Minutes of the SDC-tools Users' Group meeting 2024

Written by: Violeta Calian

Time: 10:00-12:15 (C.E.T.)
Location: online (link sent to registered participants)

Agenda
10:00 Introduction

10:05 News about tau and mu – argus, *Peter-Paul de Wolf*

10:25 From tables to be published to rtauargus inputs: an automated approach, *Julien Jamme*

10:45 Break

11:00 The SDC-tools R-packages, *Johannes Gussenbauer*

11:20 Invited talk: The risk of identity disclosure through network structure, *M.M. de Vries*

11:40-12:15 Discussion and Poll about future meetings of the User Group

Presence:
Out of the 41 registered people, 34 attended the online meeting.

Web-page, where the agenda and slides for this meeting (and previous ones) can be found: https://sdctools.github.io/UserSupport/

Introduction
The meeting started with a short welcome from the organiser and the agenda was introduced by the workpackage leader of the CoE on SDC within the STACE project.

News about tau and mu – argus, *Peter-Paul de Wolf (P.P.W.)*
The first presentation included an overview of the SDC-software and European projects during 1996-2024 as well as of the SDC-tools ensemble. New features and improvements of mu- and tau-Argus were described, with a focus on the most recent releases. The help of the members of the Tester-team was acknowledged. P.P.W. ended by enumerating the future software developments and list of priorities as well as difficulties which might occur, including lower speed in updates if no new funding is secured after 2024 when the STACE project ends.

Questions and answers:

Julien Jamme (J.J.) asked about the speed improvements and P.P.W. illustrated the order of magnitude of these changes with an example which decreased its running time from minutes to only seconds.

Michel Reiffert (M.R.) asked about the error message improvements, in relation to the fact that different types of users have different expectations about these messages. P.P.W. gave more details, concluded that the error messages should in general give more information and both M.R. and P.P.W. agreed that a good suggestion for the users should be that they raise a github-issue each time they consider an error message as incomplete.

Karina Dineen (K.D.) asked about the updating of manuals. P.P.W. explained that this is a continuous process and the most recent reformatting (quarto-book and/or compatible) ensure more flexibility about this process.

From tables to be published to rtauargus inputs: an automated approach, *Julien Jamme (J.J.)*

The new algorithm described by J.J. offers a systematic and automated method for protecting multiple linked tables by using tau-argus and a demand analysis. It is based on splitting a large set of linked tables into independent clusters of smaller ones. This amounts to solving a set of smaller problems and non-nested hierarchies according to standard methods. J.J. illustrated the algorithm with the case of 26 linked tables, showing the selected tables (smaller number than this total) which need to be protected and how the problem can be transformed into a manageable one.

Questions and answers:

P.P.W. asked about the type of table linking, whether regarding linked response variables or „overlapping“ dimension-variables. J.J. confirmed the linking types and explained how categorical, additional variables are created in order to estabilsh the „links“.

P.P.W. reminded us that tau-Argus protects linked tables and asked J.J. whether his proposal exploits this feature or solves the protection step independntly. J.J. explained that tau-Argus is called for protecting the independent components once these are built.

K.D. asked about the over-supression issues and J.J. confirmed that these are not different from the case when tau-Argus is employed.

K.D. asked about how is the problem of multiple spanning variables handled, in hierarchical context. J.J. explained that the rtauargus-package can solve very complex problems but the main condition is that the input table structure should be carefully and correctly built.

Sarah Giessing (S.G.) asked about the use of holding indicators especially when dealing with aggregated-structures. J.J. said that although this is taken into account when building the input tables, it is not yet a part of the process.

The SDC-tools R-packages, *Johannes Gussenbauer (J.G.)*

The R-packages included in the SDC-tools were presented by J.G. The progress made and details about their status on CRAN and/or github were carefully described for the main R-packages, i.e. sdcTable, sdcMicro, accompanying R-packages, i.e. cellKey, ptable, sdcHierarchies as well as sdcSpatial.

Questions and answers:

Manca Golmajer (M.G.) asked about the sdcMicro 5.7.7 - version and J.G. confirmed that this is indeed the current working version. M.G. followed with a comment about the occurence of changes in names of functions and an example from ptable-package. J.G. explained that the development github-versions might have unusual restructuring notations but that vignettes should be and are up-to-date and the CRAN versions should make clear when changes occur.

