Methods Approaches Developments

Federal Statistical Office

Information of the German Federal Statistical Office

Number 1/93

Contents

Page

The catchword

Methodology of federal statistics – Further development

European matters

Cooperation with science and research

Events

Practical anonymity of microdata – Report on a research project

Federal Statistical Office, Gustav-Stresemann-Ring 11, D-6200 Wiesbaden

Translated at the Federal Statistical Office, Wiesbaden

Published by:
Federal Statistical Office
Gustav-Stresemann-Ring 11
D-6200 Wiesbaden
Federal Republic of Germany

Published in March 1993

The catchword

Regional classifications

Official statistics require uniform categorizations and classifications for presenting their results by regions. The most significant categorization in this respect is the one by so-called administrative territorial units, with the further breakdown by local authorities, i.e. by communities, Kreise, administrative districts and Länder. An eight-digit community code number is used for classification purposes. This regional classification is published by the Federal Statistical Office together with some characteristics of the units, especially the area and the population numbers (Federal Statistical Office: Regional Classification – Official Code Numbers and Population Data of the Communities and Administrative Districts in the Federal Republic of Germany, 1990 Edition, Messrs. Metzler-Poeschel, Stuttgart).

As the breakdown by local authorities does not serve all administrative and other tasks and purposes such as scientific analyses, other administrative categorizations are also being used, e.g. postal districts, labour office districts or constituencies, as well as so-called "non-administrative territorial categorizations". The latter subdivide the entire territory by certain material criteria for specific purposes, i.e. in a functional way. For example, the territory of the Federal Republic of Germany is subdivided into 97 districts for purposes and measures of regional planning, and in 117 labour market regions for purposes of improving the economic structure of the regions. (For further information please refer to the following paper: P. Knoche/S. Köhler, "Neuere Entwicklungen in der Regionalstatistik" in the monthly review Wirtschaft und Statistik, No. 4/1992.) In contrast, regional classifications combining isolated areas according to specific criteria, i.e. the so-called regional typifications, such as the subdivision by "urban, semi-urban or rural areas", strictly speaking cannot be regarded as regional classifications.

The basis for processing and presenting statistical data by regions for the member states of the EC is provided by the Nomenclature of Territorial Units for Statistics (NUTS = Nomenclature des Unités Territoriales Statistiques). It was developed by EUROSTAT (Statistical Office of the European Communities) in cooperation with other bodies of the Commission. In contrast to other European nomenclatures, it does not have a legal foundation of its own, but a quasi-official status. For the Federal Republic, the Länder are the first level, the administrative districts the second, and the Kreise the third level of the three-level NUTS.

Below NUTS, there also is a two-level "subregional" classification called LOC. The communities are classified to the second level of this classification (LOC 2), while the LOC 1 level is not yet occupied for the Federal Republic. It may well be, however, that the communal associations existing in some federal Länder with differing denominations will be allocated to this level in due course.

There are the following peculiarities of regional classifications as compared with the standard "classifications of official statistics" such as the classifications of economic branches or of commodities:

- There is a great number of classification systems which in principle are or were developed independently and without any cooperation of official statistics.
- There is considerable heterogeneity on all levels especially with regard to the breakdown by local authorities, so that often, so to speak, the incomparable is in fact compared. For instance, the largest community according to the population number is Berlin with about 3.4 million inhabitants, the smallest being Roxförde in Saxony-Anhalt with 2 inhabitants.
- The same unit is allocated to different levels: Berlin for instance is at the same time community, Kreis and federal Land; and in the NUTS the Länder (NUTS 1) which are not subdivided into administrative districts (NUTS 2) are classified to this level (NUTS 2) as well.

Due to these peculiarities, it is necessary to prepare typifications for many purposes, e.g. by "community size classes".

Methodology of federal statistics – Further development

Register of inhabitants as a basis for household counts

In official federal statistics, information on households and the persons living in them are usually provided by the population census or the microcensus. While the population census is conducted at large intervals as a complete enumeration of the overall population, the microcensus is conducted annually as a representative household survey on a sampling basis, thus providing up-to-date information on households. The household concept underlying both surveys is based on the joint living and house-keeping of the household members. The former GDR, however, did not employ the microcensus as a survey tool; on the other hand, it would be useful to have for the transition year 1990 household data which are comparable with those of federal statistics. This is why a count of the households (number, size, composition) was made on the basis of anonymous person-related data stored in the Population

Register for Statistics (Bevölkerungsregister Statistik – BRS) of the former GDR. The data records of the BRS were derived from the Data Storage of Inhabitants (Einwohnerdatenspeicher – EDS). The computing method of household formation (household algorithm) was always performed for persons with the same main place of residence (excluding persons living in community institutions) on the basis of the stored information derived from the Personal Identification Number (Personenkennzahl – PKZ) of the person concerned – i.e. date of birth, sex – and from the marital status as well as the PKZ of the spouse or of the father and mother. The last names were not taken into account. As a further step in forming the household, the residence context could be accounted for. The information on the dwelling (the same coded dwelling number for any person living there) was derived from the reports of the housing branch of the municipal administrations and of the registration offices and was added to the data records on inhabitants. However, this was possible only for part of the person-related data records because the dwelling number as a rule was allocated only to persons having their main place of residence in a town or community of 10,000 inhabitants or more.

