Journal of Official Statistics
Vol. 9. No. 1, 1993, pp. 217–222
© Statistics Sweden

Special Aspects in Using Microdata: Data Anonymity and Non-Response Bias: Research at ZUMA

Paul Lüttinger, Heike Wirth, and Hans-J. Hippler[1]

Abstract: This contribution provides an elaborated description of the work that is done in one of the various departments within the Center for Survey Research and Methodology (ZUMA), the Department of Microdata. The first project deals with the identification risk of microdata; the second describes some results dealing with the representativity of surveys. First of all we give a short summary about the tasks of ZUMA.

Key words: ALLBUS; anonymity; identification risk; middle-class bias; microcensus; representativity; sample survey; systematic errors.

1. The Center for Survey Research and Methodology (ZUMA)

ZUMA is an institute for basic research as well as a center for the social sciences. It is part of the German Social Science Infrastructure Services (GESIS) that is made up of three independent institutes: the Central Archive (ZA) at the University of Cologne, responsible for providing social science data to academics; the Social Science Information Center (IZ) in Bonn, providing data bases for social sciences research projects and social sciences literature, and ZUMA.

ZUMA serves researchers in the social sciences and related disciplines by providing consulting services and assistance in the preparation and execution of research projects and in the analysis of data. ZUMA houses the German General Social Survey (ALLBUS) which provides data on social structure and change on a biennial basis since 1980 for researchers who otherwise would have no direct access to primary data. The sample that is interviewed consists of approximately 3000 eligible voters randomly selected from the population of Germany. The organization is also responsible for the German contribution to the International Social Survey Program (ISSP), an annual survey program established in 1985.

ZUMA also conducts its own research in social science methodology in order to provide a better basis for the execution of future social science projects. One of the current research programs, for example, deals with human judgmental processes and their application to survey interviews. The main issues covered by this research program include the validity of the estimation of behavioral frequencies, the dating of events, and the different processes involved in reporting about one's own behavior (self-

[1] Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, D-6800 Mannheim 1, Germany.

reports) or the behavior of another household member (proxy reports). The research program also addresses issues of attitude and opinion measurement, including the emergence of question order and context effects, as well as the effect of different modes of data collection (face-to-face and telephone interviews, as well as self-administered questionnaires) on cognitive processes.

Through these and other research projects, ZUMA intends to identify relevant developments and test them in a practical environment, thereby contributing new insights in central areas of methodological research.

The following section provides an elaborated description of the work that is done in one of the various departments within ZUMA, the Department of Microdata (see Papastefanou 1987).

2. The Department of Microdata

This department was established in 1987, simultaneously with the foundation of GESIS. It provides advice on the use of anonymized individual data (i.e., microdata on persons and households) collected by statistical bureaus above all at national level, especially the annual microcensus from the German Federal Statistical Office. The microcensus is the largest sample survey: One per cent of the German population is annually questioned regarding its demographic and employment characteristics. By the time Germany was reunified in 1990, the survey had reached around 250,000 households, i.e., a total of about 600,000 people (Hartmann 1989). The microcensus is a compulsory federal survey; everyone interviewed is required to provide the information asked.

The department is continually acquiring and organizing microdata into a microdata archive to meet the needs of social research as well as the requirements of data protection by adapting the data to social change and preparing it for comparative analysis. The department places particular emphasis on census data. The services of the Microdata Department include transmitting data sets to social scientists (provided there are no objections because of data protection), offering special training for junior scientists regarding knowledge and analysis of large microdata sets, and supporting investigations carried out by individual scientists, research institutions and political authorities, which are unable to collect data on their own.

3. Microdata and Identification Risk

Since the early 1980s the research community in Germany has been confronted with the problem that, because of the tightening of Statistics Laws, access to microdata has been handled in a very restrictive way by the Federal Statistical Office. According to German Law, microdata could only be released if identification was impossible (absolute anonymity). The new Federal Act on Statistics of 1987 (Article 16, paragraph 6) makes an exception from the general rule of absolute anonymity. It stipulates that microdata from official statistics may be released for scientific purposes if they are factually anonymous. Factual anonymity means that the data can be linked to the respondents only by employing an excessive amount of time, expenses and manpower. To clarify the conditions under which the factual anonymity of released microdata can be attained, a cooperative project was carried out by the University of Mannheim, the German Federal Statistical Office and ZUMA.

