# Disclosure control regulations for research results with regard to data confidentiality (output control)

Version 2.2

Last updated: October 2018

Author: Research Data and Service Centre

DEUTSCHE BUNDESBANK EUROSYSTEM

# TOC

# Foreword

Dear data users

The Deutsche Bundesbank provides researchers with transparent and cost-free access to selected microdata resources. Having access to these microdata broadens researchers' analytical reach to a considerable degree. Yet at the same time, working with such sensitive data entails significantly more stringent requirements for researchers when it comes to ensuring data protection and data confidentiality compared to working with aggregated data.

Researchers are responsible for ensuring that confidential data do not enter the public domain. Their published research results must remain completely anonymous. They must not permit any data to be traced back to individual economic agents such as banks, enterprises, individuals or households.

The following criteria and regulations aim to ensure that these requirements for research results are met at all times. They also help boost confidence among data providers (such as banks and enterprises) and microdataset producers (such as Bundesbank business units) that their data are being handled securely. This sense of confidence that data protection and confidentiality standards are being adhered to by both the Bundesbank and its data users is absolutely necessary if this public asset is to continue being made available for research purposes.

# 1 Introduction

Researchers working with Bundesbank microdata are responsible for ensuring that their calculation results do not violate data confidentiality requirements. They must ensure that their calculation results contain no data that can be traced back to individual observation units or statistical units such as banks, (non-financial) corporations, individuals or households. Results must not leave the secure Bundesbank environment until they have been completely anonymised.

For this reason, employees of the Research Data and Service Centre (RDSC) or any other Bundesbank business unit charged with performing a disclosure control in individual cases must check prior to this whether data confidentiality requirements have been ensured when such results were drafted. Should this not be the case, the calculation results cannot be disclosed.

Section 2 shows the format in which calculation results should be presented to the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases for the purpose of performing the disclosure control described herein.

This document also provides an overview of the criteria that researchers should follow when drafting calculation results so as to ensure that data are kept confidential. A distinction is made here between general criteria and dataset-specific criteria. General criteria must be followed no matter which data are being used. These general criteria are listed in section 3. A separate set of non-disclosure rules apply to certain microdata, which is why additional criteria need to be followed when working with such data. These are listed in section 4.

The criteria listed in this document do not constitute an exhaustive set of rules. Indeed, the criteria and regulations may be amended, supplemented and expanded over time. Furthermore, the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases already have the possibility of reviewing calculation results according to additional criteria, if necessary (see section 5).

Finally, the last section draws attention to a procedure drawn up by the RDSC to calculate the number of observations required for performing a disclosure control of the calculation results when using the standard analysis software STATA.

If the requirements described in section 2 or the criteria stated in sections 3 and 4 are not met, the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases have the right to reject the disclosure control of the calculation results, deny their release and call for remedial action to be taken.

Researchers who are unsure about whether their calculation results meet these requirements and criteria are advised to discuss any potentially critical points with RDSC staff. Visiting researchers should do so while their research visit is still under way. However, this discussion does not replace the disclosure control of the calculation results. Similarly, it remains the responsibility of researchers to ensure that data confidentiality is not violated when drafting calculation results.

# 2 Requirements for calculation results

## 2.1 RDSC minimum requirements for calculation results

The minimum requirements for calculation results submitted for the purpose of disclosure control with respect to the non-disclosure of individual observation units include the following.

- Results files (eg log files in STATA, usage of rdsc_log in R)
- Program codes, commented on in sufficient detail to render them comprehensible
- Reporting of all relevant information and numbers of observations for disclosure control
- Underlying frequencies in tabular format (particularly in the case of charts)
- Filing results and program codes in the designated project directory (subfolder “transfer” in your project directory, see also Agreement concerning access to microdata as part of a research project at the Deutsche Bundesbank's Research Data and Service Centre (RDSC), Annex 2, Point 9: Directory structure at Visiting Researcher workstations)
- Results and program codes will not be accepted for the output review if they are submitted in software-specific formats such as .Rdata. Plain text files must be used instead (eg .txt, .csv, .tsv., .R, .do, .log or LaTeX files). Please note that files in .doc or .docx formats are not plain text files.

## 2.2 Logging calculation results

- With the exception of the master file, every analytical file (such as the "do file" in the case of STATA) should generate a log file that completely logs the calculation results generated.
- Besides log files, further plain text files (eg .csv or LaTeX files) are also permitted. The log file must clearly state the position at which this table was created and exported.
- Log files must be editable to allow any results that are critical in terms of data confidentiality to be blocked during the disclosure control (eg .smcl files from STATA are not permitted).
- For disclosure controls with respect to the non-disclosure of individual observation units (see section 2.5), the underlying number of observations must always be specified, irrespective of how the results are presented (eg tables, charts, regressions, descriptive statistics).[1] Failure to do so will result in the calculation results not being released. In this case, calculation results will need to be revised and regenerated, together with the number of observations.

