On-site Service and Safe Output Checking in Japan

Ryo Kikuchi*, Kazuhiro Minami**

* National Statistics Center / NTT, Tokyo, Japan. kikuchi.ryo@lab.ntt.co.jp, 9h358j30qe@gmail.com

** National Statistics Center / Institute of Statistical Mathematics, Tokyo, Japan. kminami@ism.go.jp

Abstract. In Japan, we have been establishing the on-site service that allows researchers to access microdata of various public surveys at secure on-site facilities. Since the microdata contains sensitive information of the survey participants, we have been establishing output checking rules referring to those of other countries including the one on the SDC handbook.

We, however, find several remaining technical issues that make it difficult to ensure the safety of output data. Particularly, we find that to verify the safety of tabular data challenging under the presence of differential attacks. In this paper, we introduce those technical challenges in safe output checking and describe our R-based tool for statistical disclosure control, which we developed to address some of the issues and bridge a gap between a task of statistical disclosure control and that of output checking.

1 Introduction

Decision making based on solid data analysis has been increasing popular both in public and private sectors recently. Statistics offices in many countries have been trying to push this tread further by making various kinds of survey data, which has been accumulated for many years, available for secondary data analysis for policy making and academic research.

In Japan, a researcher has been able to use microdata for research purposes. However, the researcher must obtain a permission by submitting precise research objectives with a detailed plan, which usually involves a long time-consuming review process. In addition, even if the permission is obtained, only restricted number of attributes specified in the research plan will be available to the researcher. Therefore, the rigid structure of the current program puts a lot of burden to researchers and makes them difficult perform exploratory analysis that needs spontaneous access to various attributes.

To address this issue, we have been establishing the on-site service from January 2017, which is currently at a trial stage. This service allows a researcher to access

microdata of various public surveys at a secure on-site facility. A researcher visiting a facility can perform various statistical analysis on microdata interactively. While establishing the on-site service, we have to establish safety rules for output checking to protect sensitive information in microdata. We need to make sure that any sensitive information in microdata is not disclosed publicly when a researcher brings analysis results back home from the facility.

1.1 Contribution

In this paper, we introduce our safety rules for output checking that determine whether or not a researcher is allowed to bring out analysis results from the on-site facility. We mainly base our rules on the SDC handbook published from Eurostat [3] with several modifications. The two main modifications are 1) to take measures against differential attacks and 2) to ensure that suppressed cells of tabular data have a sufficient length of confidentiality intervals.

To put our safety rules in practice, we identify several remaining technical issues that make it difficult to ensure the safety of output data. Particularly, we find the process of suppressing sensitive cells and the checking of its correctness in tabular data challenging in terms of both operational labor and computational costs. We also find that there does not exist any effective technical measure against differential attacks that obtain sensitive information by taking the difference among multiple output data.

To address the issue of checking tabular data safety effectively, we consider automating the process of cell suppressions and their correctness checking for tabular data. We first examine τ-Argus as a tool for performing statistical disclosure control on tabular data automatically. However, we find that there are several functionality and usability issues in τ-Argus. Therefore, we develop a prototype tool in R that performs cell suppressions automatically and export all necessary information so that an output checker later verifies the safety of output table.

2 Philosophy in designing safety standards

In this section, we describe our philosophy for designing safety rules for output checking.

2.1 Procedure of output checking

We show the procedure of the on-site service in Figure 1 where output checking is performed at two phases. The first phase verifies the safety of output data with a well-defined set of rules, which we mainly focus on in this paper. The first check aims at the syntax check of the output without considering its semantic. For example, the first check applies the exact same set of safety rules to sampled and non-sampled data.

The second phase considers properties or semantics of microdata to determine the safety of derived output data. An output checker at this phase is usually a data manager who are knowledgeable about used microdata. For example, an output can be safe if the output is derived from sampled data or attributes that are not easily observable, even if it does not meet the rules at the first phase.