The household algorithm first links up the data records on inhabitants of one house (same coded house numbers) by utilizing the coded family relationships and then combines to family households the persons recognized as belonging together. For each person of a house, it is thus checked by comparing the relevant data (date of birth, sex, marital status) whether there are in that house persons who are relatives of the person concerned. Grouping the persons together to form family households (family nuclei) comprises the following major steps:

- Forming married couples and allocating children who are not married and do not have children of their own, by means of date of birth of the spouse and date of birth of father/mother (Married couple without or with children)
- Forming family households of divorced spouses with common children living at the (still) common main place of residence, by means of date of birth of father/mother (Divorced spouses with common children)
- Forming single parents with unmarried children, by means of date of birth of the father or mother (Single-parent household)
- Allocating further adult and directly related persons to family households (integration of single grandparents and transformation of a two-generation into a three-generation household)
- Forming households of adults not having been allocated yet and with the same reference to mother/father (Presentation as household of brothers/sisters)
- Forming one-person households (Any other adults without relationship suitable for linkage).

It was not possible to present by means of the household algorithm persons related in the collateral line (except brothers and sisters), persons related by marriage or persons not related (such as consensual unions) who are living together in one household. Such persons are always counted as persons living alone (one-person households). Persons with more than one dwelling were covered only at the place of their main residence and included only there in the ascertainment of family households.

To get still closer to the customary household concept (in the sense of living together), those person-related data records which had a dwelling number were in a further step segregated and further processed. For reasons of data protection, the original dwelling number had been replaced in an earlier step by an internal random number which however permitted to establish the connection between the dwelling and the household and thus to allocate to one household all persons having the same dwelling number. In this way it was possible to further examine in particular those cases where, by the household algorithm, at first two households had been allocated to one dwelling; these had been mostly households of one or several persons living together as partners in consensual unions with or without children. For the decision as to whether two households – which at first had been identified separately – should be combined to form one household, sex and marital status of the persons living together were taken into account, too. On the basis of these data, the original household figures, which had been overstated as compared with preceding censuses, could be adjusted. This was true in particular of the (overstated) number of one-person households, part of which – by means of the dwelling context – were grouped together to or integrated into multi-person households. The person-related data records with dwelling numbers (some 70 % of all person-related data records) were grouped together according to specific variables (e.g. household size, family type). On the basis of these data, the household figures for the entire territory of the former GDR were estimated by means of differentiated estimators per household size and federal Land.

When assessing the household figures determined with the help of registers of inhabitants, it must however be taken into account that such data are rough estimates because not all data are known which are relevant to the definition of a household. While it can be assumed for the microcensus results that all persons forming a living and house-keeping community are considered as part of one household, the registers of inhabitants do not indicate in particular which persons keep house together. The evaluation of the files of inhabitants is however mainly based on the direct kinship relations of the persons – as contained in the data records – and, in addition, largely considers the residence context of the households. It thus provides only approximate values on households.

European matters

Portrait of the Regions – A regional monograph by EUROSTAT

At the end of 1992, the Statistical Office of the European Communities (EUROSTAT) published a three-volume publication on the regions of the European Communities in cooperation with the EC Commission's Directorate General XVI, which is responsible for regional statistics. The publication has been issued in three languages (English, French and German) and is to address different target groups: in addition to the political and the administrative levels as well as enterprises and associations/chambers working in the field of management consultancy, it was also prepared for scientists at universities and research institutes, and not least of all for university students. It is of particular interest for persons or institutions dealing with regional issues or spheres of work, whether in the field of regional policy, regional science or for instance the promotion of regional economic activity.

An editorial committee with representatives of all member states had in cooperation with the EUROSTAT department responsible for regional statistics determined a common framework for characterizing the regions. While for reasons of comparability the tables and charts were prepared according to identical concepts, certain exceptions had to be made for the five new federal Länder and Berlin, as the complete basic data material, e.g. for value added, was not (yet) available for the territory of the former GDR. For the Federal Republic of Germany, each federal Land is portrayed on six pages, and each administrative district on two pages. So Bavaria and its seven administrative districts, for example, are presented on a total of 20 pages.

The text parts relating to the regions of the Federal Republic of Germany were prepared by staff members at the statistical offices of the Länder, also using common editorial guidelines, so that they follow the same system. But the contributions also comment upon the typical characteristics of each individual region, so as to give the reader an impression of the great variety of the regions within the European Communities. Included in the guidelines were, for instance, several topics to be covered and the request not to repeat the data included in the tables, but to offer additional interesting information.

Each presentation of a region covers 13 identical partial fields, from the description of the area to demographic aspects, the labour market situation, employment, education and training, the structure of the economy incl. the importance of the three large sectors agriculture, production industries, services, all the way to the state of the environment. Each presentation of a region also includes the "strong and weak points" as well as existing disparities.

Tables and charts, partly with data from Community surveys or statistics harmonized on the European level, partly from national surveys of the member states, provide the statistical basis of the publication. The quantitative data in the text parts mostly are also derived from official statistics for all member states. By the inclusion of maps and photographs in the common framework, the regional portrait is illustrated and presented with an elaborate layout.

The Portrait of the Regions may be ordered from the Office des publications officielles des Communautés européennes, 2, rue Mercier, L-2985 Luxembourg, at a price of Ecu 100 for one volume and Ecu 250 for all three volumes. As a rule, educational and training institutions will be granted a discount. Despite the time-consuming and costly work involved, it is planned to offer also the Portrait of the Community Islands, which is in preparation, at a "reasonable" price.

