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

Robustness of Clustering Methods for Identification of Potential Falsifications in Survey Data

by

NINA STORFINGER* and PETER WINKER**

No. 57

Gießen, August 2011

*Center for International Development and Environmental Research (ZEU),
Section 3
Justus-Liebig-Universität Gießen
Senckenbergstr.3
35390 Gießen
Email: Nina.Storfinger@zeu.uni-giessen.de

**Lehrstuhl für Statistik und Ökonometrie
Fachbereich Wirtschaftswissenschaften
Justus-Liebig-Universität Gießen
Licherstr. 64
35394 Gießen
Email: Peter.Winker@wirtschaft.uni-giessen.de

Robustness of Clustering Methods for Identification of Potential Falsifications in Survey Data*

Nina Storfinger and Peter Winker, Universität Giessen

August 30, 2011

Abstract

Falsifications of survey data might result in specific statistical properties of the generated data differing from those of the surveyed population. Clustering methods have been proposed to identify potential falsifications based on such indicators. As any statistical procedure, the classification might entail errors, i.e. misclassification of honest interviewers as potential falsifiers and failing to identify all falsifications as such.

Typically, the robustness of a statistical classification procedure is studied using a large number of problem instances with known allocation to the groups. However, given the sensitivity of falsifications in survey data, the access to datasets comprising correctly identified falsifications is very limited. Consequently, a bootstrap based approach is introduced and applied to assess the clustering method. This approach also allows modifying settings such as number of interviews per interviewer or share of falsifications in the dataset and to study the impact of these settings on the quality of the assignments. Results based on a small real dataset with identified falsifications are reported.

Keywords: Interviewer falsifications, cluster analysis, bootstrap method.

1 Introduction

Data quality in face-to-face interviews depends crucially on interviewers' behavior. In particular, intentional deviation from the prescribed procedures might affect data quality. The most extreme case of such misbehavior is given if interviewers falsify parts of or complete interviews. In addition to standard, but expensive procedures such as reinterviews,[1] Bredl *et al.* (2008) developed a clustering method for ex post

*We are grateful to S. Bredl for valuable comments on earlier drafts of this paper. Financial support through the DFG in project WI 2024/2-1 within SPP 1292 is gratefully acknowledged.

[1]See Forsman and Schreiner (1991, pp. 293ff) for the use of reinterviews to detect interviewer falsifications.

identification of falsifications in survey data, which is based solely on the collected data. This procedure exploits differences in specific statistical features of the data actually sampled from the population under review versus those of the data falsified by the interviewers.

While the occasional or even frequent presence of such falsifications in survey data has been reported in the literature,[2] access to data including identified falsifications is very limited.[3] Thus, the standard procedure for evaluating the proposed method, i.e., its application to a number of different publicly available datasets with different properties is precluded. Future research in cooperation with institutions running large scale surveys might result in datasets suitable for evaluation purposes. However, for the moment being, a statistical analysis has to be based on the few datasets available including identified falsifications.

Nevertheless, proposing a statistical method, such as the clustering procedure proposed by Bredl *et al.* (2008), for a highly sensitive issue such as interviewer falsification asks for an analysis of its properties prior to application. In particular, the user will be faced with two types of potential errors. First, honest interviewers might erroneously assigned to the cluster corresponding to the characteristics of falsified interviews. This misclassification might be called "false alarm". Second, an interviewer actually producing falsified data might remain undiscovered when the characteristics of the falsified data do not differ strong enough from those of honest interviewers. It will depend on the implied action which of the two types of errors might be considered worse. If the cluster containing those interviewers considered as potential falsifiers is used to concentrate follow-up calls or interviewers to this group, a "false alarm" will only reduce the efficiency of these follow-ups, while undiscovered falsifications might strongly affect the findings of further empirical analysis using the dataset.[4] However, if the results of the cluster analysis are used to decide upon payments to the interviewers or even further legal action, obviously, the rate of "false alarms" would have to be basically zero. In this paper, we will rather assume the first setting, where some "false alarms" might be accepted if this helps to reduce the share of undiscovered falsifications.

To analyse the properties of the clustering method based on a single real dataset with identified falsifications, a bootstrap method is proposed.[5] This method, de-

[2]See, e.g., Crespi (1945), Schreiner *et al.* (1988), Koch (1995), Bushery *et al.* (1999), Harrison and Krauss (2002), Diekmann (2002), and Schräpler and Wagner (2005). Bredl *et al.* (2011) provide a recent literature review on the topic.

[3]Obviously, if falsifications are identified, these are removed prior to making the dataset accessible for further analysis, e.g. in the case of the SOEP (Schäfer *et al.* 2005). In fact, the incentives to publish at least information on identified falsifications are almost not existing. Thus, most information on such cases in real surveys are of anecdotal nature.

[4]See Schräpler and Wagner (2005) for an example.

[5]For an introduction to the principles of the bootstrap method see, e.g., Efron (1982) and Chernick (2008). For early applications in the context of cluster analysis see Jain and Moreau (1987) and Peck *et al.* (1989).

scribed in more detail in Section 3 allows to generate many synthetic datasets reflecting the features of the original data and, consequently, to obtain estimates of the distribution of the results including the share of errors of the two types described above. Furthermore, it is possible to generate synthetic datasets which differ in some properties from the original data, which might affect the quality of the clustering method, e.g., the sample size or the share of falsifications in the data.

The rest of this contribution is organized as follows. In Section 2 we briefly report the clustering method for identification of falsified interviews. Furthermore, the dataset used is introduced. Section 3 provides some information on the bootstrap and its implementation for the clustering problem. It also describes the specific properties of the synthetic datasets we generate. The results of the bootstrap analysis are reported in Section 4. The conclusion from our analysis and hints for future research in the field are given in Section 5.

2 The Clustering Procedure

2.1 Idea

A central aspect of the proposed method consists in abstracting from the specific content of the questionnaires and individual interviewer characteristics (Koch 1995), and instead concentrating on particular traits or response patterns differing between real and falsified interviews. Specific indicators are selected to reflect such properties. As in Bredl *et al.* (2008), we do not consider metadata like length and date of the interview which might also provide valuable information on interviewer behavior, but which is not always available (Hood and Bushery 1997).

Given that a single indicator might not be sufficient to discriminate well enough the group of honest and suspect interviewers, following the original proposal by Bredl *et al.* (2008), a multivariate analysis is conducted. To this end, for each interviewer a set of indicators based on all data collected by this interviewer is calculated. Assuming that the indicators are chosen in a way such that the distributions of values for honest and suspect interviewers differ to some extent, a cluster analysis might be used to identify the two subgroups. However, as usual with statistical analysis, it might not be expected that this grouping turns out to be perfect and, consequently, some "false alarms" and/or undiscovered falsifications might result.

Consequently, when thinking about the practice of organizing surveys, a method like the one analyzed here might be used as a first step to initiate further checks on the interviewers assigned to the group of potential cheaters. Using, e.g., reinterviews (Schreiner *et al.* 1988) or postcard follow-ups (Hauck 1969) it will be clarified if the potential cheaters are real cheaters or only misclassified honest interviewers. To hold the costs for these further examinations at a reasonable level, the share of misclassified honest interviewers, assigned to the falsifier cluster, should be as small

as possible.

2.2 Indicators

Since the motivation for specific indicators and their measurement is described in detail elsewhere (Schnell 1991, Koch 1995, Hood and Bushery 1997, Bredl *et al.* 2008), we only briefly review the instruments used for our application.

