Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

06/2012
EN

Methodological aspects of labour market data

Data protection at the Research Data Centre

Daniela Hochfellner,
Dana Müller,
Alexandra Schmucker,
Elisabeth Roß

Bundesagentur für Arbeit

Data protection at the Research Data Centre

Daniela Hochfellner (IAB)
Dana Müller (IAB)
Alexandra Schmucker (IAB)
Elisabeth Roß (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Zusammenfassung

Forschungsdaten, die aus dem Bereich der Bundesagentur für Arbeit (BA) bzw. aus den Befragungen des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) stammen, sind für Wissenschaft und Politikberatung von zunehmend hoher Bedeutung. Zahlreiche Forschungsfragen aus der Arbeitsmarkt- und Berufsforschung lassen sich mit diesen Daten beantworten. Es handelt sich um Sozialdaten, die den Datenschutzbestimmungen des Sozialgesetzbuches X (SGB X) bzw. den Regeln der statistischen Geheimhaltung unterliegen. Das SGB X und das SGB III räumen unter bestimmten Voraussetzungen Nutzungsrechte ein. Auch Forschungsvorhaben externer Forschungsinstitute, die diese aus eigenem Antrieb oder z.B. im Auftrag des Bundesministeriums für Arbeit und Soziales (BMAS) durchführen, profitieren davon. Um der Wissenschaft Sozialdaten leichter zugänglich zu machen, wurde das Forschungsdatenzentrum der BA im IAB (FDZ) geschaffen. Daten unterschiedlicher Anonymisierungsgrade stehen dort datenschutzgerecht über standardisierte und transparente Wege zur Verfügung. Ziel dieses Artikels ist es, das Spannungsverhältnis zwischen Forschungsinteressen einerseits und Datenschutz andererseits sowie die praktische Umsetzung der ausgleichenden Maßnahmen darzustellen.

Abstract

Research data of the Federal Employment Agency as well as surveys of the Institute for Employment Research are highly relevant for the scientific community and policy consulting. These data help to find answers to various research questions regarding employment and occupational research. The legal basis for data access is mainly Section 67 of the German Social Code Book X (SGB X). Since the establishment of the Research Data Centre (FDZ) of the Federal Employment Agency (BA) in the Institute of Employment Research (IAB) social data of the BA and the IAB are accessable for researchers using standardised and transparent principles. The remainder of this paper is to discuss the trade-off between the capability of research interests and data protection as well as the satisfaction of these demands.

Keywords: data protection regulations, social data, anonymisation

Acknowledgements: We would like to thank our colleagues at the FDZ for their helpful comments and ideas, and Johanna Eberle for the technical support. In addition, we wish to thank Felix Ritchie (UK Office for National Statistics) for numerous ideas and lots of information on the subject.

1 Introduction

The importance of research data for the scientific community and policy consultation is indisputable. Demand for comprehensive datasets, supplemented by additional information from other data sources, is growing constantly. In particular process-generated data are becoming increasingly attractive for social research as a result of their advantages. In contrast to survey data, administrative data are censuses in which highly reliable information is collected, usually over long periods of time. The common problems that arise with survey data, such as non-response, panel attrition, recall lapses and errors, therefore do not occur. The Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) makes available for research purposes extensive administrative social data that are especially suitable for analyses in the field of labour market research[1]. They comprise the pool of data from the administrative processes of the BA, the data from the local authorities responsible for administrating basic social security, and the data from the social security notification procedure. All these data are merged, consolidated and prepared via the procedures of the BA Statistics Department, and are then processed further at the IAB and made available as standardised research datasets at the FDZ.

Social data, however, underlie the special protection of data privacy in the field of social security (Sozialgeheimnis) (§35 para. 1 clause 1 of German Social Code Book I), as these data constitute mandatory information that is required for calculating contribution levels and subsequent entitlements associated with social insurance (e.g. pension insurance). From a legal point of view, two conflicting constitutional principles oppose each other here: on the one hand the right to informational self-determination and on the other hand the academic freedom that is established in Germany's Basic Law (Grundgesetz [GG], Art. 5). In order to find a balance between these two aspects, a legal basis was created which permits a scientific utilisation of the data while complying with data protection laws at the same time (§75 Social Code Book X SGB X, §282 para. 7SGB III). In practice, however, this results in a conflict of aims between the data having the largest possible analysis potential and the existence of maximum data protection: the more available information there is, the greater the analysis potential is. On the other hand, as the information content increases, so too does the risk of de-anonymisation. In order to allow standardised data access for research purposes while complying with data protection legislation, the FDZ was set up in 2004 on the recommendation of the Commission to Improve the Informational Infrastructure by Co-operation of the Scientific Community and Official Statistics (Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik - KVI).

The following pages outline how the FDZ solves the conflict of aims described above in practice. To this end the FDZ uses different methods, from preparing standardised data, through the modes of access, to the monitoring of output.

[1] In addition to the administrative data, a substantial amount of data is also available from large-scale surveys, some of which are supplemented by information from the process-generated data. The same conditions apply for the use of these data as for social data.

2 Why data protection?

First of all the question arises as to why the FDZ data require protection at all. The legislation grants the scientific community the right to use social data only subject to the conditions of legally standardised data protection regulations (§§ 67ff SGB X). It is only possible to conduct research using the sensitive social data if these regulations are complied with. Data protection is thus inseparably connected with the granting of permission to use the data. The following sections address the legal basis and the task of the FDZ associated with this.

