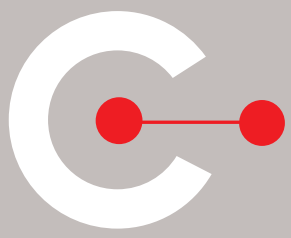
CONFI-
DENTIA-
LITY
RULES

CONFIDENTIALITY RULES



CASD

09/19

Table of contents

01.
Document description

When you are working on the data you were authorised to access, you may want to export files outside your CASD secure work environment. This process is called an "Output request". In the present document, you will find all the confidentiality rules that have to be applied to the files you want to export. The purpose of this document is to help you respect the different kinds of secrecy to apply (statistical, tax, etc.) and to make sure that all published information cannot lead to the direct or the indirect identification of either a physical person, a household or a firm.

If your export does not respect the secrecy to apply to the data in question, the CASD will inform you and will help you either mask the values in question or aggregate them in a different way in order to solve the problem (wider geographical zone, age range instead of age...).

However, **the CASD cannot, in any case whatsoever, modify the files you requested to export.**

02.
Types of outputs

In the first place, you should describe precisely the contents of your output. You can do this in the email requesting the output or in a separate text file inserted in your output. You can find a description of the process to carry out an output in the User guide available on the CASD website.

PROGRAMS

You can request an export of your programs to reuse them outside the CASD. Programs cannot contain confidential data.

REGRESSION AND ECONOMETRIC MODELS

Output requests can contain results of econometric models under SAS, Stata, R, etc. You should keep in your output known parameters and indicators related to the econometric models (estimator, R^2, pseudo R^2, confidence interval, Khi^2, etc.) and the number of observations so that the CASD can make sure that it is in fact a regression or an econometric model.

GRAPHICS AND MAPS

For graphics, such as curves, histograms, scatterplots, diagrams, etc., you should give us the information on which they were based (population and meaning of the used variables).

Other types of graphics are more complicated to review because they can contain individual information. For example, boxplots, not only can contain the maximum and the minimum, but also extreme points (the outliers) that should not be identified.

You should pay attention to factor analysis graphics representing individuals. They can contain atypical individuals that can be identified (for example a PCA on firms where the SIRET is used in the graphic as an identifier of each point).

Stata Graphics in **LIVE** Format contain the datasets that lead to their constitution. It is preferable to convert them into another format (AS-IS: a format that does not contain the datasets, pdf, jpg...).

Maps can concern very small populations (communal or infra-communal). In order to verify that the confidentiality rules are respected, you should provide us with the information on which the maps were based.

AGGREGATED DATA TABLE

In the description of the output, you should give us all the necessary information to understand and identify the types of data used: list of the variables, their meanings, frequencies of every cell, and information on the maximal contribution in the cell.

Usually confidentiality rules vary according to different type of source.

FINALISED ARTICLES

The finalised articles that you wish to export outside your CASD work environment cannot contain data concerning a small population of individuals.

In addition, in your article you should indicate that your work was made possible thanks to CASD device in the following manner:

« This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-EQPX-17 - Centre d'accès sécurisé aux données – CASD) » or in French « Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir portant la référence ANR-10-EQPX-17 (Centre d'accès sécurisé aux données – CASD) ».

03.
General rules

"HOUSEHOLD" DATA

For tables showing aggregated household data, the only imposed rule is that the direct or the indirect identification of individuals should be impossible. In practice, we consider that the statistical secrecy is respected if the knowledge of one characteristic of an individual cannot lead to the knowledge of another one with which it was crossed in the table.

Example

The following table represents the age distribution according to the matrimonial situation. It indicates that the persons aged between 50 and 59 years have all the same matrimonial status: "divorced". The statistical secrecy is not respected in this table, thus making it unpublishable. Indeed, if we know otherwise that a given individual is in the age range of 50-59 years, the table informs us as well that this individual is divorced, even though the cell crossing the levels "50 to 59 years" and "divorced" contains multiple individuals.

P.4

Matrimonial situation and age range	18-25 years	26-49 years	50-59 years	60 years and +
Married	7	27	0	30
Divorced	0	11	9	22
Other	21	12	0	4

"FIRM" DATA

For tables containing aggregated firm data, the rules to apply are the following :

- No cell should contain less than 3 units (decision of the 13 June 1980 of the general director of the Insee).
- No cell should contain information for which a sole firm contributes to more than 85% of the total (rule of the Cnis, 7 july 1960).
- No output can contain a list of SIREN/SIRET even it is associated to any other information.