First, we suppose that falsifiers operate too accurately, i.e. they tend to exhibit a lower rate of unanswered questions than accurate interviewers would do. Thus, the first indicator considered is the share of unanswered questions for all interviews conducted by a single interviewer. This indicator is called the `non-response-ratio`.

Second, semi-open questions often include the option "others" combined with an additional text field explaining this answer. The indicator `others` is defined as the share of these questions for which the interviewer selects the option "others". Given that filling in some specific content requires more effort, it is expected that falsifiers use this option less frequently.

The third indicator `extreme-answer-ratio` measures the share of all questions with ordinal answer categories, for which one of the extreme values is selected. It is expected that the falsifiers underestimate the frequency of such extreme answers. Consequently, the ratio should be lower for falsifiers as compared to honest interviewers.

Finally, an indicator related to metric data is used. It makes use of the observation known as Benford's Law (Benford 1938) that the distribution of the first digits of many metric variables – including monetary measurements – can be well approximated by a specific distribution. Given that it appears to be difficult to reproduce this distribution when falsifying questionnaires, it is assumed that the difference between the empirical distribution of first digits and the theoretical distribution (measured by a χ^2-type statistic labeled `Benford` in the following) is larger for falsified interviews. As an alternative to this standard setting, we also consider the difference in the distribution for one interviewer as compared to the distribution of all other interviewers. The corresponding χ^2-statistic will be labeled `Benford_alt` in the following.

Of course, alternative or additional indicators are conceivable for this type of analysis and are studied in complementary research (Storfinger and Opper 2011). However, the general issue of evaluating the performance of the method remains the same independent of the actually selected indicators. Thus, we concentrate here on only the same four indicators as Bredl *et al.* (2008).

2.3 Clustering

The next step of the procedure consists in clustering the indicator sets obtained for all interviewers. Here, we consider the clustering in two groups only which, in

an ideal situation, would correspond to the honest and cheating interviewers, respectively. However, future applications might also consider more than two clusters corresponding to different types of deviant behaviour by subgroups of interviewers. In fact, the number of groups might also be derived endogenously, i.e. data based using methods suggested for this purpose.

Whatever the number of clusters considered, the central idea consists in grouping interviewers such that the interviewers' indicators within a group are similar, while they differ for members of different groups. Once a grouping is obtained, the assumptions on interviewers' behavior discussed above allow to indicate which of the two clusters should correspond to the honest interviewers. While this assignment could be done manually for a small number of problem instances, the huge number of cases generated in the bootstrap procedure described below requires an automatic method.

One way to label the cluster is by taking into account the a priori assumptions about the indicator values described before. For example, the cluster showing the lower share of missing values would be considered to be more likely the cluster containing most falsifications. Consequently, for each cluster the number of indicators pointing in the direction of falsifications could be calculated. Then, in case of an uneven number of indicators, the cluster with the higher number of such signals would be assumed to be the cluster comprising the falsifications. Alternatively, one might take into account the actual standardized indicator values. For this approach, all indicators have to be defined such that smaller values should point towards potential falsifications. Then, simply summing up and assigning the label "potential falsifications" to the cluster with the smaller value seems appropriate. This method is also appropriate for an even number of indicators. It is the approach followed in our application.

To perform the actual clustering, many methods have been proposed in the literature including, e.g., k-means, or hierarchical methods such as Ward's (1963) approach, which are iterative processes aiming at reducing the distance between the elements and the respective cluster center. Alternatively, one might think about enumerating all possible assignments to two groups and select the one corresponding to the optimal value of a given objective function (Bredl *et al.* 2008). This is feasible for the original dataset given the small number of interviewers. For larger cases, some heuristic optimization approach might be implemented to approximate such a globally optimum (Winker 2001).

For the application presented here, we use both a hierarchical method, namely Ward's method and for the problem instances, for which the number of interviewers does not change as compared to the original data, also the full enumeration approach. In Ward's method the criterion for merging two clusters at any given step is the variance within the clusters. Consequently, out of all pairs of existing clusters the pair resulting in the slowest increase of the sum of in cluster variances will be merged.

The full enumeration approach uses a slightly different objective, namely the sum of pairwise distances between all elements within a cluster, which also measures the within group heterogeneity. The major difference between both methods is the sequential approach followed in Ward's method, while the full enumeration algorithm delivers the global optimum for the given objective function and number of clusters.

To evaluate the procedure, we consider the share of correctly identified interviewers. Alternatively, one might consider only the share of correctly identified cheaters, i.e. to what extent no potential falsification remains unnoticed, or the share of correctly identified honest interviewers, i.e. to what extent "false alarms" have been avoided.

2.4 The Dataset

The data used for the empirical application are from a survey of rural households in small villages in a non-OECD country.[6] The questionnaire included many metric variables as well as scale and (semi-) open questions. It was found out that five interviewers operating without supervision faked all their interviews, while the other nine interviewers conducted the interviews properly under supervision. Since one of the falsifiers produced only a very small number of interviews, this observation is left out for the further analysis. The dataset consists in total of 250 interviews.

Applying the clustering methods described before to this dataset results in a 100% assignment of all falsifiers to the cluster labeled "potential falsifications" both for Ward's method and the optimal clustering.[7] While for Ward's method, the share of "false alarms" amounts to four out of nine honest interviewers, this share is reduced to one out of nine for the optimal clustering method using the alternative indicator `Benford_alt` for the first digits.

3 Bootstrap Method

The clustering approach described in the previous section, seems to work quite well for the few problem instances we could analyze given data availability. However, the small number of these test cases does not allow to draw general conclusions on the method regarding, e.g., its expected performance over a large set of problem instances or the dependence of this performance on specific properties of the dataset. In particular, we are interested in four dimensions, the number of interviewers considered, the share of cheating interviewers, the number of interviews per interviewer and the number of questions of a specific type per interview.

[6] For more details, see Bredl *et al.* (2008).

[7] Using k-means instead of Ward's method, Bredl *et al.* (2008) report to miss one falsifier, while with the optimal clustering approach, also all falsifiers are correctly identified.

Our original dataset including nine honest interviewers and four identified cheaters and about 10 to 20 interviews per each interviewer. Thus, considering (random) subsets of this dataset would provide only a rather limited experimental setting to analyze the four dimensions mentioned above. In particular, it would not be possible to obtain some distributional information on specific performance indicators.

Consequently, we resort to a bootstrap method for constructing artificial datasets.[8] This way we are able to create synthetic data with specific settings for the number of interviewers, share of falsifiers, number of interviews per interviewer etc. Furthermore, the bootstrap method allows constructing a large number of problem instances for each setting. Thus, we are able to report distributional information on specific performance indicators. Finally, this allows to derive some conclusions regarding the dependence of the performance of the cluster method on the type of available data.[9]

Bootstrap as the related Jackknife approach belongs to the class of resampling methods. In order to obtain some statistical information on the properties of sample functions such as estimators, test statistics or – in our case – performance of clustering, a large number of problem instances is required unless an analytical method is available for deriving distributional results. If these instances cannot be empirically observed, resampling methods construct synthetic data by merging draws from the existing data. The methods differ in how the drawings are generated.

We use the simplest version of a bootstrap method. Let us assume that a synthetic set comprising data for n interviewers should be generated. Then, from the original dataset n interviewers are drawn at random with replacement, whereby every interviewer has the same probability to be drawn. If the challenge is to construct a set comprising n_1 honest and n_2 cheating interviewers, the same idea is applied first to the subset of honest interviewers to resample n_1 honest interviewers and next to the subset of cheating interviewers to resample n_2 cheating interviewers.