The main aim of this study was to test the identification risk for individual data records under realistic conditions. Then procedures were to be tested for minimizing the identification risks. Finally, on the basis of these results, recommendations were to be developed for the dissemination of factually anonymous microdata.

3.1. The identification risk of microdata

To identify a record in a released microdata file a potential invader needs "prior knowledge" consisting of individual data records overlapping with those of the microdata file at some key variables. Identification of an individual takes place when a one-to-one relationship between a record in the microdata file and the prior knowledge can be established.

There are two approaches in the literature to determine the risk of identification. One approach is to construct a statistical model that estimates the probability of successful identification (cf. Bethlehem et al. 1990). The other approach is based on simulation experiments with partly artificial data or artificial experimental set-ups (cf. Paaß and Wauschkuhn 1985). In our project an alternative approach was pursued. We studied the success rate of identification attempts under realistic conditions by using empirical data and different identification techniques that are available to an invader (Müller et al. 1991).

One of the most important and surprising results of the project was the fact that the actual risk of an individual being identified from a microdata file by using prior knowledge is much lower than is often assumed by approaches using statistical models or artificial data to estimate the identification risk (Wirth and Blien 1991). In the most riskful situation 16.7% of the individuals at risk could be correctly identified. In other situations the chances of a correct identification varied between zero and 4.5%. Mostly, the attempt at identification failed because the key variable values were not recorded identically in the microdata and the prior knowledge files (Müller et al. 1991). The differences in definitions and inquiry periods as well as the rate of coding and data errors in the microdata file and in the prior knowledge file do not worry users interested in microdata for statistical purposes. However, it does interfere with the use of data at the individual level for identification attempts (Marsh et al. 1991, Blien et al. 1992).

Though under most conditions, the probability of an individual being identified from a microdata file is very small, there remains one specific situation in which an identification attempt might be successful (Müller et al. 1991). This situation can be characterized as follows: the values of the common variables are recorded identically in the microdata file and in the prior knowledge file; the intruder knows whose data is contained in the microdata file (response knowledge); the microdata file contains fine geographical detail, and the individual the intruder is interested in belongs to a small subpopulation that can easily be isolated by a specific variable.

Even if it is very unlikely that these four conditions will coincide, measures should be taken to minimize this potential risk. Therefore the project group has listed proposals for the further release of microdata (thus for only the German microcensus and the sample survey on income and expenditure). It contains institutional and contractual solutions as well as statistical solutions (Müller et al. 1991).

3.2. Proposals for the release of microdata

Institutional or contractual solutions include the following: it is expressly forbidden to attempt to identify individuals from a microdata file; the data users are only allowed to use the data for the purposes agreed upon. Distributing the data to others is prohibited. Different needs can be negotiated.

Measures must be taken by the user to prevent both insiders and outsiders from gaining unauthorized access to the microdata.

Several statistical solutions could be employed for minimizing the remaining risk of identification. However, most of them, in particular the various forms of perturbation, reduce the utility of the microdata for multivariate estimates. Based on identification experiments and tests of different security measures, sampling and grouping remain as the two main methods of reducing the potential risk of identification.

In the microcensus, the project group differentiated between a basic sample and a regional sample. The basic sample (70% of the microcensus) contains information at two levels: household and individuals within household. It contains only broad geographical details: countries and size classes of communities. The regional sample (85% of the microcensus) contains fine geographical details that are grouped so that the smallest geographical areas identified must contain a minimum population size of 100,000. In this file, classifications such as occupation, industry, nationality, and age will be grouped where necessary to ensure that each category represents 50,000 individuals. In the case of the sample survey of income and expenditure, there will be three different subsamples (98%, 90%, 80%) with grouping for those variables for categories containing few cases, particularly at the extremes of distribution. In the future, the dissemination of microdata will be handled on the basis of these proposals (which have only been outlined here). This will prove to be a considerable advantage for the research community.