EUROSTAT Activity Programme for the statistical coverage of distributive trade in the EC member states

The Statistical Office of the European Communities (EUROSTAT) presently is developing a system for the statistical coverage of distributive trade which is to serve in all EC member states as a uniform reference framework for statistical work in the field of distributive trade. It is based on the Resolution of the EC Council of Ministers of 14 November 1989 (see Official Journal of the European Communities No. C 297 of 25 November 1989, p. 1) on intra-EC trade on the Single European Market. In this Resolution, the Council requests the Commission

- "to improve the statistical data on the distributive trade by ensuring the compatibility of these data with the Community definitions,
- to intensify – where necessary – the transmission of these data to the Statistical Office of the European Communities,
- to start soon with the sectoral programme planned for the field of distributive trade without increasing the administrative workload of the enterprises".

The activity programme for distributive trade comprises

- the monthly/quarterly collection of data on business activity (turnover, persons engaged) as well as the collection of structural data at intervals of one or several years (information on enterprises and local units; enterprise variables are, among others, turnover, purchases of commodities and services, stocks, personnel expenditure, investments, employment and trading links; for local units there are to be recorded turnover by products, employment and forms of trading), each on a sampling basis; both surveys will be based on an EC regulation; the present draft regulation provides for the introduction of the surveys in the member states within a transitional period from 1994 to 1996;
- the establishment of an enterprise panel of companies belonging neither to the finance nor the agricultural sector; it will include for the entire EC about 50,000 enterprises, among them 1,200 to 1,500 trading enterprises, and supply at short notice information on the business activity of enterprises, their strategies and policies which are not covered by other surveys;
- the provision of information on large enterprises, especially with a view to concentration developments.

Preparatory and accompanying measures are

- the compilation, standardization and dissemination of the data material already available on distributive trade, the most important instrument being the database MERCURE set up by EUROSTAT;
- conducting of pilot surveys in all member states to test statistical methods and definitions and at the same time to provide first results on the structure of retail trade in the EC; results for the Federal Republic of Germany are already available.

The methodological reference framework for these activities is the "Methodological Manual for Enterprise Statistics in the Service Sector", prepared by EUROSTAT in cooperation with the national statistical offices, which also contains a sector-specific chapter on the distributive trade and is thus to be tested.

Cooperation with science and research

Estimation from microdata according to the principle of minimum loss of information

Professor Merz (Lüneburg University) presented at the Federal Statistical Office his program for estimation from microdata according to the principle of minimum loss of information, which he had developed during his work in the Special Research Section 3 (Sonderforschungsbereich – Sfb) "Microanalytic Foundations of Social Politics" at Frankfurt University. This program enables the estimation on the basis of a sample survey by simultaneous adjustment of various, also hierarchical, variables to benchmark data from other sources, e.g. for a small sample survey of persons the simultaneous adjustment to various household and personal variables of the microcensus. The basic idea is that the new estimators ensuring the adjustment to the benchmark data should deviate as little as possible from the old estimators of the sample design. The loss of information (as derived from the entropy) is used as a measure of the deviation. This creates an optimization problem where a target function has to be minimized with collateral conditions being observed. For the nonlinear set of equations to be solved, Professor Merz developed a modified Newton procedure which accelerates convergence, so that the calculating time remains within reasonable limits. The program was used, among others, for estimation on the basis of data from the micro-simulation model of the Sfb 3 and for the socio-economic panel. The Federal Statistical Office plans to use it for estimation in connection with the 1991/92 Time Budget Survey.

Events

Seminar "Dissemination Policy for NSI Outputs"

At the end of January 1992, the Seminar "Dissemination Policy for NSI Outputs" brought together statisticians from the countries of the EC, the EFTA states, the Central and East European countries and also the former Soviet Union for an exchange of views in Wiesbaden. It was part of the Training Programme of European Statisticians, and the primary emphasis of the papers and discussions was on the subject fields public relations work and general information services, prices and costs of publications as well as on supranational issues of disseminating statistical results.

The introductory contribution on the guiding principle of the German press and public relations work "Statistics – Figures for Everyone" was followed by expositions of the common practice of releasing statistical results to the press and the general public in Great Britain and Northern Ireland. Another report about the dissemination of data material and study results provided interesting insights into the working methods of statistics in France. Further papers examined the role of the statistical institutes as service enterprises in the information society and, in this context, the market orientation of the presentation of statistical results. The representatives of the countries commented in particular on the experience acquired and the working methods used in this field. The work of information services (e.g. chambers of industry and commerce, online information offices, libraries), whose dissemination practice was also described, is of a specific nature and in this form does not apply to the statistical institutes of all countries.

Contributions on the role of EUROSTAT (Statistical Office of the EC) in the dissemination of statistical information and the observation of the international market for statistical information from the point of view of an international medium, in the present case, a big press agency, provided insights into the international dimension of disseminating statistical results and their great importance for the international information market.

Information – and this also applies to statistical information – is not generally compiled at no cost, and for this reason it cannot be offered free of charge beyond the statistical institutes' general obligation to provide information, which exists in most states. With comments on price policy and pricing, several papers touched upon aspects of the way in which prices are calculated by the statistical institutes as well as copyright issues and questions of competition in the private and public sector with regard to the dissemination of statistical information. The fact that marketing issues should not be ignored when statistical results are disseminated was stressed in the final contribution from the point of view of a marketing expert.

