

List of de-identification operations and scores for data utility and confidentiality



THE GOVERNMENT
OF THE GRAND DUCHY OF LUXEMBOURG
Ministry for Digitalisation

Government commission for data sovereignty

Updated: 01/05/2026

List of de-identification operations and scores for data utility and confidentiality

Objective of the document

This document aims to normalise operations that enable anonymisation, pseudonymisation, and aggregation, or any other data processing operations that ensure control over disclosure prior to data access and reuse in the SPE. There are two types of operations: de-identification operations and data utility and confidentiality measures. This nomenclature is to be used by the applicant to specify the appropriate measures plan.

History

Version	Date	Changes made
1.0	01-05-2026	Initial version

Contents

LIST OF DEFINITIONS.....	5
1. INTRODUCTION	7
2. DE-IDENTIFICATION TECHNIQUES.....	8
2.1 Statistical approaches.....	8
2.1.1 Stratified random sampling	8
2.1.2 Macro-aggregation of records	10
2.2 Deletion and masking.....	12
2.2.1 Vertical deletion.....	12
2.2.2 Record-level sub-setting	13
2.2.3 Horizontal deletion	15
2.2.4 Identifier Substitution	16
2.2.5 Masking	17
2.3 Generalisation	18
2.3.1 Absolute rounding.....	19
2.3.2 Relative rounding	20
2.3.3 Generalisation of Luxembourg zip codes.....	21
2.3.4 Generalisation of dates/times (datetime)	22
2.3.5 Period generalisation	23
2.3.6 Generalisation by mapping table	23
2.3.7 Data Bucketing	24
2.4 Micro-aggregation	25
2.5 Formal anonymisation and privacy models	26
2.5.1 K-anonymity model.....	26

2.5.2	L-Diversity model	31
2.5.3	T-Closeness model	34
2.5.4	β -likeness	37
2.5.5	Differential privacy.....	39
2.5.6	Synthetic data	41
3.	UTILITY AND CONFIDENTIALITY MEASURE OF DE-IDENTIFIED DATA	42
3.1	Rate of deleted or replaced values.....	43
3.2	Modified record rate	43
3.3	Number of values of the variables	43
3.4	Information loss rate	43
3.5	Retention rate.....	43
3.6	Distinct value retention rate	44
3.7	Sensitive Attribute Distortion Rate	44
3.8	Discernability Metric	44
3.9	Kolmogorov-Smirnov Distance	44
3.10	Jensen-Shannon distance	45
3.11	Hellinger distance	45
3.12	Correlation similarity	46
4.	CORRELATION SIMILARITY MEASURES THE DIFFERENCE BETWEEN THE PAIRWISE CORRELATION MATRICES OF THE ORIGINAL AND SYNTHETIC DATASETS ACROSS ALL CONTINUOUS VARIABLES. CONCLUSION	46

List of definitions

Concept	Definition
Anonymisation	<u>The process of making personal data anonymous so that the data subject to whom it relates is not or is no longer identified or identifiable, considering all means reasonably likely to be used to identify the person directly or indirectly.</u>
Sensitive variable	<u>Variables that reveal intimate or legally protected information (e.g., health, religion, political opinions).</u>
De-identification	<u>The process of reducing or removing the ability to identify an individual in a dataset.</u>
Personal data	<u>Any information relating to an identified or identifiable person (e.g., name, address, online identifier).</u>
Sensitive data	<u>A subcategory of personal data that contains sensitive variables.</u>
Direct personal identifiers	<u>Variables that directly identify a person (e.g., name, social security number, email address).</u>
Pseudonymisation	<u>The processing of personal data in such a way that it can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data is not attributed to an identified or identifiable person.</u>
Quasi-identifiers	<u>Variables that, on their own, are not sufficient to identify a person, but which can do so when combined with other variables (e.g., age, postal code, occupation) or external data sources.</u>

List of figures

Table1 : Example of using stratified random sampling	10
Table2 : Example of macro-aggregation usage	12
Table3 : Example of vertical deletion	13
4 : Example of targeting usage	14
Table5 : Example of horizontal deletion	16
Table6 : Example of use of personal identifier substitution.....	17
7 : Example of use of masking	18
Table 8 Example of source data to be protected	28
Table 9 Example of direct identifier, quasi-identifier, sensitive variable.....	29
10 : Example of k-anonymity usage.....	29
Table 11 Example of the result of applying K-anonymity.....	30
Table12 : Example of L-diversity operation	32
Table13 : Example of L-diversity with L=3.	33
Table 14 : Example of using T-closeness.....	36
Table 15 : Example of T-closeness with T=0.05.	37
Table 16: Example of using β -likeness	39
Table 17 : Example of the use of differential privacy	41
Table 18 : Properties of data minimisation techniques.	46
Table 19 : Properties of data de-identification techniques.....	47

1. Introduction

This document presents a comprehensive analysis of data de-identification techniques and their practical applications in the appropriate measures plan.

The purpose of this document is to standardize the de-identification operations that are applicable in the Intermediate Processing Environment (IPE) during their preparation before making them available in the Secure Processing Environment (SPE). It thus enables data reusers to specify de-identification operations, as well as usability and confidentiality criteria, in a precise and unambiguous manner.

The de-identification techniques proposed here are based on ISO/IEC 20889:2018[1] , as well as advanced scientific literature. This document builds upon and refines these techniques in accordance with existing implementations available in the Intermediate Processing Environment.

Each technique is described in terms of its context of use, concrete examples, and a guide to the parameters to be configured for each one.

De-identification techniques are also supplemented by standardisation of data confidentiality and utility measures. These measures will make it possible to assess the quality and relevance of the data after one or more transformations. The assessment of these criteria may be shared with applicants, after analysis of the confidentiality of these criteria by the Data Privacy Engineer.

Finally, the Section 4 provides an overview of the techniques and measures presented, highlighting their complementarity and their contribution to responsible data processing.

2. De-identification techniques

De-identification techniques aim to reduce the ability to identify or re-identify a person from a dataset, while preserving the analytical usefulness of the data. The ISO/IEC 20889:2018 standard [1] establishes an international reference framework by defining common terminology and classifying appropriate methods to effectively reduce the risk of re-identification. At the same time, the GDPR (particularly in its recitals and Articles 4 and 25) encourages the use of pseudonymisation and anonymisation as means of data minimisation, allowing personal data to be used while strengthening the protection of the rights of data subjects.

The ISO/IEC 20889:2018 standard on de-identification techniques also mentions the need to reduce data exposure by applying transformations such as deletion, generalisation, or aggregation. Although targeting or vertical deletion are not classified as de-identification techniques in ISO/IEC 20889:2018, they are considered additional *technical and organisational measures for reducing or transforming data* in order to ensure compliance and secure processing operations relating to personal data, but also to other types of protected data.

Indeed, the principle of data minimisation is one of the foundations of personal data protection, as defined by Article 5(1)(c) of the General Data Protection Regulation (GDPR⁽²⁰¹⁶⁾). This principle requires data controllers and processors to reduce the amount and granularity of data collected and used in order to limit the privacy risks to data subjects. Minimisation measures include all techniques and practices that enable this objective to be achieved, either before or in addition to de-identification mechanisms.

2.1 Statistical approaches

2.1.1 Stratified random sampling

Stratified sampling is a method of sampling from a population. This technique involves selecting data at random while ensuring a homogeneous distribution according to specific variables. Unlike the generation of synthetic data, the data obtained remains real. The objective is to comply with the criteria defined by the requester while ensuring that the different strata are representative¹.

¹ [Stratified sampling — Wikipedia](#)

Context of use

For example, a requester may wish to carry out sampling comprising X individuals per age group and socio-professional category. To do this, they must:

- Identify the variables to be taken into account (e.g., age group, socio-professional category).
- Define the desired number of individuals in each sample (X individuals).

The requester may also indicate that they wish to obtain the maximum possible number of records per sample. In this case, the system will analyse the data distribution and select the smallest number available from the combinations of values.

It is important to note that this technique may fail if the initial conditions are not met (e.g., if an age group or occupational category contains fewer than X individuals). In this case, it is necessary to define a strategy to manage these situations:

- Keep the value with a reduced number of individuals.
- Remove the value from the sample.
- Stop processing.