4. Non-Response Bias of Sample Surveys

Sample surveys conducted by federal statistical bureaus or research institutes aim to reflect the characteristics of a population as realistically as possible. The statistical distributions obtained by these surveys, however, have often been at odds with each other – even when the question content, i.e., the demographic or socio-structural characteristics the surveys were to investigate, was the same.

These statistical deviations could be caused on the one hand by random sampling errors, which can be, however, predicted and controlled. A second source of discrepancies in the data could be systematic errors, which may arise from differing conceptions as to survey methodology as well as from non-response, i.e., non-obtainability of data due to the subjects' non-cooperation, such as item non-response and unit non-response.

The problem of non-response has been encountered particularly with surveys done on a volunteer basis. Because of the voluntary nature of the surveys, the probability of acquiring incomplete data as a result of non-response varies between different target groups; therefore the distribution of those characteristics which make up the differences between these groups can be expected to vary as well.

For this reason further research at the Department of Microdata has focused on deciding to what extent disparities between the results of sample surveys and federal comprehensive total surveys (whose results can be assumed to have been assessed correctly) are caused by systematic errors. Furthermore we examined the reasons for the so-called middle-class bias which appears in non-official, voluntary sample surveys.

4.1. *Sample surveys vs. comprehensive surveys*

During the initial phase of our research, the univariate marginals of various demographic characteristics that were being investigated by the microcensus (which is a governmental sample survey in which the responses are obligatory) as well as by the ALLBUS were compared with the data obtained by comprehensive surveys with a total enumeration conducted by the Federal Bureau of Statistics (e.g., the "Volkszählung" that is, the official census; Hartmann 1990). While the results of the federal microcensus generally match the data obtained by other governmentally-conducted surveys – i.e., systematic distortions have not appeared – several characteristics investigated by the ALLBUS showed statistically significant deviations similar to those which have also appeared in the results of other non-official, voluntary surveys.

The ALLBUS, for example, obtains less of its data from elderly individuals, one-member households and lower-class individuals than do federal surveys (see also 4.2). There are special problems with missing data in the information provided by small households as a consequence of differing methods of random sampling used by the ALLBUS and the microcensus. While the microcensus is based on a selection procedure that leads to a statistically representative random sample at the individual and household level, the ALLBUS (along with other sociological surveys) ends up underestimating the existing number of small households due to its own different selection procedure.

In addition, the ALLBUS presently draws its data from private households on a collective basis only – i.e., the persons living in these households are not considered individually. To make the ALLBUS statistically representative at the individual level, the statistics obtained are weighted according to certain criteria; i.e., every household is weighted according to its number of German citizens who are eligible to vote (the so-called reduced household size). Since the distribution of this characteristic is distorted by the ALLBUS surveys (due to the sample design and problems of non-obtainability), however, this weighting procedure leads to even fewer accurate results than does the use of unweighted data.

After the statistics have been weighted, certain personal features that can be correlated with household size, such as age, sex, and marital status, tended to deviate more conspicuously from the official census data (for further information on weighting procedures for survey statistics, see Rothe, 1990, Gabler, 1991).

4.2. The middle-class bias

As mentioned before another distortion that has been observed together with the univariate approach consists of the so-called middle-class bias, which arises from the fact that ALLBUS interviews include fewer lower-class individuals; the results acquired from middle-class subjects therefore tend to weight disproportionately heavily. The middle-class bias also implies an underrepresentation of upper-class respondents. Due to the small number of participants in the samples, however, the results of our surveys cannot be used to investigate this deficiency.

Nevertheless, our research has suggested that differences in sampling methods and varying cooperativeness on the part of subjects are not the only factors that have led to such statistical deviations between sociological sample surveys and surveys (sample surveys or censuses) in which the responses are obligatory (done by the federal statistical office).

Multivariate analyses have shown that data deficiencies in survey responses like the middle-class bias can be attributed above all to the respondents' level of education and to the extent to which they are involved in the working world (Hartmann and Schimpl-Neimanns 1992), since people who are gainfully employed are harder to reach for survey purposes than are those who are not employed. In effect, the so-called middle-class bias reveals itself as an educational bias. Hartmann and Schimpl-Neimanns have provided the following summary of their findings: The description of characteristics which can be correlated to household size, level of education, or gainful employment leads to distorted marginals if these variables are not controlled. When these distorted demographic characteristics are used as independent variables, they can be applied with relatively little difficulty; however, the distorted characteristics also may lead to distorted coefficients if they are used as dependent variables.

