DEUTSCHE BUNDESBANK
EUROSYSTEM

# Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata

Technical Report 2022-01

The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Deutsche Bundesbank, Research Data and Service Centre

Jannick Blaschke
Matthias Gomolka
Christian Hirsch

Research Data and Service Centre

## Abstract

Working with confidential microdata requires researchers to focus on the equally important objectives of obtaining results and checking that these results comply with Statistical Disclosure Control (SDC) rules. However, these two objectives may lead to tensions during analysis. In this report, we outline how researchers can resolve this tension, mostly by providing examples but also by suggesting a rule of thumb.

**Keywords:** statistical disclosure control, sdc, output control, research data centre, rdc, aggregation

**Version:** 1.0

**Citation:** Blaschke, J., M. Gomolka and C. Hirsch (2022). Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata, Technical Report 2022-01 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre.[1)]

1 The authors gratefully acknowledge the contributions from our colleagues at the RDSC.

# Contents

# 1 Introduction

## 1.1 Aim of this technical report

Empirical research generally involves aggregating multiple values into a single value (e.g. by taking means or calculating sums). While the aggregation of values is necessary to compute research results, it sometimes hampers SDC.[2)] If this happens, researchers may find themselves torn between these two objectives, as both tasks need to be carried out on different datasets: one with aggregated values and one without. We present a framework that enables researchers to resolve this tension by allowing them to aggregate values while simultaneously performing SDC on the same dataset. To this end, we provide examples that resemble the structure of some of our actual datasets. Researchers interested in a more general analysis of the interplay between data manipulation and SDC should refer to Section 3.

## 1.2 A word (or two) of caution

Results that do not clear SDC cannot be taken out of the RDSC's secure environment or used in publications. Confidential microdata differ in this regard from other data sources where obtaining results is the sole objective. The purpose of this technical report is to raise awareness among researchers that working with confidential microdata means focusing on two equally important objectives: (i) obtaining results, but also (ii) checking that these results comply with SDC. Furthermore, these two objectives arise concurrently as researchers who only focus on results usually run the risk of having to make time-consuming, but mandatory alterations to code to ensure compliance, delaying access to output.

The examples provided in this technical report may serve as a starting point to satisfy both analytical and SDC requirements. However, they do not represent an exhaustive list by any means. One prerequisite for all the approaches presented in this paper is that researchers are able to use the correct commands, e.g. the right merge command, and that they clearly communicate the procedure they choose so that it can later be verified by the output checkers at the RDSC.

## 1.3 A final word on how researchers should use this report

We use examples throughout most of this report to make it easier for researchers to follow. However, researchers should not conclude that this report applies only to the datasets mentioned herein. In fact, the reverse is true. For example, Section 2 also applies to individual statements of non-financial firms (JANIS) (Becker, Biewen, Schultz, & Weisbecker (2021)) if researchers aggregate firms by sector instead of banks by security.[3)]

2 More information on the RDSC's SDC rules can be found in the Rules for visiting researchers at the RDSC, which are available at https://www.bundesbank.de/resource/blob/826176/ffc6337a19ea27359b06f2a8abe0ca7d/mL/2021-02-gastforschung-data.pdf
3 IDs in JANIS are called BBK_HM_CO_ID1 and ES03, instead of BAID and ISIN, respectively.

# 2 When does the aggregation of values hamper SDC?

## 2.1 Example 1: Aggregating values and dropping duplicates

### 2.1.1 Where does the tension arise?

Table 1 shows an extract from a dataset similar to the SHS-Base plus dataset (Blaschke, Sachs, & Yalcin (2021)), which provides information on securities holdings by individual bank on a monthly basis. Securities are identified by their ISIN and individual banks by their anonymised bank ID (BAID).[4)]

| Date | BAID | ISIN | Units | Holdings |
|---|---|---|---|---|
| 2021-01 | 1234 | DE01 | 10 | 100 |
| 2021-01 | 1234 | AT01 | 1 | 100 |
| 2021-01 | 4567 | DE01 | 20 | 200 |
| 2021-01 | 4567 | AT01 | 2 | 200 |
| 2021-01 | 8901 | DE01 | 30 | 300 |
| 2021-01 | 8901 | AT01 | 3 | 300 |
| 2021-01 | 6900 | DE01 | 40 | 400 |
| 2021-01 | 8877 | DE01 | 50 | 500 |
| 2021-01 | 8900 | DE01 | 60 | 600 |

Table 1: Modified SHS-Base plus dataset.