Once a new synthetic dataset with specific properties is built, we calculate the indicators for each interviewer and conduct the cluster analysis explained in Section 2. Then, we record, e.g., the share of correctly assigned interviewers or the share of the two types of potential misclassification. The process is continued generating a new synthetic dataset, running the cluster analysis on it and recording the results. This procedure is repeated many times. Let us denote the number of these bootstrap drawings by B. Then, typical values of B are at least 1 000 or larger. Eventually, i.e., after having computed the indicator(s) of interest for all B bootstrap samples, we can report statistical information on the distribution of these indicators such as mean and variance, but also in form of empirical distribution functions, e.g., as

[8]The label "bootstrap" has been introduced by Efron (1978) in a paper contrasting it with other resampling methods. Jain and Moreau (1987), Peck *et al.* (1989), and Rost (1995) use a bootstrap procedure for estimating the number of different clusters in a given dataset.

[9]This type of distributional information might also be used for a given dataset for inference on the number of clusters to be considered (Chernick 2008, pp. 145ff).

histogram in graphical form. Algorithm 1 summarizes the steps of the bootstrap procedure.

Algorithm 1 Pseudo-code for bootstrap procedure for cluster analysis.

```
for b = 1 to B do
    Generate synthetic data of n_1 honest and n_2 cheating interviewers
    Calculate indicators for all interviewers
    Perform clustering analysis
    Label cluster of falsifications
    Store performance of procedure for given data
end for
Summarize distribution of performance over B sets of data
```

As mentioned above, we use the bootstrap technique to analyze the effect of different modifications of the sample layout on the quality of the results. We consider four dimensions. We start with varying the total number of interviewers in the dataset and the share of honest interviewers, respectively. By doing so, we are able to find out whether the number of interviewers, given a specific share of falsifiers in the dataset, affects the quality of the results from the clustering procedure. We expect that a larger number of interviewers will make it more difficult to keep the rate of misclassification of both types low. Furthermore, we can assess whether the share of falsifiers, given a specific total number of interviewers in the data, has an impact on the outcomes. In this context we are particularly interested whether the method still performs well if the share of falsifiers becomes low given that anecdotal evidence suggests that actual prevalence of falsifications might be of the order of 5% to 10% rather than the 30% found in our original dataset. A third setting considers the number interviews per interviewer. To this end, we resample for every interviewer the specific number of interviews out of his original interviews. Since the performance of the clustering procedure depends on good estimates of the indicators, it is expected that the classification errors decline when the number of interviews per interviewer increases. Finally, the precision of the measurements for the four indicators considered might show a different sensitivity with regard to the number of observations. Thus, we also consider resampling at the level of single questions in the questionnaire. In particular, we resample from specific question types, for example scale questions, to find out how many questions of a specific type are required to render the corresponding indicator useful for the clustering procedure. Table 1 shows the type of questions considered for this step and the number of available questions per indicator in the original questionnaire.

Table 1: Type and number of questions used for specific indicators

Indicator	Number of questions/variables
Non response	59
Category "others"	13
Extreme answers	6
First digit	25

4 Results

4.1 The Design

When running the bootstrap experiments, a lot of parameters might be modified. In order to single out the effects of particular parameters, we decided to keep all parameter settings equal or close to the values for the original data, except for the specific modifications shown in Table 2. In particular, unless modified according to Table 2, the following values are used: A total of 13 interviewers, including 4 falsifiers and 9 honest ones, 59 questions per interview allowing for a non response, 13 questions with the semi-open category "others", 6 ordinal scales offering the option for extreme answers and 25 questions resulting in metric data, for which the distribution of first digits can be analyzed.

Table 2: Experimental Design for Bootstrap Runs

Dimension	Original Sample	Values for Bootstrap				Remarks
Number of interviewers/falsifiers	13/4	10/3	20/6	50/15	100/30	$\sim$ 30% falsifiers
		10/1	20/1	50/2	100/5	$\sim$ 5% falsifiers
Fraction of falsifiers among interviewers	4/13	1/13	2/13	6/13	10/13	13 interviewers
		4/40	8/40	20/40	36/40	40 interviewers
		1/100	5/100	50/100	95/100	100 interviewers
No. of interviews per interviewer	19.2 (Avg.)	5	20	50	100	4 fals./13 interv.
		5	20	50	100	1 fals./13 interv.
No. of questions						
– non response	59	0 – 100				
– category "others"	13	0 – 100				
– extreme answers	6	0 – 100				
– first digit dist.	25	0 – 100				

4.2 Number of Interviewers

We start with the analysis of the first experiment, i.e., the variation of the number of interviewers in the dataset. Figures 1 and 2 summarize the findings. For both examples, we consider 10, 20, 50, and 100 interviewer, but the share of falsifiers is about 30% for the first group of experiments as in our real dataset, while for the second group of experiments a much lower share, namely around 5% is assumed. The latter number corresponds closer to the anecdotal reports about identified falsifications in survey data. Thus, it has been our interest to check how robust the proposed methodology is when the share of falsifiers might be low and – possibly – at the same time the number of interviewers involved might be high. For this experiment, the number of interviews per interviewer is always the same as in the original dataset (in mean 19.2 interviews per interviewer).

Due to space restrictions, we only report the results obtained using `Benford_alt`. Furthermore, as for more than 20 interviewers, the optimal clustering by full enumeration is not feasible anymore, for this group of experiments, only Ward's method is employed.[10] The figures show histograms generated from the bootstrap distribution based on 5 000 replications. The left subplots provide the frequencies for the percentage of correctly identified falsifiers, while the right subplots show the corresponding findings for the correctly assigned honest interviewers.

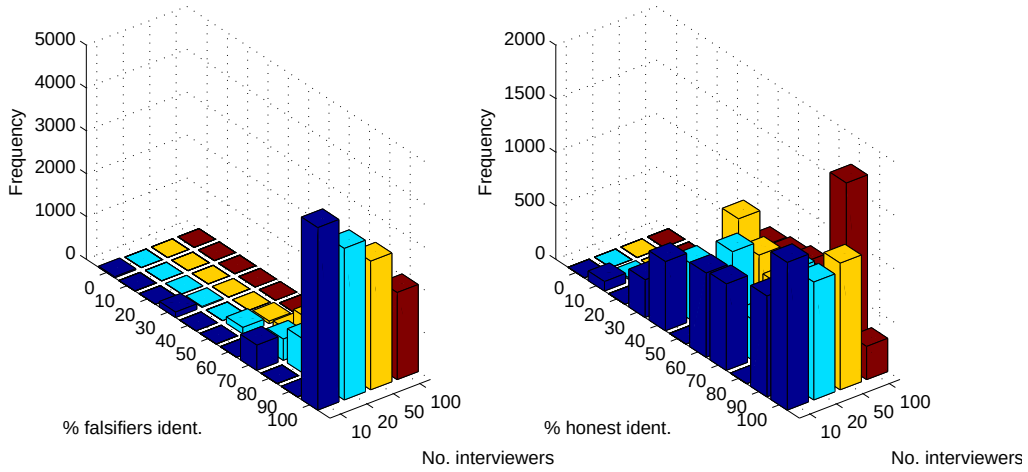

Figure 1: Performance of clustering method as function of number of interviewers for a substantial share of falsifications.