## 2.3 Use of Microsoft Excel

Microsoft Excel is available at all Visiting Researcher workstations. However, calculation results generated using Microsoft Excel which are submitted for the purpose of performing a disclosure control will not be accepted or released. Similarly, calculation results exported into an Excel document (.xls or .xlsx file) will not be accepted for the disclosure control, nor will they be released.

[1] See section 3.1 for the requirements governing the number of observations.

## 2.4 Unambiguousness of variable names and value labels

Variable names and labels of attributes within a variable must be unambiguous. This applies, in particular, to the variable names. If a variable is newly created or an existing variable is modified, the associated variable designations must, without exception, also be newly issued and included in the variable list in the syntax program header. Please use descriptive names. If (categoric) variables are newly created or modified, corresponding value labels need to be allocated to the attributes.

## 2.5 Disclosure controls of calculation results

Calculation results are to be submitted to the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases for the purpose of disclosure control with respect to the non-disclosure of individual observation units. Calculation results may not leave RDSC premises or similarly equipped Bundesbank rooms until the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases has given its written approval.

If, following a research visit, calculation results are to undergo a disclosure control and be released by the RDSC, the calculation results and all files required for the disclosure control are to be filed in the designated project directory (see section 2.1). The RDSC does not check on its own initiative whether, following a research visit, any calculation results have been filed in the project directory for disclosure control purposes. The RDSC should therefore be advised by e-mail that calculation results are to be checked and of the folder in which these results are saved. For reasons of confidentiality, please do not e-mail any calculation results to the RDSC.

The relevant program codes should also be filed in this folder alongside the calculation results. This is essential in case any critical results need to be recapitulated. If this is not possible using the program codes on file, the calculation results cannot be released. Should these program codes also be needed outside the Bundesbank, the RDSC should be notified of this. Otherwise, only the calculation results will be sent.

Researchers must ensure that the minimum requirements described in section 2.1 are met and that the time and effort which the RDSC or any other Bundesbank business unit charged with performing a disclosure control in individual cases need to expend are kept to an absolute minimum.

The regulations do not permit researchers to first submit all their calculation results for disclosure control and then only afterwards select the results that are suitable for publication (eg speech, working paper, journal article). In principle, any work conducted as part of the research project that relate to the Bundesbank's microdata should be carried out during the research visit on RDSC premises or in similarly equipped Bundesbank rooms. Besides data preparation and analysis, this also includes the selection of the calculation results required for publication.

The volume of calculation results submitted for disclosure control should remain within reasonable and logical bounds, with the quantity of calculation results needed when drafting a publication and the space they require there serving as a yardstick in this regard. The selection of calculation results submitted for disclosure control should be confined to those which appear appropriate for the final publication. Researchers should therefore conduct a review of extensive calculation results (such as those which are produced particularly during explorative analyses or when testing different model specifications in regressions) while the research visit is still under way and carefully select the results that are worthy of publication for disclosure control.

Calculation results should only be checked and released once to reduce the amount of effort for disclosure control in the RDSC. Therefore, identical calculation results must not be submitted again for disclosure control. Creating and submitting calculation results again is only allowed in duly substantiated exceptional cases. Moreover, an exact reference to the according earlier analysis has to be made.

# 3 General criteria for disclosure controls of calculation results

The criteria listed in this section should always be observed in analyses and calculations involving Bundesbank microdata. These criteria shall apply irrespective of which datasets are used in a research project.

## 3.1 Number of observations in your output

All calculation results submitted for disclosure control must be based on at least three different economic agents (institutions subject to report, other legal and natural person, legal entities or agencies such as banks (non-financial) corporations or households). Researchers must ensure that the requirement has already been fulfilled when drafting calculation results.

Therefore, the underlying number of observation units must be issued for **all** results. In the case of weighted calculation results, the unweighted number of observations shall also be shown. Usually, all the results and the associated number of observations are blocked if the results are based on fewer than three economic agents.

Bear in mind when determining the relevant number of observations that it is the number of different observation units underlying the results, rather than the sample or population size, which is key. The observation units are the units “to be protected”. Usually, the number of underlying observation units is smaller than the sample or population size in a given dataset. Care should therefore be taken to ensure that observation units are not counted more than once when the number of observations is calculated.