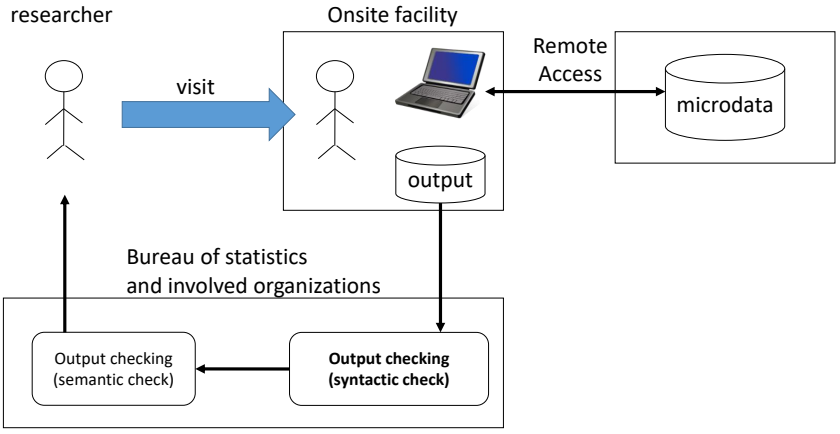

Figure 1: General picture of on-site service

2.2 Attack model

To determine appropriate safety rules, it is necessary to decide how researchers of the on-site service could be involved in information leakage. Roughly speaking, we consider two types of researchers. One is a malicious researcher who attempts to bring sensitive information out using any strategies. The other is an honest researcher leaks information by accident without any malicious intent since s/he is not familiar with disclosure control.

We decide to assume the latter type of researchers for our on-site use program. To justify the assumption of an honest researcher, we plan to provide a training program on safe output checking. More precise discussion can be found in Appendix A.

2.3 Uniformity of output checking

It should occur that researchers wish to bring not only a final output of their analysis but also intermediate outputs since a researcher may want to perform extensive analysis on the intermediate outputs but the on-site facility is not available in all day. In conclusion, we apply the same safety rules to both intermediate and final outputs. We discuss this issue more in Appendix B.

2.4 Responsibility for the prevention of privacy breach

We impose the final responsibility for preventing information disclosure from a research paper to the researchers who publish that paper. The purpose of safety rules for output checking is to catch unsafe outputs prepared by an inexperienced researcher who are unfamiliar with disclosure control. The scope of our safety rules is limited to a set of conventional statistical analysis methods producing traditional data types such as tabular data. The safety rules are not designed to detect all possible information disclosures for all types of data processing since we need to process output checking tasks in a timely manner.

2.5 Following current safety standards

Some countries have advanced in the operation of research data centers, which are similar to our on-site service. In addition, Eurostat, the statistical office of the European Union, discloses technologies relating to the safety rule in order to share knowledge and experience.

Looking at the existing safety standards, not all of them are logically derived from scientific evidences, and rule-of-thumbs, which are inductively extracted from previous examination cases, play an important role. Therefore, to make a new rule for output checking, it is reasonable to adopt rule-of-thumbs that have already been proven in each country.

We follow Eurostat's guideline [3]. However, we find several remaining technical issues, and we therefore modify the rules in Eurostat's guideline. We explain the difference in Section 3.

In addition, Eurostat [3] provides four principles: 10 units, 10 degrees of freedom, dominance rule, and group disclosure. We follow these principles and add another one; each individual value is confidential. These principles and representative information-leakage scenarios appear in Appendix C.

3 Possible outputs and checking rules

Due to the lack of space, we focus on the difference between our safety rules and the Eurostat's guideline. We can approve the following eight types of outputs: frequency table, magnitude table, regression coefficients, percentile, mode, concentration ratio, average and sum, and summary and test statistics (including index, ratio, indicator, moment of distribution, and correlation coefficient). Although these types of outputs are almost same as those of the Eurostat's guideline, some of them are merged for simplicity. For more precise difference, please see Appendix D.

Frequency and magnitude tables. Eurostat's guideline applies the principles of "10 units" and "group disclosure" to both frequency and magnitude tables, and that of "dominance rule" only to magnitude tables. Our safety rules for output checking additionally address a risk of differential attacks.

