Rules and instructions of Research Services

Contents

Contact information

Research Services:
Employee on duty — tel. +358 29 551 2758
Valtteri Valkonen — tel. +358 29 551 3431
Satu Nurmi — tel. +358 29 551 2926
Marianne Johnson — tel. +358 29 551 3777

tutkijapalvelut@stat.fi

Microsimulation
Sanni-Sandra Hellman — tel. 029 551 3210
Sampo Lappo — tel. 029 551 3922

mikrosimulointi@stat.fi

1. Introduction

Statistics Finland's Research Services offer unit-level data for scientific studies and statistical surveys. The data can be used through Statistics Finland's FIONA remote access service and in the Research Laboratory or outside Statistics Finland as sample data. By signing a pledge of secrecy the researcher commits to follow the rules and instructions of the Research Services. The rules and instructions that relate to

confidentiality, data protection, publishing of results and ending of the project apply to the research use of all unit-level data. The screening process of research results is applied both to remote access use and use in the Research Laboratory, see Section 5. The practices for remote access use of the Research Services also apply to remote use of the SISU microsimulation model. As an exception, different practices are applied to the screening of research results in microsimulation.

2. Confidentiality

The use of unit-level data is subject to a user licence. Only the person who has been granted a user licence is permitted to use the data and the data can only be used for the purpose accepted in the decision. No attempt must be made to identify the data subjects from the material. By signing the agreement applying to the use of the SISU microsimulation model or the pledge of secrecy, the researcher pledges not to disclose to anyone or use for their own or private benefit the data they have been authorised to use and are prescribed as confidential by law (unit-level personal and business data included in the research data). The obligation to maintain secrecy also concerns computer software used in statistics production and information concerning such software, the disclosure of which compromises statistical confidentiality (Statistics Act 280/2004: Section 24). The obligation to secrecy shall remain in force even after the permit for the project expires.

3. Rules of the remote access system

3.1 Logging in to the system

Remote desktop use from the researcher's workstation to the FIONA server is opened through an online service (Appendix B). The user logs in to the online service with a personal user ID and password that are to be stored carefully. In particular, the user IDs and passwords should be stored separately.

In addition, the user is identified by a text message code. The system sends a so-called flash message to the user's mobile phone number. The phones and SIM cards related to this number should also be stored carefully.

Remote access use must take place from the premises of the customer that has signed the user licence. If the user wishes to conduct work from another location this must separately be approved by the Research Services. Remote access can only be made from the IP address specified by the customer. Statistics Finland checks that they are in compliance with the terms of use.

The remote use workstation's data security updates, as well as the virus and firewall protection, must be up-to-date.

3.2 Using the system

A work folder is reserved for research projects where the project material is stored. Reading rights to the data folders are granted in accordance with the project-specific or SISU model user licence. In the SISU model, each user is reserved their own work folder to which the user can grant rights to other users of the model. The users of the microsimulation model can also share files in the remote access use environment with other users through a joint Forum folder.

The system shall only be used for the purpose mentioned in the user licence and its functionality shall not be hindered in any way. Log data are collected on the use of the system for billing, and for the maintenance and control of the system[1]. The use of the system is billed in accordance with the price list of Research Services.

The remote access is only intended to be used by the researcher who has a license for the dataset and for whom remote access has been opened. Therefore, the remote workstation must be locked when it is left unattended. The user must log out from the system when he/she no longer works with the remote access.

The data must be handled so that confidential data do not fall into the hands of outsiders. Protection can be improved, for example, by adapting work premises and protection of the screen. Data subject to user licence shall not in any way be revealed to a person that does not have a user licence to the data.

Software files and files containing results can only be sent by email via Statistics Finland's personnel from tutkijapalvelut@stat.fi. Research results, with the exception of research results from microsimulation, are screened to ensure data protection (see Section 5).

3.3 Resources

Each research project is reserved 300 GB in disk space. SISU microsimulation model users are reserved 40 GB disk space per user. If the project or the user of the SISU model requires more disk space it is charged separately. The calculation and software resources available in the system are limited and their availability depends on the number of logged in users. Saving of overlapping datasets or intermediate files is not recommended. Simultaneous heavy runs should be avoided. Individual calculation and software resources can be offered to an individual customer or project in accordance with a separate agreement.

3.4 Maintenance

Statistics Finland bears the responsibility for the maintenance of the remote access system during office hours. Problems can be reported by email to tutkijapalvelut@stat.fi or by calling Research Services +358 29 551 2758. In problem situations the first step for the researcher is to contact the contact person for remote access in his/her own institution. Statistics Finland is not responsible for the user support of software.

Statistics Finland has the right to shut down the system for maintenance reasons. Interruptions caused by maintenance are communicated in the information notifications on the start page of the remote use service and by sending an email to the contact persons of the organisations.