Repeat observations for individual observation units, especially, are an attractive feature of most Bundesbank microdata. This leads to numerous observations for one observation unit. For example, different transactions with different partners are recorded at different points in time are reported for one and the same bank or enterprise. In the example just given, the observation unit should be the bank or non-financial enterprise. It must be ensured that each observation unit is counted only once when determining the relevant number of observations for disclosure control. In this case, therefore, the number of different banks or the number of different enterprises that underlie the study must be reported. Reporting the number of observations in the dataset is not sufficient.

In some datasets, missing values are always recorded with a zero. In such datasets, zeroes are therefore initially regarded as invalid observations. Consequently, how these zeros are to be interpreted has to be considered on a case-by-case basis. For instance, if a value is based on five observation units, there could be a problem if three of the values are zero, even though they are actually missing values. This would mean the calculation results are only based on the values of two observation units.

## 3.2 Calculating the number of observations making up an aggregate

For every variable generated through aggregation, the number of observation units underlying the aggregate with valid individual values must be reported.

When generating rates or shares, it has to be ensured that the number of underlying observation units (with valid individual values) for each variable included in the calculation is sufficiently high. This means the number of observation units has to be reported for the numerator as well as the denominator.

Some datasets provide information on lower-level and higher-level units (such as shares and derivatives to the higher-level unit public limited companies, subsidiaries to the higher-level unit common parent company). If values are given for units that can be attributed in part to the same higher-level units, it is always the number of observation units at the highest level in each case that is decisive. For example, it is not just individual enterprises but also the group to which these enterprises belong that should be regarded as the unit to be protected. If values refer to a series of enterprises which all belong to a single group, these values may possibly correspond to a single figure for that group. Therefore the unit to be protected in each case should always be borne in mind when drafting calculation results.

Please note that leaving out data points with certain values (such as missing values in STATA) can also result in aggregated information (on the group of observations with that value) being given.

## 3.3 Charts

For any chart (of a relative frequency function, for example), the relevant number of observations for each datapoint in the chart must be reported, for example, by generating a suitable table (such as a table of absolute frequencies) that enables the immediate identification of the relevant number of observations for individual categories or subgroups. This means that a chart showing total bank assets for the months from January 2012 to December 2015, say, the number of banks on which the bank asset total calculation is based must be reported for each individual month. Failure to provide such a table will result in charts not being released.

Charts are to be saved in a read-only format. This is to prevent underlying value tables or case number tables being sent out. For this reason, charts may not be submitted for disclosure control as vector diagrams. The recommended file format for STATA charts is .png.

## 3.4 Stating the number of observations when giving percentiles

Where percentiles (q) of a reporting sample or subgroup are calculated, the number of observations (n) for the group in question must also be calculated, in addition to the number of observations, calculated on this basis, per percentile range (for example, at the 99th percentile (q=99) and with 1,000 observations, the number of observations per percentile range is ((1-0.99) x 1,000 = 10). In practice, a percentile value is often given as a weighted mean of two neighbouring values due to the values' indivisibility. Mathematical and practical considerations have shown that a value of 2.3 is a suitable threshold for the number of observation units per percentile range. The results for the median, therefore, have to be blocked when, for example,

$$q' = \begin{cases} q, & q \leq 50 \\ 100 - q, & q > 50 \end{cases}$$

where

$$(n+1)\frac{q'}{100} \leq 2.3$$

or

$$(n+1)q' \leq 230$$

applies.

## 3.5 Non-disclosure requirement for entire tables

In cases where calculation results are initially presented as aggregated results and the microdata are then broken down on the basis of certain criteria, the entire table shall be subject to the non-disclosure requirement. For example, a table entitled "All enterprises" and another entitled "Mechanical engineering enterprises" are generated, the difference being "Non-mechanical engineering enterprises". In this case, researchers must also create the "Non-mechanical engineering enterprises" table as this will need to be covered by the disclosure control. Otherwise, it might be possible to identify characteristics of individual observation units through differencing.

## 3.6 p%/dominance rule

Observation units differ in size. Even if a value based on three or more observation units is reported, there is the risk of major observation units being identified indirectly if they contribute a sufficiently large share to that value (dominance criteria). The possibility of indirectly identifying observation units in this way must be eliminated. If the observation units used as a basis for calculating a value include one or two observation units that are significantly larger than the remaining observation units, it must be checked whether they contribute predominantly to this value. In addition to the number of underlying observation units, in this case the largest and second-largest value should always be reported. As a rule of thumb, the largest and second-largest value together should not exceed 85% of the total value.

## 3.7 Minimum and maximum values

In essence, minimum and maximum values are microdata and are therefore to remain confidential. The exception to this is dichotomous attributes, provided that both values are given.