An example of the differential attack among multiple tables is shown in Table 1. In each table, any cell contains at least 10 units and these tables can be considered

	[0-18]	[19-25]	[26-30]
Area A	18	19	20

	[0-20]	[21-25]	[26-30]
Area A	19	18	20

Both tables show the number of individuals, where the response variables are age and area.

Table 1: Example of differential attack

as safe with respect to the 10 units principle. However, one can see that the number of individuals whose age are [19-20] is only one by differencing two tables, which contradicts the 10 units principle.

Although the risk of differential attack exists, implementing a measure against differential attacks is problematic; it is unrealistic to check the segmentation difference between the current output table and the whole tables generated in the past. A simple solution is using preliminary-defined segmentations, and other possible solution is employing some perturbation. This is still under consideration in the current trial stage.

When we suppress sensitive cells in tabular data, we need to ensure that each variable of a suppressed cell should have enough uncertainty about its value. Although Eurostat's guideline does not explicitly define this requirement, our guideline specifies a minimum threshold width of a possible value interval, such that the threshold width is 10 elements for frequency tables and 30% of a cell value for magnitude tables.

Regression coefficients. We use the safety rule that is described as principle-based model in Eurostat's guideline. Although regression coefficients are safe in rule-of-thumb if one estimated coefficient is withheld, some researchers want to the exact regression coefficients. Therefore, we use the principle-based model to allow researchers bring the coefficients out.

Mode. We add the 10 units principle in order to ensure the mode has to be computed from more than 10 units.

Index, ratio, and indicator. These are treated as one of summary statistics while they are not in Eurostat's guideline. There have been many indices, ratios, and indicators and their risk depend on how complex they are. This situation is the same as summary statistics and we therefore merge them.

4 Automation for output checking

As we describe in Section 2.5, there are several key issues in checking the safety of tabular data. In this section, we revisit those issues in detail and discuss possible solutions using τ-Argus and the prototype tool we developed in R.

4.1 Motivation of automation

It is difficult to properly perform disclosure control on tabular data manually due to two issues. The first issue is that to check the confidentiality interval of suppressed cells in tabular data requires to solve linear programming problem. After suppressing cells of less than a given threshold number of units in a table, we must perform secondary suppressions to ensure that each hidden cell has sufficiently wide range of possible values, and solve the linear programming problem to compute this range under the linear relations concerning marginal sums in the table. Although the Eurostat handbook [4] describes the requirement regarding this confidentiality interval, there is strangely no mentioning in the guideline for the checking of output [3].

The second issue is that to check the safety of a magnitude table requires an output checker to examine the original microdata from which the table is derived. Figure 2 shows a procedure for checking a dominance rule on a magnitude table. A researcher first constructs a magnitude table by aggregating values in multiple records in the microdata, and then submits that table to the output checker. However, the checker needs to access the microdata to check the dominance rule on each cell of the table because the rule refers to the values of individual units contributing to the cell. For example, verifying $p\%$ rule requires the values of the largest and the second largest ones in each cell.

The second issue brings another difficulty. To split the responsibility for checking dominance rules between a researcher and an output checker is a difficult task. A researcher is supposed to supply all necessary additional information for output checking, such as a corresponding frequency table, a set of unit values in each cell, and so on. However, this requires a lot of work on researchers. On the other hand, if an output checker needs to check the safety of a magnitude table itself, he essentially needs to recompute that same magnitude table while producing auxiliary information that is necessary to verify the rule.

We, therefore, explore the possibility of using output checking tools that automate many of the output checking tasks discussed in this section.

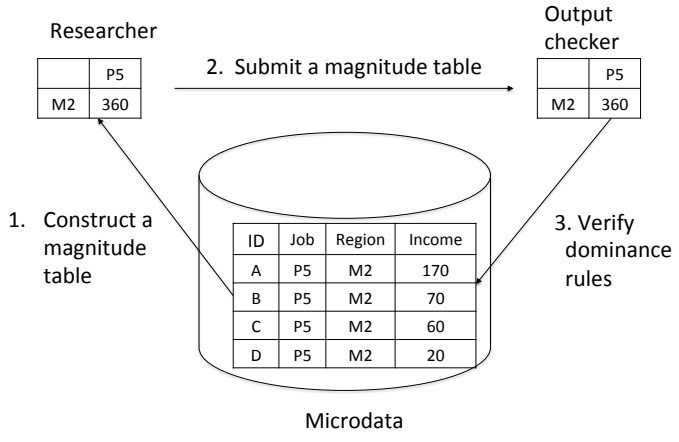

Figure 2: Procedure for dominance rule checking. The output checker needs to access the microdata to obtain the values of four contributing units, A, B, C, and D to verify the dominance rule.