Parameters

Name of technique	Stratified random sampling
Parameters	
ID	Description
VARIABLE_LIST	<u>List of variables that must be taken into account for sampling.</u>
STRATEGY	<u>One of the following values:</u> <u>COMPUTE_MAX: the largest possible number of individuals</u> <u>KEEP_UNSATISFACTORY: keeps unsatisfactory values</u> <u>REMOVE_UNSATISFACTORY: removes unsatisfactory values</u> <u>STOP_PROCESS: stops processing in case of failure</u>
COHORTE_COUNT	<u>The number of individuals per sample. This parameter is not taken into account in a strategy based on the largest possible number of individuals per sample.</u>

Example of use

To obtain a dataset comprising 1,000 individuals per age group and socio-professional category, insert and adapt the following operation:

Table1 : Example of using stratified random sampling

Process ID	TRT-EXAMPLE-1_1
Description	Sampling by age group and socio-professional category
Technique Name	Stratified random sampling
Parameters	
ID	Name
VARIABLE_LIST	AGE, OCCUPATION
STRATEGY	STOP_PROCESS
COHORT_COUNT	1000

2.1.2 Macro-aggregation of records

Macro-aggregation groups records of a dataset together based on the values of one or more fields. This technique aims to reduce the granularity of data in a dataset to limit the risk of re-identification while preserving its statistical utility.

In the context of de-identification practices, two aggregation methods are distinguished in the ISO/IEC 20889:2018 standard[1] :

- Macro-aggregation, described in this section, operates at a more global level by grouping data according to grouping variables and calculating aggregates for other variables, thus producing macro-data suitable for reports or consolidated analysis.
- Micro-aggregation (Section 2.4), which consists of grouping records into small homogeneous sets and replacing individual values with a representative value (mean, median, etc.) in order to preserve the statistical structure while masking the exact data.

The dataset resulting from macro-aggregation contains two types of variables:

- Variables whose values allow records to be grouped together (e.g., locality, sector).
- Aggregated variables (mean, sum, etc.) for the same set of grouped variable values.

Context of use

The aggregated data (macro level) resulting from macro-aggregation is often used for reports or overall analyses. Example: calculating the average age by locality or the total income per municipality.

The objective of macro-aggregation is to reduce the granularity of the data by moving from an individual level to an aggregated level (e.g., by region or category).

In this operation, in the resulting dataset:

- The grouping variables will be preserved.
- The transformation applied to the aggregated variables will be detailed in the processing description.
- Other variables will not be retained in the resulting dataset.

Parameters

Name of the technique	Macro-aggregation		
Parameters			
ID	Description		
AGGREGATION_VARIABLES	List of variables on which aggregation is to be performed. These variables will be present in the resulting dataset.		
AGGREGATED_VARIABLES	List all variables that must be retained in the dataset in addition to the aggregation variables. For each variable to be retained, specify the aggregation method to be used. For a variable to be aggregated, several OUTPUTs can be specified. For example, for a numeric variable, you can define an output with the average value, an output with the maximum value, etc. The set is specified in the form of a table:		
	INPUT	OUTPUT	Method
	Identifier of the variable to be aggregated	Identifier of the variable containing the aggregated values.	Method from the following list: SUM MEAN MAX MEDIAN MIN MODE PERCENTILE STANDARD DEVIATION COUNT

Example of use

For a dataset containing people's ages by locality, where we want to obtain an aggregation of ages across localities, the grouping variable will be locality, and the aggregated variable will be age. We can

specify several aggregation methods for a variable (for example, mean, minimum, and maximum). The Appropriate Measures Plan will then contain the following description:

Table2 : Example of macro-aggregation usage

Process ID	TRT-EXAMPLE-1_2		
Description	Aggregation of age by locality		
Technique Name	Macro-aggregation		
Settings			
ID	Name		
AGGREGATION_VARIABLES	LOCALITY		
AGGREGATED_VARIABLES	INPUT	OUTPUT	Method
	AGE	MEAN_AGE	MEAN
	AGE	MAX_AGE	MAX
	AGE	AGE_MIN	MIN

2.2 Deletion and masking

The deletion technique is a de-identification method that involves completely removing certain variables from a dataset to reduce the risk of re-identification. Its purpose is to eliminate direct identifiers (name, address, social security number) or quasi-identifiers (date of birth, postal code) to prevent individuals from being identified from the remaining data. This approach is classified as a "destructive" technique because it deletes information rather than transforming it. The deletion technique is particularly suitable when the deleted variables or records are not essential to the analysis [1]. However, there are several types of deletion.

2.2.1 Vertical deletion

Vertical deletion allows you to completely delete a variable (i.e., a column in tabular data) from a dataset.

Context of use

This measure aims both to comply with the principle of data minimisation provided for in the GDPR (Article 5(1)(c)) and to reduce the amount of information available in the SPE as much as possible in order to:

- Limit the risks of re-identification.
- Reduce the exposure of sensitive data.
- Optimise processing performance.

Vertical deletion must therefore be used systematically for all variables that are not useful for subsequent analysis by the Data Scientist in the Secure Processing Environment.

To delete a variable, the requester must consider two scenarios:

- Either the variable is not used in de-identification processing, in which case the requester asks the holder to delete it in order to make the data available.
- Or the variable is used in de-identification or cross-referencing processing, in which case the applicant must explicitly specify a deletion process in the list of de-identification processes (i.e., after use in the IPE).

Parameters

Technique name	Vertical deletion
Parameters	
ID	Description
VARIABLE_LIST	List of variables to be deleted.

Example

A study aims to target the population aged over 50 years, but does not need age in this context. The "AGE" variable must be considered in the Appropriate Measurement Plan for targeting, but this variable must be deleted from the target dataset. In this context, the variable will be marked as "to be used" in the "source description" section; the deletion process will then be specified in the list of processes as follows:

Table3 : Example of vertical deletion

Process ID	TRT-EXAMPLE-2_1
Description	Deletion of unnecessary variables
Technique Name	Vertical deletion
Parameters	
ID	Value
VARIABLE_LIST	AGE

2.2.2 Record-level sub-setting

Record-level sub-setting consists of retaining only specific records. Records are retained according to a predefined record retention rule.

Context of use

Record-level sub-setting Targeting is a functional process that involves selecting a subset of individuals or items from a dataset, based on defined criteria, to apply a specific action for a particular purpose. Targeting is necessary to match the data made available to the business purpose and processing strategy. This technique allows only records that comply with a specified rule to be retained in a dataset. The goal is to:

- Identify a relevant population for a specific purpose.
- Optimize the effectiveness of actions by reducing the scope to what is necessary.

Record-level sub-setting will therefore be used to target the elements to be displayed in the target dataset. This technique should be used to modify the geographic, temporal, or individual category scope of the target dataset in relation to that of the source datasets. It should be noted that the scope can also be modified when crossing multiple source datasets.

The context of use of record-level sub-setting is based on:

- A domain selection criterion such as demographic, behavioural, geographic, temporal, etc.
- An explicit purpose, where each targeting must be justified by a legitimate and documented purpose.
- The impact on individuals' rights, which may lead to risks of discrimination or profiling and requires a Data Protection Impact Assessment (DPIA).
- Risks relating to otherwise protected data.

Parameters

Name of technique	Record-level subsetting
Parameters	
ID	Description
RETENTION_RULE	<u>Record retention rule</u> <u>The rule can be applied to several variables in the dataset.</u> <u>It is written as a Boolean formula using the following operators:</u> ▪ <u>Boolean operators: OR, AND, NOT</u> ▪ <u>Mathematical operators: = < > ≤ ≥ ≈</u> ▪ <u>Set operators: IN (xx,yy,zz)</u> ▪ <u>Basic types: numbers, Boolean (True, False), String.</u>

Example of use

A study wants to target the population aged older than 50 years. The variable "date of birth" must be considered in the appropriate measurement plan to carry out the targeting.

4 : Example of targeting usage

Process ID	TRT-EXAMPLE-2_2
Description	Targeting people over 50
Technique Name	Record-level sub-setting
Parameters	
ID	Value
RETENTION_RULE	AGE > 50

2.2.3 Horizontal deletion

Horizontal deletion or filtering refers to the complete deletion of records (rows) from a dataset. Records are deleted according to a predefined deletion rule.

Context of use

Like record-level sub-setting, horizontal deletion aims to limit:

- Limiting the volume of data to what is strictly necessary for the purpose of processing.
- Reducing the risks associated with storing or using irrelevant data.
- Prepare a target dataset consistent with business needs.

This technique can also be used to delete outliers, which are, in statistics, observations that are distant from other observations in a dataset. Such data present significant risks of individualisation in a dataset.

This technique can be used in the context of outliers known a priori in the source datasets, either because they have been detected by the data holders or because they have been identified as such during the risk analysis performed on the data to be made available.

Parameters

Name of the technique	Horizontal deletion by rule
Parameters	
ID	Description
DELETION_RULE	<u>Record deletion rule</u> <u>The rule can be applied to several variables in the dataset.</u> <u>It is written as a Boolean formula using the following operators:</u> ▪ <u>Boolean operators: OR, AND, NOT</u> ▪ <u>Mathematical operators: = < > ≥ ≈</u> ▪ <u>Set operators: IN (xx,yy,zz)</u> ▪ <u>Basic types: Numbers, Boolean (True, False), String</u>

Example of use

Risk analysis has shown that the population aged over 90 years presents risks of re-identification. All data on persons over 90 years of age are deleted from the dataset. The Appropriate Measures Plan, therefore, contains the following operation:

Table5 : Example of horizontal deletion

Process ID	TRT-EXAMPLE-2_2
Description	Deletion of data relating to persons over the age of 50
Technique Name	Horizontal deletion
Parameters	
ID	Value
DELETION_RULE	AGE > 50

2.2.4 Identifier Substitution

This technique involves replacing the values of a variable, often corresponding to a Personal Identifier (PID), with a consistent artificial identifier, known as a pseudonym. The value of the PID is replaced with its encrypted version. This technique is often referred to as pseudonymisation of personal identifiers.