Suppose you want to calculate the sum of holdings for each ISIN over all banks because your analysis is based on ISINs instead of banks. For example, for ISIN AT01 this will amount to $600(100 + 200 + 300)$ . After calculating the new variable, it is quite natural for researchers to delete duplicates with respect to the ISIN before proceeding with their analysis.[5)] Table 2 presents the dataset after summing and dropping duplicates. However, this makes it impossible to perform SDC for the sum holdings variable on the dataset as it leads to incorrect results.

| Date | BAID | ISIN | Holdings | Units | Sum holdings |
|---|---|---|---|---|---|
| 2021-01 | 1234 | DE01 | 100 | 10 | 2,100 |
| 2021-01 | 1234 | AT01 | 100 | 1 | 600 |

Table 2: Modified SHS-Base plus dataset after summing and dropping duplicates.

Basing SDC on Table 2 will yield the following results:

1. One bank is behind the sum holdings for ISIN DE01 and AT01, respectively.
2. Concentration in both cases is 100% because there is only one bank behind each ISIN.

Both results are incorrect.[6)] Remember that the sum holdings variable in the case of DE01 (AT01) is based on holdings of six (three) banks. Similarly, the concentration of the two largest holdings

**4** Note that for this example we have made some modifications to the SHS-Base plus dataset to illustrate our point. The units variable is not contained in the real SHS-Base plus dataset. Since July 2020, the ID variable BAID has been called BAID_DOM. And needless to say, the data points are not real either.

**5** This is certainly true for Stata users. We acknowledge that users of R and Python may use different programming strategies. Insofar as they delete duplicates, the examples in this report are applicable to R and Python as well.

**6** Researchers obviously understand why the results are incorrect. The reason is that it is somewhat random which duplicate you keep. Table 2 would look similar with respect to sum holdings if we had kept the observations for BAID 5678 instead of 1234.

combined needs to be calculated based on the holdings variable, as sum holdings is the aggregate of this variable. Therefore, concentration comes to 52.38% (83.33%)[7] for ISIN DE01 (AT01).

In this example, the results for ISIN DE01 would pass SDC, while those for ISIN AT01 would not. This is because compliance with SDC requires results intended for release to be based on at least five different observational units (see Principle O.1.3 "Adherence to minimum sample size" in Research Data and Service Centre (2021)). Furthermore, the combined share of the two largest observational units must not exceed 85% (see Principle O.1.4, "Adherence to p% or dominance rule" in Research Data and Service Centre (2021)). The tension in this example arises because researchers would work with a dataset similar to Table 2 to obtain their results but then have to perform SDC on a dataset similar to Table 1.

### 2.1.2 How can the tension be resolved?

Note that in our example we went back to the dataset before summing and dropping duplicates to perform SDC. Let us now look at the dataset before dropping duplicates, as shown in Table 3 below. Note that Tables 1 and 3 present similar information, the only difference being that Table 3 contains an extra column for the sum holdings variable. Therefore, Table 3 offers a path to resolve the tension between obtaining results and performing SDC.

| Date | BAID | ISIN | Holdings | Units | Sum holdings |
|---|---|---|---|---|---|
| 2021-01 | 1234 | DE01 | 100 | 10 | 2,100 |
| 2021-01 | 1234 | AT01 | 100 | 1 | 600 |
| 2021-01 | 4567 | DE01 | 200 | 20 | 2,100 |
| 2021-01 | 4567 | AT01 | 200 | 2 | 600 |
| 2021-01 | 8901 | DE01 | 300 | 30 | 2,100 |
| 2021-01 | 8901 | AT01 | 300 | 3 | 600 |
| 2021-01 | 6900 | DE01 | 400 | 40 | 2,100 |
| 2021-01 | 8877 | DE01 | 500 | 50 | 2,100 |
| 2021-01 | 8900 | DE01 | 600 | 60 | 2,100 |

Table 3: Modified SHS-Base plus dataset before dropping duplicates.

Researchers have two options, both based on the dataset in Table 3.

– Option 1: Perform SDC after aggregation but before dropping duplicates.
– Option 2: Keep duplicates throughout analysis.

Both options take advantage of the fact that, in this example, SDC could be performed on the dataset in Table 1. They differ, however, with respect to the way in which researchers could obtain results after performing SDC. Option 1 calls for SDC before any results are obtained and states that if all individual aggregations pass SDC, then all results obtained later based on those individual aggregations will pass SDC as well. In the example, aggregations are based on individual ISINs. Therefore, if the sum holdings variable for each individual ISIN passes SDC, then all future results will do so as well. DE01 would pass this test, but AT01 would not, because the sum holdings variable for AT01 is based on three distinct BAIDs only. Therefore, option 1 is not applicable to our example, and researchers would have to turn to option 2.