4. Rules of the Research Laboratory

4.1 Use of the identity card and electronic access key

Researchers that work in the Research Laboratory are issued an identity card with a photograph which must always be worn when on the premises of Statistics Finland.

[1] Register description TK-00-473-08

Researchers must contact a member of the personnel of Statistics Finland if the card has been mislaid or forgotten.

Unless otherwise agreed, researchers may only work in the Research Laboratory during normal office hours between 8 am and 4.15 pm.

Researchers are issued an electronic access key for moving about the premises of Statistics Finland. On arrival at Statistics Finland, researchers must sign in by pressing the “sisään” (in) button and on leaving the premises sign out by pressing the “ulos“ (out) button on the time clock. Respectively, when leaving for lunch, researchers should press the “ulos” (out) button and on returning from lunch the “sisään” (in) button. Forgotten sign ins and outs are reported by email to tutkijapalvelut@stat.fi.

4.2 Working in the Research Laboratory

The researcher logs in to the FIONA remote access system from the workstation in the laboratory. The researcher logs in with a personal user id and password. These should be stored securely. The same rules and instructions apply for researchers using the FIONA-system from the Research Laboratory as for those logging in from their own workstations (see chapter 3)

The workstation must be locked whenever it is left temporarily, and switched off at the end of the day. In exceptional cases and based on separate agreement, runs can be left open on the workstation. Reservations for a researcher space can be made by telephone (+358 29 551 3493 and 2926) or email (tutkijapalvelut@stat.fi). The researcher space is invoiced in accordance with the price list of Research Services.

Software files and files containing results can only be sent by email via the Research Services' personnel from tutkijapalvelut@stat.fi. Research results are screened to ensure data protection (see Section 5).

5. Data protection and output checking process

According to the obligation to maintain secrecy, the researcher must ensure that the research results contain no unit-level data or possibility of their disclosure. Research Services applies an output checking process of output produced by the researcher from the data to ensure the implementation of data protection in the research results. The researcher shall ensure that the output sent for checking meets the data protection rules listed in Section 5.1. The output must be clearly interpretable. In tabular output, the number of observations in each cells must be visible, as well as the number of observations used in calculating estimates and model parameters. If necessary, Statistics Finland's personnel can provide additional information concerning data protection.

5.1. Data protection rules for statistical output

The following is a more detailed description of the data protection rules for different types of output.

Frequency and magnitude tables

Output published by the researcher must follow Statistics Finland's policies for protection of tabulated data (Appendix A).

The principle rule in the protection of business data is that each cell or group must contain at least three (unweighted) observations. As an exception, the threshold for GVC/international sourcing questionnaire data is five observations. In addition, the dominance rule[2] (1,75) is to be applied when the business data is less than 15 month from the reference date. The dominance rule is always to be applied to data on international trade of services. When protecting establishment-specific data, enterprise level protection must also be ensured, so each cell must contain establishments of at least three different enterprises. Likewise, group level protection must be considered in business data that contain data on group relationships. When it comes to commodity data (products, raw material and supplies from the statistics on industrial production) the number of enterprises is confidential for all production titles.

In the protection of personal data, a cell-specific threshold rule of three observations is applied and special attention is paid to the sensitivity of variables to be tabulated. In combined employer-employee data, both personal and business data must be protected, so each cell must contain employees from at least three different enterprises. The own-account workers that appear in business data are subject to same protection rules that are applied to other business data.

For some data, especially data provided by other governmental agencies, there might be other rules for data protection. These differing protection measures are stated in the licence to use the data.

Descriptive statistics

Maxima and minima usually refer to only one observation. If this observation can be identified, the maximum or minimum cannot be published.

Percentiles or centiles (excl. minimum and maximum) form a special case of tables where the number of observations in each percentile interval correspond with the cell frequencies. If these numbers exceed the threshold value three, percentiles can be published.

Modes can be published if (nearly) all observations in the data do not get the same value.

Means, ratios and higher moments of distribution (e.g. variance) can be published if at least three observations have been used in their calculation.

Other numerical output types

Indicies, correlation coefficients and test statistics (t, F, X^2, etc.) can usually be published if enough observations (at least ten) have been used in the calculations.

Complete regression model can be published if the model is based on enough observations and the model in not on one unit (e.g. time series on one company). Individual coefficients of the model can usually always be published.

---

[2] Information about the dominance rule can be found in Section *3 Concepts and methods connected to the statistical disclosure control of tabulated data* of Appendix A.

Images

Like numerical output, graphs shall not disclose data of a single observation. Graphs based on the data are permitted if a single image point does not disclose the underlying individual observation.