4.2 Using τ-Argus

τ-Argus [2] is a software for performing disclosure control on tabular data. τ-Argus is mainly designed for researchers with strong background on statistical disclosure control to produce safe tabular data interactively. Since τ-Argus supports several variants of dominance rules, such as the threshold rule, (n, k)-rule, $p\%$ rule, for magnitude tables, a researcher can explore parameter spaces of safety rules using τ-Argus to produce desirable safe tabular data.

However, if we use τ-Argus in our on-site service, an output checker needs to perform output checking on tabular data, which is prepared by a researcher using τ-Argus. It is crucial for τ-Argus to export a suppressed table with all auxiliary information that is necessary for an output checker to verify its safety. We, therefore, evaluate the feasibility of using τ-Argus in the framework of our on-site service by considering the following output checking scenario.

1. A researcher generates a table with an analytic tool of his choice and exports that table to a file of a standard format (e.g., csv).

2. A researcher converts that file into a file of another format so that τ-Argus can read it.

3. A researcher imports the input table to τ-Argus and performs cell suppressions on that table and exports a suppressed table and explanatory materials to files for output checking.

4. An output checker examines those files to decide whether or not the table suppressed by τ-Argus is safe according to our safety rules.

At step 2, τ-Argus requires a researcher to define metadata as well as a table of a certain format defined by the metadata. We show an example of metadata in Appendix E.

If we use τ-Argus in this scenario, we found several functional limitations and usability issues as follows:

- Input and output files of τ-Argus have specific formats. It is not reasonable to expect researchers to prepare such input files of the specific format comforming to the metadata.

- The interface of τ-Argus is difficult for a researcher who is not familiar with disclosure control. τ-Argus has been designed for experts of disclosure control. Therefore, it provides many options that are difficult for ordinary researchers to understand and use them properly.

- As far as we explored, τ-Argus does not support the group disclosure. We apply the group disclosure to both frequency and magnitude tables. Therefore, the lack of group disclosure is critical to our on-site service.

The first issue can be addressed by preparing a tool for converting the specific format of τ-Argus into a common format. The second issue can be also avoided by generating a batch file that enables a researcher to suppress the table via τ-Argus without interacting with τ-Argus to choose security parameters, which is suggested in [6]. However, we need to wait τ-Argus to be extended to support functionality of group disclosure prevention to resolve the third issue.

Whereas the above limitations and difficulties, τ-Argus has several advantages. It can manage a table of more than two dimensions. In addition, a researcher can use some sophisticated optimization algorithms for secondary suppression while choosing the best cost function that suites the needs of the researcher. We will continue to explore the possibility to use of τ-Argus for output checking in the near future.

4.3 Using tailor-made commands in R

We develop a prototype tool in R to address the issues of τ-Argus. We provide several functions for primary and secondary suppressions in R, and, thus, as long as a researcher produces tabular data in R, we do not have any issue of converting data format of tabular data. Our tool, which is a set of R functions, does not provide any GUI interface as τ-Argus does, but it exports all necessary information to files so that an output checker later verifies the safety of submitted tables.

We first introduce the function *suppressFT*, which performs primary and secondary suppressions on a frequency table. Figure 3 shows the functionality of the function *suppressFT*, which takes as inputs an original table and four security threshold parameters for unit frequency, row sum dominance ratio, column sum dominance ratio, and confidentiality interval, and outputs a secondary suppressed table and the confidentiality intervals of the suppressed cell variables. The function also exports the same information to the file so that an output checker can refer them to check the safety of the suppressed table.