MIXED SOURCES

Mixed sources are either the results of combinations (merges) between statistical surveys and administrative data or sources containing financial and economic information (firms) and private facts and behaviour information (households) at the same time.

When working on this type of source, the principle is simple: the rules to take into consideration are the accumulation of the rules applied to statistical surveys and the rules applied to administrative data. For example, the tax and social incomes survey is a combination of statistical surveys, results of the employment survey, fiscal data and of data provided by the Caisse D'allocations familiales (family allowance funds).

DATA FROM FISCAL SOURCES

- Fiscal firm data: same rules as the rules applied for firms.
- Fiscal household data: a threshold of 11 individuals per cell.

PRIMARY SECRECY

Cells that do not respect confidentiality rules, which are therefore masked, form what we call the primary secrecy. It is simple to manage: for each source, specific rules for each different source are applied.

Example

Firms' categories	Number of firms
TPE	7
PME	2
Big firms	10

Firms' categories	Number of firms
TPE	7
PME	s
Big firms	10

SECONDARY SECRECY

The secondary secrecy is more complex to manage. It concerns the tables with margins and it impedes the reconstitution of masked cells of the primary secrecy by addition or subtraction.

Situations that can lead to the secondary secrecy are the following:

Margins and hierarchal variables

Because of margins present in tables or on the Internet, it is sometime possible to retrieve the value of the masked cells of the primary secrecy. To solve this problem, two cells should be masked.

Example

Firms' category	Number of Firms
TPE	7
PME	2
Big Firms	10
Total	19

Firms' category	Number of Firms
TPE	s
PME	s
Big Firms	10
Total	**19**

Hierarchal variables concern mainly geographical variables or nomenclatures, they should be treated like margins and inserted in the table.

Example

These two tables cannot be published separately because they have additive relations between them. If a user wishes to export the first table showing the number of firms in Bretagne, according to the confidentiality rules applied to firms data, he should mask the cell showing the number of firms in Finistère.

WARNING: in this example, it is imperative that the second table permitting to find the masked cells of the first table by showing the number of firms at the regional level, cannot be exported in a later output or diffused in any other way. You can anticipate any breach of confidentiality by masking two cells in the first table.

Department	Number of firms
Morbihan	8
Finistère	2
Côtes-D'Armor	9
Ille-et-Vilaine	6

+

Region	Number of firms
Basse-Normandie	20
Bretagne	25
Pays de la Loire	28

↓

Region	Number of firms
Morbihan	8
Finistère	S
Côtes-D'Armor	9
Ille-et-Vilaine	S

One no-null level

Having only one no null level in a table leads to the identification of an individual or a firm characteristic by another one. To avoid this problem, it is necessary to have two non-null levels.

Example

After statistical treatment, if we have a table indicating that the persons who are between 20 and 34 years old are all in the same professional situation (Jobseekers), statistical secrecy is no longer respected. In fact, if, in any other way, we know that a given individual is in the age range of 20-34 years, this table informs us also that this person is necessary looking for a job, even if the cell crossing the level "20-34" with the level "Jobseeker" contains multiple individuals (7 in this case).

Professional situation and age range	Employee	Jobseeker	Non-employee	Other
20-34 years	0	7	0	0
35-49 years	6	5	5	8
50-64 years	4	5	6	5

↓

Professional situation and age range	Employee	Jobseeker	Non-employee	Other
20-34 years	0	s	s	0
35-49 years	6	5	5	8
50-64 years	4	5	6	5

CONTROL FILES

To verify the results of amount variables, you should also provide us with a separate control file that contains all the information that needs to be exported, as well as additional columns indicating the maximum and the percentage that this maximum represent in the amount variable.

This control file will be removed before the output is sent.