Invited talk:

The risk of identity disclosure through network structure, *M.M. de Vries (M.M.V.)*

This invited talk is based on the related paper https://unece.org/sites/default/files/2023-08/SDC2023_S6_4_Netherlands_deVries_D.pdf by M.M. de Vries , R.G. de Jong, M.P.J. van der Loo, P.-P. de Wolf , F.W. Takes. The presentation and the paper address an important and less studied question, namely if, when and how it is possible to estimate the probability of a database attack-scenario. This is in contrast to the most frequently studied probability of disclosure, conditional on a certain attack. The problem is made better-defined by describing the most likely condition with direct impact on the probability of creating an attack, i.e. acquiring network data type of knowledge.

Statistics Netherlands has recently developed population-scale network data where nodes are persons and links represent various real-world connections (family, household, work, school, and geographical connections). In addition, they have developed an anonymity measure where it is assumed that an attacker has certain prior knowledge about the network structure surrounding a node. The presentation showed experimental results from a hackathon and discussed several related SDC-open questions.

Questions and answers:

An interesting discussion with some of the participants (Violeta Calian, Arndis Vilhjálmsdóttir, M.M.V) followed, regarding the rather limited literature concerning the calculation/estimation of likelihood of attack-scenarios and even ideas about trying to propose unusual methods to achieve it, such as crowd sourcing/gamifying the goal.

Discussion and Poll about future meetings of the User Group

Questions from the Users' Group members

Two questions were proposed by Martha Düker (M.D.):

- ***1. Methods to reduce discloser costs***

"We are aware that different secondary and primary disclosure algorithms lead to higher or lower discloser costs. We use the 85-dominance-rule for primary suppression and Modular in Tau-Argus for secondary suppression. We are also aware about the case that how the table is structured lead to different discloser costs. At the moment we mostly try to use the deepest possible hierarchy of a data room/table available and know that from the primary suppression resulting from lower areas of the hierarchy are leading to a high amount of secondary suppressed cells in higher hierarchical levels.

So, one of our plans is to make the hierarchy a little bit less deep to reduce the suppression costs, since it is not in our hands to change the primary suppression rule.

> *Besides those 3 methods to reduce costs (choice of primary suppression algorithm, choice of secondary suppression algorithm and structure of the table/ data room) do you have any other methods or tips to reduce discloser costs?"*

- ***2. Methods to deal with new suppression patterns resulting from data revisions***

"We are aware that we can use the history files to „save" the pattern of the suppressed cells before the data revisions to take care that those cells are also suppressed in the revised table. And only add the new primary and secondary suppressions resulting from the revised data.

> *Do you have any other ideas/ methods to deal with this issue?"*

Answers

Regarding question 1, both S.G. and P.P.W. agreed that restructuring the tables and taking into account the structure of the table to be published is a correct approach. P.P.W. discussed further the issue of defining the „disclosure costs" in terms of cell values and number of cells to be supressed, and pointed out the trade-off between user and sdc-expert needs/goals. He also mentioned the linked table cases, which can be handled by tau-Argus.

Regarding question 2, M.G. shared her experience about revisions and suppression patterns, namely that they check whether new suppression is needed by using their Visual Basic Macros, since using tau-Argus for this would be too much.

M.R. explained his views about this question, i.e. that: one identifies new primary suppressions, use tau-Argus, compare with the previous case (usually differencies are not too big). Tau-Argus should be able to run. A possible problem could be the frozen cell but this is much better handled by the most recent tau-Argus without issues. One should at least try and re-run. P.P.W. pointed out that one needs to make sure that the costs for the previous suppression steps were low.

Additional questions/discussions

Organisers asked the participants about the benefit of the discussions within the Users' Group and acknowledged that the meetings in person have stimulated more active exchange of opinions and information than the online. However, the online meetings are still useful. They could take place and continue in the future, even without further financing if some of us offer to take turns in organising them.

K.D. mentioned that indeed, our discussions have been useful. She commented on the issue of not always having sample-data available in order to illustrate and/or test SDC methods and software.

P.P.W. exemplified with tau-Argus which does have example-data although not linked tables type. He also reminded us that for the Business Statistics course, there has been an example-data with linked tables and attached exercises, with instructions. The repositories/manuals usually contain some examples as well.