Figure 4 shows that the suppression process consists of multiple steps. At a high level, the process is divided into two stages: the primary suppression process and secondary suppression process. The primary suppression process is further broken down into three steps: frequency checking, row sum dominance checking, and column sum dominance checking. An original table is incrementally suppressed at each step of the primary suppression process, and then the primary suppressed table is further suppressed such that the confidentiality interval of each suppressed cell is greater than the specified threshold interval.

We take a greedy approach for secondary cell suppressions because we prefer the efficiency of the algorithm to the optimality of a solution among all cell suppression patterns. The algorithm starts with a primary suppressed table. We compute the

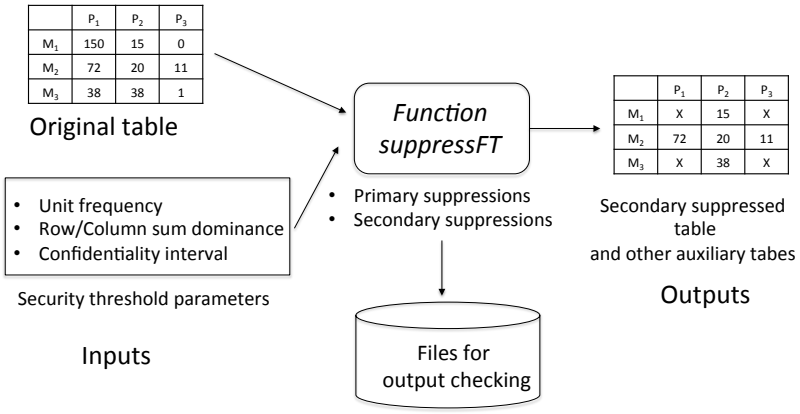

Figure 3: Function *suppressFT*

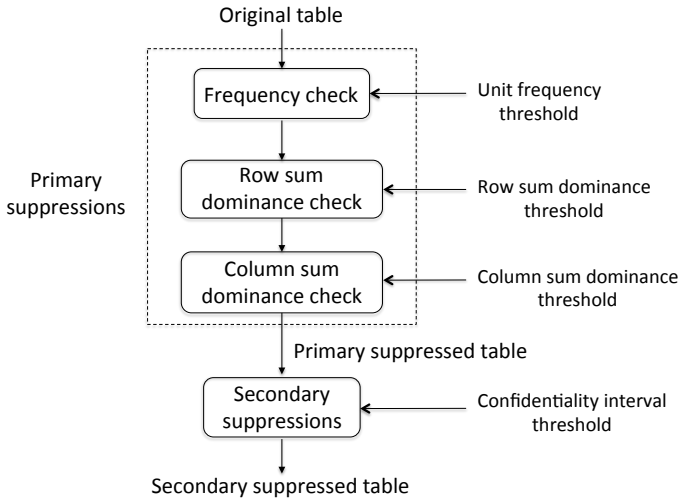

Figure 4: Incremental suppressions on a frequency table. The three steps in the dotted rectangle consist of the primary suppression stage. After the primary suppressions are performed, the suppressed table is further suppressed at the secondary suppression stage.

maximum and minimum values of each suppressed cell and derives a confidentiality interval. We repeatedly suppress a cell one by one until all the suppressed cells have sufficient width of confidentiality intervals. When we pick a new cell to be suppressed, we use the heuristic of choosing the cell of the minimum value either a row or column with the least number of suppressed cells. Since our tool can handle a manually suppressed table by a researcher, we believe that our tool serves the needs of researchers in many realistic situations.

To suppress magnitude tables requires more involved procedure as shown in Figure 5. A magnitude table is primary suppressed based on dominance rule (e.g., $(2, k)$-dominance rule, $p\%$ rule, etc.). Separately, the corresponding frequency table must be suppressed as well, and the cell positions of suppressed cells in the frequency table must be incorporated into the result of primary suppressions for the magnitude

table. We finally perform secondary cell suppressions on this merged magnitude table and obtain the final result.