2.1 Legal background

Since the decision of the Federal Constitutional Court regarding the census[2], the constitutional basis of data protection is recognised in the right to informational self-determination. According to this, everyone can in principle personally determine the use and disclosure of his or her personal data. Any restrictions to this are only permitted on the basis of statutory provisions. The data protection regulations safeguard this right by permitting certain intrusions and simultaneously setting limits. What must be emphasised is the key term "personal data", which encompasses every piece of information about an identified or identifiable natural person ("data subject"). The protection of social data contained in the 2nd chapter of Social Code Book X (SGB X) is more narrowly defined than the protection of personal data according to the Federal Data Protection Act (Bundesdatenschutzgesetz - BDSG). The reason for this is the fact that social data[3] are not collected from data subjects on a voluntary basis but comprise mandatory information. For example, anyone applying for unemployment benefit only receives it if he or she provides personal details. The data subjects are under a legal obligation to accept the processing of their personal data. Firms, too, are obliged to reveal information about themselves - in the context of the social security notification procedure.

In accordance with § 35 Social Code Book I (SGB I), social data underlie the principle of data privacy in the field of social security (Sozialgeheimnis)[4]. Prohibition with an authorisation proviso applies. This means that everything that is not expressly permitted by law with regard to handling social data is prohibited and everything that is prohibited constitutes, without exception, an administrative or criminal offence. Every case of use, storage, transmission and disclosure therefore requires a legally recognised justification. Unlike in

[2] The decision taken by the Federal Constitutional Court in 1983 regarding the census reads ("Census verdict" BVerfGE 65, 1): "(...) Under the modern conditions of data processing, the right to the free development of one's personality assumes the protection of the individual against the unrestricted collection, storage, use and disclosure of his or her personal data. This protection is therefore covered by the fundamental right of Art. 2 para. 1 in connection with Art. 1 para. 1 of the German Basic Law (Grundgesetz - GG). In this respect the fundamental right guarantees the authority of the individual to decide in principle him or herself about the disclosure and use of his or her personal data."

[3] "Social data are particulars regarding personal or factual circumstances of an identified or identifiable natural person (data subject) that are collected, processed or used by an authority as mentioned in Section 35 of Social Code Book I with regard to its duties in accordance with this code of law" (§ 67 para. 1 SGB X).

[4] "Everyone has a right to the social data pertaining to his person (§ 67 para. 1 SGB X) not being collected, processed or used by the social security agencies without authorisation (data privacy in the field of social security - Sozialgeheimnis) (. . .)" (§ 35 para. 1 SGB I).

the Federal Data Protection Act, in the Social Code sensitive data include not only personal information but also trade and business secrets [5].

In order to create a balance between the constitutionally guaranteed academic freedom[6] on the one hand and the constitutionally established right to informational self-determination on the other hand, the legislator specified conditions for access to social data by the research community in § 282 para. 7 SGB III and § 75 SGB X.

- In § 282 para. 7 of SGB III it is laid down that the Federal Employment Agency (Bundesagentur für Arbeit - BA) may transmit (factually) anonymous[7] data to external research institutions for the purpose of employment research. These so-called scientific use files (SUF) contain microdata which have been aggregated in such a way that de-anonymisation would only be possible with a disproportionate amount of time, expense and effort.[8] The SUFs can be used to provide answers to a multitude of research questions. Although the risk of de-anonymisation is very small, the use of these datasets is linked to certain conditions (see Chap. 3).

- As the analysis potential of the SUFs is restricted by the anonymisation measures, certain questions can no longer be answered using these data. The FDZ has therefore created weakly anonymised datasets that are provided at special separate workplaces for guest researchers[9] or can be analysed by means of remote execution[10]. In legal terms, access to social data at these guest researcher workplaces constitutes a "transmission of data" (§ 67 para. 6 No. 3b SGB X) and therefore requires authorisation by the Federal Ministry of Labour and Social Affairs (Bundesministerium für Arbeit und Soziales) in accordance with § 75 of the German SGB X. The FDZ has developed a standardised procedure for this (see Chap. 3)[11].

2.2 Data protection as a task of the FDZ

One of the key tasks of the FDZ besides compiling and documenting research data is the practical implementation of so-called statistical disclosure control (see Ritchie (2011)). This

5 "Trade and business secrets are equal to social data" (§ 35 para. 4 Social Code Book I SGB I).

6 "Art and science, research and teaching are free. Freedom of teaching does not absolve from loyalty to the constitution." (Art. 5 para. 3 German Basic Law (GG))

7 "'Anonymisation is the modifying of social data in such a way that the particulars about personal or factual circumstances can no longer be attributed to an identified or identifiable natural person or that this can only be done with a disproportionate amount of time, expense and effort." (§ 67 para. 8 SGB X)

8 "For the purpose of scientific projects, the Federal Statistical Office and the statistical offices of the Länder may transmit individual data to institutions of higher education or other institutions entrusted with tasks of independent scientific research if the individual data can only be attributed to a person with a disproportionate amount of time, expense and effort, and if the recipients are public officers, persons specially sworn in for public service or persons obligated according to section 7." (§ 16 para. 6 Federal Statistics Law BStatG)

9 The computers for guest researchers at the FDZ are configurated PCs that have no access to the Internet and do not permit the transfer of data to external storage media or printers.

10 For this, researchers prepare evaluation programs on the basis of test data. At the FDZ the evaluations are conducted using the original data and the results are sent to the researcher after verification of compliance with data protection legislation.

11 § 75 SGB X regulates the transmission of social data to third parties in general. In addition to the standardised data access via on-site use which is offered by the FDZ, there is still the possibility of an individual project-specific data transfer via the Federal Employment Agency subject to a charge.

is understood as safeguarding the confidentiality of information about statistical units, e.g. individuals or businesses. In order to guarantee this, the risk of identifying an individual from a piece of information must be checked before the information is released. It is not only information such as names, addresses or social security numbers that is regarded as highly risky but also characteristics or combinations of characteristics that make it possible to identify an individual indirectly. Although it can be assumed that researchers are not at all interested in identifying individuals or firms, statistical disclosure control must also be observed when transmitting research data. The aim of this is to prevent the publication[12] of information with which third parties would be able to recognise certain individuals or firms, or data subjects can identify themselves. It must be taken into account in this context that the data subjects themselves as well as third parties could possess additional knowledge that makes it possible to identify individuals in the data.

