A step-by-step process to deal with the protection of a set of tabular data

Julien Jamme[1], Clara Baudry[2]

[1]*Insee, France*

[2]*Insee, France*

Abstract

Before releasing a set of tabular data to the public, NSIs have to minimize the risk of disclosure of confidential information. The people responsible for protecting sets of tabular data using a suppressive method are confronted with many issues that require a certain level of expertise. However they struggle to reach this level of expertise since they protect tabular data only once a year. Even though tools such as Tau-Argus protect tables efficiently, they do not eliminate all the practical difficulties. In fact, one of many challenges for producers is to understand that the set of tables they want to release is not the same as the set of tables they need to protect. For example, a producer might want to release a table that breaks the population down by region and another table that breaks the same population down by municipality. In this case, the two tables must be merged into a single one in order to take into account the hierarchical structure present between municipalities and regions. The step-by-step process described in this paper is inspired by the roadmap suggested by Hundepool et al. in the Handbook on Statistical Disclosure Control, but specialized in the tabular data protection.

At Insee's statistical methods department, a team is developing a step-by-step process to harmonize methodology, and has been implementing tools (such as the *rtauargus* package) to reduce the level of expertise required to protect a set of tables.

Keywords: Disclosure Control, tabular data, quality

1. Introduction

The dissemination of tabular data is subject to protective measures against the disclosure of confidential information, primarily through suppressive methods. While data protection is an integral part of the production process (downstream in the chain), the integration of this protection phase is not self-evident in practice.

Indeed, this step is singular because it is most dependent on the form the dissemination will take. The stages of collection, cleansing, handling of non-response, calibration, etc., are performed independently of the form of dissemination. Distributing one table or another will not radically change the form and substance of the data processing operations. On the contrary, data protection – regardless of the method used – depends on what is actually disseminated or not. With suppressive methods, this dependence is total: removing or adding a table, a variable, a breakdown, or even just a modality of dissemination can lead to changes in the implementation of processing and the results obtained. Moreover, even annual dissemination can vary from one year to another. The removal of a variable or the addition of others is frequent even for the most enduring and stable sources. Thus, integrating this step into the current production process is genuinely a difficult task.

Furthermore, a statistical institute rarely limits its dissemination to a few tables per source. To meet the needs of all potential users, it is not uncommon to disseminate hundreds of tables from a single source. Dissemination sometimes occurs at various times during the year, via different mediums (direct availability on the internet or more specialized or public publications). However, when using suppressive methods, all disseminated tables should be managed simultaneously to ensure consistent and optimal data protection. This requires the producer to have a very clear idea of the dissemination details sometimes several months before the tables are actually published.

The main challenge, however, remains the complexity of accomplishing data protection when employing suppressive methods. The numerous tables to be disseminated contain many interrelations that must be taken into account (See Jamme & Rastout, 2023, for an example). Moreover, analyzing these links is a specific and sometimes complex task that requires expertise and experience to successfully carry out the protection task. This task is also made more difficult by the inherent limitations of the methods and tools at our disposal.

Faced with these organizational, problem conceptualization, expertise acquisition, and

tooling challenges, the prevailing sentiment among producers at this stage of the process is one of feverishness, indicating a certain lack of confidence and control over the mask production process. Such unease among producers is such that the production burden of the most complex cases has long been delegated to a small team in the statistical methods department of the Insee.

Such unease among producers led to the burden of production of the most complex cases to be delegated to a small team of Insee's methodologists. As it is not the role of such a department to internalize a production burden over time, we present here the efforts made by this team to lighten it, with the ultimate ambition of giving the producer complete control over the protection of their tabular data.

Firstly, the methodological approach followed to review and improve the production mask process is presented. Secondly, this whole process is quickly described in a second time. Then, the three main steps which have been particularly targeted by our work – the metadata file, the analysis and the secondary suppression step – will be introduced.

2. Methodological approach

Starting in 2021, the team responsible for the production of confidentiality masks within the statistical methods department of Insee sought ways to lighten the burden, with the ultimate goal of transferring it to the producer. This medium-term objective necessitates a complete overhaul of the entire data protection process to regain control over all its stages. The approach involves describing the process stages, generalizing the problems encountered at each of these stages, automating tasks wherever possible, and disseminating these practices. This work is conducted concurrently with production tasks, with a constant focus on continuous improvement.