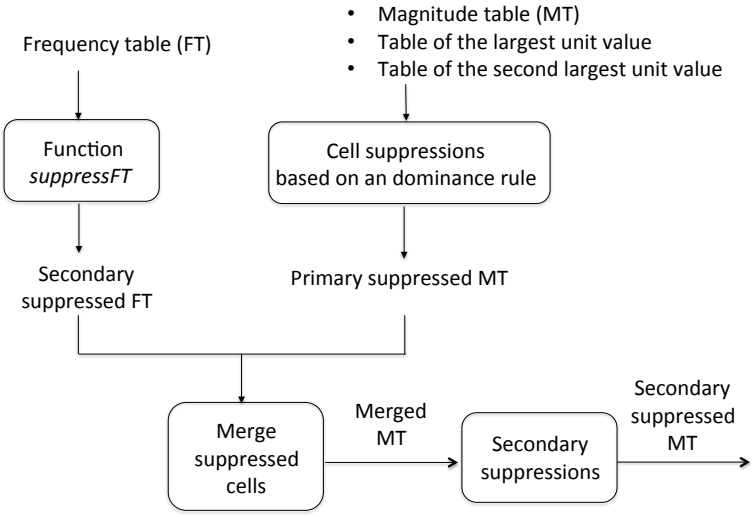

Figure 5: Procedure for suppressing a magnitude table.

5 Future work for the second checking

As we explained in Section 2.1, we conduct two-phase output checking. At the second stage of output checking, we consider semantics and other characteristics of output data to determine its safety. We believe that we should pay particular attention to the following issues: rules for applying dominance rule and relaxing safety rules by sampling rate. We discuss them in Appendix F.

6 Conclusion

We discuss safety rules and output checking procedures for the on-site service in Japan. We identified several technical issues in Eurostat's guideline and make necessary modifications to address the issues of differential attacks and the verification of confidentiality intervals for suppressed cells in tabular data.

To put our safety rules in practice, we explore the possibility of using a software tool that performs both statistical disclosure control and output checking on tabular data automatically. Since we find several pitfalls in using τ-Argus for our on-site service, we developed a prototype tool in R to achieve an efficient transition from statistical disclosure control on tabular data to its output checking.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP15K00195 and JP16H02013.

References

[1] FAQ: Frequently asked questions. `http://neon.vb.cbs.nl/casc/FAQ.htm#WhichSensRule`.

[2] τ-Argus homepage. `http://neon.vb.cbs.nl/casc/tau.htm`.

[3] M. Brandt, L. Franconi, C. Guerke, A. Hundepool, M. Lucarelli, J. Mol, F. Ritchie, G. Seri, and R. Welpton. Guidelines for the checking of output based on microdata research.

[4] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. S. Nordholt, G. Seri, and P.-P. D. Wolf. Handbook on statistical disclosure control.

[5] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125. IEEE Computer Society, 2008.

[6] A. T. Staggemeier, P. Lowthian, and G. Lee. Applying Tau-Argus to Super-CROSS tables: A practical example using the uk business register unit data. In *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2007.

[7] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

A The attack model

As we describe in Section 2.2, we decide to assume honest researchers for our on-site use program since handling a malicious researcher by safe output checking is unrealistic from both technical and practice viewpoints. For example, if a malicious researcher secretly embeds sensitive information into cells of a table in a hidden way, detecting such modifications requires an output checker to essentially recompute all the outputs resulting in unacceptable burden to the output checker. Furthermore, there is no technical measure to prevent a researcher from keeping sensitive information, which he come across during the analysis, in his mind even after exiting from the facility.

To justify our "honest" researcher model, we plan to take a combination of non-technical measures; we provide a training program on safe output checking, obtain

a non-disclosure agreement from researchers, certify a punishment to a researcher when an improper disclosure of sensitive information occurs.

Rather than honest researchers, we consider a reader of research papers including analysis results on microdata as a possible adversary. Such an adversarial reader attempts to infer personal sensitive information by re-identifying a survey participant contributing to certain analysis data. Such a malicious reader usually tries to associate an analysis result in the paper with some external knowledge, which he obtains in another way.

B Uniformity of output checking

Although intermediate outputs tend to include much more detailed information than the final result, we do not relax our safety rules for intermediate outputs based on the following two reasons:

- We cannot trust researchers in performing appropriate disclosure control on intermediate outputs such that they satisfy the safety standards of the final product.
- Once a researcher takes out intermediate outputs from the on-site facility, there is no reliable way to track them down to make sure that the final output for publication satisfies the conditions of the safety rules.