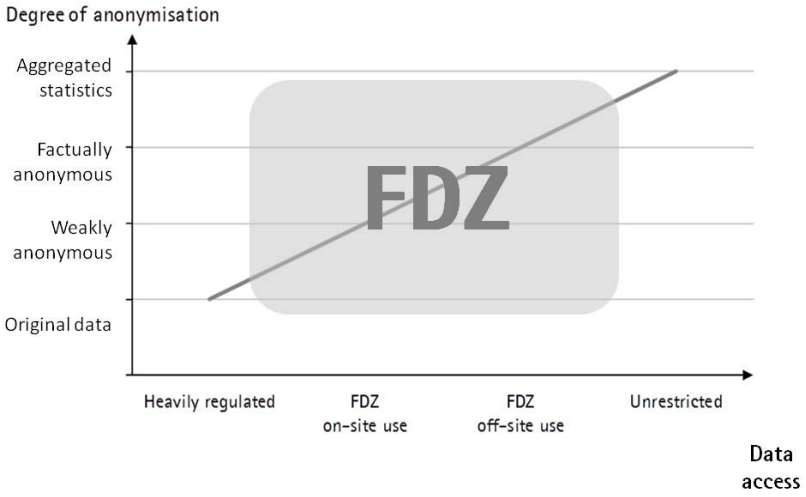
An example can illustrate this problem: a scientific publication shows the mean income of employed female dentists by district. Due to a 2% sample, the calculation is based on only one or two individuals for many districts, especially for those with a small population. As the profession of a dentist is generally practised on a self-employed basis, it is quite possible that there is actually only one employed female dentist in a district and by coincidence she is included in the sample. Generally neither the researchers nor the staff of the FDZ will possess this additional knowledge of there being only one employed female dentist in the district, but the dentist concerned and the inhabitants of the district may know it. It is therefore possible for the dentist to identify herself and for third parties to find out her income. A similar problem exists if there are only two dentists and both of them are contained in the sample. Here, third parties are not able to see the individual dentist's income, but the two dentists concerned are each able to calculate the other dentist's income using their knowledge about their own income. For firms the risk of deanonymisation is far greater, as additional information about firms is easy to access. Especially large enterprises can be identified easily via the details regarding industry and location.

[12] Publication also includes making information available to unauthorised third parties.

3 The FDZ data protection portfolio

The task of the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) is to safeguard the anonymity of the statistical units. This is always associated with the aggregation level of the information that is to be protected. Generally speaking, the spectrum of degrees of anonymisation ranges from the original data to strongly aggregated statistics. The mode of data access depends on how strongly the data have been anonymised. In some cases, for example, aggregated statistics can be published freely on the Internet, whereas the original data are only transmitted following detailed verification and only if absolutely necessary. This correlation is made clear in Figure1.

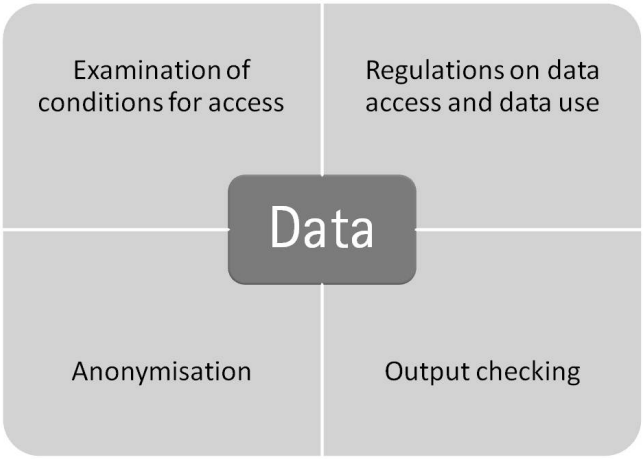
Figure 1: Degree of anonymisation and data access



Source: own representation

From the figure above it becomes evident that there is a broad range both between original data and aggregated statistics and between unrestricted and heavily regulated data access. These aspects determine the scope for action of the institutions holding the data. As the FDZ makes available neither original data nor aggregated statistics, it has a somewhat smaller radius of action, within which it can, however, respond flexibly to different requirements. For instance, the FDZ offers its microdatasets as factually anonymous scientific use files (SUF), which the users can analyse on the premises of their research institution. Alternatively it is also possible to use weakly anonymous data, which contain detailed information, in the context of research visits to the FDZ, or via remote execution. Generally speaking, the higher the level of anonymisation, the more flexible the data access is. The more sensitive the information is, the more strongly regulated the data access is. For all modes of data access the FDZ ensures data security by means of various procedures that were developed in collaboration with the legal department of the IAB. In order to coordinate this, the FDZ works in accordance with a portfolio approach following Lane/Heus/Mulcahy (2008). The four main fields of the FDZ portfolio are shown in Figure 2.

Figure 2: Four-field portfolio



Source: own representation

Basically we distinguish between measures implemented before the data are used, and those that take place following data use. Prior to the use of the data, first the conditions for access are checked and the conditions of data use are stipulated by contract. Second, the microdata are rendered anonymous in such a way that guarantees data protection. In addition, the results of analyses conducted using the weakly anonymous data are checked to verify compliance with data protection legislation. The details of the individual aspects are outlined below and summarised in Table 11 in the Appendix.

3.1 Examination of conditions for access

In accordance with the legal regulations (Chap. 2), the use of FDZ data is linked to certain conditions. In order to clarify whether these conditions are met, a request for data access must first be submitted. The following formal requirements are then checked by the FDZ: first, the data must be required for a research project in the field of employment research. The most important aspect is whether it is absolutely necessary to use these data. Proof must be provided that the research objective cannot be achieved with any other data that are more easily available (e.g. aggregate data). In addition to the formal examination, a description of the research project is also required, which serves to check whether the research project is feasible in terms of content in relation to the data applied for. Here the staff of the FDZ advise the applicant with regard to the analysis potential and quality aspects of the data.