Describing the process allows for its segmentation into several steps, each with its inputs and outputs, tasks to be completed, and its own challenges. The idea is to create process similar to what Hundepool et al., 2010, proposed in their "Roadmap to releasing a microdata file". *Generalizing* involves proposing a unified approach to perform a task or solve a problem regardless of the form of the request, the source, or the data producer. This entails transcending the vast diversity encountered by confidentiality experts. Automating entails finding ways to implement a tool that performs various process tasks while reducing the level

of expertise required to execute them. Lastly, dissemination involves sharing the methodology, concepts, and tools developed by the expert team with potential users and engaging with their feedback to refine the methodology, clarify concepts, and continuously improve tools.

Of course, this work builds upon the efforts of previous Insee teams that have tackled the most complex issues related to confidentiality mask production. Furthermore, the fundamental concepts and general approach to protecting tabular data using suppressive methods are outlined in the chapter 5 of the *Handbook on Statistical Disclosure Control* (Hundepool et al., 2010). Finally, this endeavor would not be feasible without existing tools to carry out the most technical tasks. This notably includes the Tau-Argus software (de Wolf et al., 2023), within which a set of algorithms is implemented, among other functionalities, to apply suppressive methods to tabular data. The idea is to leverage this wealth of experience and materials to push boundaries further and to reduce the difficulty of managing the protection process considered in this paper.

In Desai et al., 2016, "the best way of managing access to sensitive data" is described by the identification of five safes (projects, individuals, data, access parameters, and outputs). Drawing from our experience, a sixth one could be suggested: the safety provided by mastering the process of producing confidentiality masks. Indeed, improving the quality of the process and making protection easier to manage contribute to overall data safety.

3. The whole process

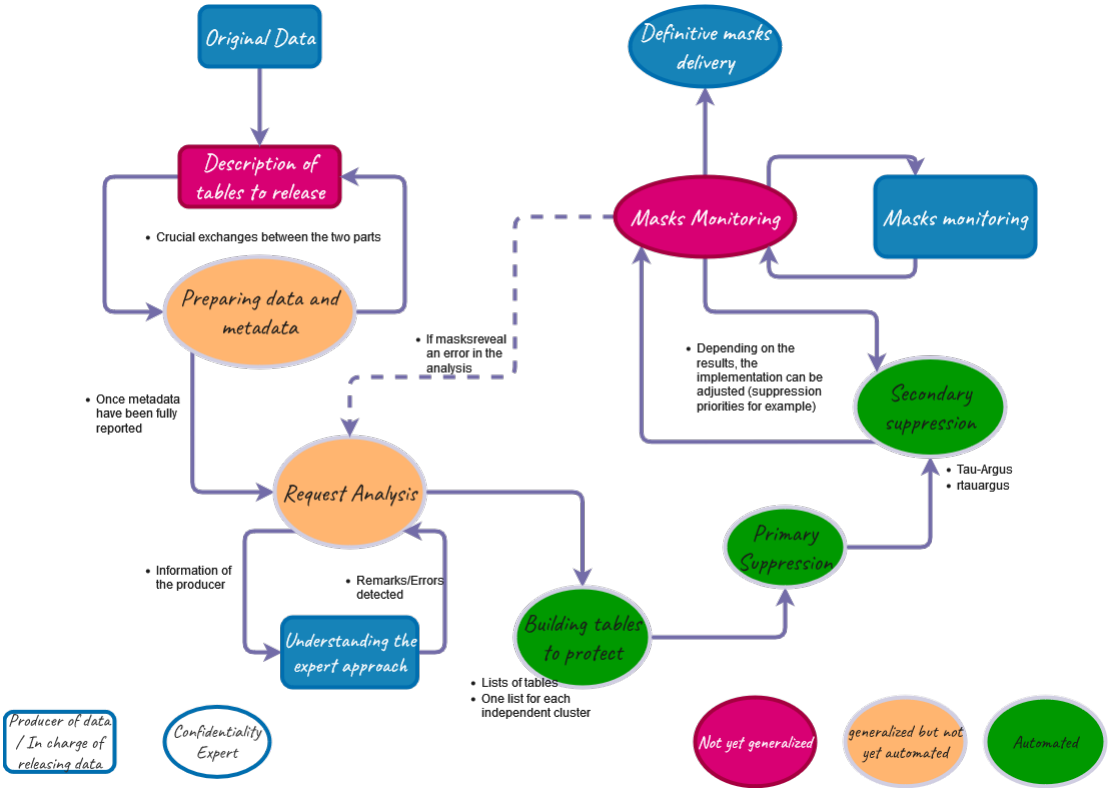
We consider the case of a data producer who fully controls the dissemination of their tabulated data. This is the working framework in which suppressive methods are most relevant. After completing the pure production phase – including collection, cleansing, handling of non-response, calibration, etc. – the producer prepares the dissemination of numerous tables, which must be protected in advance against the risks of disclosure. Given the complexity of the dissemination, they turn to the privacy experts to produce confidentiality masks. In the following, a “request” refers to the release which the producer asks the expert to protect.