C Five principles

We describe five basic principles based on which all safety rules are defined below.

Each individual value is confidential The first principle states that we should never disclose any attribute value of an individual included in microdata. There is a possibility that the household income of a certain individual is disclosed from a table including household incomes. Normally, if the identity information is removed from the table, it is difficult for an adversary to re-identify the record in that table corresponding to the individual. However, if the disclosed household income is the highest income in a small local area, such a disclose becomes a real risk. Some neighbors may know the person of the highest income in that area from observable information on the lifestyle of residence (e.g., owned cars). In this case, the accurate income of the individual's household can be revealed to the neighbors.

Even if a released value is not an extreme value such as the maximum and the minimum, a combination of several attributes of the same person, such as age, sex, and job, can identify a single individual. In fact, there are several incidents in which such privacy breach occurs [7, 5]. Therefore, even if a single attribute value of a person itself seems to be safe, we have to consider a risk of combining multiple

attributes. Since it is difficult to determine which attributes could be part of a set of identifying attributes, We treat *any* attribute value of each individual as confidential.

10 units The second principle states that all output must be an aggregate of at least 10 units (unweighted) contributing to the aggregate value. This is the same principle in [3]. We represent a data analysis task conducted by a researcher by a n-array function $y = f(x_0, \ldots, x_{n-1})$, which takes $x_0, \ldots, x_{n-1}$ as inputs and outputs y. Inputs $x_0, \ldots, x_{n-1}$ and y corresponds to the attribute values of individual units in microdata and the output to be released, respectively. Here we consider a risk of an adversary obtaining y inversely determine x_i for some $0 \leq i < n - 1$.

Intuitively speaking, the greater the number of input units, the better security we achieve, since the number of unknown values among x_i that the adversary does not know increases leading to more uncertainty about the value of x_i. On the other hand, less number of inputs provides more detailed output with better data utility and a researcher can bring more detailed outputs out from the facility. Therefore, we have to choose the right "threshold" to take a balance between security and usability. Although each country uses a different threshold value for the minimum number of input data, we use 10 as a threshold following the Eurostat's guideline.

11 degrees of freedom The third principle states that the degree of freedom, which is defined as "the number of units" minuses "number of parameters" and "other restrictions in the model," must be greater than 11. This principal is applicable to statistical analysis results such as linear regression. Although the aim of this third principle is same as the third one, we employ not 10 but 11 as a threshold. In some cases, it is difficult to determine whether an output should satisfy which number of principles, 10 units or 11 degrees of freedom. Since the degree of freedom of simple statistics is the number of units minus 1, we pick 11 rather than 10 so that we can use the same threshold value (i.e., 10) for the second or third principles.

Dominance rule The fourth principal states that we must prevent the largest contributor of an aggregate value from occupying more than a threshold ratio of the aggregate value. This principle is mainly applicable to cells in a magnitude table and similar aggregate outputs. In this principle, we consider the risk of having $f(x_0, \ldots, x_{n-1}) \approx x_i$ for some $i \in \{0, \ldots, n-1\}$. To prohibit such a risk of dominating units, we use 50% as a threshold ratio, which is specified by $(1, 50)$ rule.

We currently use $(1, 50)$ rule but will use p% rule, which is supposed to provide better security, in the future. The p% rule assumes the second largest contributor is an adversary. Let x_0 is the value of the largest contributor of the value X and x_1 is the one of the second. The adversary estimates the value of the largest contributor as $X - x_1$ and $(X - x_1) - x_0$ is the difference. If the difference is larger than p % of

x_0, X is regarded as safe. Since the p% rule addresses the concrete attack scenario above, we believe that p%-rule is more preferable to (n, k) rule as suggested in [1].

Group disclosure The fifth principle states that we must prevent a situation that an individual unit belongs to a certain group with high probability. If all the members in a group have the same sensitive value, to learn just that an individual belongs to that group reveals the sensitive information about that individual. To prevent the issue of such a *group disclosure*, we define the safety rule for tabular data requiring that no cell can contain more than 90% of the whole units in its row or column, for example.