Example

CONTROL FILE

	Firm number	Maximum value	Total amount	Percentage of the maximum
Bretagne	**139**	**1 668**	**27 800**	**6%**
Morbihan	82	1 590	19 882	8%
Finistère	19	590	3 476	17%
Côtes-d'Armor	36	1 103	2 567	43%
Ille-et-Vilaine	2	1 312	1 875	70%
Picardie	**99**	**825**	**20 643**	**4%**
Oise	67	825	13 750	6%
Aisne	5	722	821	88%
Somme	27	790	6072	13%

OUTPUT FILE

	Firm number	Total amount
Bretagne	**139**	**27 800**
Morbihan	82	19 882
Finistère	S3	S4
Côtes-d'Armor	36	2 567
Ille-et-Vilaine	S1	S2
Picardie	**99**	**20 643**
Oise	67	13 750
Aisne	5	S9
Somme	27	S10

The reason behind the masked values:

S1	**Number of firms is less than 3**
S2	**Information related to S1**
S3	**To avoid recalculating S1 from the total amount**
S4	**To avoid recalculating S2 from the total amount**
S9	**The percentage of the maximum is higher than 85%**
S10	**To avoid recalculating S9 from the total amount**

04.

Detailed rules for each source

INSEE DATA

Secrecy rules related to INSEE sources are described in the statistical secrecy guideline of INSEE:

https://www.insee.fr/en/statistiques/fichier/2388575/guide-secret-en.pdf

DADS (Annual declaration of social data)

No published table can contain any information leading to the direct or indirect identification of either an employee or a company.

- Tables about place of residence (« household » vision):
  - No cell can contain less than 5 employees
  - No cell can contain a sole employee contributing for more than 80% of the workforce
- Tables about place of work ("company" vision), in addition to the two previous rules, the following rules concerning firm data are added:
  - No cell can contain less than 3 firms or establishments
  - No cell can contain a sole enterprise or establishment contributing for more than 85% of the total

CLAP (Local knowledge of the production system)

CLAP data belong to « firm » data. Indicators subject to statistical secrecy are salaries and remuneration. The rules to apply are the following:

- No cell can contain less than 3 units (a unit is either an enterprise or an establishment)
- No cell can contain a sole unit contributing to more than 85% of the total
- No cell can contain less than 5 employees

FARE (File approaching the results of the Elaboration of annual statistics of companies) / FICUS (Unified Corporate Statistics System)

FARE and FICUS data are mixed sources, they are both "statistical and tax" data at the same time. The confidentiality rules to apply in this case are the sum of rules applied to statistical surveys and tax data.

- No cell can contain less than 3 units
- No cell can contain 1 unit contributing to more than 85% of the total
- As for sole proprietorship, no cell can contain less than 11 units

SINE (Information system for new firms)

SINE is also a « firm » source. The confidentiality rules to apply are the following:

- No result can concern less than 3 units per cell
- No data in which a sole company contributes to more than 85% of the total value

Furthermore, rates of survival should not be calculated on a population containing less than 20 enterprises. This minimum threshold of 20 firms is also required for zoning and particular regroupings.

The population census

The rules of publishing data issued from the population census evolved with the Bylaw of 19 July 2007 relative to the population census results publication. This bylaw replaced the one of 22 may 1998, which have been modified on the 8th of April 2002, and which was relative to the population census results publication.

For all variables, cells tables should contain at least 4 observed units (before weighting).

- At least 10 for a 40% poll
- At least 16 for a 25% poll
- At least 20 for a 20% poll

Concerning "sensible" variables, specific geographical thresholds should be respected. Sensible variables are the following: current nationality (or nationality at birth), place of birth, former place of residency, arrival year (or duration) to France. Notions of immigrants and French Citizens by acquisition are included in the range of sensible variables until the census of 1999.

The geographical thresholds for these variables are the following:

1999 and before	From 2006 (annual census)
Communes with more than 5 000 residents	Communes with more than 5 000 residents
Threshold of 10 000 residents for the arrondissements, employment zones, urban airs, urban units (or their **regroupings**) and zones of urban public policies or their regroupings	Threshold of **5 000 residents** for the arrondissements, employment zones, urban airs, urban units and zones of urban policies
Infra-communal zones resulting from the combination of 3 neighbourhoods (a fixed zone resulting from the division of the commune into geographical zones containing around 2 000 residents)	
Department for the year (or the duration) of arrival	**Department** for the year (or duration) of arrival

For output concerning sensible variables of the Census, and in order to verify it, you should provide us with the geographical scale on which your data are based. In the case of data aggregated at the commune level, you should provide us with the inhabitants number in each commune or group of commune for all concerned variables.