Figure 1 presents an ideal schema outlining the key steps of the process involving the interaction between two types of experts:

- "*The producer*": This is the data expert who defines the specifications of his/her request (tables to be disseminated, confidentiality rules, and possibly the choice of dissemination priorities). Steps specifically involving the producer are depicted in rectangular forms in Figure 1.
- "*The expert*": This refers to the confidentiality expert who handles the production of confidentiality masks when the producer's request is complex. Steps that the confidentiality expert must undertake are represented in elliptical forms in Figure 1.

A constructive dialogue between these two actors is crucial for an accurate mask production. However, each party has its own jargon, and the risks of misunderstanding or miscommunication can compromise the quality of the outcome. Therefore, exchanges should not be limited to initial data delivery and final mask delivery stages. Ensuring stages of exchange and interaction between the two stakeholders is essential for a smooth progress of work. Hence, the overall process is not linear: feedback loops are anticipated at various points to ensure a mutual understanding of the problems encountered and the proposed solutions.

Figure 1: Confidentiality masks production process



There are three key moments in the process, which are described in more details below.

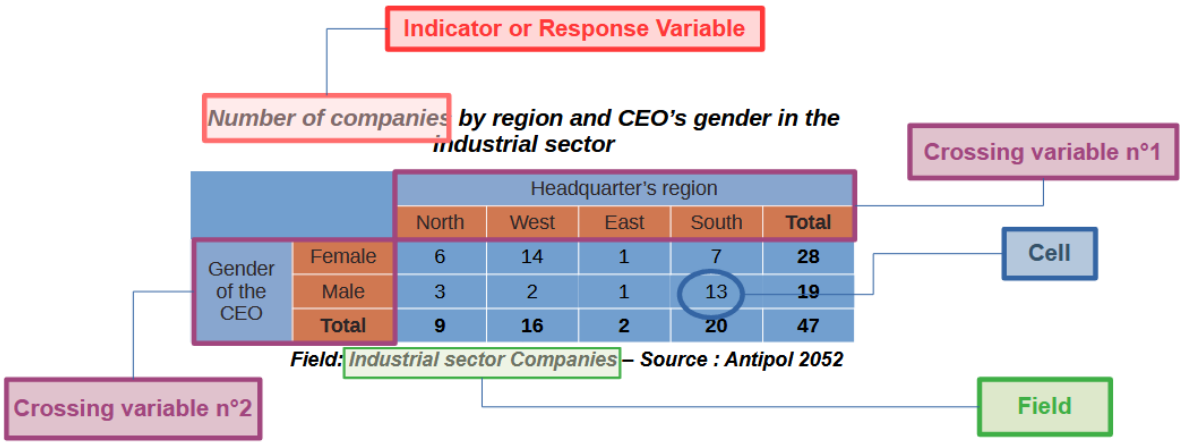
4. Building relevant metadata

The first key step requires the producer to fully and accurately describe their request, and the expert to comprehend it thoroughly.