Starting with Figure 1 we see that the probability of correctly identifying all falsifiers is high when the number of interviewers – and consequently the number of

[10] As results for the last two experiments indicate some advantages of the optimal clustering method, future research will be directed to use heuristic optimization tools to obtain good approximations to the optimal clustering also for larger problem instances.

falsifiers – is small, but tends to decrease with a growing number of interviewers. However, also for the largest problem instance with 100 interviewers including 30 falsifiers, in about 40% of the bootstrap replications, still 100% of the falsifiers are correctly identified, while for more than 70% of the bootstrap replications this share is above 80% and never falls short of 60%. Similarly, for the honest interviewers, the share of false alarms tends to slightly increase when more interviewers are included. In fact, for the largest problem instance with 100 interviewers, no false alarms are generated only in about 6% of the cases, but for 40% of the cases the share of false alarms turns out to be smaller than 10%.

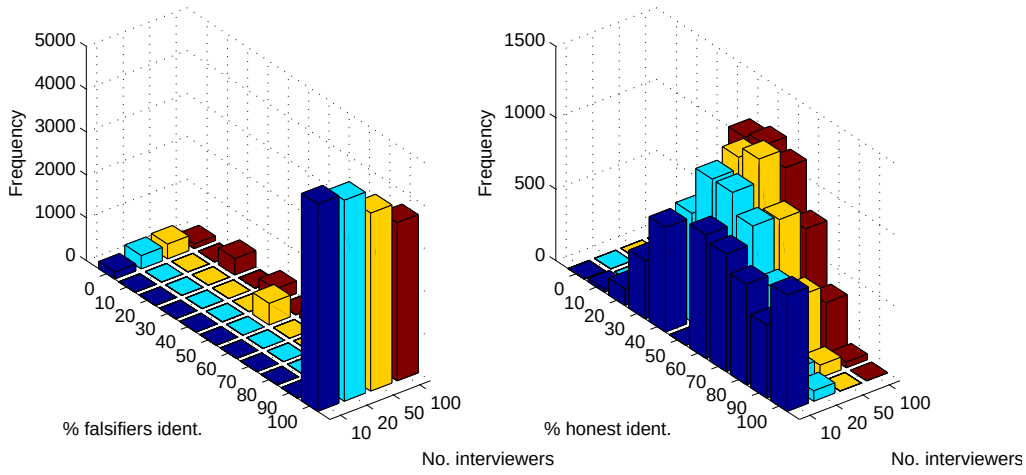

Figure 2: Performance of clustering method as function of number of interviewers for a low share of falsifications.

The findings are similar with regard to the identified falsifiers in Figure 2 when the share of falsifiers is pretty low. Still, even when increasing the number of interviewers, which should make it more difficult to spot falsifiers, the shares of correctly identified falsifiers are often close to 100%. The results, however, are much less impressive for the right hand part of the figure, indicating that the low rate of missed falsifiers comes as the cost of a high rate of false alarms, in particular, when the number of interviewers grows.

4.3 Share of Falsifications

In the previous subsection, we already considered two experiments with different shares of falsified cases, but the focus there was on the influence of the total amount of interviewers on the detection frequencies. In this step, we examine whether the clustering method still performs satisfying when varying the share of falsifiers in the dataset, ranging from a very small number ($\sim 1-8\%$) to a very high number ($\sim$ 77 – 95%). As Table 2 shows, the first setting corresponds to the original

number of interviewers (13), which will be increased to 40 and 100 in the remaining two settings. As for the first experiment, the number of interviews per interviewer is always set to the same value as in the original dataset, i.e. with a mean of 19.2 interviews per interviewer.

Given that the optimal clustering could only be applied to the case with 13 interviewers, we restrict ourselves to using Ward's method for clustering in this experiment. Furthermore, based on the previous findings, only results for the alternative Benford indicator are presented.[11]

Figure 3 reports the results for the experiments with 13 interviewers. In the left plot, histograms for the percentage of correctly assigned falsifiers are shown for the different shares of falsifiers ($1/13 \sim 8\%$ to $10/13 \sim 77\%$). The right plot provides the corresponding results for the correctly identified honest interviewers. Overall, it can be summarized that the percentage of correctly assigned falsifiers is highest for very small shares of falsifiers. Although this share of correctly assigned falsifiers decreases with an increasing share of falsifications up to about 50%, it still remains substantial. For even higher shares of falsifications, the frequency of identifying all of them correctly increases again. In this case, also the risk of finding only very few falsifiers increases as the whole cluster might be wrongly assigned in some cases. However, the case of extremely high shares of falsifiers might be considered less relevant for most real applications.

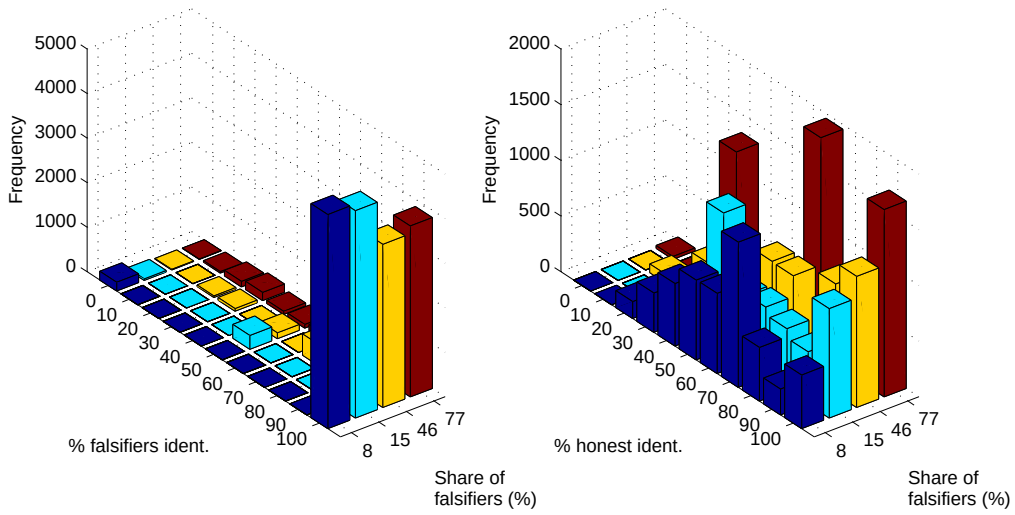

Figure 3: Performance of clustering method as function of share of falsifiers (13 interviewers).

This trend with regard to the share of correctly assigned falsifiers is reflected

[11]The non reported results when using the original Benford indicator are overall only slightly inferior.

by the share of correctly assigned honest interviewers, which tends to increase with higher shares of falsifications. The right part of Figure 3 shows that only 10% of the bootstrap replications identify 100% of the honest interviewers if the share of falsifiers is just 8% (1 out of 13), while in 35% of the replications all three honest interviewers are correctly assigned if the dataset includes ten falsifiers (77%).

As the qualitative findings for the other two sets of experiments with 40 and 100 interviewers are qualitatively similar, we do not report the figures to save space.[12] In both settings the probability to detect all or most of the falsifiers decreases with an increasing share of falsifiers up to 50% and increases again for even higher shares of falsifiers. Corresponding to the results of the previous section we also recognize that the percentage of correctly assigned falsifiers decreases while increasing the number of interviewers (40 and 100 interviewers) throughout all specific shares of falsifiers. For example, in the setting with 100 interviewers and 95% falsifiers only in 50% of the bootstrap replications all falsifiers could be identified. However, in virtually all cases at least 70% of the falsifiers are still identified in this most demanding setting.

The qualitative results are also similar with regard to the frequency of detecting the honest interviewers, i.e., avoiding false alarms. We only find that these frequencies tend to be lower for all shares of falsifiers considered compared to the situation with 13 interviewers, which reflects the higher complexity of the task to cluster 40 or 100 interviewers correctly as compared to just 13. Consequently, e.g., in the setting with 100 interviewers the identification of all honest interviewers succeeds never for small shares of falsifiers, i.e., below 50%. However, even in these most difficult cases, typically far more than 50% of the honest interviewers are correctly assigned.