7 Calculated as 1,100/2,100 and 500/600, respectively.

Recall that option 1 states that if all individual aggregations pass SDC (in the case above, for DE01 and AT01), then all results obtained later based on these individual aggregations will pass SDC as well. Since this is not the case here, researchers have to switch to option 2, which performs SDC on all results.

Option 2 states that researchers must not drop duplicates after aggregating values. The idea here is to be able to switch between a "results dataset" and an "SDC dataset." One approach would be to introduce a dummy equal to one if the row is the first observation for each ISIN, and zero otherwise. Table 4 presents this dataset. Thus, filtering on $dummy == 1$ results in the dataset in Table 2 (the "results dataset"), while not filtering at all leads to the dataset in Table 3, which is Table 1 plus one column for the sum holdings variable (the "SDC dataset"). We would recommend dropping all variables that are not essential to the analysis. In this example, we dropped the units variable to keep the dataset as small as possible, because the proposed datasets can become quite large and unwieldy. More ideas on how to work with large datasets at the RDSC can be found in Gomolka, Blaschke, & Hirsch (2021).

| Date | BAID | ISIN | Holdings | Sum holdings | Dummy |
|---|---|---|---|---|---|
| 2021-01 | 1234 | DE01 | 100 | 2,100 | 1 |
| 2021-01 | 1234 | AT01 | 100 | 600 | 1 |
| 2021-01 | 4567 | DE01 | 200 | 2,100 | 0 |
| 2021-01 | 4567 | AT01 | 200 | 600 | 0 |
| 2021-01 | 8901 | DE01 | 300 | 2,100 | 0 |
| 2021-01 | 8901 | AT01 | 300 | 600 | 0 |
| 2021-01 | 6900 | DE01 | 400 | 2,100 | 0 |
| 2021-01 | 8877 | DE01 | 500 | 2,100 | 0 |
| 2021-01 | 8900 | DE01 | 600 | 2,100 | 0 |

Table 4: Modified SHS-Base plus dataset before dropping duplicates with dummy.

## 2.2 Example 2: Aggregating values to allow datasets to be combined

### 2.2.1 Where does the tension arise?

As a second example, consider a researcher who wants to analyse the correlation between bank size and trading activity. His hypothesis is that larger banks on average trade larger amounts of securities. To analyse his question he needs to combine total assets data from the "Monthly Balance Sheet Statistics (BISTA)" dataset (Gomolka, Schäfer, & Stahl (2021)) with trading volume information from the "Markets in Financial Instruments Directive (MiFID)" dataset (Cagala, Gomolka, Krüger, & Sachs (2021)). We present the BISTA dataset in Table 5 and the MiFID dataset in Table 6.[8)] The SDC variables, i.e. the variables containing the IDs of entities that need to be protected, are "BAID" and "Counterparty." As before, the datasets used here have been modified for illustrative purposes and contain no real data points.

It is important to understand the similarities between this example and the one shown above. Combining these two datasets to analyse the research question also involves some form of aggregation, because it is necessary to calculate mean trading volumes per BAID per year. Aggregating

8 For simplification, we only show BISTA year-end values in our examples, i.e. year = 2019 refers to 2019-12-31.

| BAID | Year | Total assets |
|---|---|---|
| 1234 | 2019 | 1,000 |
| 5678 | 2019 | 2,000 |
| 8877 | 2019 | 3,000 |

Table 5: Modified BISTA dataset.

| BAID | Counterparty | Time | ISIN | Trading volume |
|---|---|---|---|---|
| 1234 | 1 | 2019-01-27 | DE01 | 100 |
| 1234 | 2 | 2019-05-16 | AT01 | 200 |
| 5678 | 3 | 2019-03-09 | DE01 | 300 |
| 5678 | 4 | 2019-12-12 | AT01 | 400 |
| 8877 | 5 | 2019-09-01 | DE01 | 500 |
| 8877 | 6 | 2019-10-01 | AT01 | 600 |

Table 6: Modified MiFID dataset.

variables to enable datasets to be combined is quite a common source of tension between the objectives of obtaining results and performing SDC.

It is instructive to look at this issue from a more technical angle. Note that both datasets in the example share the same identifier for banks – BAID – but differ with respect to the frequency with which each bank appears in the dataset. In the BISTA banks appear on an annual basis, while in the MiFID dataset trades are recorded on a daily basis.