1.1 Inputs and outputs

The process begins when the producer approaches the expert with their request. Typically, this presentation takes the form of a descriptive list of tables they wish to disseminate. The objective of this step is to produce a comprehensive description of these tables, which must also be relevant in terms of confidentiality considerations. This complete description is like a metadata file for all the remaining process. At least, this file should contain the essential elements describing a table, as presented in Figure 2.

Figure 2: Components to describe a table



1.2 Challenges

The informational needs of the expert may not necessarily align with how a producer naturally describes their request and tables. Many pieces of information that the producer deems useful to mention may have no impact on confidentiality management, while others essential for this purpose may be omitted at first. Moreover, all the pieces of information describing a table (Figure 2) are necessary but not sufficient in general to handle the protection. Indeed, additional information may be required, particularly information that helps

determine the types of links between the tables. In the case where two tables have no relationship, they can be protected independently. The challenge for the experts is to communicate their needs, for example by guiding discussions through a series of questions to obtain this important additional information to establish the existence of different types of links between the various elements of the tables:

- Possible links *between cross-tabulation variables*, allowing for the deduction of (nested or non-nested) hierarchical relationships between these variables.

- Possible links *between response variables*, often taking the form of an equation.

- Possible links *between fields*: these fields may complement each other, generating a relationship between tables to consider during processing.

- Possible links *between certain modalities of the same variable*, allowing for potential (nested or non-nested) hierarchical relationships within the same variable.

1.3 Generalization

Listing all the necessary requirements to describe the request in a relevant way has allowed for the generalization of this key stage's approach. Each producer is thus asked for the same information. This step is about to minimize the heterogeneity of requests upstream of the actual processing. However, it is only by dealing with multiple requests from various sources and formulated by different stakeholders that gradual generalization should be able to achieve its ultimate goal: finding the right metadata format that includes and presents all the necessary and sufficient information in a single document for the application of suppressive methods. In the very simple case of the table shown in the Figure 2, the metadata file describing it can be as minimal as presented in Table 1. Hence, the metadata file takes the form of a spreadsheet, where each row describes a variable in a given table. But the list of potential fields to fill in is greater in most cases : the information about hierarchies could be mentioned, if relevant, for each of the following fields : " Crossing Variables ", " Response Variables " and " Field "[1]. Ideally, for comprehensive information, the

[1]A complete list of fields to fit with most cases would be the following one : Table number, Field, Field Hierarchy, Response Variable (RV), RV hierarchy, Crossing Variable (CV), CV hierarchy, CV Total Code. But an extensive list could be probably necessary in other cases: for example, in case of different data years, it would be relevant to mention this information too.

metadata file should be accompanied by descriptions of the various hierarchies, nested or not.

We think that this file can be fully generalized to fit all protection requests. For the moment, it is filled in by the confidentiality experts team during the first phase of exchanges with the producer. The communication has to be good to collect all the needed information. And a good metadata file is very helpful for the analysis step which follows. In the future, one have to think about if a producer, expert of its data but not of confidentiality treatment, could be required to fill in by him/herself. It would be the final stage before giving back to producers the mastering of the masks production process.

Table 1 Minimal metadata file describing the table in Figure 2

Table	*Field*	*Response Variable (RV)*	*Crossing Variable (CV)*	*CV Total Code*
T1	Companies of Industrial Sector	Frequencies	HQ's Region	Total
T1	Companies of Industrial Sector	Frequencies	CEO's Gender	Total

To facilitate subsequent analysis, each table will be described with the following formalism :

$$\mathrm{RV}^{hrc_{RV}} \underset{Field^{hrc_F}}{\otimes} \left\{ \mathrm{CV1}_{tot_1}^{hrc_1} \times \mathrm{CV2}_{tot_2}^{hrc_2} \right\}$$

Except the field, all the potential information needed about a table is mentioned here :

- At the left side of the $\otimes$ sign, the response variable (the indicator) ;
- At the right side of the $\otimes$ sign and between braces, the crossing variables ;
- Below the $\otimes$ sign, the field ;
- For each variable, the superscript refers to any hierarchy to which the variable belongs ;
- For each crossing variable, the subscript text refers to the name of the modality that is the sum of the others.