Graphs are sent for checking like tables, documented clearly and precisely. Graph formats that are suitable for checking include:

Bitmap formats
- PNG (Portable Networks Graphics)
- BMP (Bitmap)
- JPEG (Joint Photographic Experts Group)
- TIFF (Tagged Image File Format)

Vector formats
- EPS (Encapsulated PostScript)
- PS (PostScript)
- PDF (Portable Document Format)
- SVG (Scalable Vector Graphics)
- WMF/EMF (Windows Metafile)

In Stata software, above-described graph formats can be created with the graph export command. In SPSS software, the graph format can be selected in the Export output function. In R software, information about the drawing function is available with the help command (grDevices). Certain graph types, like Stata's gph files, include by default in the graph file the data used for making the graph, which means that they are not necessarily suitable to be transferred out from the remote access system.

It is more difficult to identify an individual observation from a graph if the graph is based on sample data. Many business data are, however, total population data, so the data of an outlier enterprise, for example, can be more easily identified. Even if the data are not total data of all enterprises in Finland they may be total data in terms of a less exhaustive sub-population, like a particular field of industry.

Bar charts and other graphs used to present classified data are typically accepted for publication as long as each category has enough observations. These types of data can usually also be presented in table format and Statistics Finland's rules for protection of tabulated data (Appendix A) can be directly applied to these.

Sometimes distribution charts contain deviating observations or outliers that could disclose the unit-level data. Distributions, histograms or cumulative distribution functions that have been adjusted or are presented at a sufficiently crude scale are allowed. Statistical software's drawing functions often automatically mark outliers in, for example, box plots, and these should usually be removed from published graphs.

Scatter plots are typically used to show the values of two continuous variables, which means that they are trickier in terms of data protection than the graphs described above. For scatter plots, special attention should be paid to the nature of the data, for example, the size of the sample in relation to the sensitivity of the data and occurrence of outliers.

Clearly forbidden graphs include graphs that present the values of outliers or scatter plots from which one can deduce, for example, the data of the largest enterprise in the sector.

5.2 Output checking process in practice

The output checking processes applied to output from research data sets and to output from the use of the microsimulation model differ. Research results produced in remote access and Research Laboratory are checked before the output are released, and you cannot transfer files to your own workstation from the remote access environment. The transfer of the output can only be made upon separate request by email. In the SISU microsimulation, the user transfers output files directly to his/her own workstation without any pre-checking.

Output checking process in research projects

Research results and output produced in remote access and Research Laboratory are checked to ensure data protection. All output files sent to checking must meet the same criteria as tables and graphs intended for publication. For example from log files, only necessary sections and sections intended for publication should be sent to checking. The files are transferred to output checking by copying them to the output folder (...\out) and sending a checking request to the email address tutkijapalvelut@stat.fi. The number and size of output files must be kept reasonable (max. size 2Mb).

After the output checking, the results are sent to the researcher's email address. One to two working days should be reserved for the output checking.

Output checking process in microsimulation

In the microsimulation remote access environment, research results can be published without pre-checking. Thus the researcher can transfer output files containing research results from the remote access environment to his/her own workstation.

Every user has his/her own personal email folder (Mail) in the remote access environment through which files can be transferred to the user's own workstation. The transfers takes place by copying the desired files (from the User, Forum or Admin folders) to the user's personal Mail folder. After about two minutes of the copying the file is automatically transferred both to the user's personal and Statistics Finland's microsimulation (mikrosimulointi@stat.fi) email. A separate email message is sent on every file copied to the Mail folder with the file copied to the folder attached to the message. The email shows the name of the transferred file and the sending date. The size of the attached file can at most be 1 megabyte (Mb).

Statistics Finland checks the transferred files from the microsimulation email afterwards. The researcher is obliged to follow the instructions and rules applying to the remote access system of Research Services when it comes to transferable data. These include protection of research results, limitations related to file sizes, etc. interpretability of data and publication of results (Section 6). *In particular, special attention should be paid to the data transferred in the microsimulation environment not, even by mistake, containing unit-level data or any possibility for such data to be disclosed.*

6. Publication of results

Researchers pledge to publish their research results only in a form in which no individual enterprise's or person's data can be identified. In order to ensure this, the research reports and publications can be demanded for screening before the results are published. Researchers should note that adequate time (one to two weeks) must be allowed for the screening of data protection. You should agree on screening of research results in advance with the personnel of the Research Services.

When the results are published, Statistics Finland must be quoted as the source.

7. When the project ends

The material generated in remote access use and Research Laboratory projects are removed when the user licence of the project or the SISU model expires unless a separate agreement has been made on storing the material. The data released outside Statistics Finland must be returned/destroyed once the permission on their use has expired. All copies taken and intermediate files formed of the data must also be destroyed. Statistics Finland must be notified about their destruction.

8. Sanctions

If the researcher or customer breaches the agreements or instructions on remote access use, the remote access connection of the researcher or of the whole organisation is shut down. The connection is reopened if the customer presents an acceptable written justification for the reason of the breach and the actions taken to prevent any future breach.

The sanctions for breaching the obligation of secrecy are set out in Chapter 38, Section 1 and 2 of the Penal Code. The obligation of secrecy also applies to researchers. The punishment is a fine or imprisonment for at most one year.