To combine the two datasets, researchers either need to produce duplicates for the BISTA variables and keep all the information in the MiFID dataset, or they have to aggregate the MiFID variables so that there is a 1:1 match with the BISTA. Either way, it will be a challenge to perform SDC on the combined dataset. In the following, we will proceed by aggregating the MiFID data to meet the demands of the research question.

### 2.2.2 How can the tension be resolved?

Table 7 shows what the MiFID dataset from Table 5 looks like after aggregation of all the variables by BAID and deletion of all the resulting duplicates.

| BAID | Counterparty | Time | ISIN | Trading volume | Mean trading volume |
|---|---|---|---|---|---|
| 1234 | 1 | 2019-01-27 | DE01 | 100 | 150 |
| 5678 | 3 | 2019-03-09 | DE01 | 300 | 350 |
| 8877 | 5 | 2019-09-01 | DE01 | 500 | 550 |

Table 7: Modified MiFID dataset after calculating mean trading volume and dropping duplicates.

Table 7 is similar to Table 2, which presents the modified SHS-Base plus dataset after summing and dropping duplicates. Therefore, the same logic applies. Performing SDC on this dataset yields incorrect results.

In addition, there is a second layer of complexity when combining datasets. In our experience, researchers often first modify the MiFID data, as in Table 8, before combining them with the BISTA data. This makes sense, as combining the two datasets requires unique combinations of BAID and

Year (extracted from the time variable in the MiFID dataset). All other variables are dropped, as we recommended above, to preserve space. The combined dataset is presented in Table 9.

| BAID | Year | Mean trading volume |
|---|---|---|
| 1234 | 2019 | 150 |
| 5678 | 2019 | 350 |
| 8877 | 2019 | 550 |

Table 8: Common modifications to the MiFID dataset before combining it with the BISTA.

| BAID | Year | Total assets | Mean trading volume |
|---|---|---|---|
| 1234 | 2019 | 1,000 | 150 |
| 5678 | 2019 | 2,000 | 350 |
| 8877 | 2019 | 3,000 | 550 |

Table 9: Combined MiFID and BISTA dataset for analysis.

As this dataset is sufficient to obtain results but not to perform SDC (as discussed above), there are again two options to choose from. As option 1 will not work here, researchers must switch instead to option 2, which states that the duplicates need to be left in. This leaves the dataset looking as follows:[9)]

| BAID | Time | Total assets | Trading volume | Mean trading volume | Dummy |
|---|---|---|---|---|---|
| 1234 | 2019 | 1,000 | 100 | 150 | 1 |
| 1234 | 2019 | 1,000 | 200 | 150 | 0 |
| 5678 | 2019 | 2,000 | 300 | 350 | 1 |
| 5678 | 2019 | 2,000 | 400 | 350 | 0 |
| 8877 | 2019 | 3,000 | 500 | 550 | 1 |
| 8877 | 2019 | 3,000 | 600 | 550 | 0 |

Table 10: Combined BISTA/MiFID dataset with duplicates.

Researchers will use this new dataset to obtain results and perform SDC. However, what about the second SDC variable "Counterparty?" In the MiFID dataset, both "BAID" and "Counterparty" contain entities that have to be protected. Therefore, as an additional layer of complexity in this example, you will also need to carry this variable into the dataset in which you perform SDC. Table 11 presents the resulting dataset.

| BAID | Counterparty | Time | Total assets | Trading volume | Mean trading volume | Dummy |
|---|---|---|---|---|---|---|
| 1234 | 1 | 2019 | 1,000 | 100 | 150 | 1 |
| 1234 | 2 | 2019 | 1,000 | 200 | 150 | 0 |
| 5678 | 3 | 2019 | 2,000 | 300 | 350 | 1 |
| 5678 | 4 | 2019 | 2,000 | 400 | 350 | 0 |
| 8877 | 5 | 2019 | 3,000 | 500 | 550 | 1 |
| 8877 | 6 | 2019 | 3,000 | 600 | 550 | 0 |

Table 11: Combined BISTA/MiFID dataset with duplicates and counterparty.

**9** There are at least two different approaches to writing code that produces the resulting dataset. The first approach starts by combining BISTA (from Table 5) and MiFID (from Table 6) before calculating the mean trading volume variable on this combined dataset. Note that after the merge this approach is similar to the one described in Section 2.2.2, in that duplicates are not deleted after aggregating. The second approach starts from Table 9 and merges MiFID information (i.e. Table 5) back into this table. Let us call the result Table X. If researchers follow this approach, they have to show that the difference between the mean trading volume variable calculated from Table 9 and the mean trading volume variable calculated from the information in Table X is zero for all rows in Table X.