Context of use

The de-identification of personal identifiers is mandatory when making data available in the SPE. The preferred technique is vertical deletion. However, it may be considered that these personal identifiers be processed by substitution.

In this operation, all the variables in the dataset are retained. Only the values in the PID variables are replaced by pseudonyms.

A replacement may be deterministic in a particular context, meaning that the same pseudonym is consistently assigned to the same PID. In the context of the DGA, only the following contexts are permitted:

- None: no context defined, the pseudonym will always be different.
- Dataset: the pseudonym will be the same for the same PID in the same dataset. It will be different for different datasets.
- Project: the pseudonym will be the same for the same PID in any dataset in the project.

If a target variable name is specified, the source variable is deleted, and the target variable is added. It is prohibited to keep both the source variable and the target variable in the dataset resulting from this transformation.

Parameters

Technique name	Identifier substitution
Parameters	
ID	Description

VARIABLE	<u>Variable on which the identifier substitution is to be made.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the substituted values (pseudonyms).</u>
CONTEXT	<u>NONE: substitution will be random</u> <u>DATASET: the pseudonym will be the same within the dataset but will be different for two different datasets.</u> <u>PROJECT: the pseudonym will be the same for all data sets in the project.</u>

Example of use

Table6 : Example of use of personal identifier substitution

Process ID	TRT-EXAMPLE-2_4
Description	Encryption of employee numbers S2_RESID_MAT
Technique Name	Substitution of personal identifiers
Parameters	
ID	Name
VARIABLE	S2_RESID_MAT
TARGET_VARIABLE	S2_RESID_PSEU
CONTEXT	PROJECT

2.2.5 Masking

The masking-based deletion technique involves replacing all or part of an variable's values with neutral characters (e.g., "***" or "XXXX") or symbols in order to render the information unreadable while preserving the structure of the values. The purpose of masking is to prevent the disclosure of sensitive or identifying information while maintaining the consistency of the format for further processing. This method is classified as a non-destructive technique because it does not delete the variable but alters its content.

Context of use

Masking is particularly useful in environments where data must be partially visible for operational reasons, such as displaying the last four digits of a credit card. This approach balances confidentiality and utility but must be applied with clear rules to avoid indirect re-identification risks.

Parameters

Name of the technique	Substitution by masking
Parameters	
ID	Description

VARIABLE	<u>Variable on which the substitution is to be performed.</u>
SELECTION_RULE (OPTION)	<u>Rule for selecting records to be modified</u> <u>The rule can be applied to several variables in the dataset.</u> <u>It is written in the form of a Boolean formula using the following operators:</u> ▪ <u>Boolean operators: OR, AND, NOT</u> ▪ <u>Mathematical operators: = < > ≥ ≈</u> ▪ <u>Set operators: IN (xx,yy,zz)</u> ▪ <u>Basic types: Numbers, Boolean (True, False), String</u> <u>If the parameter is not specified, the subrule will be applied to all records.</u>
SUBSTITUTION_RULE	<u>Value substitution rule</u> <u>For example:</u> - <u>Replace the first 5 characters with XXXXX</u> - <u>Assign the value '*'</u>

Example of use

Only keep the first 4 characters corresponding to the bank in the IBAN.

7 : Example of use of masking

Process ID	TRT-EXAMPLE-2_5
Description	IBAN code processing
Technical Name	Substitution by masking
Parameters	
ID	Name
VARIABLE	IBAN
SELECTION_RULE	
SUBSTITUTION_RULE	Keep only the first 4 characters

2.3 Generalisation

Generalisation is a de-identification technique that involves replacing specific values with more general or less specific values (e.g., converting continuous data to intervals, grouping detailed values into categories). This method is recognized by ISO/IEC 20889:2018[1] as a way of reducing data precision to lower the risk of re-identification. The objectives of generalisation are:

- To reduce data granularity in order to protect privacy.
- Maintain a balance between confidentiality and utility by preserving the statistical structure (distribution, proportions).

Use cases for generalisation include:

- Reducing granularity for confidentiality, which involves transforming precise values such as age, income, and location into categories or intervals to limit the risk of re-identification.
- Preparation of data for statistical analysis in order to create homogeneous classes (e.g., age groups, geographic areas) to facilitate calculations and visualisations.
- Reporting: Publishing aggregated data (e.g., by region or bracket).
- Business segmentation, which involves grouping professions, socioeconomic categories, or areas to simplify analysis and decision-making.

2.3.1 Absolute rounding

Rounds a given number to the nearest inferior specified multiple.

Context of use

All personal or protected numerical variables.

Parameters

Name of the technique	Absolute rounding
Parameters	
ID	Description
VARIABLE	<u>Identifier of the variable to be transformed.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
ROUNDING_INCREMENT	<u>Increment to be considered in the following formula describing the result of the operation:</u> $Z = \text{round}(x, m) = \text{round}(x/m) \cdot m$ Where: - m is the value of the ROUNDING_INCREMENT e.g. 10, 100... - x : the initial value - Z : the resulting value

Exemple

Following technics allows to round AGE to multiple of 5:

Nom de la technique	Rounding absolu
---------------------	-----------------

Paramètres	
ID	Description
VARIABLE	<u>AGE</u>
TARGET_VARIABLE	<u>AGE_ARRONDI</u>
ROUDING_INCREMENT	5

Transformation of values:

Initial value	Transformed value
7	5
13	10
46	45
27	25
72	70

2.3.2 Relative rounding

A generalisation technique also consists of taking into account rounding based on one hundredth or one thousandth of the maximum and minimum values.

Context of use

All personal or protected numerical variables.

Parameters

Name of the technique	Relative rounding
Parameters	
ID	Description
VARIABLE	<u>Identifier of the variable to be transformed.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
ROUNDING_UNIT	<u>Will be expressed as a power of 10. For example, rounding to the nearest hundredth will be expressed by the number 2.</u> $Z = \text{round}(x,p) = \text{round} (10^p * (x - V_{\min}) / (V_{\max} - V_{\min}))$ <u>Where:</u> <u>P: the percentile</u> <u>Z: the transformed value</u>

	• <u>x: the initial value</u> • <u>Vmin = the minimum value of all values of the variable</u> • <u>Vmax = the maximum value of all values of the variable</u>
--	---

Example

This example allows to round the consumption to percentage of the interval between max and min value of the consumption:

Nom de la technique	Relative rounding
Paramètres	
ID	Description
VARIABLE	<u>CONSOMMATION</u>
TARGET_VARIABLE	<u>CONSOMMATION_ARRONDIE</u>
ROUDING_UNIT	2

Transformation of values:

Initial value	Transformed value
20	3
5	0
7	0
24	3
27	4
320	63
300	60
505	100

2.3.3 Generalisation of Luxembourg zip codes

Luxembourg postal codes have a fine granularity, sometimes corresponding to a single street. They are therefore generally considered to be quasi-identifiers. It is therefore advisable to generalise them:

- To the first digit for a larger area division
- To the first two digits of the postal code for fine granularity.

Context of use

Any Luxembourg postal code variable.

Parameters

Name of the technique	Postal code generalisation
Parameters	
ID	Description
VARIABLE	<u>Identifier of the variable to be transformed.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
ROUNDING_UNIT	<u>1 or 2 depending on the desired granularity.</u>

2.3.4 Generalisation of dates/times (datetime)

Date/time generalisation will be performed according to several levels of generalisation: years, months, days, hours, minutes, or seconds.

Context of use

Any date variable.

Parameters

Name of the technique	Date generalisation
Parameters	
ID	Description
VARIABLE	<u>Identifier of the variable to be transformed.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
ROUNDING_UNIT	<u>YEAR or MONTH or DAY or HOUR or MIN or SEC depending on the desired granularity.</u>

2.3.5 Period generalisation

A generalisation that allows a period in the form of two dates, "From" and "To," to be transformed into a number of years, months, etc.

Context of use

Any pair of dates or times that define a period.

Parameters

Name of the technique	Period generalisation
Parameters	
ID	Description
FROM_VARIABLE	<u>Variable corresponding to the start of the period.</u>
TO_VARIABLE	<u>Variable corresponding to the end of the period.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
ROUNDING_UNIT	<u>YEAR or MONTH or DAY or HOUR or MIN or SEC depending on the desired granularity.</u>

2.3.6 Generalisation by mapping table

Generalisation by mapping table is one of the generalisation techniques classified in the ISO/IEC 20889:2018 standard[1]. A value generalisation is a method in which a table defines the more general value for each value of the variable to be generalised. Generalisation by mapping table involves using a predefined table that associates each possible value with an aggregated value (e.g., correspondence of postal codes to municipalities, age groups to categories).