The main results of the papers and discussions of this seminar may be summarized as follows:

- Statistical results serving as a support of democratic societies must be reliable, objective and presented rapidly and in a comprehensible form, if a modern democracy is to work properly. Therefore, biasing in any direction cannot be accepted.
- Across national borders, aspects of the internationalization of disseminating statistical results are gaining in importance. In this context, training programmes for statisticians, standardization problems of data collection and presentation, the exchange of experience and publications also with East European countries are of particular interest. Over the last few years, a marked increase in demand for statistical information could be observed in all countries.

An exhibition of printed publications of the participating countries, supplemented by general information material for public relations work, offered an interesting framework for this event, which was regarded as positive by all participants.

Practical anonymity of microdata

Report on a research project

Preliminary remarks

After having been briefly outlined in number 1/1992, the project and in particular its results and the methods applied will be described in a more elaborate and detailed way in the present article.

The confidentiality of individual statistical data has always been a fundamental principle of the system of federal statistics. It protects the individual from revealing his personal and material situation and serves to maintain mutual trust between respondents and statistical offices. It is an essential prerequisite for the respondents' preparedness to answer questions and the reliability of the information given, and consequently the quality and validity of statistical results.

The Federal Constitutional Court, too, underlined the great importance of statistical confidentiality in its judgement on the population census. It regards the confidentiality of individual data not only as constitutive of the proper functioning of federal statistics, but also as indispensable for protecting the right of informational self-determination[1].

Pursuant to the principle of confidentiality, statistical offices may pass on individual information supplied by a respondent only in a form in which it definitely cannot be related to him, unless an exception is explicitly stipulated by the legislator. To comply with the condition of "absolute anonymity", individual information as a rule must be altered in such a way that this will be detrimental to the informational contents required for scientific purposes. For this reason, research institutions insisted on being granted a "privilege for science" allowing them for research purposes to get individual information from official statistics in "practically" instead of "absolutely" anonymized form.

The legislator has complied with this request. The Federal Statistics Law (BStatG) of 22 January 1987, Art. 16, Para. 6 admits that, under certain conditions, individual data be transmitted to institutions of higher education or other

[1] See Decisions of the Federal Constitutional Court, Vol. 65, p. 49 ff.

institutions entrusted with tasks of independent scientific research provided these data can be allocated only by employing an excessive amount of time, expenses and manpower.

The concept of practical anonymity as it is called is based on a definition of the European Science Foundation. When introducing this concept, the legislator did not specify how to ensure the practical anonymity of a given data stock. Therefore it was necessary to define the "criterion of excessiveness" in Art. 16, Para. 6 of the BStatG more concretely for the transmission practice of statistical offices.

To this end, a research project was planned and carried out jointly by Mannheim University, the Centre of Surveys, Methods and Analyses (ZUMA) in Mannheim and the Federal Statistical Office. The direction of the project was entrusted to Professor Walter Müller (Chair of Methods of Empirical Social Research and Applied Sociology at Mannheim University). The project, the results of which are presented in this article, was funded by the Federal Ministry for Research and Technology (BMFT) and accompanied by a project committee comprising representatives of the BMFT, the Data Protection Commissioners of the Federation and the Länder, the Federal Statistical Office and the statistical offices of the Länder as well as scientists.

Project objectives

The goal of the project was to develop concrete criteria for the concept of practical anonymity and to put them to the test under practical conditions. In this context, rational calculations, i. e. weighing the advantage of a successful reidentification and the cost incurred by an "investigator" were assumed to play a role in deciding upon disclosure attempts. The purpose was to provide information about the practical aspects of the "excessiveness" of time, expenses and manpower involved under the conditions prevailing in science and research. In economic terms, this was a cost-benefit analysis. With the help of the scenario method, a wide range of potentially realistic situations was covered by assuming different motives, disclosure strategies, additional knowledge, and disclosure methods and procedures.

Another aim was to use the results obtained in the scientific analysis to derive concrete recommendations for official statistics concerning anonymization and other protective measures for passing on individual data sets to research institutions.

Approach and methods used

The basic hypothesis is that a so-called investigator, in his capacity as a scientist illegally tries to reallocate anonymized individual data of official statistics to the persons concerned. To this end, he matches the official anonymized microdata file (AF) supplied to him with a non-official, non-anonymous identification file (IF) as it is called. The latter is also referred to as additional knowledge. During the matching procedure, he checks the values of the two files' coinciding variables for correspondence or great similarity. Provided data sets of the two files correspond to each other or are sufficiently similar, the investigator will be in a position to relate identifying elements such as names and addresses of the IF data sets concerned to the relevant AF data sets. Thus he would obtain additional information on individuals regarding variables included only in the AF, but not in the IF.

However, a reidentification as a result of such an allocation could be achieved only if a data set were related to only one rather than several data sets of the other file, i.e. if statistical doubles did not exist there with regard to coinciding variables. Biunique, i.e. one-to-one allocations will hence be required. But even in this case, a reidentification will not be guaranteed unless at least one of the two files comprises the complete population involved. If both files are only sample surveys, there may still exist a double of a biunique data set outside the files. This problem, however, will be reduced if the investigator is informed about the participation of IF persons in the AF survey, i.e. has response knowledge as it is called. And even if all these preconditions are fulfilled, reidentification attempts may nevertheless fail in practice because values of coinciding variables in biunique data sets may be incompatibly presented in the two files.