5. References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure Risk for Microdata. Statistica Neerlandica, 46, 1, 69–82.

Gabler, S. (1991). Eine allgemeine Formel zur Anpassung an Randtabellen. ZUMA-Nachrichten 29, 29–44.

Hartmann, P. (1989). Der Mikrozensus als Datenquelle für die Sozialwissenschaften. ZUMA-Nachrichten 24, 6–26.

Hartmann, P.H. (1990). Wie repräsentativ sind Bevölkerungsumfragen. Ein Vergleich des ALLBUS und des Mikrozensus. ZUMA-Nachrichten 26, 7–31.

Hartmann, P.H. and Schimpl-Neimanns, B. (1992). Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44, 2, 315–340.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991). The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society, 154, 305–340.

Müller, W., Blien, U., Knoche, P., and Wirth, H. (1991). Die faktische Anonymität von Mikrodaten. Band 19 der Schriftenreihe Forum der Bundesstatistik, ed.: Statistisches Bundesamt. Stuttgart: Metzler-Poeschl.

Papastefanou, G. (1987). Zentrum für Mikrodaten – eine neue Abteilung von ZUMA. ZUMA-Nachrichten 21, 20–31.

Paaß, G. and Wauschkuhn, U. (1985). Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von Anonymisierten Einzelangaben, R. Oldenbourg, München, Wien.

Rothe, G. (1990). Wie (un)wichtig sind Gewichtungen? Eine Untersuchung am ALLBUS 1986. ZUMA-Nachrichten 26, 31–56.

Wirth, H. and Blien, U. (1991). Empirische Überprüfung der Anonymität von Mikrozensusdaten. Ed.: Glatzer, W., 25. Deutscher Soziologentag 1990. Die Modernisierung moderner Gesellschaften 1, 230–232.

Received March 1992
Revised October 1992

Journal of Official Statistics
Vol. 9. No. 1, 1993, pp. 223–231
© Statistics Sweden

SCPR's Role in British Social Survey Research

Barry Hedges[1]

Abstract: After a brief description of SCPR, issues discussed include the relationship between theory and practice; tendering as a means of placing contracts; staffing and training; technological change; policy analysis; and ethical issues.

Key words: Research organisation; policy research; qualitative research; telephone surveys; computer-assisted interviewing.

1. Introduction

Social and Community Planning Research, more often known as SCPR, was founded at the end of the 1960s to fill what appeared to be a conspicuous gap in research in Britain at that time. There was no major survey-based organisation, outside government, that specialised solely in social research. In the United States, there were notably successful examples of this genre, such as the university-based Institute for Social Research, Ann Arbor, Michigan. In Britain, in contrast, there was no major organisation within the academic sector. Public sector surveys were undertaken by the Government Social Survey (now the Social Survey Division of the Office of Population Censuses and Surveys), by market research organisations and by relatively small academic or policy units that for the most part did not maintain full resources for carrying out surveys but either subcontracted the work or set up the necessary machinery ad hoc.

SCPR apart, the description above fits social survey research in Britain more or less equally well in 1932 as in 1969. No other major new survey organisation has emerged to specialise in social research. But there have been large changes over this period, during the second half of which Britain went through the social and ideological upheaval of Thatcherism. The use of social surveys, both by social scientists and by government, has continued to grow more or less unchecked by changes in government or by social and cultural trends, although there have been shifts of emphasis from time to time. Early fears that the Thatcher administration might be less convinced than previous administrations of the value of survey research proved unfounded.

Surveys have become more sophisticated as well as more familiar. There has been a continual growth in their complexity. One contributory factor is the technological advance that has made possible many things previously out of reach. Another iś an increasingly ambitious range of information requirements, due to a more sophisticated perception of the value of surveys and their uses for policy and social scientific purposes.

[1] Social and Community Planning Research, 35 Northampton Square, London EC1V 0AX, England.