Context of use

Generalisation by mapping table is used when specific values need to be replaced by predefined categories in order to reduce data granularity and limit the risk of re-identification. This technique is particularly suitable for:

- Sensitive categorical or numerical variables (e.g., postal codes, ages, occupations).
- Creating consistent classes according to business or regulatory rules (e.g., grouping postal codes into geographic areas, ages into ranges).
- Maintaining analytical consistency: mapping tables ensure uniform and reproducible transformation.

Potential use cases for generalisation using mapping tables include:

- Segmentation (e.g., grouping occupations into socio-professional categories).
- Statistical analysis (e.g., transforming ages into brackets for histograms).

- Regulatory reporting (e.g., grouping areas to anonymise geographic data).

Parameters

Name of technique	Generalisation by mapping table
Parameters	
ID	Description
VARIABLE	<u>Variable that must be generalised.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
File	<u>A CSV file used to map the value to the generalisation to be obtained. The CSV file will be attached to the request.</u> <u>The CSV file can have several columns. The first column will always correspond to the value to be generalised.</u> <u>The second column will contain the generalised value with the finest granularity. The following columns may define a higher level.</u>

2.3.7 Data Bucketing

Data Bucketing, also known as "data binning," aggregates numerical data into intervals.

Context of use

All personal or protected numerical variables.

Data **bucketing** will be used to:

- Reduce the complexity of continuous numerical data,
- Create categories or intervals (called "bins" or "classes"),
- Facilitate analysis or visualisation (e.g., in histograms).

Parameters

Name of the technique	Data bucketing
Parameters	
ID	Description
VARIABLE	<u>Variable whose values must be transformed.</u>
TARGET_VARIABLE	<u>Name of the variable that will contain the transformed values.</u>
METHOD	<u>FIXED_EDGES: defines the set of values framing the bins in the form of a table.</u> <u>FIXED_SIZE: Defines the size of the bins to be created.</u>

	<u>FIXED_NUMBER: The number of bins is fixed, and all bins will be the same size. The requester must specify this number.</u>
--	---

2.4 Micro-aggregation

Micro-aggregation is a technique for de-identifying numerical data. It consists of grouping records into homogeneous sets (strata) of a minimum size k , then replacing individual values within each group with a representative value, such as the mean, median, centroid, or interval. Thus, all records within a group share the same aggregated values. The aim is to reduce the risk of re-identification while preserving the statistical properties of the dataset[1] [2].

Context of use

Using micro-aggregation on individual numerical data (e.g., income, age) makes it possible to:

- Largely preserving key statistical properties: Raw values are replaced with the mean of small, homogeneous groups, ensuring the global mean. However, intra-group variability is reduced (a smoothing effect), which may slightly alter variance or other distributional properties.
- Enable the use of data for statistical analysis and predictive modelling, as essential information remains intact.
- Ensure a good balance between confidentiality (by reducing the risk of re-identification) and data accuracy.

It should be noted that the difference between micro-aggregation and macro-aggregation lies mainly in the level of aggregation and the purpose. Micro-aggregation involves grouping individual records into small homogeneous groups of a minimum size (k) and replacing the original values with a representative value, usually the mean or median of the group. This technique aims to reduce the risk of re-identification while preserving the fine statistical structure of individual data, as defined in ISO 20889[1]. Conversely, macro-aggregation operates at a more global level: it groups data according to grouping variables (e.g., locality or sector) and calculates aggregates for other variables, thus producing macro-data. This approach removes individual granularity. In summary, micro-aggregation partially protects confidentiality while maintaining a certain level of analytical granularity, whereas macro-aggregation consolidates data into aggregated information.

Parameters

Name of technique	Micro-aggregation	
Parameters		
ID	Description	
K	<u>Sets the minimum group size.</u>	
AGGREGATED_VARIABLES	<u>List all the variables that must be kept in the dataset. For each variable to be kept, specify the aggregation method to be used.</u>	

	For a variable to be aggregated, several OUTPUTs can be specified. For example, for a numeric variable, you can define an output with the average value, an output with the maximum value, etc. The set is specified in the form of a table:		
	INPUT	OUTPUT	Method
	Identifier of the variable to be aggregated	Identifier of the aggregated variable.	Method from the following list: SUM MEAN MEDIAN
DISTANCE_DEFINITION	For each non-numeric variable, a definition of distances can be made. This definition will take the form of a distance matrix. This distance matrix will be attached in CSV format to the Appropriate Measurement Plan.		
	VARIABLE	Matrix	
	Non-numeric variable identifier	Name of the file containing the distance matrix between the variable values.	

2.5 Formal anonymisation and privacy models

Anonymisation consists of applying de-identification techniques to reduce the risk of re-identifying individuals in a dataset. A formal privacy measurement model applied to anonymisation enables assessing this risk and, in some cases, providing mathematical guarantees. Such a model considers the use-case context and are based on parameters derived from empirical analysis, including risks associated with re-identification through isolation, linkage, or inference. Three models are well known in scientific literature and ISO 20889:2018: K-anonymity, L-diversity, and T-closeness.

2.5.1 K-anonymity model

K-anonymity consists of grouping information in such a way that each identifier (person) cannot be distinguished from others in a group of at least K-1 individuals [3] [4]. This reduces the risk of identifying a specific person (the probability is limited to 1/K). However, this model is insufficient to prevent certain deductions or inference attempts from the remaining data.

Context of use

K-anonymity is particularly applicable in cases where:

- The data contains quasi-identifiers (e.g., age, zip code, gender) that, when combined, can identify an individual.

- The goal is to preserve the data structure for simple statistical analyses (e.g., averages, frequencies) without disclosing individual information.
- Organisations want a simple, interpretable mechanism to ensure that no individual is identified in the dataset: each record must be indistinguishable from at least K individuals

K-anonymity is suitable for tabular datasets where variables can be grouped or generalised to form equivalence classes, while maintaining minimal utility for descriptive analyses.

In practice, to apply the K-anonymity method, it is essential to:

1. Identify quasi-identifiers: Quasi-identifiers are variables that, when combined, can lead to the re-identification of individuals.
2. Define the K threshold: Specify the minimum value of K (size of equivalence classes) at which anonymisation is considered satisfactory.
3. Specify the transformation for each variable: each quasi-identifier must be processed using an appropriate technique, such as:
 - a. Generalisation (e.g., grouping ages into ranges)
 - b. Masking (e.g., replacing part of the postal code with asterisks)
 - c. Deletion (e.g., removing gender if not necessary)

Techniques such as generalisation (e.g. grouping ages into ranges) or masking (e.g. replacing part of a postal code with asterisks) are applied iteratively to increase the granularity of values when the k-anonymity condition is not met.

The main challenge of this technique is identifying the optimal K parameter that best suits the analysis requirements. A very high K value can significantly alter the data, reducing its usefulness for data analysis in the SPE. In this context, it is recommended to use a K value as recommended in the literature, and to supplement the appropriate measures plan with measures of the utility of the datasets after transformation to ensure the data remain useful.

Parameters

Name of technique	K-anonymity
Parameters	
ID	Description
VARIABLE_LIST_IDENT	List of variables constituting identifiers
VARIABLE_LIST_QUASI_IDENT	List of variables constituting quasi-identifiers
THRESHOLD_K	The threshold corresponding to k in the k-anonymity model below which the solution is not acceptable. In best practices, $K \geq 5$ for basic protection. $K =$ from 10 to 20 for better protection in large data sets.
Transformation 1	

VARIABLE QUASI_IDENT	<u>Variable (quasi-identifier) on which the technique can be used.</u>
TRANSFORMATION	<u>Technique to be used.</u>
Transformation 2	
VARIABLE QUASI_IDENT	<u>Variable (quasi-identifier) on which the technique can be used.</u>
TRANSFORMATION	<u>Technique to be used.</u>
....	

Example

Table 8 shows an example of a dataset containing direct identifiers (e.g., PatientID, Name, Email), quasi-identifiers (e.g., Age, Postal Code, Gender), and a sensitive variable (Disease).