Depending on the type of data access, further conditions must also be met: when applying for a SUF, proof must be provided that the institution conducting the research is an independent scientific research institution. Furthermore, the applicant must explain in a data security concept that adequate technical and organisational measures exist in the institution for storing and processing the data safely. When applying for a research visit in order to use the weakly anonymous data, the research project also has to be of public interest. After the request for data access has been checked by the FDZ as regards contents, it is

submitted to the Federal Ministry of Labour and Social Affairs for approval. The conditions for access are less stringent in cases where the researchers use only remote execution, as they do not have direct access to the microdata in this case. Nonetheless, here, too, the data may only be used for a scientific research project related to employment research or to the social security system[13].

3.2 Regulations on data access and data use

After the data request has been checked and approved, data use agreements are concluded in which the conditions for using the data are regulated. The key principles of the limitation of data use to specific purposes, the time limitation and the specification of the data to be accessed and the individuals entitled to access are defined in all of the data use agreements. Consequently, the use of the data requested is only permitted for a specific project with defined contents within the period stipulated in the agreement. Furthermore, the individuals who are entitled to access the data are also specified. This group of persons is to be kept as small as possible. In addition, the data use agreements contain bans on disclosing the data to third parties, linking the data with other microdata and de-anonymisation.

The data use agreements differ depending on the type of data access. For instance, in data use agreements for SUFs the data security concept of the research institution is an additional component of the agreement. Furthermore, the research institute is obliged to delete all microdata after the end of the contract period and, if applicable, to return to the FDZ any data carriers on which the data were transmitted. The data use agreement for on-site use contains guidelines regarding conduct during the research visit. As when weakly anonymous data are used, data protection is additionally ensured by subjecting the results to statistical disclosure control (see Chap. 3.4), the data users are bound by contract to refrain from recalculating their results. This means that they are not allowed to recreate values of a table which were deleted during the process of statistical disclosure control by comparing them with previously reviewed values from earlier output or other similar procedures[14]. In addition to the conditions for using the data, all the data use agreements also contain information regarding penalties for misuse.

3.3 Anonymisation

Besides the restrictions on data access, data protection is already taken into account during the data preparation process. One important step aimed at safeguarding the data is the drawing of a sample. This alone reduces the risk of de-anonymisation substantially. When, as is the case, for instance, with the Sample of Integrated Labour Market Biographies (Stichprobe der Integrierten Arbeitsmarktbiografien - SIAB), there is a sampling probability

[13] Detailed information as to what is asked in the requests for data access as well as information on how to prepare the requests can be found on the FDZ website under `http://fdz.iab.de/de/FDZ_Data_Access.aspx`.

[14] Weakly anonymous data can either be analysed during research visits to the FDZ or by means of remote execution. A data use agreement for on-site use always also covers data use by means of remote execution.

of 1:50 and someone believes they have identified a person, there is the possibility that the population includes another 49 individuals with the same characteristics. In order to identify a person definitely in the sample it is therefore necessary to possess the additional knowledge that either the person's characteristics are unique in the population or that the supposedly identified person is contained in the sample. In addition to drawing samples, the FDZ also uses other anonymisation methods. Three levels of anonymisation are distinguished for this:

- weakly anonymous
- factually anonymous or
- absolutely anonymous.

In the case of the weakly anonymous data, identifiers such as name, address, social security number or establishment number are deleted and the attributes of some particularly sensitive variables, for example nationality, are aggregated[15]. As the risk of de-anonymisation is considerably higher in the case of large enterprises or industry leaders as a result of detailed information about the economic activity and location, only the groups of economic activity (3-digit level of the classification of economic activities) and the federal state (Bundesland) are provided as a standard. In justified cases, however, it is also possible to use information at the 5-digit level (sub-class of economic activity) and about the district (Kreis). These sensitive variables generally serve either for merging aggregated statistics at this level (e.g. unemployment rates by district) or for creating separate regional or industry-specific groups that are not included in the given classifications. Very specific analyses of certain sub-classes of economic activity and/or small regional units are frequently not approved due to the very high risk of de-anonymisation.

Scientific use files are factually anonymous microdata whose information content has been reduced to the extent that de-anonymisation would only be possible with a disproportionate amount of time, expense and effort[16]. Here it is often necessary to decide which variables should be aggregated. If, for example, detailed regional information is to be retained in the employment data, other variables (such as establishment information) have to be strongly aggregated or even deleted instead. In general the recommendations made by Müller et al. (1991) are taken into account when generating the SUFs.

Aggregate data in which it is impossible to identify either individuals or establishments - including large enterprises and industry leaders - are regarded as absolutely anonymous. A table of results from a microdataset need not automatically be absolutely anonymous. If, for example, individual cells contain only one person, then even an aggregated table is not anonymous. As problems of this kind frequently arise when analysing weakly anonymous

[15] Weak anonymisation therefore goes one step further than the pseudonymisation of data. "Pseudonymisation is the replacement of a name or other identification characteristics by an indicator with the purpose of precluding the identification of the data subject or of making this significantly more difficult." (§ 67 para. 8a SGB X)

[16] "Anonymisation is the modifying of social data in such a way that the particulars about personal or factual circumstances can no longer be attributed to an identified or identifiable natural person or that this can only be done with a disproportionate amount of time, expense and effort." (§ 67 para. 8 SGB X)

data, these results have to be subjected to an output check after the analysis (see Chap. 3.4). Furthermore, so-called campus files are regarded as absolutely anonymous. Campus files are microdatasets that prevent any of the individuals or firms contained from being identified by means of information reduction and data modification procedures. As a result of this major intervention, however, these files are no longer suitable for content-specific analyses but only serve for teaching survey, data management and analysis techniques at universities and research institutes (see Kirchner/Gschwind (2011)).