9. Responsibilities

All institutions, that sign an agreement on remote access to research data, are to name contact persons for administrative and technical questions. Their task is to instruct new remote access users, follow information that Statistics Finland sends out regarding the FIONA-system and take part in information and training sessions concerning the use of the system as well as together with Statistics Finland try to find solutions to possible problems the users are encountering. The contact persons are to familiarize the researchers dealing with the data with the remote access environment. This should be done in order to ensure that the researchers have sufficient technical capacity to deal with the data material and that they do not inadvertently compromise the functioning of the system or the security of the data. The contact persons must inform Statistics Finland of any changes in data protection or data security within the organization which are of significance to the remote access use.

The person who signs the agreement on remote access use of research data on behalf of the institution/organization is responsible of the researchers working on the

FIONA-system in the institution. He/she is also responsible for that the working spaces provided for the researchers are suitable for research and for remote access use so that the privacy of the data is not compromised. The person approves the researchers that wish to use the remote connection from the institution by co-signing their remote access commitment. He/she informs Research Services if there is a change in the contact persons of the institution.

The researchers are responsible for observing their pledge of secrecy as well as the rules stated in this document. The researchers are to inform Research Services about the termination of their project as well as about changes in their work environment. When changing jobs the researcher should send in a new remote access commitment if he/she wishes to continue using remote access.

Appendices

APPENDIX A: Guidelines on the protection of tabulated data formed from research data

APPENDIX B: Instructions for logging in and using the remote access system

Appendix A: Guidelines on the protection of tabulated data formed from research data

1. Purpose of the guidelines

These guidelines were compiled based on Statistics Finland's own guidelines on the protection of tabulated business and personal data (TK-00-270-13 and TK-00-271-13). By means of these guidelines, Statistics Finland seeks to promote responsible ways to operate in statistical disclosure control (SDC) issues and to ease the application of the acts and principles in the publication of statistical tables produced from research data.

All researchers using Statistics Finland's research data and publishing tabulated business or personal data should familiarise themselves with these guidelines and the underlying acts and principles of statistical ethics.

2. Application of the rules

In the Statistics Act (280/2004), provisions are given in Section 13 concerning the release of data collected by Statistics Finland for research purposes. The following is stated in the preamble of that Section[3]:

"When releasing data, the protection of data concerning personal data and business or professional secrets must be ensured case-specifically by practical measures, such as by requiring sufficient data protection measures and by attending to the provision of required data supervision and monitoring concerning the use of data. – – Because the end results of scientific research are usually public, it should always be separately made sure in connection with their publication that it would not be possible to identify the individual statistical units on which the research is based from the public end result of the research."

Based on the above, it is necessary to attend to the protection of data suppliers' privacy and business and professional secrecy in the end results of those scientific surveys where data released by Statistics Finland have been used. Researchers must take into consideration these SDC guidelines when planning and compiling tabulated publications on their research results. In its part, Statistics Finland sees to the implementation of SDC in them through its output checking process.

3. Concepts and methods connected to the statistical disclosure control of tabulated data

Tabulated data refer to statistics where unit-level data have been aggregated and arranged in table format. The statistical unit of tabulated personal statistics is such as a private individual, family, household or household-dwelling unit. The statistical unit of tabulated business statistics is such as an enterprise, establishment or group. The same SDC practices are applied to own-account worker data in tabulated business statistics as to other business data.

Tabulated data can be either

[3] Government proposal to Parliament for the acts amending the Statistics Act and Sections 2 and 3 of the Act on rural industry statistics (HE 154/2012).

- A ***frequency table*** where the cell values are the numbers of statistical units belonging to the cell, or
- A ***magnitude table,*** where the cell values are sums, means or other corresponding statistical figures of some variable to be tabulated (e.g. turnover), or
- Combinations of the above where both cell frequencies and magnitude data are visible.

Magnitude tables are clearly more common in presenting business data, while personal data are more often given as frequency tables.

Tabulated data are subject to a ***disclosure risk*** if there is a risk of disclosure for some statistical unit in the table. ***Sensitivity rules*** are used in defining cell-specific disclosure risk. The most common sensitivity rules are:

- The ***threshold rule***, by which a cell is sensitive if it contains fewer statistical units than the predetermined threshold value.
- The ***dominance rule***, i.e. the ***(n,k) rule***, according to which a cell is sensitive if its *n* largest units contribute more than *k* % to the cell total. When using the dominance rule, parameters *n* and *k* must be defined numerically to ensure equal treatment of the enterprises the statistics concern.
- The ***p% rule***, by which a cell is sensitive if the estimate for the value of the biggest statistical unit calculated based on the total cell value differs at most by *p* per cent from the correct value.

More than one sensitivity rule can also be used side by side. Then a cell is sensitive if it is sensitive according to at least one sensitivity rule.

Cell suppression or ***changing the classification*** are generally used protection method for tables.