Overall, it can be summarized that the falsifiers are most often correctly assigned if their share is low. Although this probability decreases with an increasing share of falsifications up to 50%,[13] it still remains substantial. On the other hand, it becomes obvious, that there will be a substantial amount of false alarms in particular if the actual share of falsifiers is low.

4.4 Number of Interviews per Interviewer

In our third experiment, we modify the number of interviews per interviewer. For this setting we resample for each interviewer 5, 20, 50, and 100 interviews out of all conducted interviews. Additionally, we vary the share of falsifiers (about 30% and 5%) to consider the influence of the share of falsifiers on the performance of the clustering method. Given that we consider only the same number of interviewers as in the original dataset (13), the optimal clustering method can be applied, which

[12] These graphs as well as the versions for the standard Benford indicator are available on request from the authors.

[13] A similar pattern of decreasing ability to detect unusual cases is reported by Karabatsos (2003) in the context of aberrant response detection.

results in somewhat better results.[14] We also limit the reports to the results obtained including the `Benford_alt` indicator for the metric variables.

Figure 4 shows the findings for the first group of experiments using a high share of falsifiers. The boxplots in the left part of the figure display the share of the correctly assigned falsifiers, while the boxplots in the right part show the corresponding results for the correctly identified honest interviewers. It can be seen that the share of correctly assigned falsifiers varies between 50% and 100%, whereat the lower quartile is already at 75% and the median at 100%, i.e., in more than 50% of the cases, all falsifiers are correctly identified. This finding holds already for a very small number of interviews and does not change when increasing the number of interviews per interviewer. It might be concluded that the number of interviews has no strong effect on the frequency of correctly assigning the falsifiers in this experimental setting.

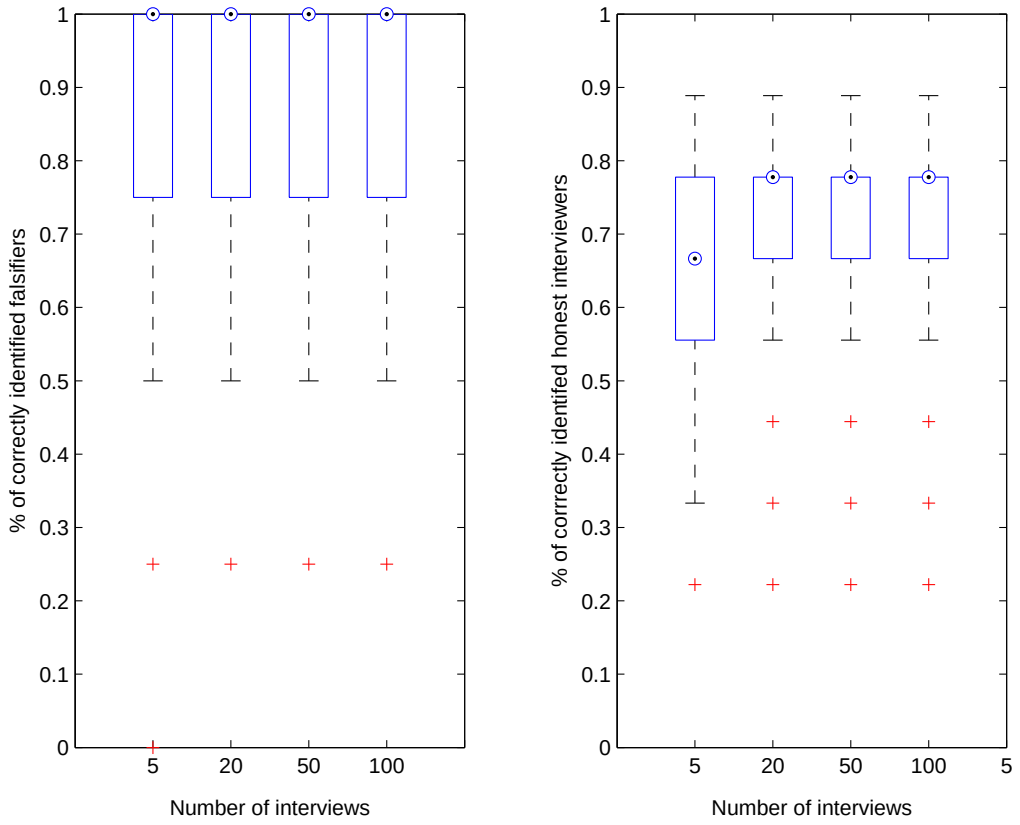

Figure 4: Performance of clustering method as function of number of interviews per interviewer (4 falsifiers).

A similar result is obtained for the correctly identified honest interviewers in the right part of Figure 4. Except when using the smallest number of interviews per

[14]The results for Ward's method are not presented to save space, but are available on request.

interviewer, the lower quartile is already above 65% and the median above 75%, i.e., in more than half of the cases more than three out of four honest interviewers are correctly assigned. When only five interviews per interviewer are available, the lower quartile goes down to about 55% and the median shrinks to 65%. Thus, one might conclude that in order to avoid a too high number of false alarms, i.e., honest interviewers wrongly assigned to the cluster containing the falsifiers, the number of interviews should not be too small. A number of the order of 20 appears sufficient in our setting. Obviously, the minimum number of interviews required also depends on the number of questions contained in an interview which will be analyzed in the following subsection.

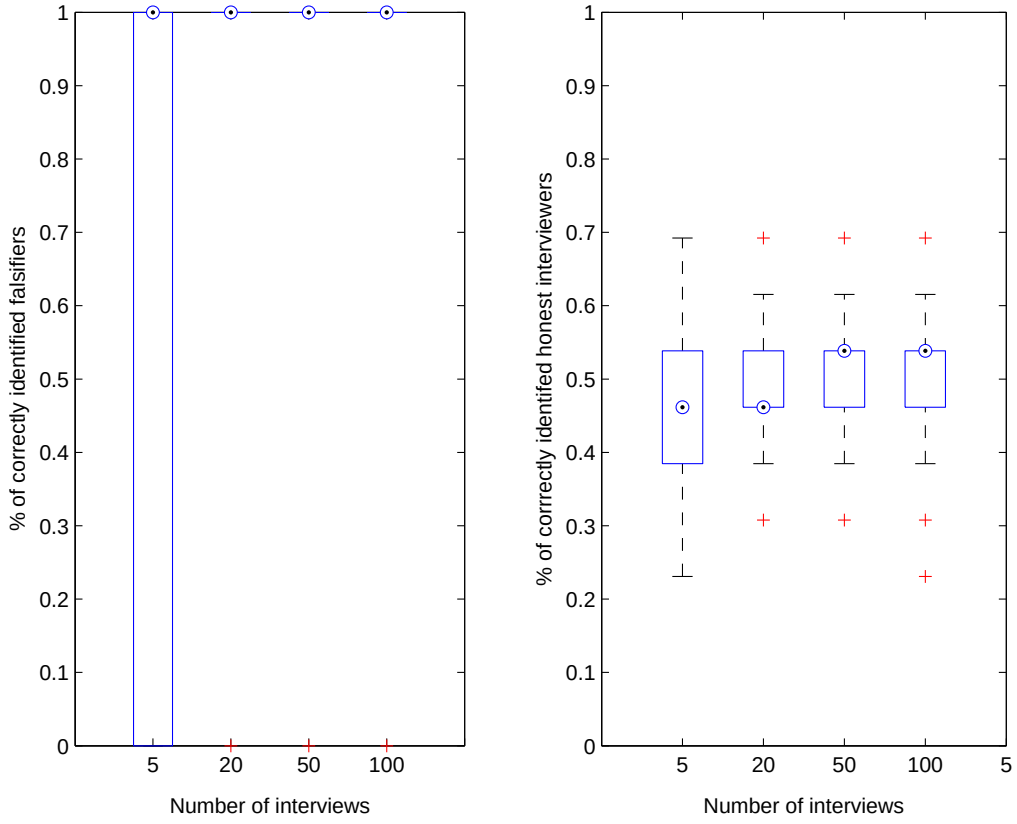

Figure 5: Performance of clustering method as function of number of interviews per interviewer (1 falisifier).