However, it is possible to show the average of the three observation units with the lowest value and the three units with the highest value as an alternative to the minimum and maximum values of a given variable. Bear in mind during calculation that all six of these observation units should be different.

## 3.8 Dummy variables

If the mean value of a dummy variable encoded with 0 and 1 is given, bear in mind that this value corresponds to a ratio. It shows the proportion of observations with the characteristic of 1 among all the observations. A mean value of 0.7 for this kind of dummy variable means that, in 70% of cases, the variable has the characteristic of 1 and in 30% of cases it has the characteristic of 0. Therefore, the same rules apply when giving mean values for a

dummy variable as when calculating ratios (see section 3.2). At least three different observation units must have the characteristic of 0, and at least three different observation units must have the characteristic of 1.

Dummy variables can be included in regressions even if fewer than three observation units have the characteristic of 0 or 1. In this case, however, the regression coefficients for these dummy variables will be blocked and not be released. By contrast, the remaining regression coefficients can be released. Each dummy variable in a regression must therefore be reviewed to see whether the variable has the characteristic of 0 or 1 for a sufficient number of observation units. This can be checked in STATA immediately following regression with the command "`nobsreg`" (see section 6). This information is only required for dummy variables that relate directly to the observation unit. In all other cases, there is no need for this information.

### 3.9 Identifiers

Neither the do files nor the calculation results that are submitted for disclosure control may contain identifiers (eg BAID, company numbers, etc).[2]

### 3.10 Previous results

In addition to the submitted calculation results, disclosure control shall also include the calculation results that researchers have received previously. Since such previous calculation results can increase the informative value of the calculation results submitted for disclosure control, it may be necessary to put additional blocks on the data.

### 3.11 Analyses at regional level

When working with regional information or carrying out an analysis at the regional level, it should be noted that regional information can increase the re-identification risk. There are therefore special requirements with regard to data confidentiality and the output review in such cases. Evaluations and, in particular, descriptive analyses of individual regions will not be accepted for the output review if the regional units are smaller than the federal states (NUTS1). This applies irrespective of whether three or more observation units are present for the regional unit in question. Calculation results can only be cleared if it is not possible to trace these to a certain region and if it is not discernible from the results which region is being examined. Even in this case, the number of underlying observation units must be sufficiently large and the dominance criterion must not be violated.

## 4 Disclosure control of calculation results: additional criteria for specific datasets

When working with most datasets, it is sufficient to follow the general criteria. However, some datasets are subject to special non-disclosure guidelines. In such cases, the criteria specific to this dataset must be followed in addition to the general criteria described in section 3. These special criteria are shown for the individual datasets below.

[2] See section 4.2 on how to handle ISINs.

## 4.1 External position of banks (AUSTA – "*Auslandsstatus*")

- A bank (parent) can run several legally dependent branches (foreign branches) and/or legally independent subsidiaries (foreign subsidiaries) in a country. Therefore, the number of corresponding parents must be given for analysis at the level of foreign branches and subsidiaries (see also section 3.2).
- In many countries, there are only a small number of parents and branches / subsidiaries. If aggregated values are reported for these countries, a single bank might turn out to have a very high share. Hence the dominance criteria needs to be investigated as described in section 3.6.

## 4.2 Securities holdings statistics (SHS base plus)

- The unit subject to protection is the reporting institution and not the security, ie securities-based evaluations are not a problem. It is also generally not a problem when ISINs are reported in the calculation results.
- Calculation results that include a combination of depositor sector 1221/1222/1223, ISIN and bank are critical. These concern treasury securities held by reporting institutions. It is possible to determine who issued the security using the ISIN. This makes it immediately clear which bank is concerned here.
- It is possible to input information on the securities from other or external sources via the ISIN. Visiting researchers can be supplied with a list of the ISINs included in the dataset for this purpose. On account of the problem described in point 2, the merging of names of securities is not permitted.

## 4.3 Credit register of loans of €1.0 million or more

- Credit relationships between lenders and borrowers are shown in the credit register. Every value reported must therefore relate to at least three different lenders and three different borrowers.
- If microdata are aggregated, the number of underlying lenders and borrowers must also be reported.
- The credit register of loans of €1.0 million or more also contains information on superordinated lenders (lender group) and borrowers (borrower unit). They should likewise be considered units to be protected.