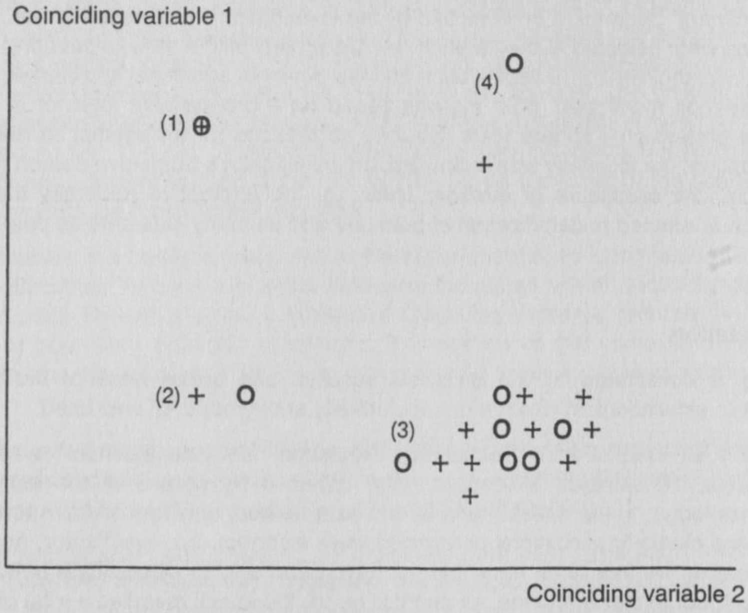
Practical aspects of this kind were an essential reason for checking in this project the risk and cost of reidentifications on the basis of empirical data. The core of the work therefore consisted in simulating allocations between real data files. An important element was the possibility of having a data trustee check in official statistics the number of persons simultaneously covered in both files and the number of them allocated correctly or wrongly or not allocated at all.

Simulations were based on two different data sources used as IF. On the one hand, "Kürschners Deutscher Gelehrtenkalender 1987" was used for this purpose and on the other, a comprehensive social science study served as an example of a data source which, though not being available to the general public, would be quite available to a scientist working in the field of empirical social sciences. The above data sources were selected because, after a detailed analysis of the potential additional knowledge, they belonged to those sources which were most likely to endanger the anonymity of the 1987 microcensus of North Rhine-Westphalia used as AF in both cases. The reason was a relatively large number of variables coinciding with the AF and their partly high information content which, as mentioned above, reduces the investigator's problem regarding doubles. The "Gelehrtenkalender" (Register of Scholars) contains about 10 coinciding variables incl. informative data such as the year of birth, profession, branch of economic activity and subject of the most recent university degree. The number of data sets of this IF, i.e. the number of scholars – mainly university teachers – in North Rhine-Westphalia amounted to approximately 8,000. The social science survey which was a representative population sample survey, even contained about 35 coinciding variables incl. not only personal but also household variables and information about the children of respondents. The kind and number of coinciding variables were particularly to represent the potential knowledge of an investigator about persons associated with him, such as colleagues or neighbours of whom he could know that they participated in the microcensus and therefore are included in the AF. The IF size, i.e. the number of data sets for North Rhine-Westphalia included in this study was 2,685.

Besides the two different data sources serving as IF, two also differing allocation algorithms were used. On the one hand, a simple matching procedure and on the other, a complex method developed by Paaß and Wauschkuhn[2] and based on discriminatory analysis were applied. The simple matching procedure just checks the two data sets to be compared for identical values of the coinciding variables, while the algorithm based on discriminatory analysis calculates the probability for data sets to belong on certain assumptions to the same person even though the values of the given coinciding variables may not be identical. It then does or does not perform allocations based on a lower limit of probability chosen by the investigator.

The basic idea of the method of discriminatory analysis and its difference as compared with the simple matching procedure is illustrated by the following chart[3].

DIFFERENT SITUATIONS OF ALLOCATION



+ Case in the anonymous data file.

O Case in the identification file.

(1), (2), (3), (4): Allocation situations; description in the text.

Federal Statistical Office 92 0165

For the sake of simplicity, not more than two coinciding variables are assumed. The simple method results in an allocation only in situation (1) which is the only situation with an identity of two cases of the two files. However, the algorithm based on discriminatory analysis would possibly produce an additional allocation in situation (2). Due to

[2] See Paaß, G./Wauschkuhn, U., "Datenzugang, Datenschutz und Anonymisierung", Munich, Vienna, 1985.

[3] See Müller, W./Blien, U. et al., "Die faktische Anonymität von Mikrodaten", Vol. 19 of the series "Forum der Bundesstatistik" issued by the Federal Statistical Office, Wiesbaden, 1991, p. 47.

the little difference, i.e. the great similarity of the two cases and the fact that they are clearly isolated from the other cases, there is a relatively high probability that they concern one and the same person whose coinciding variables were presented as incompatible in the two files. In situation (4), however, the difference between the two cases might be too big for a sufficiently high probability of correspondence though there is a clear isolation from the other cases as well. Finally, there is just the reverse situation shown in (3). Because of the local accumulation of cases the distance between which is small, the probability of two of them belonging together is rather low. The chart shows in particular that the joint distribution of coinciding variables plays an important role in the procedure developed by Paaß and Wauschkuhn for calculating the probability of two data sets belonging together, i.e. of their belonging to the same "class" in terms of discriminatory analysis. By the way, the measure used for the difference between two data sets was the Mahalanobis distance[4].