3.4 Output checking

As the weakly anonymous data are still social data and there is thus still a residual risk of de-anonymisation on the basis of tables of results, it is necessary to check the output despite the data use agreement. In order to be able to check the results as quickly and efficiently as possible, the evaluation programs have to be in accordance with certain guidelines laid down by the FDZ. However, this should not limit the researchers in their use of analysis methods. But this means that the output checks can not be integrated entirely into a standardised and automated procedure. Instead, the statistical disclosure reviews always have to refer to the individual case at hand. This involves the staff at the FDZ viewing all results before they are published. The review is conducted in accordance with certain criteria and rules. Among experts there are generally accepted rules that are always to be used[17]. The following chapters describe how this statistical disclosure control should be put into practice.

[17] One example of this is ESSnet. This is an international project of the European Statistical System, which deals with all areas that can be associated with data protection, such as the publication of standards that should be observed in statistical disclosure control. (`http://neon.vb.cbs.nl/casc/handbook.htm`)

4 Statistical disclosure control at the FDZ

All results generated on the basis of the weakly anonymous data are subjected to statistical disclosure control before being transmitted to the data users. The effort involved in checking the results depends on the output files that have to be checked.

4.1 Theoretical differentiation of analysis results

The results produced can be classified as "safe" or "unsafe" according to their contents. This differentiation is used to deduce how high the risk is of the data material being de-anonymised. In the case of results that are classified as safe it is assumed that there is no risk of the data material being de-anonymised. In contrast, there is a residual risk of de-anonymisation in the case of analysis outputs classed as unsafe. For this reason, values are deleted from theses analysis results until the tables of results are absolutely anonymous. The purpose of the statistical disclosure control is therefore to transform analysis outputs that are classified as "unsafe" into completely "safe" outputs. The output is classified on the basis of the following aspects in accordance with Brandt et al. (2010):

- the data material used
- the type of analyses conducted
- restrictions of the data material to certain variables or inclusion of certain variables
- data transformations applied

If the output has been classified as safe with regard to the data protection measures of the FDZ on the basis of this classification, nothing is deleted from the results. An example of output that is generally safe is coefficients of multivariate estimates for large populations. In most cases, however, the analyses are results that can not be classified directly as "safe", which is the case in particular with descriptive evaluations. Outputs that are regarded as unsafe in principle are those that display

- statistical indicators, such as means,
- individual data points, e.g. in a scatter plot (these may permit conclusions to be drawn about an individual),
- the percentiles and
- the number of observations.

The classification into "safe" and "unsafe" outputs does not mean that safe outputs are transmitted without being checked beforehand, but only that no values are deleted in "safe" outputs. There are guidelines for structuring programs, which the guest researchers should observe in order to make it easier to classify the analysis outputs.

4.2 Preconditions for the feasibility of statistical disclosure control

The tests for statistical disclosure control do not only involve checking that each individual table contains sufficient case numbers, but examines the entire process of data preparation and analysis. The output is not checked in isolation but with reference to the research project. For this it is necessary that the programs are documented comprehensibly and in detail. The following criteria should be met during programming for a program to be in accordance with the FDZ guidelines[18]:

1. Detailed documentation of the analysis steps:
   In order to be able to conduct the outcome checks the FDZ staff have to be able to orientate themselves in the programs and to understand the programming at least as a whole. The program documentation guarantees this.

2. Program files:
   The analyses must be conducted via a program file by entering the corresponding program code. The program file must be structured in such a way that the program codes are also contained in the output files as this makes the sequence of the individual program steps clear. The data preparation generally has to be conducted in Stata. Other software packages can be used for further analyses.

3. Storing output files:
   So that it is clear which program file generates which output file, a correspondingly named output file must be created for every program file.

4. Creating a master file:
   The master file is created in order to start all the programs that belong to one analysis program at once both in the case of remote execution and after on-site use. The content of the master file is therefore the program calls for the individual programs in the correct sequence. The contents of the called programs should also be explained briefly in the master file.

As it is generally difficult to understand the analyses if the guidelines are not observed, the FDZ reserves the right to delete the analysis results entirely if there is any doubt at all with regard to compliance with data protection regulations.

4.3 FDZ guidelines for checking the analysis results

The aim of the routines followed is to modify the results by deleting values so that no risk of re-identification remains. At the FDZ there are general criteria for checking output. When deleting data a distinction is basically made between:

- primary suppression, which prevents the identification of information in a cell of a table,

[18] Further information can be found at: `http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF`.

- secondary suppression, which prevents the identification of information via subtotals and/or marginal totals and
- dominance suppression[19], which stops any identification of dominant firms.

There is no full automatic statistical disclosure control. However, the FDZ manages with a specially developed program script which, for selected Stata commands, scans the output files for low case numbers and deletes them accordingly. As the automatic disclosure limitation review only checks standard outputs, all additional evaluations are checked manually and values deleted where necessary. The script is adapted and developed continuously. The following sections look briefly at the standardised deletions. Results based on fewer than 20 observations are classified as critical and therefore deleted. This minimum requirement applies for both establishment data and personal data. This threshold was selected for the following reason: the FDZ checks every output independently and does not compare the results with those of the previous analyses. Furthermore, the researchers are bound by contract not to re-calculate values from the individual analysis results sent to them.

4.3.1 Statistical indicators

At first sight statistical indicators do not permit conclusions to be drawn regarding the case numbers on which they are based. However, this does not mean that displaying statistical indicators, such as means, is not to be regarded as problematic. Here, too, the principle applies that the indicators mentioned are only classified as safe when the calculation basis comprises at least 20 observations. One special case is the displaying of statistical indicators in the case of dummies. With binary coded variables their values are distributed between just two categories. Even when the total number of observations of a dummy variable is more than 20, it is possible, due to a skewed distribution, for only three individuals to fall into one of the categories. In this case the analysis is regarded as unsafe even though it is not possible to conclude from the total number of observations that one category contains very few cases, as this can be calculated easily using the mean. In order to be able to identify and test these cases, it is necessary to display not only the number of cases but also the minimum, the maximum and the standard deviation whenever means are displayed. As the FDZ program script only recognises the standard outputs of indicators, in special cases the frequencies of the two categories must be calculated afterwards when using dummy variables and if necessary the statistical indicators must be deleted.