- Cell suppression includes primary suppression of cells with a risk of disclosure (sensitive cells) and secondary suppression. Secondary suppression ensures that the values of primarily suppressed cells cannot be disclosed by means of table row or column totals.

  Suppression can also be made specifically for each row. If only a small number of statistical units belong to a particular row total of the table (fewer than the used threshold value), the row is suppressed in total without regard to the number of statistical units in its different cells.
- By changing the classification, the cells with a risk of disclosure are removed from the table by combining the categories certain variables in the table. In practice, changing the classification usually means that the whole classification becomes less detailed.

Adjusting the values of cells with a disclosure risk can also be used as a protection method for tables. Such methods are rounding and replacing the original cell value with an approximate random number.

4. Recommendations on the protection of tabulated personal data

The disclosure risk of persons, individual families or household-dwelling units included in table data must always be assessed when planning tables and before pub-

lishing data. SDC measures should be directed so that the disclosure risk is sufficiently small but without unnecessarily losing information from the data as a result of protection. Account should primarily be taken of the right of the statistical unit to data protection, but at the same time, remembering that society and people have a right to reliable statistical data needed for social decision-making and planning.

4.1 Assessment of the disclosure risk relating to a table

The disclosure risk relating to a table specifies the necessity for SDC measures. When defining the disclosure risk, the following are considered:

- **Variables containing sensitive data in the table,**
- **Small cell and category frequencies (threshold rule),**
- Size of the population,
- Number of variables, and
- Size and location accuracy of the statistical area.

When assessing the protection need of data, it may also be important whether the table values are relative or absolute, whether the data concern one year or whether they are sums or means of several years, and whether the population used in compilation of statistics is a certain special population group (e.g. foreigners, offenders, police, unemployed or high-income earners).

When assessing the disclosure risk, particular attention should be paid to sensitive data, which, according to the Personal Data Act, are sensitive data[4] describing a person's:

- Race or ethnic origin;
- The social, political or religious affiliation or trade-union membership of a person;
- A criminal act, punishment or other criminal sanction;
- The state of health, illness or handicap of a person or the treatment or other comparable measures directed at the person;
- The sexual preferences or sex life of a person; or
- The social welfare needs of a person or the benefits, support or other social welfare assistance received by the person;

and other sensitive data such as:

- Cause of death;
- Language, nationality, origin or country of birth;
- Income, debts and wealth, and
- Main type of activity, occupational status, rare occupation or other variable describing socio-economic group.

If a disclosure risk exists, the personal data in the table must be sufficiently protected.

4.2 Protection recommendations

The disclosure risk directed to the table always concerns certain cells. By defining these cells and the categories including them and by protecting them with a suitable protection method, the disclosure risk of the table can be lowered to an acceptable level.

Use of the threshold rule

[4] Personal Data Act (523/1999), Section 11

Cells containing a disclosure risk, or so-called sensitive cells in the table are determined with the help of the threshold value. In more exact determination of the threshold value, account should be taken of the same matters as when assessing the disclosure risk. In row-specific application, the recommended threshold value is at least ten statistical units and in cell-specific application, at least three statistical units.

Protection methods

The protection of cells (and at the same time the table) can be made by changing the structure of the table, by suppressing individual cell values or whole rows, or by adjusting cell values by rounding, for instance.

If the table has many sensitive cells or the sensitive cells are centred in a few categories, the protection should be made by changing the structure of the table. The structure of the table can be changed by revising the classifications of variables or by controlling the number of variables.

The number of table variables must be the lower the smaller area the statistics concern. In small area statistics, it is advisable to avoid cross-tabulation of several variables and rather publish one-way distributions.

If the table has only a few sensitive cells, protection by cell suppression is recommended. Suppression can be done either cell-specifically or row-specifically, i.e. by suppressing all the cells of the row.

Complementary recommendations

Multidimensional tables. If a table containing personal data has three or more variables, of which at least one variable is sensitive or the area level is smaller than region, the disclosure risk is very probable.

Size of the statistical area. In statistics concerning bigger areas, such as statistics on province level or major regions, data protection measures need hardly ever be taken. In these tables, the population and classifications should also be selected so that the table will not unnecessarily have small cell frequencies.

When the size of the statistical area gets smaller, the number of units to be included in the statistics also decreases. Municipality-based statistics can also include a disclosure risk when the population of the area is small. In statistics on areas smaller than municipalities, the disclosure risk is always possible.

Sample statistics. Sampling has an effect on the disclosure risk. The risk is bigger if statistical data are produced using the whole population as data than if estimates concerning the whole population are produced from the sample data by means of sample weights. However, the disclosure risk of personal tables formed from sample data should also be assessed. To ensure confidentiality and the quality of estimates, the threshold rule should be used in sample-based statistics in defining the sensitivity of cells. The threshold value can be smaller, however, than used in corresponding statistics based on total population data.