J.J. suggested that the new tau-Argus manuals should include more example-data and how to deal with that. He pointed out the rtauargus package has some example-data.

M.G. asked whether the tau-argus-modular experienced any problems in applications. P.P.W. commented on the way to correctly deal with bogus levels and modular splits, i.e. that one needs to have all the bogus levels at same safety level.

Poll regarding the future of the Users' Group

The following questions were part of a slido-poll:

- Should we continue the SDC-UG meetings, once a year, after the end of the STACE project (August 2024)?

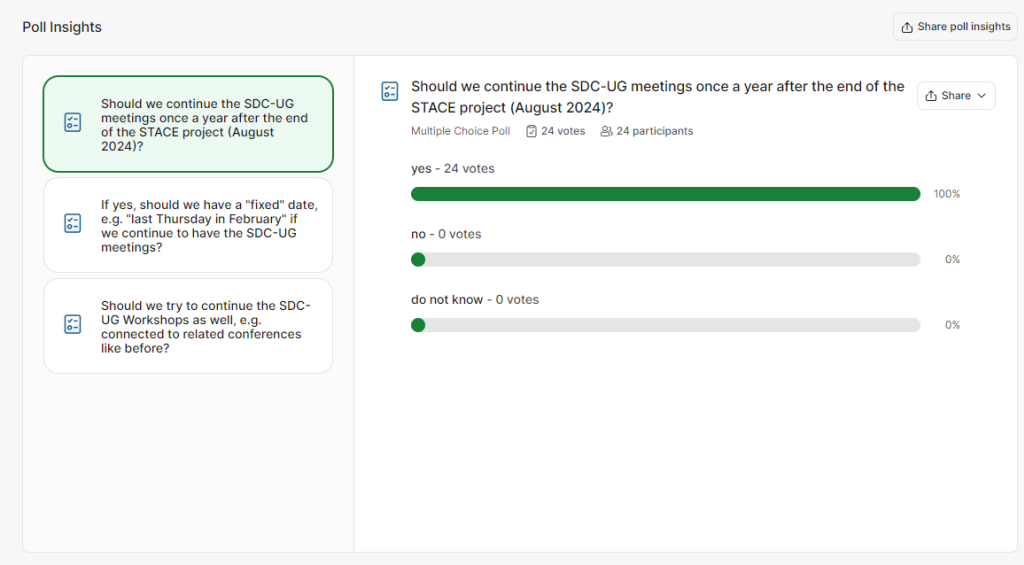
- If yes, should we have a fixed date, e.g. "last Thursday in February" for the SDC-UG meetings?

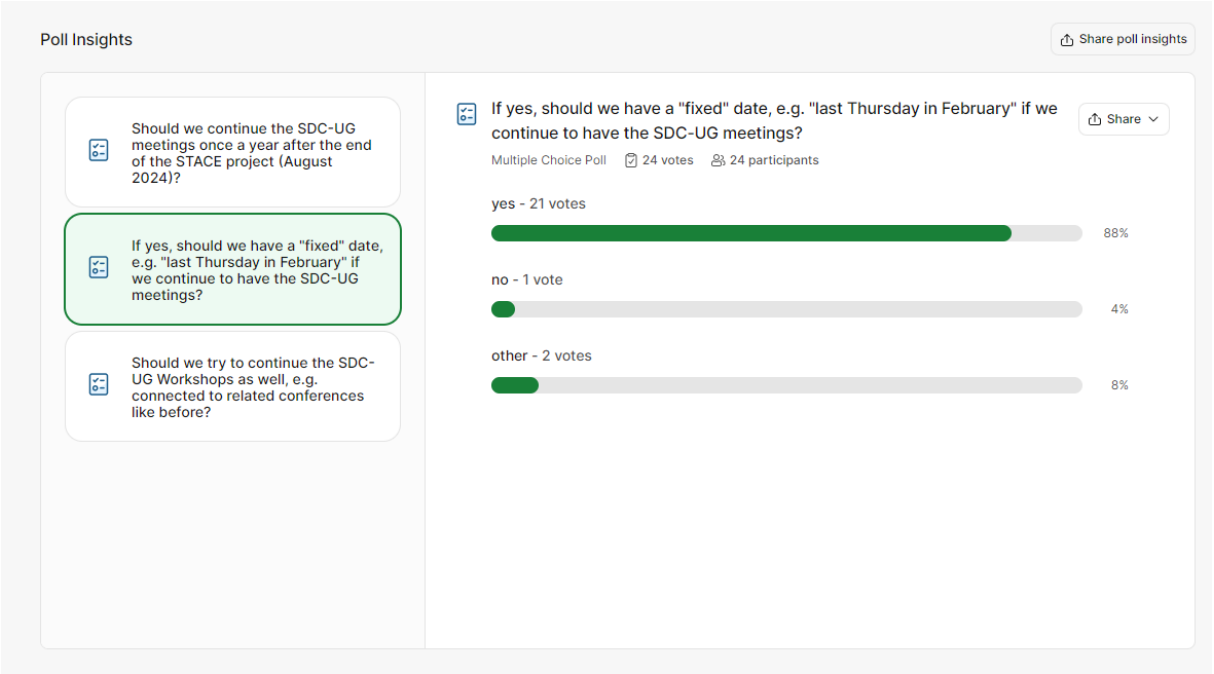
- Should we try to continue the SDC-UG Workshops as well, e.g. connected to related conferences, as before?