Table 8 Example of source data to be protected

PatientID	Name	Email	Age	Zip	Gender	Illness
P001	<u>A.M</u>	<u>a.m@example.com</u>	<u>26</u>	<u>12345</u>	<u>F</u>	<u>Flu</u>
P002	<u>B.L</u>	<u>b.l@example.com</u>	<u>27</u>	<u>12346</u>	<u>M</u>	<u>Cancer</u>
P003	<u>C.B</u>	<u>c.b@example.com</u>	<u>33</u>	<u>12347</u>	<u>F</u>	<u>Diabetes</u>
P004	<u>D.P</u>	<u>d.p@example.com</u>	<u>35</u>	<u>12345</u>	<u>M</u>	<u>Asthma</u>
P005	<u>E.D</u>	<u>e.d@example.com</u>	<u>38</u>	<u>12346</u>	<u>F</u>	<u>Flu</u>
P006	<u>F.R</u>	<u>f.r@example.com</u>	<u>39</u>	<u>12347</u>	<u>M</u>	<u>Diabetes</u>
P007	<u>G.M</u>	<u>g.m@example.com</u>	<u>42</u>	<u>45610</u>	<u>F</u>	<u>Flu</u>
P008	<u>H.L</u>	<u>h.l@example.com</u>	<u>44</u>	<u>45611</u>	<u>M</u>	<u>Asthma</u>
P009	<u>I.L</u>	<u>i.l@example.com</u>	<u>45</u>	<u>45612</u>	<u>F</u>	<u>Cancer</u>
P010	<u>J.M</u>	<u>j.m@example.com</u>	<u>48</u>	<u>45610</u>	<u>M</u>	<u>Diabetes</u>
P011	<u>K.H</u>	<u>k.h@example.com</u>	<u>51</u>	<u>45611</u>	<u>M</u>	<u>Flu</u>
P012	<u>L.F</u>	<u>l.f@example.com</u>	<u>58</u>	<u>45612</u>	<u>F</u>	<u>Cancer</u>
P013	<u>M.G</u>	<u>m.g@example.com</u>	<u>61</u>	<u>78920</u>	<u>F</u>	<u>Flu</u>
P014	<u>N.R</u>	<u>n.r@example.com</u>	<u>63</u>	<u>78921</u>	<u>M</u>	<u>Cancer</u>
P015	<u>O.C</u>	<u>o.c@example.com</u>	<u>66</u>	<u>78922</u>	<u>F</u>	<u>Asthma</u>
P016	<u>P</u>	<u>p.s@example.com</u>	<u>68</u>	<u>78920</u>	<u>M</u>	<u>Diabetes</u>

P017	<u>Q R</u>	<u>g.r@example.com</u>	<u>70</u>	<u>78921</u>	<u>M</u>	<u>Flu</u>
P018	<u>S C</u>	<u>s.c@example.com</u>	<u>72</u>	<u>78922</u>	<u>F</u>	<u>Asthma</u>

Table 9 classifies the columns of Table 8 into direct identifiers, quasi-identifiers, and sensitive variables useful for K-anonymity.

Table 9 Example of direct identifier, quasi-identifier, sensitive variable.

Type	Columns	Definition	Processing
Direct identifier	<u>PatientID,</u> <u>Name,</u> <u>Email</u>	<u>Allows a person to be identified on their own (e.g., name, email).</u>	<u>Deletion</u> <u>Direct identifiers are deleted by default.</u>
Quasi-identifier	<u>Age, Postal code,</u> <u>Gender</u>	<u>Can re-identify in combination (e.g., age + postal code).</u>	<u>Generalisation/masking/deletion</u>
Sensitive variable	<u>Disease</u>	<u>Information whose disclosure creates a risk (health).</u>	<u>Retained for analysis. Sensitive variables are not considered by K-anonymity but are controlled via L-diversity/T-closeness that are further defined in Sections 3.5.2 and 3.5.3, respectively.</u>

Based on the example at Table 8, an instantiation of the parameters to be provided to specify the k-anonymity operation is shown in 10:

10 : Example of k-anonymity usage.

Name of the technique	K-anonymity
Parameters	
ID	Description
VARIABLE_LIST_IDENT	PatientID, Last Name and First Name, Email.
VARIABLE_LIST_QUASI_IDENTIFIER	Age, Postal code, Gender
THRESHOLD_K	5
Transformation 1	
VARIABLE	Zip code
TRANSFORMATION	Generalisation of postal codes
ROUNDING_UNIT	2 then 1
Transformation 2	

VARIABLE	AGE
TRANSFORMATION	Grouping into intervals
METHOD	FIXED_SIZE = 10
Transformation 3	
VARIABLE	Gender
TRANSFORMATION	MASKING
SUBSTITUTION_RULE	Replace with an '*'

Table 11 shows the transformation of quasi-identifiers (age groups, masking of postal codes, removal of gender) to create equivalence classes (groups) containing at least K individuals (here $K=5$). This approach reduces the risk of re-identification by "dropping" each individual into a sufficiently large group.

Table 11 Example of the result of applying K-anonymity.

Age ranges	Postal code_Masked	Gender	Disease
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Flu</u>
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Cancer</u>
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Diabetes</u>
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Asthma</u>
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Flu</u>
<u>25–39</u>	<u>123**</u>	<u>*</u>	<u>Diabetes</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Flu</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Asthma</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Cancer</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Diabetes</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Flu</u>
<u>40–59</u>	<u>456**</u>	<u>*</u>	<u>Cancer</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Flu</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Cancer</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Asthma</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Diabetes</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Flu</u>
<u>60–74</u>	<u>789**</u>	<u>*</u>	<u>Asthma</u>

2.5.2 L-Diversity model

L-diversity is an improvement on the K-anonymity method, designed for datasets with low variability in sensitive variables [5] [6]. It aims to prevent overly easy inferences by ensuring that each group (equivalence class) contains at least L different values for the sensitive variable (e.g., disease). In practice, this means that even if two people share the same quasi-identifiers (age, zip code), inferring their sensitive information remains difficult because multiple plausible sensitive values exist within the group.

Context of use

L-diversity is used to strengthen the protection offered by K-anonymity in tabular datasets. It is used when:

- The data contains sensitive variables (e.g., disease, income, opinion), and we want to avoid inference attacks.
- Each equivalence group formed by K-anonymity must have sufficient diversity in sensitive values so that an attacker cannot deduce the exact information about an individual, even if they know the individual's group.
- The goal is to limit the risks associated with homogeneous distributions (e.g., a group where all individuals have the same disease), which make K-anonymity insufficient.

L-diversity is suitable for tabular datasets with sensitive variables, ensuring that within each anonymised group, there are at least L distinct values for these variables, thereby reducing the risk of inference.

To implement L-diversity, it is recommended to:

- Identify quasi-identifiers: Quasi-identifiers are variables that, when combined, can lead to re-identification (e.g., age, zip code, gender).
- Identify sensitive variables: Determine which columns contain sensitive information (e.g., illness) and measure their variability within each class.
- Set the K-anonymity threshold of K before L (optional): In practice, L-diversity is typically applied in combination with K-anonymity. The dataset is first transformed to satisfy the K-anonymity criterion, ensuring that each class contains at least K records, individuals, and then L-diversity is reinforced by requiring that each of these classes contains at least L distinct values to comply with the minimum diversity L for the sensitive variable.
- Define the L threshold (optional): Specify the minimum value of L (number of distinct values of the sensitive variable per class) at which anonymisation is considered satisfactory. Applying L-diversity alone increases diversity among sensitive variables but does not guarantee the minimum class size.
- Specify quasi-identifier transformations (optional): To achieve both the K and L thresholds, each quasi-identifier can be processed using an appropriate technique:

- Generalisation: (e.g., grouping ages into ranges)
- Masking: (e.g., replacing part of the postal code with asterisks)
- Deletion: (e.g., removing gender if not necessary)

Parameters

Name of technique	L-diversity
Parameters	
ID	Description
VARIABLE_LIST_IDENT	List of variables constituting identifiers
VARIABLE_LIST_QUASI_IDENT	List of variables constituting the quasi-identifier
VARIABLE_LIST_SENSITIVE	List of sensitive variables to be considered
THRESHOLD_K	The threshold corresponding to k in the k-anonymity model below which the solution is not acceptable. In best practices, $K \geq 5$ for basic protection. K = from 10 to 20 for better protection in large data sets.
THRESHOLD_L	The threshold corresponding to L in the L-diversity model below which the solution is not acceptable. $L \geq 2$ or 3 for minimum diversity. L = 5 or more for sensitive variables.
Transformation 1	
VARIABLE QUASI_IDENT	Variable (quasi-identifier) on which the technique can be used.
TRANSFORMATION	Technique to be used.
Transformation 2	
VARIABLE	Variable (quasi-identifier) on which the technique can be used.
TRANSFORMATION	Technique to be used.

