raises, ceteris paribus, the probability that an interviewer has cheated. This is a contradiction to our above mentioned hypotheses. One possible explanation might be that the effect of the others ratio is already captured by the non-response ratio. The correlation coefficient between both variables is quite high with a value of 0.71 and the value of the coefficient related to the other ratio is very close to zero.

4 Conclusion

Survey data are potentially affected by cheating interviewers. Interviewer cheating is a non-negligible problem as it can cause severe biases. Even a small num-

ber of fabricated interviews might seriously impair the results of further empirical analysis. Many of the existing methods dealing with the identification of fabricated interviews are derived from the survey design - like reinterviews - or are based on the application of one single indicator such as Benford's law.

We extend previous approaches by combining several indicators derived directly from the survey data by means of cluster analysis and discriminant analysis. The former serves as the 'at risk' interviewer identifying tool as it necessitates no a priori information on which interviewers cheated and which were honest. The latter requires this information and is used to validate our assumptions on interviewer cheating and to reveal how well the two groups of interviewers can be separated by the indicator variables employed in the cluster analysis. The four indicator variables we use are the non response ratio, the proportion of 'extreme' ordinally scaled answers, the proportion of answers where the item 'others' including an alternative was selected and the χ^2-value obtained by comparing the leading digit distribution of the data produced by an interviewer with Benford's distribution. We find the χ^2-value to be independent from sample size.

To check the success of our approach in identifying 'at risk' interviewers we apply it to a dataset which was partly fabricated by cheating interviewers. The fact that we know the cheaters from the outset allows us to evaluate the results of the cluster analysis and to conduct the discriminant analysis. Different types of cluster analyses are conducted. All of them lead to the identification of a 'cheater' cluster, with the non-response ratio and the others ratio being the clearest indicators. We are not able to identify cheaters perfectly. However, in all cases the 'cheater cluster' contains a much higher share of cheaters 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.

The discriminant analysis reveals that our indicator variables allow for a very precise separation of the cheater group and the honest interviewer group. Three coefficients are in line with the hypotheses on interviewer cheating. The coefficient associated with the others ratio has an unexpected sign, which is possibly caused by the high correlation between this indicator and the non-response ratio.

To further explore the usefulness of our approach it would be interesting to observe its performance when applied to larger datasets. Additionally, larger datasets might allow the construction of additional indicators for the cluster analysis. We also intend to pursue the analysis in an experiment setting.

References

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

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

Bushery, J., J. Reichert, K. Albright, and J. Rossiter (1999). Using date and time stamps to detect interviewer falsification. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 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.

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 M. Bushery (1997). Getting more bang from the reinterviewer buck: Identifying 'at risk' interviewers. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 820–824.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics 4*(1/4), 39–40.

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.

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., J. Schräpler, K. Müller, and G. Wagner (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 G. Wagner (2003). Identification, charcteristics 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., K. Pennie, and J. Newbrough (1988). Interviewer falsification in census bureau surveys. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 491–496.

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

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