Case as the statistical unit. When compiling statistics on cases (e.g. criminal cases, traffic accidents) it is not necessarily a question of personal statistics. The SDC recommendations of personal statistics should, however, be applied to case statistics if an individual person can be identified from the statistics or characteristics of that person can be disclosed.

5. Recommendations on the protection of tabulated business data

Protection of tabulated business statistics can be made in different ways depending on various factors connected to the compilation of statistics. The following presents a three-step hierarchy for business data protection, to which all implementation modes of SDC can be grouped:

1. In situations where exact disclosure is a sensitive matter, the use of the ***threshold rule*** is sufficient. The threshold rule is the default rule. The threshold value always has to be at least three.
2. When approximate disclosure of business data is a sensitive matter, the ***dominance rule*** or the ***p% rule*** must be used as the sensitivity rule. However, using the dominance rule or the p% rule must be restricted to recent statistical data and the threshold value rule must always be used alongside them. Data are recent as long as their disclosure has an impact on the market situation or the activity of an individual enterprise. The time limit for the recentness of data and use of the dominance rule is 15 months from the reference time. The threshold value rule must be used for older data than these.
3. If protection can be made by ***not disseminating the identity and number of data suppliers***, this is recommended. Examples of this are estimates calculated from sample data, in connection with which data on the statistical units belonging to the sample are not published.

A statistical table need not be protected if no disclosure risk is directed to it or the data contained in it are prescribed by law as public. For example, data describing the activity of central and local government authorities and production of public services are mainly public[5]. Old data of over 25 years concerning enterprises are public.[6]

Complementary recommendations

Time dimension. The time dimension must be taken into account when deciding the SDC procedures of business data. The relevance of business data decreases considerably the more time has passed from the reference time of the statistics to the publication time. Very topical short-term statistics must be protected against approximate disclosure as well.

Cell frequencies. Without endangering confidentiality, even small cell frequencies can be published in magnitude tables even if the actual cell value is protected. On the other hand, not disseminating the number of statistical units is a way in applicable circumstances by which the data on the response variable can be made public.

Statistics on establishments. When protecting establishment data, enterprise-level protection must also be ensured. When defining the disclosure risk of a cell, attention should be paid to both the number of establishments in the cell and the number of enterprises to which the establishments belong.

Sample-based statistics. Sampling alone is not necessarily a sufficient protection method, because in sample-based statistics the largest enterprises are usually included and the biggest interest and disclosure risk specifically concerns large enterprises. The disclosure risk of business tables formed from sample data should also be

[5] Statistics Act (280/2004), Section 12
[6] Act on the Openness of Government Activities (No 621/1999), Section 31 If it is a question of an individual own-account worker (personal data), its term of secrecy is 50 years from the death of that person.

evaluated. The estimates and index point figures enabling disclosure must be protected. An estimate or point figure may enable disclosure if data from only a few enterprises are used in its calculation. Then the reliability of the estimate also suffers.

Appendix B: Logging in to and using the remote access system

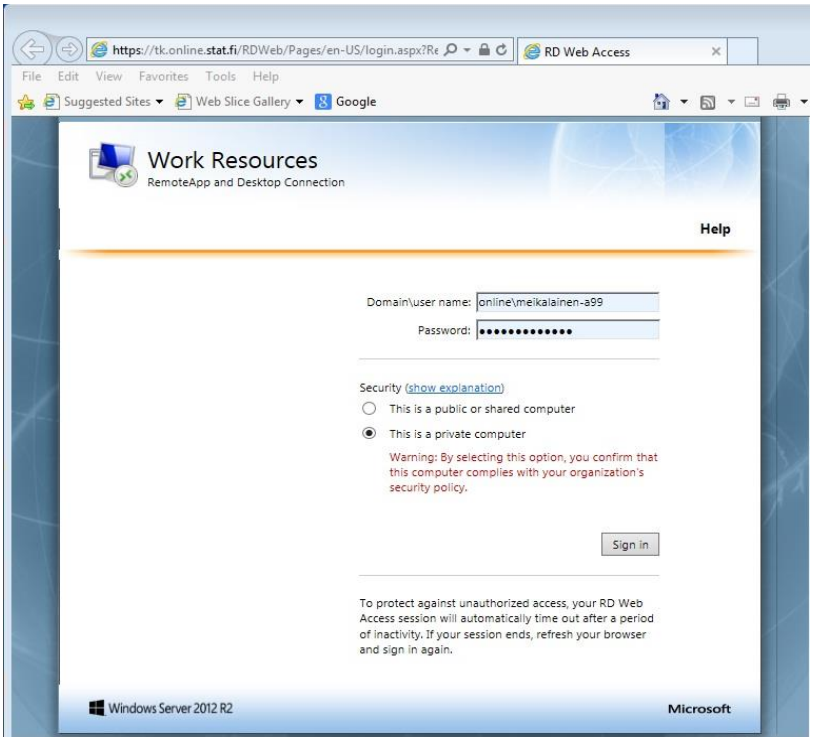
This document describes how to log in to and use Statistics Finland's remote access system for research data. Logging in to the remote access environment differs somewhat depending on what web browser is being used. Guidelines for how to use the SISU-microsimulation model can be found in the handbook for the model.