Example

In the context of the sensitive dataset in the previous section, the L-diversity technique can be used as follows:

Table12 : Example of L-diversity operation

Name of technique	L-diversity
Parameters	
ID	Description
VARIABLE_LIST_IDENT	PatientID, Last Name and First Name, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Postal code, Gender

SENSITIVE_VARIABLE_LIST	<u>Illness</u>
THRESHOLD_K	<u>5</u>
THRESHOLD_L	<u>3</u>
Transformation 1	
VARIABLE	<u>Postal code</u>
TRANSFORMATION	<u>Masking</u>
Transformation 2	
VARIABLE	<u>AGE</u>
TRANSFORMATION	<u>Group ages into ranges</u>
Transformation 3	
VARIABLE	<u>Gender</u>
TRANSFORM	<u>Deletion. Remove gender if not necessary.</u>

Table13 shows the transformation of quasi-identifiers based on the sensitive variable. To do this, each group must contain several different values for the sensitive variable, making any deterministic inference impossible. L-diversity counts the number of distinct categories of the sensitive variable in each class. In our case, all classes have $L = 4$ distinct values for the "Disease" variable, while the threshold is set to $L = 3$, ensuring compliance. The difference between the exampleTable 11 and Table13 lies in the age groups. The age groups have been adjusted to comply with the $L = 3$ threshold.

Table13 : Example of L-diversity with $L=3$.

Age groups	Postal code_Hidden	Gender	Illness
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Flu</u>
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Cancer</u>
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Diabetes</u>
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Asthma</u>
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Flu</u>
<u>20-59</u>	<u>123**</u>	<u>*</u>	<u>Diabetes</u>
<u>20-59</u>	<u>456**</u>	<u>*</u>	<u>Flu</u>
<u>20-59</u>	<u>456**</u>	<u>*</u>	<u>Asthma</u>
<u>20-59</u>	<u>456**</u>	<u>*</u>	<u>Cancer</u>
<u>20-59</u>	<u>456**</u>	<u>*</u>	<u>Diabetes</u>

20–59	456**	*	Flu
20–59	456**	*	Cancer
60–79	789**	*	Flu
60–79	789**	*	Cancer
60–79	789**	*	Asthma
60–79	789**	*	Diabetes
60–79	789**	*	Flu
60–79	789**	*	Asthma

2.5.3 T-Closeness model

T-closeness is an advanced technique that improves L-diversity for datasets where sensitive variables are weakly diversified or unevenly distributed. Its objective is to limit the risks of statistical inference: each group must have a distribution of sensitive values close to the one observed in the overall dataset [7]. This proximity is measured by a distance and must not exceed a threshold T. In practice, this ensures that the anonymised data remains representative of the original dataset, while protecting privacy. However, this method may reduce data utility by removing certain correlations between quasi-identifiers and sensitive variables.

Context of use

T-closeness is used to reinforce the protection offered by K-anonymity and L-diversity in tabular datasets. It is used when:

- The dataset contains sensitive variables (e.g., disease, income), and there is a need to protect against inference attacks based on the distribution of sensitive values.
- Each equivalence group formed by K-anonymity must have a distribution of sensitive variables that is close to the overall dataset distribution, so that an attacker cannot deduce information with high probability.
- There is a risk of homogeneous or biased equivalence classes, which make K-anonymity and L-diversity insufficient.

T-closeness applies to tabular datasets containing sensitive variables. It ensures that the distribution of sensitive values in each group (formed by K-anonymity) remains close to the overall distribution, within a maximum distance threshold T, which limits the risks of statistical inference.

To implement T-closeness, it is recommended to:

- Identify quasi-identifiers: Quasi-identifiers are variables that, when combined, can lead to re-identification (e.g., age, zip code, gender).

- Identify sensitive variables: Determine which columns contain sensitive information (e.g., illness) and calculate their overall distribution.
- Set the K-anonymity threshold K before T (optional): In practice, T-closeness is often applied in combination with K-anonymity. The dataset is first transformed to ensure that each class contains at least K individuals (K-anonymity criterion), and then the distribution of sensitive variables within each class is required to be close to the overall distribution.
- Set the K-anonymity threshold K and L-diversity threshold L before T (optional): In practice, T-closeness is often applied in addition to K-anonymity and L-diversity. We start by ensuring that each class contains at least K individuals (K-anonymity criterion), then that these classes respect the minimum diversity L for the sensitive variable, before imposing distribution closeness (T-closeness criterion).
- Define the T threshold: Specify the maximum distance between the distribution of sensitive variables in a class and the overall distribution. The lower the T is, the stronger the protection against inference.
- Specify the transformations of quasi-identifiers (optional): To achieve both the K and T thresholds, each quasi-identifier can be processed using an appropriate technique:
 - Generalisation: (e.g., grouping ages into ranges)
 - Masking: (e.g., replacing part of the postal code with asterisks)
 - Deletion: (e.g., removing gender if not necessary)

Parameters

Name of technique	T-closeness
Parameters	
ID	Description
VARIABLE_LIST_IDENT	List of variables constituting identifiers
VARIABLE_LIST_QUASI_IDENTIFIER	List of variables constituting the quasi-identifier
VARIABLE_LIST_SENSITIVE	List of sensitive variables to be considered
THRESHOLD_K	The threshold corresponding to k in the k-anonymity model below which the solution is not acceptable. In best practices, $K \geq 5$ for basic protection. K = from 10 to 20 for better protection in large data sets.
THRESHOLD_L	The threshold corresponding to L in the L-diversity model below which the solution is not acceptable. $L \geq 2$ or 3 for minimum diversity. L = 5 or more for sensitive variables.
THRESHOLD_T	The threshold corresponding to T in the T-closeness model above which the solution is not acceptable. t ≤ 0.2 (20%) for moderate confidentiality. t ≤ 0.05 – 0.1 for high confidentiality in sensitive datasets.

Transformation 1	
VARIABLE	<u>Variable (quasi-identifier) on which the technique can be used.</u>
TRANSFORMATION	<u>Technique to be used.</u>
Transformation 2	
VARIABLE	<u>Variable (quasi-identifier) on which the technique can be used.</u>
TRANSFORMATION	<u>Technique to be used.</u>
....	

Example

Table 14 : Example of using T-closeness

Name of technique	T-closeness
Parameters	
ID	Description
VARIABLE_LIST_IDENT	<u>PatientID, Last Name and First Name, Email.</u>
VARIABLE_LIST_QUASI_IDENTIFIER	<u>Age, Postal code, Gender</u>
THRESHOLD_K	<u>5</u>
THRESHOLD_T	<u>0.05</u>
Transformation 1	
VARIABLE	<u>Postal code</u>
TRANSFORM	<u>Masking</u>
Transformation 2	
VARIABLE	<u>AGE</u>
TRANSFORMATION	<u>Group ages into ranges</u>
Transformation 3	
VARIABLE	<u>Gender</u>
TRANSFORM	<u>Delete. Remove gender if not necessary.</u>

Table 15 illustrates the transformation of quasi-identifiers based on the sensitive variable. To do this, each group must have a distribution of sensitive values that closely matches the overall distribution,

thereby greatly reducing the risk of inference. T-closeness measures the distance between these distributions and requires it to be less than the fixed T threshold. In our case, all classes meet $T \leq 0.05$, which ensures compliance.

Table 15 : Example of T-closeness with $T=0.05$.

Age groups	Zip code_Hidden	Gender	Illness
20–59	123**	*	Flu
20–59	123**	*	Cancer
20–59	123**	*	Diabetes
20–59	123**	*	Asthma
20–59	123**	*	Flu
20–59	123**	*	Diabetes
20–59	456**	*	Flu
20–59	456**	*	Asthma
20–59	456**	*	Cancer
20–59	456**	*	Diabetes
20–59	456**	*	Flu
20–59	456**	*	Cancer
60–79	789**	*	Flu
60–79	789**	*	Cancer
60–79	789**	*	Asthma
60–79	789**	*	Diabetes
60–79	789**	*	Flu
60–79	789**	*	Asthma

2.5.4 β -likeness

β -likeness, introduced in [8], is a method that is not part of ISO 20889:2018 but aims to protect the sensitivity of data published after anonymisation (e.g., medical data or survey data). Its goal is to prevent an attacker from becoming too confident about knowing sensitive information about a person after seeing the anonymised data. This is to ensure the robustness of confidentiality; even if the attacker knows the a priori distribution of sensitive variables, they cannot significantly improve their

prediction for each value. If, prior to publication, the probability that a person has a certain characteristic (e.g., a disease) was low, publication should not increase that probability excessively. The β parameter sets this limit. Thus, even if someone analyses the data, they will not be much more certain than before.

β -likeness is an alternative to t-closeness; it aims to reduce inference risk by maintaining distributional similarity.

Context of use

The β -likeness technique can be used in the following cases:

- Publication of pseudonymised data without revealing sensitive information.
- Application of methods by *generalising* or *altering* sensitive values to limit the risk of inference in order to achieve a defined level of β -likeness while preserving data quality.
- Strengthen control over disclosed sensitive microdata (stricter than t-closeness or l-diversity).
- Sensitive values are unbalanced, rare, or highly informative.
- Situations where data protection must be granular, preventing the exposure of a sensitive value from becoming too likely for each individual value, even in subgroups.