## 4.4 Microdatabase direct investment (MiDi)

- The MiDi is a dataset showing investment relationships, ie each row in the MiDi corresponds to an investment relationship reported in a given year. It is therefore by no means sufficient to count only the number of rows, even if an investment relationship exists for only one year. This is because it is usual for more than one line to belong to an enterprise (all the enterprise's investment relationships above the reporting thresholds). When determining the number of observation units, ensure that each enterprise is counted only once, even if it carries out several investments, and that multiple counting does not occur.
- The MiDi also implicitly contains information on group structures and therefore information on superordinated enterprises. The superordinated enterprises should likewise be considered units to be protected. In the case of outward investment, for example, this applies to the German reporting agent (the “num”) as well as the German parent of the reporting agent, if applicable (the “nui”). This means that it is not enough for a calculated aggregate to be based on more than three foreign subsidiaries. Instead, the first step should be to count the number of the German reporting agents (num) underlying the aggregate and give it as the number of observations. If a German parent of the reporting agent exists

(nui), then "nui" instead of "num" must be included to calculate the number of observations.[3] Conversely, in the case of inward investment, protection should be given to the following: num (German reporting agent), nu4 (foreign parent), noa (group parent of foreign parent, if any).

- In many countries, there are only a small number of parents and subsidiaries. If aggregated values are reported for these countries, a single enterprise unit that is to be protected (eg a German parent in the case of outward investment) might turn out to have a very high share. It is therefore quite possible for an individual unit to account for more than 85% of the aggregate, and this needs to be investigated if there are any grounds for suspicion. This makes it necessary to check the risk of re-identification according to the dominance criterion. Therefore, in addition to the aggregated value, the highest and second-highest value included in the aggregated value must be reported as well.

## 4.5 Statistics on international trade in services (SITS)

- The SITS dataset is concerned with services transactions, ie each row depicts a transaction reported for a given date (year and month). Because an enterprise will usually carry out more than one transaction, it is quite possible for more than one row to contain observations for one and the same enterprise. The number of rows showing one and the same month therefore does not match the number of enterprises. When determining the number of observations, make sure that an enterprise is counted only once and not multiple times, even if it carries out more than one transaction.
- In many countries, only a small number of enterprises carry out services transactions. If aggregated values are reported for these countries, a single enterprise might turn out to have a very high share. It is therefore quite possible for an individual unit to account for 85% or more of an aggregate, and this needs to be investigated if there are any grounds for suspicion. This makes it necessary to check the risk of re-identification according to the dominance criterion. Therefore, in addition to the aggregated value, the two highest values included in the aggregated value must be reported as well.

## 4.6 Corporate balance sheets from the refinancing operation (USTAN – *Unternehmensbilanzen*)

- Some enterprises report several sets of financial statements a year. Therefore, when determining the number of observations, make sure that these enterprises are counted only once and not multiple times.

# 5 Further criteria

The RDSC or any another Bundesbank business unit charged with performing disclosure control may, insofar as it is deemed necessary, review the calculation results against additional criteria.

[3] In practice, this can be done, for example, by generating in a first step an auxiliary variable containing the value of "num". In the next step, this auxiliary variable is replaced by the value of "nui" if "nui" contains a valid value and is not missing. Finally, this auxiliary variable is used to compute the relevant number of observation units for disclosure control instead of "num". If you work with a panel data set, please keep in mind that a reporting agent could have different parents across time. A higher number of different parents than different reporting agents in your dataset can be consider as a clear hint. In this case, the auxiliary variable leads to misleading results. It is possible to take this problem into account eg by determining the number of observational units based on the auxiliary variable for each point in time.

# 6 Determining the relevant number of observations in STATA

The number of observations can be calculated in various ways using STATA. However, the RDSC created two STATA commands `nobsreg` and `nobsdes` for this purpose. `nobsreg.ado` and `nobsdes.ado` can be found in the ado_library together with the associated .hlp files and on our websites.[4] If you would like to use one of these commands, please refer to the .hlp file and the detailed examples it contains. Only a rough outline of the procedure is given here.

`nobsreg` without any additional specification directly after a regression command reports the number of different identifiers (IDs) underlying the regression. It also checks whether a variable has values unequal to zero for only one or two IDs. If so, a warning message is issued. `nobsreg` therefore seeks to be economical in its output.

In the descriptive observation (including charts), `nobsdes ID Variable [if] [, options]` determines the number of different IDs with valid observations for a given variable and checks whether the dominance criteria is violated. The name of the ID and the variables of interest are required, and in that order. There are also several options. If categories are named using `by()` , a table with the number of different IDs with valid observations is displayed as a default setting without any additional variable being generated. However, for subsequent use, a variable with zeros and ones can also be given, the sum of which gives the number of different IDs with valid observations.

[4] https://www.bundesbank.de/resource/blob/604786/c9c14eade63b7fb3f85f38233bbacaa8/mL/number-of-observations-data.zip