1. *Logging in to the remote access environment when using IE or FIREFOX browsers*

1. Contact Statistics Finland Research Services' Internet page at (Statistics Finland-> Products and services -> Research data-> Micro data-> Remote Access use) http://tilastokeskus.fi/tup/mikroaineistot/etakaytto_en.html
2. Next, select Logging in to the remote access system by choosing
   - the FIONA-system, CSC-environment. All new projects run on this system.

   You can also log in straight using the address https://fiona.stat.fi

3. The first time you log in you are to allow (Allow) the installation of the add-on. ” This webpage wants to run the following add-on: ”Microsoft Remote Desktop Services Web Access Control…”
4. For signing in you give the domain, which is online, and your User name together with the project number, for example, online\meikalai-a99. Your User name and Password have been sent to you from Statistics Finland. Click ”This is a private computer” before hitting the Sign in –tab.

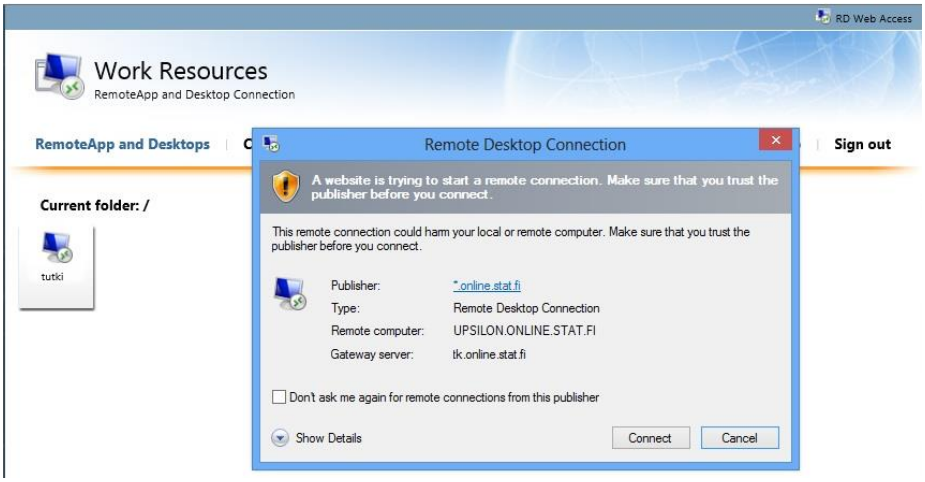


5. The Work Resources –page shows the server groups. Clicking the Tutki –icon you can use the remote servers Alfa, Beta, Gamma, Delta. Zeta and Eta, They have 4 CPU/server.

   Those that have requested the use of the SAS software can choose the Tutki-SAS –icon which opens the remote servers that host SAS 9.4 (Penta, Heksa, Hepta and Okta. 4 cpu/server).

   The Mikrosim –icon leads those using the remote access system for microsimulation to the remote desktop for microsimulation.

   You will be alerted about an unknown server Upsilon. It distributes the load of the sessions. You can tick "Don´t ask me again…" and press Connect.

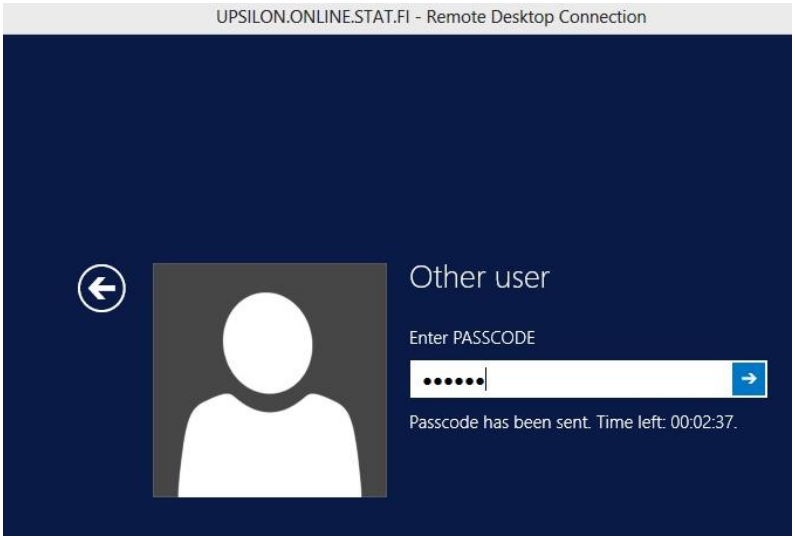


6. If you are using the Firefox-browser, there will at this point be an extra inquiry the first time you log in:
   "You are opening the file cpub-tutki-tutki-CmsRdsh.rdp. What should be done to the