In addition to allocations with formally anonymous data, i.e. using as AF microcensus data only adjusted by eliminating identifying elements, additional anonymization measures were tested. To avoid data distortion, however, only the methods of presenting the values of variables in less detail and drawing sub-samples were used. Presenting the values of variables in less detail reduces the information contained in coinciding variables, thus increasing the problem of doubles for an investigator. Transmitting only a sub-sample of the microcensus reduces the information available to an investigator from his potential response knowledge because with a sub-sample, it is despite an obvious participation usually not certain whether the data sets of the persons concerned are also included in this reduced sample. This further aggravates the investigator's problem regarding doubles.

However, simulated allocations should empirically provide information not only about rates of success and failure of potential reidentification attempts, but also the expenses involved, considering BStatG, Art. 16, Para. 6, which refers to "employing an excessive amount of time, expenses and manpower". These expenses could then be compared with the costs of alternative methods of data acquisition, such as a survey made by an investigator on his own. Accordingly, the costs incurred were also recorded in greater detail during the simulation process.

Apart from these scenarios involving the simulation of allocations, three other situations were examined under consideration of the empirical results obtained. The imputed IF was the residents' register, data offered by an address trader and data on "prominent" persons to be compiled by the investigator from press reports, archives and the like. The total of five scenarios were selected at the beginning of the project with a view to covering the broadest spectrum possible, after the potential motives for reidentification and the potential additional knowledge serving as IF had been discussed. In the scenarios mentioned, both motives based on a professional logic of action in social sciences and others outside the professional sphere were included as reasons for an attempt to reidentify anonymized individual material. In addition, the following was examined: an investigator's purposive search for AF data of specific persons selected by him, the disclosure of arbitrary units, i.e. the attempt to reidentify the AF data of persons at random, and finally the attempted reidentification of both one and as many data sets as possible.

Simulation results and interpretation

The core of the project results is constituted by the empirical success and failure rates of the simulation of allocations. The data below refer to simulations in which a merely formally anonymous AF was used.

Regarding the scenario in which an excerpt from "Kürschners Deutscher Gelehrtenkalender" was used as IF comprising about 8,000 data sets, 14 biunique allocations were obtained by means of the simple matching procedure[5] without response knowledge. In the check made by the data trustee, only four of them turned out to be correct. Consequently, 71 % of the biunique allocations concerned were incorrect. An investigator, however, would not have been in a position to realize this fact since he would not have been able to differentiate between correctly and incorrectly allocated cases without further enquiries. As another result, the check revealed a total of 53 data sets simultaneously included in both files. On the assumption that all 53 participants were known, which is rather unlikely, the number of correct biunique allocations rose from four to nine because of the lack of doubles. Apart from a single one-many allocation, however, 43 data sets remained without any allocation at all as they were incompatible. Incorrect allocations did not occur when the participants were known. It should however be noted that the correct biunique allocations were obtained only by successively and alternatively coding the two possibly pertinent values of four coinciding variables and adding up the "success numbers" of the resulting 16 alternative combinations.

Costs were mainly incurred for coding the IF plain-language information in terms of the microcensus code, transferring the data to machine-readable media, developing the matching algorithm, implementing it in a computer system and carrying out the matching procedure there[6]. All in all, costs amounting to approximately DM 60,000 were incurred with regard to the above-mentioned 14 biunique allocations, ten of which were incorrect.

[4] See footnote 3, p. 73 ff.

[5] See footnote 3, p. 266 ff. for a detailed description of the results of this constellation.

[6] See footnote 3, p. 280 ff. for the costs.

Applying in the same scenario the more complex procedure of Paaß and Wauschkuhn based on discriminatory analysis was even more expensive[7]. The entire costs resulting from it were approximately DM 260,000. A substantial share of this amount was spent on adapting the algorithm to the concrete conditions of the experiment. Originally, the procedure had been developed for experiments with an IF synthetically generated from the AF[8]. And due to the more complex nature of the algorithm, the actual costs of computation were clearly above those of the simple matching procedure. However, it would be even more expensive for an investigator to newly develop a complex algorithm of this kind. With at most three correct biunique allocations (depending on the version used), the method based on discriminatory analysis was less successful than the simple algorithm[9] despite the higher costs involved.

For this reason, exclusively the simple matching procedure was tested with the second scenario[10]. Though not less than 35 coinciding variables were available then, not a single correct biunique allocation was achieved for one of the 2,685 data sets of the excerpt from the social science study functioning as IF. A total of 10 persons were included in both the AF and the IF. None of their data sets was correlated correctly so as to obtain a biunique allocation. This was even the case when it was assumed that the participants were known to the investigator, and thus was due to incompatibilities between the two data files. The total costs amounted to approximately DM 30,000, and it should be noted that apart from the AF, the IF was available in machine-readable form as well[11].

All in all, the empirical results of both scenarios reflect the strong effect incompatibilities of the data files to be matched have in practice on the probability of successful reidentification attempts. It is considerably lower than the probability that would result from a theoretical computation considering only the frequency of doubles while neglecting the above natural protection. It should be stressed, however, that with an increasing number of coinciding variables, the number of doubles becomes smaller while the number of incompatible data sets rises.