Hence, the table described above can be written as follows :

$$\mathrm{Freq} \underset{Ind.Sect}{\otimes} \{\mathrm{Region}_{total} \times \mathrm{Gender}_{total}\}$$

5. Analysis

1.4 Inputs and outputs

This step can be highly complex. A high-quality metadata file is the ideal input to facilitate analysis. The objective of the analysis is to provide an accurate description of the various independent clusters of tables to be protected. This description will enable the construction of the tables and the application of suppressive methods for each of these clusters, independently. This step anticipates the needs of the tool that is used to apply the suppressive methods : Tau-Argus, here.

1.5 Generalization

The formation of these sets can be generalized through an approach that involves detecting the existing links between tables: whenever a table or a set of tables has no connection with another, it can be processed separately. The various types of links listed above can be identified based on the information gathered in the metadata. Each link helps describe, step by step, the tables that will need to be protected.

As all steps of the analysis cannot be detailed here, a small example, using the formalization presented in the previous section, is displayed. Let's imagine that a producer wants to release 6 tables sharing the same field : the companies of a given area that cook and sell pizzas. Here are the tables described in the simplest way[2] :

$$\begin{cases} \text{to\_margarita} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_margarita} \otimes \{\text{NUTS3} \times \text{SIZE}\} \\ \text{to\_calzone} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_calzone} \otimes \{\text{NUTS3} \times \text{SIZE}\} \\ \text{to\_pizzas} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_pizzas} \otimes \{\text{NUTS3} \times \text{SIZE}\} \end{cases}$$

" to " denotes the turnover in making pizzas. " NUTS2 " and " NUTS3 " are nested european geographical areas and " SIZE " is a categorical variable of the size of the companies. The first step is to add the additional information for each variable (hierarchies and total codes).

[2]The field of each table is not mentioned here since it is the same for all these tables.

$$\begin{cases} \text{to\_margarita}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_margarita}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_calzone}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_calzone}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_pizzas}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_pizzas}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \end{cases}$$

Here " pizzas " is the name of the relation between " pizzas " as a total of " margarita " and " calzone " ; " nuts " is the name of the nested hierarchy including " NUTS2 " and " NUTS3 " areas. As " NUTS2 " and " NUTS3 " are included in the same hierarchy, the tables can be merged as followed :

$$\begin{cases} \text{to\_margarita}^{pizzas} \otimes \left\{ \text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_calzone}^{pizzas} \otimes \left\{ \text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\ \text{to\_pizzas}^{pizzas} \otimes \left\{ \text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \end{cases}$$

where the " NUTS " variable includes all the information of both " NUTS2 " and " NUTS3 " previous variables.

Finally, as the response variables are also linked together (pizzas = margarita + calzone), the six original tables described by the producer can be summed up in only one table :

$$\text{to} \otimes \left\{ \text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all} \times \text{PIZZAS}_{pizzas}^{(h)} \right\}$$

The third crossing variable is a categorical one taking three different values (“pizzas”, " margarita” and “calzone”). The " (h) " exponent denotes a holding variable. The table with three crossing variables is the only table that actually needs to be protected in order to protect all six tables initially presented in the request.