SIASP (System for Information on Civil Servants)

The publishing of statistical results based on the SIASP data should conform to the rules depicted by statistical secrecy and protection of individual data related texts. In particular, no table meant to be published, should lead to the direct or the indirect identification of either an employee or an establishment:

Tables about place of residence:

- No cell can contain less than 5 employees
- No cell can contain a sole employee contributing to more than 80% of the workforce

Tables about place of work, **in addition to the two previous rules,** the following rules must be applied:

- No cell can contain less than 3 establishments
- No cell can contain a sole establishment contributing to more than 85% of the total

Exception: for data related to the state civil service, at the national, regional or departmental level, it is not always possible to find three establishments with a Siret number. In consequence, this rule will not be applied only in this context.

FIDELI (Demographic files on housings and individuals)

The rules for FIDELI are the following:

- For communes with 5 000 residents or more, the minimal geographical scale for producing tables is the commune or the IRIS.
- For communes with less than 5 000 residents, the minimal geographical scale for producing tables is the Établissement public de coopération intercommunale (EPCI: the public institution of intercommunal cooperation). However, the EPCI where the total population of communes of less than 5000 residents is

less than 2000 residents, cannot lead to the production of tables.

- For results that concern Urban policy priority neighbourhoods (QPV), the minimal geographical scale for producing results is the region.
- **For all published results,** no cell can contain less than 11 individuals.

For output concerning FIDELI Data, and in order to verify it, you should provide us with the geographical scale on which your data are based. In the case of the data being aggregated at the commune or at the EPCI level, you should provide us with the inhabitants number in each commune for all concerned variables.

DARES DATA – MINISTRY OF LABOUR

The rules for DARES's data are the following:

- For individual/household data : no cell can contain less than 5 individuals
- For firm data : no cell can contain less than 5 individuals and no cell can contain a single enterprise contributing for more than 85% of the total

DEPP DATA – MINISTRY OF EDUCATION

The rules for the DEPP's data are the following:

- For individual data : no cell can contain less than 10 individuals
- For aggregated data at the establishment level (secondary school, high school, etc...): no cell can contain less than 10 establishments

SSP DATA – MINISTRY OF AGRICULTURE

Agriculture exploitations are considered as firms and are subject to the same confidentiality rules than firm data, while parcels are assimilated to establishments:

- No cell can contain less than 3 firms/agriculture exploitations (the statistical secrecy is applied to exploitations and not on parcels)
- No cell can contain data in which a sole agriculture is contributing for more than 85% of the total.

SDSE DATA– MINISTRY OF JUSTICE

A minimum of 5 observations in each cell of the table.

CIFRE DATA OF THE MESRI – MINISTRY OF RESEARCH

The rules for CIFRE data of the MESRI are the following:

- In the case of individual data, it is forbidden to publish information, which can lead to a person direct or indirect identification. In addition, even if we cannot identify a person, it is forbidden to give an information about him. These rules limit the preciseness of available published information. Strict rules are defined specifically for the population census. No output can contain 10 observations concerning individuals
- In the case of firm data, no published result can concern less than 3 enterprises and no data can concern a single company contributing for more than 80% of the obtained value. However, diffusion of lists extracted from the repertory of enterprises or establishments mentioning economic activity, a range of frequencies and range of turnovers is allowed.
- No edition of lists providing the name, the address or any other individual characteristics of beneficiary students or enterprises.

CEREQ DATA

A minimum of 5 observations in each cell of the table.

BPI FRANCE DATA

A minimum of 10 observations in each cell of the table.

ANIL DATA

A minimum of 50 observations in each cell of the table.

ODR DATA

Information based on a geographical scale which is smaller than the canton cannot be published. A minimum of 3 observations per cell should be respected as well.

For output concerning ODR Data, and in order to verify it, you should provide us with the geographical scale on which your data are based.

MSA DATA

The rules of MSA's data are the following:

- No cell can contain less than 5 units
- No cell can contain data toward which a single non-employee contributor represents more than 85% of the total
- The knowledge of an individual characteristic cannot lead to the knowledge of another one with which it is crossed in the same table.

ACOSS DATA

A minimum of 10 observations in each cell of the table.

CNAF DATA

A minimum of 5 observations in each cell of the table.


C/\SD