Conclusions and recommendations

The results of this project based on real data show that statements concerning the issue of 'reidentification risks', which have so far been based on purely theoretical considerations or experiments with synthetical data sets, must be relativized. Many of these statements resulting from anonymization research have been based on the assumption of complete and perfect information. Such imperfections as incompatibilities of most different kinds, e.g. changes over time, differing definitions, etc. have not been taken into account. The present results however clearly show that data incompatibilities of this kind in practice would lead to a noticeably smaller number of reidentifications than has been assumed so far. The risk of reidentification hence was overestimated since contrary to what has often been assumed, it is not sufficient to assess the risk exclusively on the basis of the actual or supposed number of unique cases in a microdata file (uniqueness concept).

The lack of doubles is a necessary but, due to the above-mentioned incompatibilities, not a sufficient prerequisite for correct reidentification. To be in a position to assess the actual risk of reidentification, two contrary factors must be taken into account. Though a growing number of coinciding variables reduces the number of doubles, thus raising the number of potentially biunique allocations, it increases at the same time the probability of incompatibilities occurring between the AF and IF which, in turn, will lead to an increasing number of non-allocations or incorrect allocations.

Apart from the question of incompatibilities, discussions have often neglected the fact that it is not the absolute number of transferred survey variables which is essential regarding the risk of reidentification, but always only the number of those variables which are part of the IF and AF intersection.

New knowledge has also been obtained by quantifying the costs and effort involved for a potential investigator. Whatever motives and strategies were assumed, the result has always been that there are by far more economical alternatives of acquiring information than attempting to reidentify anonymized individual data. Besides, these alternatives have the advantage of being legal. The above assessment and calculations do not only apply to the allocation procedure of Paaß and Wauschkuhn which is rather complex in methodological terms and very costly, but also to the simple reidentification procedures. In the case of simple procedures, the most important cost factor is not the software or the computing time, but the availability of a comprehensive address-related identification file with faultless coinciding variables that are deeply broken down and correspond exactly to the variables of the official survey in terms of time and contents. Only if such data files became accessible to scientists easily and at low cost, new investigations of the costs and efforts involved and possibly also new or additional anonymization measures would be required.

[7] See footnote 3, p. 84 ff. and p. 310 ff. for the costs.

[8] See footnote 3, p. 59 ff.

[9] See footnote 3, p. 295 ff. for a detailed description of results.

[10] See footnote 3, p. 329 ff. for a detailed description of results.

[11] See footnote 3, p. 349 ff. for costs.

Against the backdrop of these results, the potentially most risky data constellation with respect to an attempted reidentification of practically anonymized individual data material can be characterized as follows[12]:

- The AF contains very detailed regional information so that only few members of this specific subpopulation live in the regional units concerned (detailed regional breakdown).
- A person searched for in the AF belongs to a very small subpopulation which can be delimited by a specific variable (detailed breakdown of all other variables).
- The investigator knows that the person he is in search of and has or can obtain information about, is included in the AF (knowledge of participation).
- The variables of the person concerned are recorded in the AF just in the way the investigator presumes (compatibility).

To consider even this data constellation, which certainly is very rare, and to practically ensure that individual data passed on to research institutions cannot be reidentified, the project council adopted unanimously the general precautions and measures specifically relating to data files as indicated in the table below. Its recommendations were based on the transmission of data of the sample survey of income and expenditure (EVS) and the microcensus[13].

The majority of the general precautions concerned have already become part of the BStatG. In the microcensus, a differentiation between a basic and a regional file was made in order to make scientists benefit from the opportunity of regionalizing the results of this survey, which was further enhanced by the new sample design. Passing on regional variables in a detailed breakdown involving a relatively great identification potential requires a less detailed breakdown of all other variables as compared with the basic file which is regionalized down to the federal Länder level only.

Implementing the results in practical data transmission

To ensure that the results of the research project are uniformly implemented by the statistical offices of both the Federation and the Länder in practical data transmission, the Federal Statistical Office, closely following the project recommendations, worked out guidelines for transmitting microdata sets of the EVS and microcensus pursuant to Art. 16, Para. 6, BStatG. These guidelines were approved by the heads of the statistical offices in the form of a joint catalogue of measures for practical anonymization.

The Federal Statistical Office additionally drafted a Standard Agreement on Transmitting Practically Anonymized Statistical Microdata, which integrated the general precautions recommended. Besides, an internal regulation stipulates how the other basic conditions laid down by the legislator will be guaranteed. Thus for instance, the scientific nature of a project must be established in a project outline. In cases of doubt, proof must be given of the independent scientific nature of the research work by submitting the standing rules, statutes or similar documents concerned.

The project, its results and plans to implement them in practical data transmission were presented to representatives of science and research interested in this issue at a conference convened by ZUMA in December 1991. However, the project also met with great interest beyond the borders of the Federal Republic of Germany. Selected parts of the study were presented at both the International Symposium on Statistical Disclosure Avoidance convened by the Netherlands Central Bureau of Statistics at Voorburg and the International Seminar on Statistical Confidentiality jointly convened by EUROSTAT (Statistical Office of the European Communities) and ISI (International Statistical Institute) in September 1992. The detailed project report entitled "Die faktische Anonymität von Mikrodaten" (The Practical Anonymity of Microdata) was published in volume 19 of the series "Forum der Bundesstatistik" issued by the Federal Statistical Office in 1991.