To implement β -likeness, it is recommended to:

- Define what you want to protect: Decide what analyses or queries will be performed on the data (e.g., aggregate statistics or predictive modelling).
- The confidentiality level (β) will be set in the same way as the previous thresholds in the other metrics, after empirical testing. The values of β are in the range [0.1, 1]. Small values sacrifice utility but offer strong protection, while larger values have the opposite effect.

Parameters

Name of the technique	β -likeness
Parameters	
ID	Description
VARIABLE_LIST_IDENT	List of variables constituting identifiers
VARIABLE_LIST_QUASI_IDE NT	List of variables constituting the quasi-identifier
VARIABLE_LIST_SENSITIVE	List of sensitive variables to be considered
THRESHOLD_K (Option)	The threshold corresponding to k in the k-anonymity model below which the solution is not acceptable. In best practices, $K \geq 5$ for basic protection. $K =$ from 10 to 20 for better protection in large data sets. K is optional but recommended, as it guarantees k-anonymity for quasi-identifiers in addition to β-likeness.

β	The threshold corresponding to β in β-likeness. Small values ([0.1-0.2]) sacrifice utility but offer strong protection, while larger values have the opposite effect. The implicit rule for the β threshold is that for each sensitive value, its probability in a group must not increase by more than the β threshold relative to its overall probability.
---------	--

Example

Table 16: Example of using β -likeness

Name of technique	β -likeness
Parameters	
ID	Description
VARIABLE_LIST_IDENT	<u>PatientID, Last Name and First Name, Email.</u>
VARIABLE_LIST_QUASI_IDENT	<u>Age, Postal code, Gender</u>
SENSITIVE_VARIABLE_LIST	<u>Illness</u>
THRESHOLD_K (Option)	<u>5</u>
β	<u>0.15</u>
Transformation 1	
VARIABLE	<u>Postal code</u>
TRANSFORM	<u>Masking</u>
Transformation 2	
VARIABLE	<u>AGE</u>
TRANSFORMATION	<u>Group ages into ranges</u>
Transformation 3	
VARIABLE	<u>Gender</u>
TRANSFORM	<u>Delete. Remove gender if not necessary.</u>

2.5.5 Differential privacy

Differential privacy is a mathematical framework that provides formal guarantees for protecting individuals in a dataset. Unlike traditional anonymisation approaches, it does not merely aim to

prevent re-identification; rather, it seeks to reduce the influence of an individual's data on the outcome of analyses. This is achieved by adding calibrated noise to query responses so that the presence or absence of an individual has minimal effect on the results, controlled by a privacy parameter ϵ . Lower values of ϵ offer stronger privacy guarantees but reduce the accuracy of the target data, while higher values improve accuracy of target data but decrease privacy. This approach provides robust protection against many attacks, including those that leverage external information. However, striking an appropriate balance between privacy and utility requires advanced techniques and can be computationally expensive [9].

Context of use

Differential privacy is particularly useful when queries involve quasi-identifiers or sensitive variables (e.g., age, income, health status) and one wishes to publish statistics or aggregate values without revealing individual information.

Differential privacy is particularly applicable in situations where:

- Aggregated statistics must be published (e.g., censuses, public health indicators) without revealing individual information.
- Data is used for machine learning analysis or research purposes while complying with regulations such as the GDPR.
- Inter-client collaborations (e.g., healthcare, insurance) require cross-analysis without disclosing raw data.
- The dataset primarily consists of numerical data (continuous or discrete), which is the most common data type for the differential privacy method. For categorical data, variable encoding is necessary (one-hot², ordinal encoding³).

To implement differential privacy, it is recommended to:

- Define the protection scope: Decide what analyses or queries will be performed on the data (e.g., aggregate statistics or predictive models).
- Set the privacy parameter (ϵ):
 - Small ϵ (e.g., 0.1) = strong privacy guarantees but less accurate results due to higher noise.

² https://fr.wikipedia.org/wiki/Encodage_one-hot

³ <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html>

- Large ϵ (e.g., 10) = higher accuracy but weaker privacy guarantees.

Parameters

Name of the technique	Differential privacy
Parameters	
ID	Description
VARIABLE_LIST_IDENT	List of variables constituting identifiers
VARIABLE_LIST_QUASI_IDE NT	List of variables constituting the quasi-identifier
VARIABLE_LIST_SENSITIVE	List of sensitive variables to be considered
ϵ	The threshold corresponding to ϵ in differential privacy above which the solution is not acceptable. • [0.1 – 0.5]: high confidentiality, greater distortion. • [1 – 3]: common practice • > 5: low privacy but accurate dataset.

Example

Table 17: Example of the use of differential privacy

Name of technique	Differential privacy
Parameters	
ID	Description
VARIABLE_LIST_IDENT	PatientID, Last Name and First Name, Email.
VARIABLE_LIST_QUASI_IDE NT	Age, Postal code, Occupation
SENSITIVE_VARIABLE_LIST	Medical diagnosis, Annual income
ϵ	1

2.5.6 Synthetic data

Synthetic data refers to artificially generated data that aims to preserve the statistical properties and relationships of the original dataset without containing any actual individual data.

Avatarisation

The principle of the avatarisation method is similar to that of complete anonymisation techniques, but it is based on the creation of synthetic records from real data. Avatarisation uses proximity-based

methods such as k-nearest neighbours (a machine learning method) to preserve the data structure while diluting individual information. This process generates avatars for each record: artificial individuals that reproduce the statistical characteristics of the original data without containing any personal information.

Context of use

Avatarised data is particularly useful in situations where:

- Training AI models without the risk of re-identification.
- Aggregate statistics must be published (e.g., censuses, public health indicators) without revealing individual information.

Data is used for machine learning (AI models) or research while complying with regulations such as the GDPR.

In the context of generating synthetic data through avatarisation, certain approaches rely on mechanisms that adjust the level of information masking while applying targeted transformations to the most exposed areas of data. It is therefore important for the requester to define clear utility criteria for the data: the hidden rate and local cloaking.

The objective is to ensure robust confidentiality without compromising the overall quality of the data, maintaining its relevance for analytical and predictive purposes.

Parameters

Name of the technique	Avatarisation	
Parameters		
ID	Description	
K	<u>Represents the number of closest neighbours used to ensure confidentiality during the avatarisation process.</u>	
Seed (optional)	<u>A random integer used to initialize the random number generator in the avatarisation algorithm. By setting a value for seed, the same result is obtained for the same data set and the same parameters.</u>	

3. Utility and confidentiality measure of de-identified data

Data de-identification is an essential step in reducing privacy risks. However, these techniques often result in a decrease in the analytical value of the data. To ensure de-identified data remains usable, it is necessary to assess its utility using quality indicators. These indicators enable assessment of the

extent to which the dataset retains the characteristics required to meet the requester's objectives while complying with the confidentiality requirements defined by ISO/IEC 20889:2018.

Quality indicators measure the statistical similarity between the original actual data and the de-identified version. If the de-identified and actual data are statistically similar, they are considered to be of high quality. These indicators are intended as targets to be achieved, as it is not always possible to achieve optimal quality (100%) for all de-identified datasets.

3.1 Rate of deleted or replaced values

For a given variable, it calculates the number of values deleted or replaced by a transformation with respect to the number of original values. This measure can be used for any transformation technique. In the context of a synthetic dataset, this measure shows the proportion of real data in the target dataset.

3.2 Modified record rate

This indicator represents the proportion of records in the original dataset that have been modified after applying a transformation (e.g., deletion, perturbation, aggregation). The modified record rate can be applied to all de-identification methods to quantify the extent of record modifications or variable changes. In the context of a synthetic dataset, this measure allows you to see the proportion of real data in the target dataset.

3.3 Number of values of the variables

The number of values as a percentage of the combinations of distinct values for a set of variables before and after de-identification. Used on quasi-identifiers, this measure is a measure of confidentiality. In anonymisation techniques, this measure is calculated by the average size of the equivalence classes formed by the quasi-identifiers during anonymisation.

3.4 Information loss rate

Information loss rate is a quantitative measure used to assess the extent to which useful information is degraded, distorted, or removed during the de-identification process. It reflects how much of the dataset's original informational value is lost due to transformation techniques. The information loss

rate is expressed as:

$$\text{Information Loss Rate} = 1 - \frac{\text{Information retained after deidentification}}{\text{Total original information}}$$

3.5 Retention rate

Retention rate is a performance metric used in data de-identification and privacy-enhancing workflows to evaluate how much of the original dataset remains usable after de-identification techniques have been applied. The retention rate is expressed as:

$$\text{Retention Rate} = \frac{\text{Number of records retained after deidentification}}{\text{Total original number of records}} \times 100$$

3.6 Distinct value retention rate

The distinct value retention rate is the percentage of a variable's distinct values after de-identification relative to its unique values before de-identification.