In the second group of experiments run with regard to the number of interviews per interviewer, a low share of falsifiers (one out of thirteen) was assumed corresponding to a share of about 8% of cheating interviewers, which might be considered a more realistic value. Figure 5 showing the results is again organized in two parts: the boxplots for the share of correctly identified falsifiers are displayed

on the left side, and on the right the boxplots for the share of correctly identified honest interviewers are presented.

For this setting, a slightly higher influence of the number of interviews per interviewer on the performance of the clustering method is found, especially for the identification of honest interviewers. Looking first at the identification of the one falsifier in these bootstrap samples, he is almost always (lower quartil being at 100%) detected as soon as at least 20 interviews per interviewer are available. Only when using the smallest number of interviews, the lower quartile shrinks to zero, i.e., in more than 25% of the cases the one falsifier is not detected, while the median is still at 100%. The results are less impressive when considering the share of correctly assigned honest interviews. As the right part of the figure shows, the median increases from about 45%, when using five interviews, to only about 55%, when using 50 or 100 interviews per interviewer. Hence, the higher the number of interviews per interviewer, the higher is the share of correctly assigned honest interviewers. We have to admit that the realization of 50 interviews is not to be expected in most cases, but the findings reported above indicate that 20 interviews per interviewer might be already sufficient if the primary interest is in detecting the one falsifier.

4.5 Number of Questions per Question Type

The results for the last group of design modifications are summarized in Figure 6 with `F/F` standing for "the share of identified falsifiers" and `H/H` standing for " the share of correctly assigned honest interviewers". For these designs, only the number of questions per interview of a particular type of questions are modified, i.e., the number of interviews per interviewer and the total number of interviewers as well as the share of falsifications consequently correspond with the numbers in the original dataset.

Each questions type corresponds to one of the indicators used in the analysis. Thus, increasing the number of questions of this type – holding all other factors constant – should improve the precision of the estimates of the corresponding indicator and, consequently, the performance of the clustering method in separating the two types of interviewers. Again, only the results using the optimal clustering method and `Benford_alt` are reported. The other results are qualitatively similar.

As expected, we find for all four indicators that increasing the number of questions contributing to the calculation of these indicators increases the probability of a correct assignment to the two clusters. However, it turns out that this effect is most pronounced for up to 20 questions per questionnaire for the first three types of indicators relying on binary information, while for the distribution of first digits some further improvements can be found even for still higher numbers. This might have been expected given the higher informational content of this indicator. Overall, the findings for this setting seem to confirm that the available numbers of questions per interview in the original dataset might be considered as sufficient

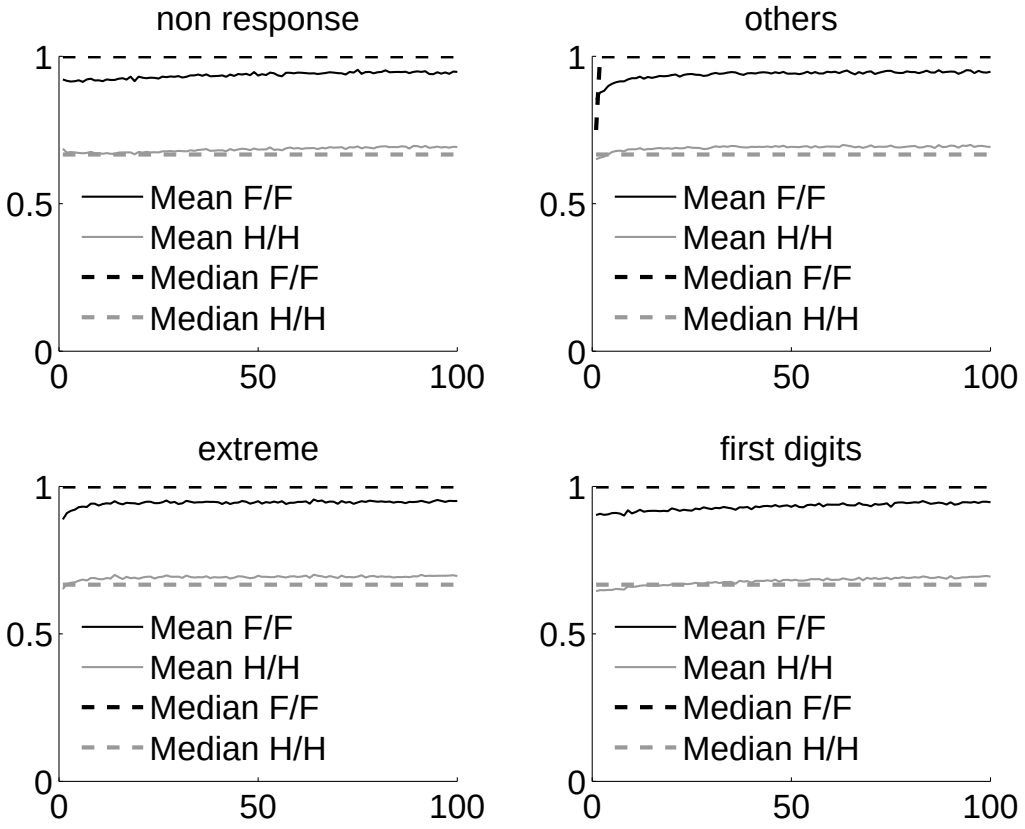

Figure 6: Performance of clustering method as function of number of questions.

with the sole exception of the alternative Benford indicator, for which more than the 25 available metric entries per questionnaire would help to improve the overall performance, though only to a small extent.

5 Conclusion and Outlook

Although there exists substantial anecdotal evidence on falsifications in survey data, few datasets with ex-post known falsfications are accessible which could be used to analyze methods for data based detection of falsified data. Consequently, in order to assess the performance of such methods on a broader database with different characteristics, a bootstrap analysis is proposed based on an available dataset with identified falsifiers.

The robustness of a data driven clustering method is considered with regard to several features of the dataset such as number of interviewers, share of falsifiers, number of interviews per interviewer and number of questions of particular question types. The results indicate that the promising results reported by Bredl *et al.* (2008)

might not be considered as pure chance or statistical artefact, but seem to reflect an actual convincing performance of the proposed clustering method. In fact, it is possible to identify most of the falsifiers as soon as the available information is not too limited, i.e., for more than five interviews per interviewer and at least about 20 questions per interview for each of the question types used to construct indicators for the clustering analysis. However, it has to be taken into account that the number of false positives, i.e., honest interviewers erroneously assigned to the cluster containing the falsifiers, might be substantial for some settings, e.g., very low shares of falsifiers and limited data per interviewer.