[19] The risk of an establishment being identified in the analysis increases, for example, when detailed information on economic activities or regional information on a small scale are used. The results are checked for cases of dominance using the following measures: the minimum number of cases is also set at 20 units for the number of establishments when details concerning the number of employees from establishment data are shown. The sole use of samples instead of populations and the monitoring of the program steps also ensure that no dominance cases can be identified. In addition, detailed information regarding economic activity and region are only made available with special justification. As a standard only data at federal state level and the 3-digit codes of the classifications of economic activities are provided (see Chap. 3.3). In the dominance suppression we follow the guidelines of the Federal Employment Agency (Bundesagentur für Arbeit (2012)).

4.3.2 Percentiles

When displaying percentiles, care must be taken to ensure that each percentile contains at least 20 observations. In the case of a detailed output (1% percentiles), a total of at least 2000 observations therefore have to go into the output in order to guarantee compliance with the data protection guidelines of the FDZ. The basic principle is that the more information one wishes to obtain, the more observations have to be available for the entire distribution:

- At least 20 observations for releasing means (exception: dummies, see Chap. 4.3.2)
- At least 40 observations for releasing 50% percentiles
- At least 80 observations for releasing 25% or 75% percentiles
- At least 200 observations for releasing 10% or 90% percentiles
- At least 400 observations for releasing 5% or 95% percentiles
- At least 2000 observations for releasing 1% or 99% percentiles

4.3.3 Weights

In the case of weighted outputs, the statistical disclosure control is always conducted on the basis of the unweighted values. The output files must permit the weighted output to be attributed clearly to the corresponding unweighted output. Deletions in the unweighted tables are transferred to the corresponding weighted table. If the unweighted output is missing, the weighted table is deleted completely.

4.3.4 Graphs

Checking and transmitting graphs is an additional service provided by the FDZ. [20] In principle only graphs that have been created using the analysis programs can be published. As a result of the time-consuming disclosure control procedures for graphs, they should only be created if it is not possible to create them later on from the values in the output files. For each graph the number of observations underlying the individual values depicted must be indicated. The case number threshold of at least 20 observations also applies for graphs. Scatter plots, for example, are therefore characterised by a high risk of de-anonymisation as there are very probably fewer than 20 observations behind the individual data points displayed.

[20] We do not pass currently graphs on to users because the statistical disclosure control involved in high effort. Therefore, all the information needed to create graphs has to be recorded in a table in the output. Afterwards the users can recreate the graphs. There is a working tool with examples for our users available. (`http://doku.iab.de/fdz/access/Vorgaben_DAFE_EN.PDF`)

4.3.5 File formats

Statistics programs generally provide their users with the possibility to save the analysis results as separate files in different formats (e.g. LaTeX or ASCII). As checking these files would always have to take into account the corresponding program and output files, the complexity and the time required for this at the FDZ would increase enormously. That would hinder a prompt transmission of the results. For this reason the results created in this way must be integrated into the original output file directly below the corresponding analysis results[21]. For example, with Stata ado files it is possible to display results in LaTeX codes. These are only transmitted if they can be found in the log file directly below the corresponding Stata tables. In the case of descriptive tables they are deleted entirely as soon as a cell shows an insufficiently large number of cases. As additional time and effort is involved in reviewing the LaTeX codes, the users are asked to have only the results that they need for their publication displayed using LaTeX codes.

4.3.6 Transmission of aggregated data files

It is possible to create aggregated data files from the weakly anonymous microdata and to have them transmitted. As checking aggregated data is time-consuming, there are also certain rules for transmitting aggregated data. It is necessary to speak to somebody at the FDZ before generating aggregated data files. During this talk it must be clarified which level of aggregation is to be used, which variables are contained in the data file and in which aggregation state (total, mean etc.). So that the data can be checked, for each aggregated variable an additional variable must be created that contains the number of cases underlying the aggregated value. As an aggregated data file is only transmitted once per project, the FDZ recommends that researchers create these files on-site.

Of course there are other analysis methods besides the data queries mentioned. As the aim of this paper is to present the basic procedure followed by the FDZ when conducting output checking, we dispense with explanations regarding other analysis possibilities at this point. The following examples are intended to illustrate this procedure.

4.4 Examples

Finally, on the basis of some examples we show what information is removed or not released in the statistical disclosure control. The examples were created using the test data[22]of the IAB Establishment Panel. First of all, primary suppression is illustrated, which involves deleting information within a table. Table 1 is the original table and Table 2 follows

[21] In Stata this can be done, for example, using the command "type PATH".

[22] The test data of the IAB Establishment Panel are intended to enable researchers to write and test analysis programs prior to remote data access. The test data were generated by drawing a subsample and performing data masking while simultaneously retaining important data structures. This renders the test data ineligible for analysis.

statistical disclosure control. Example 1 shows the number of establishments with and without a works council for different establishment size classes in eastern Germany.

As already mentioned, the threshold for values to be deleted is smaller than 20. In the example this pertains to the value 16. In order to retain as much information as possible, the marginal totals are generally left and the associated value, 142 in this case, is anonymised. However, this is not yet sufficient to rule out the possibility of re-identification, as the deleted values may be inferred via the marginal totals and the remaining values in the table. Two further values therefore have to be removed. It is not necessary to remove the entire values in the case of multi-digit numbers, it is enough to delete the last full digit. There is no rule governing which two values are deleted next. This decision lies with the person checking the results at the FDZ. Generally an attempt is made to remove the next smallest value. In the example at hand this is the value 39 and accordingly the value 547.