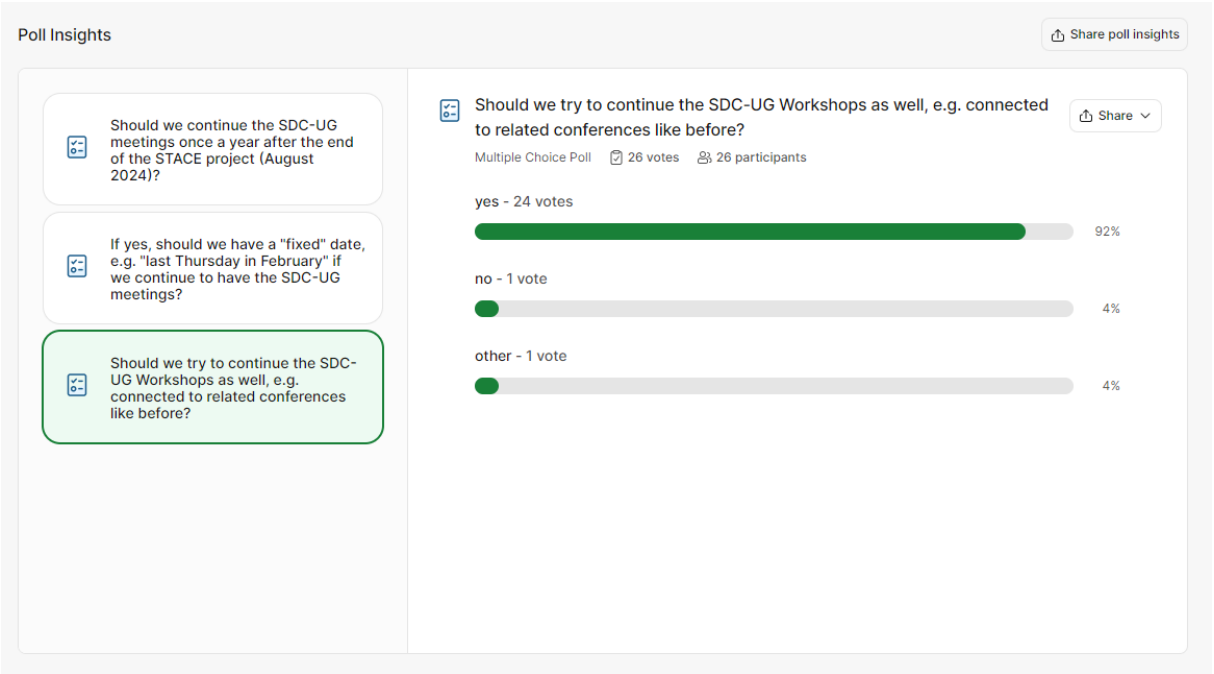
* *Do you have suggestions about how to organize these, if there is no financing project? (e.g. taking turns, on voluntary bases)*

Answers:

All the questions were answered in the affirmative, as follows:







We concluded with the optimistic promise of future interactions via the SDC-tools Users‘ Group.