Future research in this field will have to address several issues. First, as it turned out that the optimal clustering often results in solutions outperforming those obtained by classical clustering methods such as Ward's method, it will be aimed at enabling the use of this method also for larger problem instances. Given that a full enumeration of all potential clusters will not be feasible anymore in this case, we will resort to heuristic optimization methods to obtain at least high quality approximations of the optimal cluster.[15] Second, additional datasets will be used as base for the bootstrap analysis. Given the limited access to real datasets with identified falsifiers, we will also resort to data obtained from experiments (Menold *et al.* 2011). Third, for other datasets, additional indicators can be used and might help to improve the performance of the clustering method.[16] Of course, the bootstrap method can also be used to identify those indicators which are most useful for identifying falsifiers. Finally, while data obtained from honest interviewers might be considered as coming from a unique dataset, the falsified data might also be split up in more than one cluster, e.g., falsifications by experienced and unexperienced interviewers, respectively. The bootstrap methodology presented can also be adjusted for this more general case.

References

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

Bredl, S., N. Storfinger and N. Menold (2011). A literature review of methods to detect fabricated survey data. Discussion Paper 56. ZEU. Giessen.

Bredl, S., P. Winker and K. Kötschau (2008). A statistical approach to detect cheating interviewers. Discussion Paper 39. ZEU. Giessen.

[15] See Chipman and Winker (2005) for a related application in optimal aggregation of time series data.

[16] Storfinger and Opper (2011) provide an overview on indicators which might be considered in this context.

Bushery, J.M., 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.

Chernick, M.R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley. Hoboken, NJ. 2nd Ed.

Chipman, J.S. and P. Winker (2005). Optimal aggregation of linear time series models. *Computational Statistics and Data Analysis* **49**(2), 311–331.

Crespi, L.P. (1945). The cheater problem in polling. *The Public Opinion Quarterly* **9**(4), 431–445.

Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. ITA manu:scripts 02_04. Institute of Technology Assessment (ITA).

Efron, B. (1978). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics* **7**(1), 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans*. Vol. 38 of *CBMS-NSF Monographs*. Society of Industrial and Applied Mathematics.

Forsman, G. and I. Schreiner (1991). The design and analysis of reinterview: An overview. In: *Measurement Errors in Surveys* (P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, Eds.). pp. 279–301. Wiley. Chichester.

Harrison, D.E. and S.I. Krauss (2002). Interviewer cheating: Implications for research on entrepreneurship in africa. *Journal of Developmental Entrepreneurship* **7**(3), 319–330.

Hauck, M. (1969). Is survey postcard verification effective?. *Public Opinion Quarterly* **33**, 117–120.

Hood, C.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.

Jain, A.K. and J.V. Moreau (1987). Bootstrap technique in cluster analysis. *Pattern Recognition* **20**(5), 547–568.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. **16**(4), 277–298.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS1994. *ZUMA-Nachrichten* **36**, 89–105.

Menold, N., N. Storfinger and P. Winker (2011). Development of a method for ex-post identification of falsifications in survey data. In: *New Techniques and Technologies for Statistics (NTTS) Conference Proceedings*. http://www.cros-portal.eu/content/s9-paper-4-ntts-2011-s9.

Peck, R., L. Fisher and J. Van Ness (1989). Approximate confidence intervals for the number of clusters. *Journal of the American Statistical Association* **84**(405), 184–191.

Rost, D. (1995). A simulation study of the weighted k-means cluster procedure. *Journal of Statistical Computation and Simulation* **53**, 51–63.

Schäfer, C., J.-P. Schräpler, K.R. Müller and G.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.-P. and G.G. Wagner (2005). Characteristics and impact of faked interviews in surveys - an analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv* **89**(1), 7–20.

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.

Storfinger, N. and M. Opper (2011). Datenbasierte Indikatoren für potentiell abweichendes Interviewerverhalten. Discussion Paper 58. ZEU. Giessen.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**(301), 236–244.

Winker, P. (2001). *Optimization Heuristics in Econometrics: Applications of Threshold Accepting*. Wiley. Chichester.

Bisherige Veröffentlichungen in der Discussion Papers-Reihe

No. 1 HERRMANN, R., KRAMB, M. C., MÖNNICH, Ch. (12.2000): Tariff Rate Quotas and the Economic Impacts of Agricultural Trade Liberalization in the WTO. (etwas revidierte Fassung erschienen in: "International Advances in Economic Research", Vol. 7 (2001), Nr. 1, S. 1-19.)

No. 2 BOHNET, A., SCHRATZENSTALLER, M. (01.2001): Der Einfluss der Globalisierung auf staatliche Handlungsspielräume und die Zielverwirklichungsmöglichkeiten gesellschaftlicher Gruppen.
(erschienen in: "List-Forum für Wirtschafts- und Finanzpolitik", Bd. 27(2001), H. 1, S. 1-21.)

No. 3 KRAMB, M. C. (03.2001): Die Entscheidungen des "Dispute Settlement"-Verfahrens der WTO im Hormonstreit zwischen der EU und den USA – Impli kationen für den zukünftigen Umgang mit dem SPS-Abkommen.
(überarbeitete Fassung erschienen in: "Agrarwirtschaft", Jg. 50, H. 3, S. 153-157.)

No. 4 CHEN, J., GEMMER, M., TONG, J., KING, L., METZLER, M. (08.2001): Visualisation of Historical Flood and Drought Information (1100-1940) for the Middle Reaches of the Yangtze River Valley, P.R. China.
(erschienen in: Wu et al. (eds) Flood Defence '2002, Beijing, New York 2002, pp. 802-808.)

No. 5 SCHROETER, Ch. (11.2001): Consumer Attitudes towards Food Safety Risks Associated with Meat Processing.
(geänderte und gekürzte Fassung ist erschienen unter Christiane SCHROETER, Karen P. PENNER, John A. FOX unter dem Titel "Consumer Perceptions of Three Food Safety Interventions Related to Meat Processing" in "Dairy, Food and Environmental Sanitation", Vol. 21, No. 7, S. 570-581.)

No. 6 MÖNNICH, Ch. (12.2001): Zollkontingente im Agrarsektor: Wie viel Liberalisie rungsfortschritt? Ergebnisse und Diskussion einer Auswertung der EU-Daten.
(gekürzte Fassung erschienen in BROCKMEIER, M., ISERMEYER, F., von CRA MON-TAUBADEL, S. (Hrsg.), Liberalisierung des Weltagrarhandels - Strategien und Konsequenzen. "Schriften der Gesellschaft für Wirtschafts- und Sozialwissenschaften des Landbaues e.V.", Bd. 37(2002), S. 51-59.)

No. 7 RUBIOLO, M. (01.2002): EU and Latin America: Biregionalism in a Globalizing World?

No. 8 GAST, M. (02.2002): Zollkontingente bei US-amerikanischen Käseimporten. (gekürzte Fassung erschienen in: "Agrarwirtschaft", Jg. 51, H. 4, S. 192-202.)

No. 9 BISCHOFF, I. (08.2002): Efficiency-enhancing Effects of Private and Collective Enterprises in Transitional China.

No. 10 KÖTSCHAU, K. M., PAWLOWSKI, I., SCHMITZ, P. M. (01.2003): Die Policy Ana lysis Matrix (PAM) als Instrument zur Messung von Wettbewerbsfähigkeit und Politikeinfluss - Zwischen Theorie und Praxis: Das Fallbeispiel einer ukraini schen Molkerei.

No. 11 HERRMANN, R., MÖSER A. (06.2003): Price Variability or Rigidity in the Food-retailing Sector? Theoretical Analysis and Evidence from German Scanner Data.

No. 12 TROUCHINE, A. (07.2003): Trinkwasserversorgung und Armut in Kasachstan: Aktueller Zustand und Wechselwirkungen.