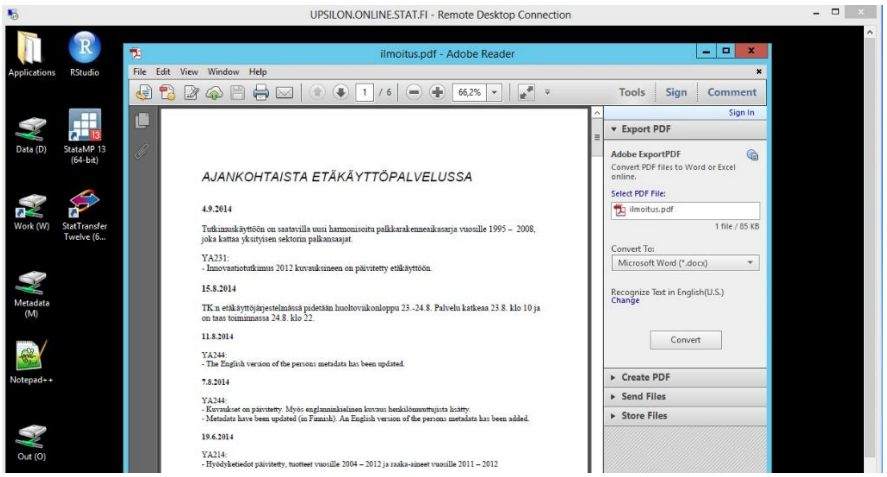
file?" Choose "Open with the program: Remote Desktop Connection" (default) and tick the alternative " Do this as default for this type of files"

7. If you are using the Firefox browser, you will at this point be prompted once again for your User ID. Use the same project number as at point 4.

8. The next step is the SMS Passcode –authentication. The code received as a flash message to your mobile phone is entered under the "Enter Passcode" prompt.



9. After this the remote desktop will open.



2. General working instructions

The remote access system operates on ten Window servers. During in-logging users are directed to the server that has less loading at the time of logging in. Work is carried out as on a normal Windows workstation. Projects are named with consecutive numbering starting with the example project a01.

All material is stored in the work directory W:\a01. The directory is backed up once a day which helps secure the data if the disk server breaks down. The following disks are visible to the user:

D: Data — contains the Research Services' ready-made datasets, reading rights depend on the user licence.

M: Metadata contains data descriptions.

O:\a01 Out directory for transferring results.

W:\a01 Users work directory.

Burdening the system can disrupt other users' work. The server slows down noticeably if the system runs out of memory (due to the use of swap memory). Therefore, the users should avoid unnecessary memory use.

In problem situations, contact the contact person of your organisation. The contact persons are responsible for user support for the software. Statistics Finland's support is only responsible for the functioning of the software and user IDs.

Data and software

Research data are in SAS 7 format (.sas7bdat). The data can be transferred into the desired format using the Stat/Transfer software. The following software are available in the system:

R
Python Anaconda + components
Stata
SAS
SPSS Statistics (limited number of users)
Octave
Stat/Transfer
Graphviz
Kdiff
Libre Office (word processing and tabulation)
Lyx, Latex, MikTeX
Notepad++
Pnet, Pajek
Rstan, Rtools, Rstudio, TramineR (connected to R-program)
Strawberry
QGIS

Statistics Finland does not offer support for the applications in use.

Stata software

Stata 12 , 13 and 14 have been installed in the system. The software being 64 bit means that it can read large amounts of data into memory. However, reserving too much memory results in the system slowing down, which means that at most 4,000 MB of memory can be reserved for Stata.

Finishing work

Log Off

When you finish your work for the day log off the system. Logging off releases the system resources.

Select the Logoff icon on the desktop (or in the start menu).

Close the browser window where you see the log in page.

Disconnect

If you take a short break while working, the remote connection can be disconnected without logging off. You can return to the same session by logging in again. The connection is closed by closing the remote access window (or from the start menu by selecting Shut Down and Disconnect). Disconnect leaves the session open, which means that the resources are not released (software licences and memory reserved for the user). Thus, you must always remember to log off at the end of the day.

The remote desktop connection locks if it remains idle for ten minutes. The connection is closed if it remains idle for 30 minutes. The software are not shut down. By opening the connection again, work can continue.

Transferring of results

All research results are transferred out from the system by administrators. Special consideration shall be applied to sending results for output checking. You should avoid sending results for output checking in small batches. The contents of the tables and graphs that are meant to be checked should be in the same format as they will be published.

Ensure that you follow tabulation rules (see instructions and their appendices).

Document result tabulation carefully in the files to be checked. Every table should be self-explanatory as in journal articles. The number of observations must be visible in each group.

Move the files to the folder O:\a01 (example project).

Send the result output checking request to tutkijapalvelut@stat.fi (include the serial number of the project).

The results are sent to the researcher's email address after the output has been checked.