D Difference of possible outputs

In Eurostat's guideline, average, indices, ratios, and indicators are treated as the same way. However, it is difficult to separate these values with summary statistics, which is in different way. The difference between these values and summary statistics is whether or not applying the dominance rule. In this paper, we apply the dominance rule with an average and sum but does not with the other values. This is because applying the dominance rule heavily depends on the "complexity" of computing the values and it is difficult to manage them in the same way.

E Required information for τ-Argus

We give an example of τ-Argus's input of magnitude table. Table 2 is the magnitude table we want to input. Table 3, 4, and 5 are the tables of frequency, the largest contributor, and the second largest contributor, which are required information for output checking.

	a	b	TOTAL
A	40	50	90
B	20	30	50
TOTAL	60	80	280

Table 2: Magnitude table

	a	b	TOTAL
A	10	10	20
B	10	10	20
TOTAL	20	20	40

Table 3: Frequency table

	a	b	TOTAL
A	10	12	12
B	8	7	8
TOTAL	10	12	12

Table 4: The largest contributor

	a	b	TOTAL
A	5	8	10
B	5	7	7
TOTAL	8	8	10

Table 5: Second largest contributor

τ-Argus requires meta-data as well as input-data for specifying the format of the input-data. An example of input-data and meta-data for τ-Argus is Table 6 and 7. For example, the first line specifies that each value in the input-data should be separated by a comma, and the lines from the second to fourth specifies that S, U, and P are used as special characters to represent a cell is regarded as safe/unsafe/protected in default. In the input-data, one line represents one cell. Each line contains two explanatory variables, frequency, value of magnitude table, the largest contributor, the second largest contributor, and default act of the corresponding cell in order.

```
A,a,10,40,10,5,s
A,b,10,50,12,8,s
A,TOTAL,20,90,12,10,s
B,a,10,20,8,5,s
B,b,10,30,7,7,s
B,TOTAL,20,50,8,7,s
TOTAL,a,20,90,10,8,s
TOTAL,b,20,80,12,8,s
TOTAL,TOTAL,40,280,12,10,s
```

Table 6: Input-data for τ-Argus

```
<SEPARATOR> ","
<SAFE> "S"
<UNSAFE> "U"
<PROTECT> "P"
Vname
    <RECODEABLE>
    <TOTCODE> "TOTAL"
Class "x"
    <RECODEABLE>
    <TOTCODE> "TOTAL"
N
    <FREQUENCY>
VAR
    <NUMERIC>
TOP1
    <MAXSCORE>
TOP2
    <MAXSCORE>
StatusVar
    <STATUS>
```

Table 7: Meta-data for τ-Argus

F Details of future work for the second checking

Rules for applying dominance rule

As we discuss in Section 4.1, to check a dominance rule on tabular data involves a lot of technical and procedural problems. It is desirable to skip the task of checking a dominance rule if possible. However, to decide whether or not to apply the dominance rule to tabular data, we have to specify criteria based on scientific evidences.

One candidate of criteria is whether or not categorical attributes of a table are easily observable. Although a table sometimes contains attributes (e.g., a couple with a large age gap) that are easily observable, we expect that there are many non-observable attributes (e.g., precise salary, clinical history, etc.) as well. It might be possible to skip a task of checking the dominance rule if the table use the attributes

that are difficult to be observed.

Another candidate is whether or not an adversary can recognize who is the second largest contributor. The dominance rule, especially p% rule, assumes an adversary who is the second largest contributor. However, in some cases, the assumption may not hold due to some reasons, (e.g., microdata is sampled from population data). Therefore, it may be possible to skip a task of checking the dominance rule if it is difficult for the second contributor to convince s/he is in fact the second.

Relaxing safety rules by sampling rate

We uniformly apply the principles of 10 units / 11 degrees of freedom for all microdata. However, the risk of cells less than 10 units depends on whether or not microdata has been sampled. In other words, sample unique does not imply population unique. There have been many studies about the gap of two "uniques" and reflecting those is a future work.