Example 1: Eastern Germany

	Works council		
Number of employees	Yes	No	Total
1 1-4	43	1,380	1,423
2 5-9	39	547	586
3 10-19	89	487	576
4 20-49	250	590	840
5 50-99	255	245	500
6 100-199	290	110	400
7 200-499	283	65	348
8 500-999	142	16	158
Total	1,391	3,440	4,831

Table 1: before

	Works council		
Number of employees	Yes	No	Total
1 1-4	43	1,380	1,423
2 5-9	3*	54*	586
3 10-19	89	487	576
4 20-49	250	590	840
5 50-99	255	245	500
6 100-199	290	110	400
7 200-499	283	65	348
8 500-999	14*	/	158
Total	1,391	3,440	4,831

Table 2: after

The deletion of values can extend to other tables if information for certain variables is depicted in a differentiated way. Example 1, for instance, contains information that only applies to eastern Germany. If the same information is depicted for western Germany and for the country as a whole, these tables must not be considered independently of one another. A re-identification of previously deleted values is otherwise possible by calculating differences (total - west = east). In example 2 the values in Table 4 that are in the same position as those deleted in example 1 are therefore also deleted here. Table 5 contains the values for Germany as a whole and remains unchanged because it is no longer possible to recalculate values.

Example 2:

Western Germany

Western Germany

Number of employees	Works council		
	Yes	No	Total
1 1-4	64	2,461	2,525
2 5-9	54	847	901
3 10-19	130	762	892
4 20-49	364	853	1,217
5 50-99	365	370	735
6 100-199	391	165	556
7 200-499	402	90	492
8 500-999	198	22	220
Total	1,968	5,570	7,538

Table 3: before

Number of employees	Works council		
	Yes	No	Total
1 1-4	64	2,461	2,525
2 5-9	5*	84*	901
3 10-19	130	762	892
4 20-49	364	853	1,217
5 50-99	365	370	735
6 100-199	391	165	556
7 200-499	402	90	492
8 500-999	19*	2*	220
Total	1,968	5,570	7,538

Table 4: after

Germany

Number of employees	Works council		
	Yes	No	Total
1 1-4	107	3,841	3,948
2 5-9	93	1,394	1,487
3 10-19	219	1,249	1,468
4 20-49	614	1,443	2,057
5 50-99	620	615	1,235
6 100-199	681	275	956
7 200-499	685	155	840
8 500-999	340	38	378
Total	3,359	9,010	12,369

Table 5: before

Number of employees	Works council		
	Yes	No	Total
1 1-4	107	3,841	3,948
2 5-9	93	1,394	1,487
3 10-19	219	1,249	1,468
4 20-49	614	1,443	2,057
5 50-99	620	615	1,235
6 100-199	681	275	956
7 200-499	685	155	840
8 500-999	340	38	378
Total	3,359	9,010	12,369

Table 6: after

When displaying tables the researchers must always indicate the number of establishments, as it is of vital importance for statistical disclosure control to know how many establishments are behind the result. Example 3 illustrates the problem. Column 2 (sum) contains the number of trainees retained after completion of training in selected branches of economic activity, column 3 (N) shows the number of corresponding establishments. It is possible that only a small number of establishments are behind a sufficiently large number of cases of trainees. If this is the case, both values have to be deleted. This also prevents large enterprises from being identified.

Example 3:

r90b	sum	N
1 agricult./hunting/forestry	13	10
2 mining/quarrying	1	1
3 elec./gas/water supply	118	21
4 manufacture food/beverages	130	32
5 manufacture textiles/leather	5	4
6 manuf. wooden prod's/paper	35	13
7 manuf. chem./pharmaceut.	78	14
8 manuf. rubber/plastic prod's	164	23
9 manuf. glass/stone products	20	12
10 manuf. basic metals	138	21
Total	702	151

Table 7: before

r90b	sum	N
1 agricult./hunting/forestry	/	/
2 mining/quarrying	/	/
3 elec./gas/water supply	118	21
4 manufacture food/beverages	130	32
5 manufacture textiles/leather	/	/
6 manuf. wooden prod's/paper	/	/
7 manuf. chem./pharmaceut.	/	/
8 anuf. rubber/plastic prod's	164	23
9 manuf. glass/stone products	/	/
10 manuf. basic metals	138	21
Total	702	151

Table 8: after

With statistics for selected variables, the mean is checked in the case of dummy variables, for example, as there is the possibility of small values being re-identified here. In example 4 a mean of 0.085 is shown for variable r61 (trainee positions offered: yes/no). This is equivalent to a percentage distribution of 8.57% for variable attribute 1. By multiplying the number of cases by the mean (140 x 0.0857143) it is possible to calculate that 12 establishments have the value 1.

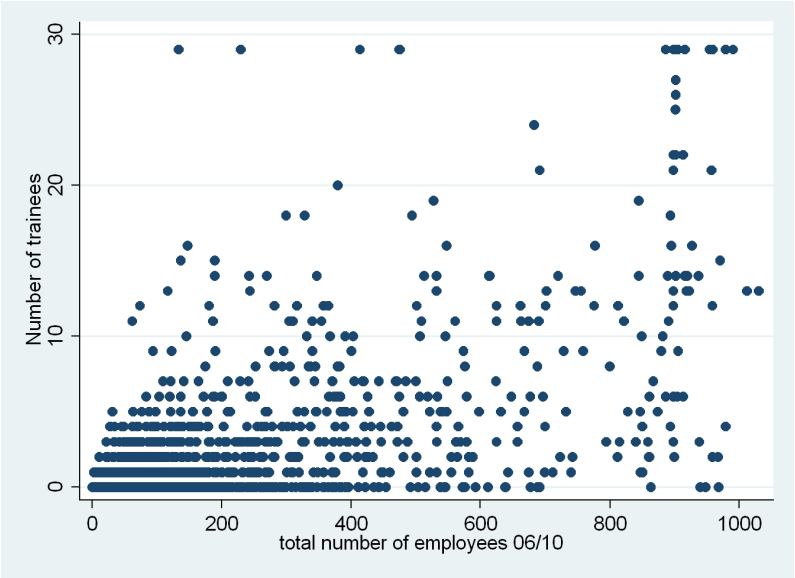
Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	.0857143	.2809469	0	1
r62a	73	2.219178	2.340742	1	15