[12] See footnote 3, p. 435.

[13] See footnote 3, p. 440 ff.

Recommendations concerning anonymization measures for transmitting practically anonymized microdata of the microcensus and the sample survey of income and expenditure

General precautions:

Contract with the recipient of practically anonymized data stipulating the following:

- appropriate technical and organizational measures to supervise the use of data;
- contractual penalty in the case of a reidentification attempt;
- restriction of use to the scientific project specified;
- no transmission of data to third parties;
- deletion or return of data upon completion of the scientific project;
- treatment of data excerpts or copies like originals;
- no enquiries regarding the local implementation of sample designs.

Confidentiality of the local implementation of the sample designs by official statistics.

Data are not arranged systematically.

Anonymization measures for the microcensus basic file:

Regional data of only little detail (only federal Land and type of settlement structure or larger community size class):

- it must not be feasible to identify an individual community of less than 500,000 inhabitants;
- a community type comprising several communities must not have less than 400,000 inhabitants in any of the Länder.

It must be impossible to identify any nationality or group of nationalities of less than 50,000 members living in the Federal Republic of Germany.

If required, the values of all other variables must be presented in less detail and in such a way that each variable value shown in the univariate marginal distribution comprises for the Federal Republic of Germany at least 5,000 cases.

Only the data of a 70 % subsample are passed on.

Anonymization measures for the microcensus regional file:

It must not be possible to identify a regional unit of less than 100,000 inhabitants by combining regional classifications.

Less detailed analysis of such variables as profession, branch of industry, nationality and age to ensure that values of variables are not presented if

- the share in the basic population of the Federal Republic of Germany is less than 50,000 inhabitants,
- the microdata file comprises less than three cases per regional unit transmitted (excl. subsampling).

If required, all other variables are aggregated in such a way that each variable value presented comprises at least 5,000 cases of the basic population of the Federal Republic of Germany.

Only the data of a subsample of (at least) 85 % are passed on.

Anonymization measures for the sample survey of income and expenditure:

Regional data like with the microcensus basic file.

Nationality like with the microcensus basic file.

Less detail of the values of variables:

- As regards "visible" variables or variables remaining stable over time (e.g. year of birth, status in occupation or ownership of conspicuous consumer goods):
  Less detail of the values of variables like with the microcensus basic file.
- As regards variables that are little known in public or rather instable, but for which a large number of values are recorded (particularly income, property and expenditure):
  The five lowest and the five highest values of a variable are presented as mean values. The other values of the lowest and highest deciles of the distribution of such a variable are blurred with a random error of up to plus or minus one per cent of the variable value concerned.

Subsample depending on the number of survey parts transmitted:

98 %:	household and personal variables + 1 survey part
90 %:	household and personal variables + 2 survey parts
80 %:	household and personal variables + 3 survey parts

Foreign-Language Publications

English

Survey of German Federal Statistics

The "Survey of German Federal Statistics" is the most important compendium of information on federal statistics. The present edition primarily comprises updated summary contributions on the organization of federal statistics, their legal foundations, tasks and objectives as well as their implementation, on public relations work and the cooperation with international organizations.

Published at irregular intervals.

Present and Future Tasks of Official Statistics

Non-recurrent publication.

Statistical Compass

This brochure presents a selection of major benchmark figures from all subject fields along with comparative figures for back years.

Annual publication.

Foreign Trade according to the Standard International Trade Classification (SITC-Rev. 3) – Special Trade – until 1987 SITC-Rev. 2

This publication comprises the foreign trade figures according to the SITC-Rev. 3 with data by countries of origin/destination.

Annual publication.

Studies on Statistics

Published at irregular intervals. Issues which are still available:

No.	Title
36	Statistical Information System of the Federation
37	Surveys and Registers
38	Indices of Production and Productivity
39	Concentration Statistics
40	Kind-of-Activity Units in Mining and Manufacturing
41	Dissemination of Statistical Information
42	Indices of Orders Received and Unfilled Orders
43	Calendar Adjustment of Time Series
44	Information Campaign for the Population Census 1987

French

Aperçu de la statistique fédérale allemande

Cette édition abrégée de 1976 a été préparée surtout à l'intention des utilisateurs désireux de se renseigner sur les grandes lignes des activités statistiques plutôt que sur tous les détails. Elle contient donc de la version intégrale l'ensemble des textes décrivant les buts, les bases, les méthodes et les résultats de la statistique fédérale.

Publié à intervalles irréguliers.

Boussole des chiffres

Cette brochure comprend une sélection des principaux chiffres de référence de tous les domaines ainsi que des chiffres comparatifs pour des années antérieures.

Publication annuelle.

Spanish

Guía Estadística

Este folleto contiene una selección de datos importantes en todos los campos así como los datos comparativos de los años anteriores.

Publicación anual.

Trilingual

Trilingual List of Statistical Terms (German – English – French)

Non-recurrent publication.

List of Major International Abbreviations (German – English – French)

Published at irregular intervals.

The publications of the Federal Statistical Office may be obtained direct from the publishers Metzler-Poeschel Verlag, Delivery: Messrs. Hermann Leins, Postfach 11 52, D-7408 Kusterdingen. A detailed list of publications may be ordered from the Federal Statistical Office.