No. 13 WANG, R.; GIESE, E.; GAO, Q. (08.2003): Seespiegelschwankungen des Bosten-Sees (VR China).

No. 14 BECKER, S.; GEMMER, M.; JIANG, T.; KE, CH.. (08.2003): 20th Century Precipitation Trends in the Yangtze River Catchment.

No. 15 GEMMER, M.; BECKER, S.; JIANG, T (11. 2003): Detection and Visualisation of Climate Trends in China.

No. 16 MÖNNICH, Ch. (12.2003): Tariff Rate Quotas: Does Administration Matter?

No. 17 GIESE, E.; MOßIG. I. (03.2004) Klimawandel in Zentralasien

No. 18 GIESE, E.; SEHRING, J. TROUCHINE, A. (05.2004) Zwischenstaatliche Wassernutzungskonflikte in Zentralasien

No. 19 DIKICH, A. N. (09.2004)
Gletscherwasserressourcen der Issyk-Kul-Region (Kirgistan), ihr gegenwärtiger und zukünftiger Zustand

No. 20 Christiansen, Th.; Schöner, U. (11.2004)
Irrigation Areas and Irrigation Water Consumption in the Upper Ili Catchment, NW-China

No. 21 NARIMANIDZE, E. et al. (04.2005)
Bergbaubedingte Schwermetallbelastungen von Böden und Nutzpflanzen in einem Bewässerungsgebiet südlich von Tiflis/Georgien - Ausmaß, ökologische Bedeutung, Sanierungsstrategien

No. 22 ROMANOVSKIJ, V.V.; KUZ'MIČENOK, V.A. (06.2005)
Ursachen und Auswirkungen der Seespiegelschwankungen des Issyk-Kul' in jüngerer Zeit

No. 23 ZITZMANN, K.; TROUCHINE, A. (07.2005)
Die Landwirtschaft Zentralasiens im Transformationsprozess
(nicht mehr lieferbar!)

No. 24 SEHRING, J. (08.2005)
Water User Associations (WUAs) in Kyrgyzstan -
A Case Study on Institutional Reform in Local Irrigation Management

No. 25 GIESE, E., MAMATKANOV, D. M. und WANG, R. (08.2005)
Wasserressourcen und Wassernutzung im Flussbecken des Tarim (Autonome Region Xinjiang / VR China)

No. 26 MOSSIG, I., RYBSKY, D. (08.2005)
Die Erwärmung bodennaher Luftschichten in Zentralasien. Zur Problematik der Bestimmung von Trends und Langzeitkorrelationen

No. 27 GAST, M.: (09.2005)
Determinants of Foreign Direct Investment of OECD Countries 1991-2001

No. 28 GIESE, E., TROUCHINE, A. (01.2006)
Aktuelle Probleme der Energiewirtschaft und Energiepolitik in Zentralasien

No. 29 SEHRING, J. (06.2006)
The Politics of Irrigation Reform in Tajikistan

No. 30 LANGENOHL, A. / WESTPHAL, K. (11.2006)
Comparing and Inter-Relating the European Union and the Russian Fede ration. Viewpoints from an international and interdisciplinary students' project

No. 31 WEBER, S./ ANDERS, S. (3.2007)
Price Rigidity and Market Power in German Retailing

No. 32 GAVARDASHVILI, G. / SCHAEFER, M. / KING, L. (8.2007)
Debris Flows at the River Mletis Khevi (Greater Caucasus Mountains, Georgia) and its Assessment Methods

No. 33 TEUBER, R. (5.2007)
Geographical Indications of Origin as a Tool of Product Differentiation – The Case of Coffee D

No. 34 DOSTAJ, Ž. D. (in Zusammenarbeit mit E. Giese und W. Hagg) (6.2007)
Wasserressourcen und deren Nutzung im Ili-Balchaš Becken

No. 35 FLATAU, J./ Hart, V. / KAVALLARI, A./ SCHMITZ, P.M. (7.2007)
Supply Chain Analysis of Olive Oil in Germany

No. 36 HART, V. / KAVALLARI, A. / SCHMITZ, P.M. / WRONKA, T. (7.2007)
Supply Chain Analysis of Fresh Fruit and Vegetables in Germany

No. 37 MÖSER, N. (7.2008)
Analyse der Präferenzen russischer Fachbesucher für ausgewählte Messeleistungen mit Hilfe der Choice-Based Conjoint-Analyse

No. 38 BISCHOFF, I. / EGBERT, H. (8.2008)
Bandwagon voting or false-consensus effect in voting experiments? First results and methodological limits

No. 39 BREDL, S. / WINKER, P. / KÖTSCHAU, K. (12.2008)
A Statistical Approach to Detect Cheating Interviewers

No. 40 HERRMANN, R. / MÖSER, A./ WEBER, S. (01.2009)
Grocery Retailing in Poland: Development and Foreign Direct Investment

No. 41 HERRMANN, R. / MÖSER, A./ WEBER, S. (02.2009)
Grocery Retailing in Germany: Situation, Development and Pricing Strategies

No. 42 GÖCKE, M. (05.2009)
Efficiency Wages and Negotiated Profit-Sharing under Uncertainty

No. 43 KRAMB, M. / HERRMANN, R. (05/2009)
Wie wirken gemeldete SPS-Maßnahmen? Ein Gravitationsmodell des Rindfleischhandels der EU

No. 44 BREDL,S. (10/2009)
Migration, Remittances and Educational Outcomes: the Case of Haiti

No. 45 BELKE, A. / GÖCKE, M. / GUENTHER, M. (11/2009)
When Does It Hurt? The Exchange Rate "Pain Threshold" for German Exports

No. 46 EGBERT, H. / FISCHER, G. / BREDL, S. (12/2009)
Advertisements or Friends? Formal and Informal Recruitment Methods in Tanzania

No. 47 RAKHIMOV, M. (01/2010)
The European Union and Central Asia: Challenges and Prospects of Cooperation

No. 48 NAJMITDINOV, A (01/2010)
Central Asian integration as a way of guaranteeing regional security and economic growth feasibility and prospects

No. 49 JETPYSPAEVA, Y (03/2010)
Central Asia: Changing Politics. Case of Kazakhstan

No. 50 JONBOBOEV , S. (03/2010)
Humanities in Transition: Liberation of Knowledge in Central Asia and possible Impact of European Union

No. 51 KULAKHMETOVA, , A. (03/2010)
Protection Mechanisms and services for young Workers in Central Asia and the European Union

No. 52 MOMOSHEVA, S. (03/2010)
The EU strategy for Central Asia and Kyrgyzstan foreign policy

No. 53 EGBERT, H. / FISCHER, G./ BREDL, S. (06/2010)
Different Background – Similar Strategies: Recruitment in Tanzanian-African and Tanzanian-Asian

No. 54 GÖNSCH, I. (11/2010)
Determinants of Primary School Enrolment in Haiti and the Dominican Republic

No. 55 GÖNSCH, I. / GRÄF, S.: (05/2011)
Education for All and for Life? An Introduction into Primary School Education in Senegal

No. 56 BREDL, S. / STORFINGER, N. / MENOLD, N. (08/2011)
A Literature Review of Methods to Detect Fabricated Survey Data

No. 57 STORFINGER, N. / WINKER, P. (08/2011)
Robustness of Clustering Methods for Identification of Potential Falsifications in Survey Data

Stand: 30. August 2011

Die Diskussionsbeiträge können auf der Homepage des ZEU
http://www.uni-giessen.de/zeu
im Menü „Forschung“, „Veröffentlichungen“ kostenlos heruntergeladen werden.