Table 9: before

Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	/	/	/	/
r62a	73	2.219178	2.340742	1	15

Table 10: after

The final example addresses the transmission of graphs (see at chapter 4.3.4). Graphs can only be transmitted if no fewer than 20 establishments can be attributed to individual data points. The checking of graphs is performed analogously to that of tables. In the example below, the graph would not be released as each data point stands for one establishment.



5 Outlook

The FDZ is constantly working on improving the provision of data to researchers in Germany and abroad in compliance with data protection legislation. For example, in the context of the externally funded project "RDC in RDC" ("Projekt FDZ in FDZ" - PFiFF), the data held by the FDZ can be analysed on-site not only in Nuremberg, but also at the Research Data Centres of the Statistical Offices of the Länder in the German cities of Berlin, Bremen, Dresden and Düsseldorf, and at the Michigan Center on the Demography of Aging (MICDA) in the Institute for Social Research (ISR) at the University of Michigan. In addition to that, work is being carried out to facilitate access to micro-data for researchers across Europe in the EU project Data without Boundaries (DwB). At the FDZ we are currently in the process of automising job submission and are working on the use of the JoSuA software[23] provided by the International Data Service Center (IDSC) of the Institute for the Study of Labor (IZA).

[23] Further information about JoSuA can be found at: `http://idsc.iza.org/josua`

References

Brandt, Maurice/Franconi, Luisa/Guerke, Christopher/Hundepool, Anco/Lucarelli, Maurizio/Mol, Jan/Ritchie, Felix/Seri, Giovanni/Welpton, Richard. Guidelines for the checking of output based on microdata research. Final report of ESSnet sub-group on output SDC 2010

Bundesagentur für Arbeit. Statistische Geheimhaltung: Rechtliche Grundlagen und fachliche Regelungen der Statistik der Bundesagentur für Arbeit. März 2012, abgerufen am 18.05.2012 ⟨URL: http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Statistische-Geheimhaltung/Generische-Publikationen/Statistische-Geheimhaltung.pdf⟩

Bundesstatistikgesetz (BStatG) – Gesetz über die Statistik für Bundeszwecke vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 3 des Gesetzes vom 7. September 2007 (BGBl. I S. 2246).

BVerfG. Urteil v. 15.12.1983, Az. 1 BvR 209, 269, 362, 420, 440, 484/83.

Kirchner, Antje/Gschwind, Lutz. Panel Arbeitsmarkt und soziale Sicherung - Die PASS Campus Files. Datensätze für den Einsatz in der wissenschaftlichen Lehre. FDZ-Methodenreport 06/2011 2011

Lane, Julia/Heus, Pascal/Mulcahy, Tim. Data Acess in a Cyber World: Making Use of Cyberinfrastructure. Transactions on Data Privacy 2008

Müller, Walter/Blien, Uwe/Knoche, Peter/Wirth, Heike. Die faktische Anonymität von Mikrodaten. Stuttgart: Metzler-Poeschel. 1991

Ritchie, Felix. Statistical disclosure detection and control in a research environment. WISERD DATA RESOURCES 006 2011

SGB X. Zehntes Buch Sozialgesetzbuch – Sozialverwaltungsverfahren und Sozialdatenschutz – (SGB X), in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), zuletzt geändert durch Entscheidung des Bundesverfassungsgerichts vom 23. November 2010 (BGBl. I S. 1718).

Sozialgesetzbuch (SGB) Erstes Buch (I) – Allgemeiner Teil (SGB I) vom 11. Dezember 1975 (BGBl. I S. 3015), zuletzt geändert durch Artikel 110 Absatz 5 des Gesetzes über die weitere Bereinigung von Bundesrecht vom 8. Dezember 2010 (BGBl. I S. 1864.

Sozialgesetzbuch (SGB) Drittes Buch (III) – Arbeitsförderung (Artikel 1 des Gesetzes vom 24. März 1997, BGBl. I S. 594), zuletzt geändert durch Artikel 12 Absatz 8 des Gesetzes vom 24. März 2011 (BGBl. I S. 453).

Table 11: FDZ Portfolio

	General	factually anonymous data	weakly anonymous data
Conditions	Scientific research and necessity of the data	Labour market research	Research in the field of social security
		Independent scientific research institution	-
		Data security concept	-
		-	Public interest
		-	Approval by Federal Ministry for Labour and Social Affairs
Access and use	Limitation of use to specific purpose Limited period of time	Guarantee of data security	-
		-	Guidelines for on-site use
	Ban on disclosure to third parties, on merging with other microdata and on de-anonymisation	Deletion of microdata at end of project	-
	Restriction of user group	-	Ban on re-calculating deleted values
Anonymisation	Sampling Deletion of original identifiers LDeletion or aggregation of sensitive variables	Deletion or aggregation of further variables	-
Output checking	-	-	Absolute anonymisation of output

Imprint

FDZ-Methodenreport 6/2012

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2012/MR_06-12_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Alexandra Schmucker,
Phone: +49 (0)911 / 179-1762
Email: alexandra.schmucker@iab.de

Dana Müller,
Phone: +49 (0)911 / 179-2409
Email: dana.mueller@iab.de

Forschungsdatenzentrum,
Regensburger Str. 104
D - 90478 Nürnberg