3.7 Sensitive Attribute Distortion Rate

Sensitive attribute distortion rate measures how much useful information is lost for a specific sensitive attribute after de-identification. To calculate it, we look at every row in the original dataset, which contains a total of N rows, and check whether a row was removed during processing or whether its sensitive attribute was changed so that it no longer matches the most common value within its group (called an equivalence class). Each row that meets either of these conditions is counted as a 1, while all other rows are counted as 0. The Sensitive attribute distortion rate is then obtained by adding up all these values and dividing by the total number of rows. A higher rate indicates more information was lost because more rows were removed or altered, while a lower rate means the dataset preserved more of its original useful information.

3.8 Discernability Metric

The discernability metric sums the squares of the sizes of the equivalence classes formed during anonymisation and adds a penalty for row suppression when using the K-anonymity, L-diversity, or t-closeness algorithms [10].

3.9 Kolmogorov-Smirnov Distance

The Kolmogorov–Smirnov (KS) test is a nonparametric statistical method that measures the difference between two cumulative distributions: the original data and the de-identified data. KS calculates the maximum distance between the empirical distribution functions of the two samples. The smaller this distance, the more similar the distributions are⁴. This test is primarily applied to continuous variables (e.g., age, income, temperature) because it is based on ordered distributions. However, KS is not

⁴ [Kolmogorov–Smirnov test — Wikipedia](#)

suitable for categorical variables. Therefore, metrics such as χ^2 ⁵, or Jensen-Shannon divergence are preferred.

3.10 Jensen-Shannon distance

The Jensen–Shannon distance (JSD) is an information-theoretic measure of similarity between two probability distributions. It is derived from Kullback–Leibler divergence⁶ but has two advantages: it is symmetric and bounded between 0 (identical distributions) and 1 (completely different distributions). JSD is particularly well-suited to categorical data or continuous variables discretised into classes (e.g., postal codes, product categories, or income grouped into brackets), as it compares distributions in histogram or probability form. For raw continuous data, it is necessary to group them into intervals before calculation⁷.

3.11 Hellinger distance

Hellinger distance⁸ measures the difference between the original distribution and the avatar distribution. This metric returns the average of the Hellinger distances calculated for each variable. Hellinger distance varies between 0 and 1, where 0 represents no difference between the two distributions, and 1 represents a large difference between the two distributions.

⁵ [Chi-square test — Wikipedia](#)

⁶ [Kullback–Leibler divergence - Wikipedia](#)

⁷ [Jensen–Shannon divergence - Wikipedia](#)

⁸ [Hellinger distance — Wikipedia](#)

3.12 Correlation similarity

4. Correlation similarity measures the difference between the pairwise correlation matrices of the original and synthetic datasets across all continuous variables. Conclusion

Table 18 summarises the main characteristics of the minimisation techniques described and indicates the risks that can be mitigated by each of them, in accordance with the recommendations of ISO 20889:2018 and the scientific literature [1] [2] [3].

Table 18 : Properties of data minimisation techniques.

Technique name	Data reliability at the record level	Applicable to value types	Applicable to variable types	Reduces the risk of		
				Isolation	Link	Inference
<u>Minimisation techniques</u>						
Targeting	<u>Yes</u>	<u>All</u>	<u>All</u>	<u>No</u>	<u>No</u>	<u>No</u>
Stratified random sampling	<u>Yes</u>	<u>Categorical (strata), continuous (values within each stratum)</u>	<u>Quasi-identifiers and sensitive</u>	<u>No</u>	<u>No</u>	<u>No</u>

Table 19 summarises the main characteristics of the de-identification techniques described in this document. This table also indicates the risks associated with each technique, in accordance with the recommendations of ISO 20889:2018[5] and the scientific literature [6] [7]. The definitions of the indices in the table are:

- Index 1: Unless K-anonymity is implemented using micro-aggregation.
- Indices 2 and 3 are described based on the literature. 2: Isolation: Protected (inherited from K-Anonymity). Linkage: Improved (because diversity of sensitive values reduces the probability of exact linkage). Inference: Reduced (because there are several possible values for the sensitive variable). 3: Isolation: Protected (inherits from K-Anonymity). Linkage: Protected (because distribution is controlled). Inference: Very well reduced.
- 4: De-identification metrics should be measured to confirm risk reduction. N/A: Not Applicable

Table 19 : Properties of data de-identification techniques.

Name of technique	Data utility at the record level	Applicable to value types	Applicable to Variable types	Reduces the risk of		
				Isolation	Link	Inference
Statistical tools						
Sampling						
Aggregation	<u>N/A</u>	<u>Continuous, discrete</u>	<u>All</u>	<u>Yes</u>	<u>Yes</u>	<u>Yes</u>
Delete	<u>Yes</u>					
Vertical/horizontal deletion	<u>Yes</u>	<u>N/A</u>	<u>N/A</u>	<u>Partially</u>	<u>Partially</u>	<u>Partially</u>
Masking	<u>Yes</u>	<u>Categorical</u>	<u>Direct identifiers</u>	<u>Yes</u>	<u>Partially</u>	<u>No</u>
Generalisation	<u>Yes</u>	<u>All. Everything is subject to interpretation.</u>	<u>Direct identifiers</u>			
Rounding	<u>Yes</u>	<u>Continuous</u>	<u>Direct identifiers</u>	<u>No</u>	<u>Partially</u>	<u>Partially</u>
Generalisation of postal codes/dates/period/correspondence table	<u>Yes</u>	<u>Continuous, ordinal</u>	<u>Direct identifiers, Quasi-identifiers and</u>	<u>No</u>	<u>Partially</u>	<u>Partially</u>
Pseudonymisation	<u>Yes</u>	<u>Categorical</u>	<u>Direct identifiers</u>	<u>No</u>	<u>Partially</u>	<u>No</u>
Randomisation	<u>No</u>		<u>Identification of variables.</u>			
Micro aggregation	<u>No</u>	<u>Continuous</u>	<u>Indirect identifiers and all variables.</u>	<u>No</u>	<u>Partially</u>	<u>Partially</u>
<u>Formal models</u>						

K-anonymity	<u>Yes¹</u>	<u>All</u>	<u>Quasi- identifiers</u>	<u>Yes</u>	<u>Partially</u>	<u>No</u>
L-diversity	<u>Yes²</u>	<u>All</u>	<u>Quasi- identifiers and sensitive</u>	<u>Yes²</u>	<u>Yes²</u>	<u>Yes²</u>
T-closeness	<u>Yes³</u>	<u>All</u>	<u>Quasi- identifiers and sensitive</u>	<u>Yes³</u>	<u>Yes³</u>	<u>Yes³</u>
Differential confidentiality	<u>No</u>	<u>All</u>	<u>Variable identificati on</u>	<u>Yes</u>	<u>Yes</u>	<u>Partially</u>
β -likeness	<u>Yes</u>	<u>All</u>	<u>Quasi- identifiers and sensitive</u>	<u>No</u>	<u>No</u>	<u>Yes</u>
Summary data						
Avatarisation	<u>Yes</u>	<u>All</u>	<u>Identifiers, quasi- identifiers, and sensitive data</u>	<u>Yes⁴</u>	<u>Yes⁴</u>	<u>Yes⁴</u>

Bibliography

- [1] ISO/IEC 20889:2018, "Privacy enhancing data de-identification terminology and classification of techniques," <https://www.iso.org/standard/69373.html>, 2018.
- [2] J. Domingo-Ferrer, "Micro-aggregation," In: LIU, L., ÖZSU, M.T. (eds) *Encyclopaedia of Database Systems*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_1496, 2009.
- [3] S. L. Samarati P., "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalisation and suppression.," *Technical Report SRI-CSL-98-04, SRI International*, 1998.
- [4] S. L., "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* vol. 10 no. 5, pp. pp.557-570, 2002.
- [5] G. J. K. D. V. M. Machanavajjhala A., "L-diversity: Privacy beyond k-anonymity," *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pp. 24-24, 2006.
- [6] K. D. G. J. V. M. Machanavajjhala A., "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data* vol. 1, no. 1, 2007.
- [7] K. S. S. A. Ganta S., "Composition attacks and auxiliary information in data privacy," *arXiv: 0803.0032 [cs.DB]*, March 2008 <https://arxiv.org/pdf/0803.0032.pdf>.
- [8] P. K. Jianneng Cao, "Publishing Microdata with a Robust Privacy Guarantee," *Proc. VLDB Endow.* 5(11), pp. 1388-1399, 2012.
- [9] J., D. D. a. L. N. Near, "Guidelines for Evaluating Differential Privacy Guarantees, Special Publication (NIST SP)," *National Institute of Standards and Technology, Gaithersburg, MD*, [online], <https://doi.org/10.6028/NIST.SP.800-226>, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958909 (Accessed December 8, 2025), 2025.
- [10] R. A. Roberto J. Bayardo Jr., "Data Privacy through Optimal k-Anonymisation.," *ICDE*, pp. 217-228, 2005.