1.6 Automation

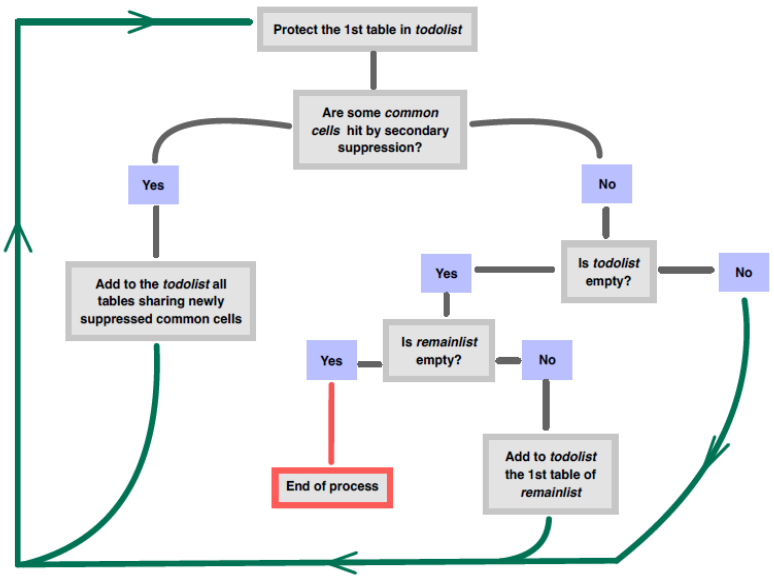
Automating this stage is the challenge Insee’s methodologists are currently working on. The automation is a great promise in terms of efficiency and productivity gains. It could lead to producers becoming autonomous in managing the confidentiality of their tables in a closer future. The automation of cases such as those presented above is already implemented and our tool, which is still in the testing phase, can also handle non-nested hierarchies.

6. Secondary suppression

The secondary suppression is the last very technical step of the process, but the first which the team focused on, thus the most advanced one on the road to automating protection management. As mentioned above, Tau-Argus is an efficient and reknown tool to handle this stage (de Wolf et al., 2023). But, it does not manage as many linked tables (sometimes hundreds of them) as encountered in some requests. For this reason, a small algorithm was developed in the *rtauargus* R package (Berrard et al., 2024) to automate the management of an indefinite number of linked tables. All the information needed to handle this stage is actually provided by the previous steps (list of tables, hierarchies and total codes).

The automation of this step is the main gain in productivity for the team because the suppression step of any request is now implemented with the same code as any other one and with the same standards. No need to write a particular and complex algorithm each time. The algorithm presented in the Figure 3 sends each table one by one to Tau-Argus to apply secondary suppression. It handles the relevance of the suppression pattern between linked tables by efficiently taking into account the suppression occurred on common cells. This technique necessarily leads to some over-suppression compared to an ideal situation where all tables could be taken into account during the optimization program implemented in Tau-Argus. But, this ideal does not exist for the cases we are talking about.

Figure 3: Algorithm implemented in rtauargus to handle protection of linked tables in coordination with Tau-Argus



7. Conclusion

The paper presents a comprehensive approach to managing the protection of tabular data during dissemination, addressing the challenges faced by Insee, and maybe other statistical institutes. The long term goal of the team is to empower data producers to take control of the confidentiality masking process, but the short term one is to reduce the burden on the expert team. The paper proposes a step-by-step methodology to address all the challenges encountered, including describing the process, generalizing problem-solving approaches, automating tasks, and disseminating best practices. The approach involves collaboration between data producers and confidentiality experts, with an emphasis on communication and mutual understanding. The efforts to automate the process, particularly in analyzing the request and managing secondary suppression, led to improve efficiency and reproducibility of the tasks, and productivity of the team. Even if producers have not yet taken back control of data protection, it is now conceivable that this will happen in the near future.

Acknowledgment

Clément Guillo (Insee) and Nathanaël Rastout (Insee) for all the work done altogether.

References

Desai, T., Ritchie, F., & Welpton, R. (2016). Five Safes : Designing data access for research. *Economics Working Paper Series*, *1601*, 1-27.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G., & De Wolf, P.-P. (2010). *Handbook on Statistical Disclosure Control* (Version 1.2). ESSNet SDC.

Jamme J., & Rastout, N. (2023). About protecting multiple linked tables with a suppressive approach : An illustration with the ICT survey. Presentation to NTTS conference, Brussels.

Berrard P-Y., Jamme J., Rastout N., Beroud F., Pointet J., Pomel W. & Socard A-R. (2024), *rtauargus, run Tau-Argus from R*, Paris, Insee, 2024, R package version 1.2.0, https://github.com/inseefrlab/rtauargus

De Wolf P-P., Hundepool A., Gießing S., Salazar J-J. & Castro J., Tau-Argus, Statistics Netherland, version 4.2.4, 2023, https://github.com/sdcTools/tauargus/releases/tag/v4.2.4.2