# 3 A rule of thumb to resolve the problem

The two examples presented above are a good way of raising awareness when the two objectives of researchers – obtaining results, and performing SDC – call for different datasets, which is a challenge researchers face when working with confidential microdata. Building on the examples provided above, we can offer researchers the following simple rule of thumb:

**When researchers use a summary function[10] in their analysis, they need to retain all the duplicates in the dataset (i.e. not use the "duplicates drop" command in Stata). Otherwise, they need (i) two separate datasets for obtaining results and performing SDC and need to show that (ii) both datasets are equivalent.**

Note that this is a rule of thumb. As with any such general guidance, there are exceptions to the rule. Two such exceptions are as follows:

1. This rule holds only for summary functions that generate a new variable. By contrast, using a summary function without generating a new variable leaves the underlying dataset intact. Therefore, researchers could obtain results and perform SDC on the same dataset.
2. This rule of thumb also does not hold if researchers calculate their summary function over all SDC variables[11] in the dataset (e.g. aggregating over BAID and counterparty variables in the MiFID dataset).

The rule of thumb above follows from the observation that both examples have some features in common:

1. Data handling involves aggregating values in a variable.
2. Values are aggregated without taking into account (all) SDC variables.
3. Aggregating over values in variables creates duplicates which are then dropped.
4. The solution always demands that any duplicates should not be dropped.
5. SDC can be performed on the original dataset.

**10** See the glossary for a definition.
**11** See the glossary for a definition.

# 4 Conclusion

We provide two examples that highlight the importance of already bearing SDC in mind when preparing data for analysis. This is because the aggregation of values will sometimes impair a researcher's ability to perform SDC, which creates tension between the objectives of obtaining results and performing SDC. Tables 4 and 11 illustrate how this tension might make a researcher want to drop duplicates in order to obtain results. However, the resulting datasets would be insufficient to perform SDC.

# Glossary

## Key variables

One or more variables (combined) enable the distinct identification of each cell of the other columns. As these variables are the key to identifying the other cells, we refer to them as "**key variables**." We refer to the key variables in their entirety as the "**set of key variables**." Database terminology may also call the set the "compound key" or "unit of observation."

## SDC variables

Variables containing the IDs of sensitive entities (e.g. SYSNR, BAID or real names) are referred to as "**SDC variables**." As the classification of SDC variables is purely context-related and depends on whether the entities contained therein need to be protected or not, the RDSC specifies SDC variables for each original dataset in its data reports. These reports identify SDC variables as such for each individual dataset, which is why a variable may be an SDC variable for one original dataset but not for another.

## Summary functions

Some functions may calculate results based on multiple values from multiple rows, which leads to an aggregation of information. Examples include sums or means across several rows. We refer to these functions as "**summary functions**."

## References

Becker, T., Biewen, E., Schultz, S., & Weisbecker, M. (2021). *Individual financial statements of non-financial firms (JANIS) 1997-2020, Data Report 2021-24 – Metadata Version 8*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/882750/712ad07b1d840472cfef119b24fd977d/mL/2021-24-janis-data.pdf

Blaschke, J., Sachs, C., & Yalcin, E. (2021). *Securities Holdings Statistics Base plus, Data Report 2021-18 – Metadata Version 4-1*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/875904/3102d22febaadcb0a9ba4d05df2db28e/mL/2021-18-shsb-data.pdf

Cagala, T., Gomolka, M., Krüger, M., & Sachs, K. (2021). *Markets in Financial Instruments Directive (MiFID), Data Report 2021-13 – Metadata ID 1.0*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/634886/3414886ff73afe8ce02a009084b812e5/mL/2021-13-mifid-data.pdf

Gomolka, M., Blaschke, J., & Hirsch, C. (2021). *Working with large data at the RDSC, Technical Report 2021-04 – Version 1.1*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/623988/62b8c17881d63bccc61efd11af0b47db/mL/2021-04-large-data-data.pdf

Gomolka, M., Schäfer, M., & Stahl, H. (2021). *Monthly Balance Sheet Statistics (BISTA), Data Report 2021-10 – Metadata Version BISTA-Doc-v3-0*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/747114/29fb691901403642127353c7935116ef/mL/2021-10-bista-data.pdf

Research Data and Service Centre. (2021). *Rules for visiting researchers at the RDSC, Technical Report 2021-02 - Version 1-0*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/826176/ffc6337a19ea27359b06f2a8abe0ca7d/mL/2021-02-gastforschung-data.pdf