Nomenclature des opérations de désidentification et de mesures d'utilisabilité et de confidentialité des données

Objectif du document

Ce document vise à normaliser les opérations de permettant d'anonymiser, de pseudonymiser et d'agrégation ou tout autre opération de traitements des données qui permettent de s'assurer du contrôle de la divulgation préalablement à l'accès aux données et leur réutilisation dans l'ETS. Ces opérations sont de deux types : des opérations de désidentification et des mesures d'utilisabilité et de confidentialité des données. Cette nomenclature est à utiliser par le demandeur pour spécifier le plan de Mesures Appropriées.

Historique

Version	Date	Modifications effectuées
1.0	13-02-2026	Version initiale

Contents

Liste des définitions

Concept	Définition
Anonymisation	Le processus consistant à rendre anonymes des données à caractère personnel de telle sorte que la personne concernée à laquelle celles-ci se rapportent ne soit pas ou plus identifiée ou identifiable, compte tenu de l'ensemble des moyens raisonnablement susceptibles d'être utilisés pour identifier la personne physique directement ou indirectement.
Variable sensible	Variables révélant des informations intimes ou protégées par la loi (ex. santé, religion, opinions politiques).
Désidentification	Processus visant à réduire ou supprimer la capacité d'identifier une personne dans un jeu de données.
Données personnelles	Toute information se rapportant à une personne physique identifiée ou identifiable (ex. nom, adresse, identifiant en ligne).
Données sensibles	Sous-catégorie des données personnelles qui présente des variables sensibles.
Identifiants personnels directs	Variables qui permettent d'identifier directement une personne (ex. nom, numéro de sécurité sociale, email).
Pseudonymisation	Le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.
Quasi-identifiants	Variables qui, seuls, ne suffisent pas à identifier une personne, mais qui peuvent le faire lorsqu'ils sont combinés (ex. âge, code postal, profession).

Liste des figures

1. Introduction

Ce document présente une analyse exhaustive des techniques de désidentification des données, ainsi que leurs applications pratiques dans le plan de mesures appropriés.

Ce document a pour objectif de normaliser les opérations de désidentification qui sont applicables dans l'Environnement de Traitement Intermédiaire (ETI) lors de leur préparation avant mise à disposition dans l'Environnement de Traitement Sécurisé (ETS). Il permet ainsi au réutilisateurs des données de pouvoir spécifier les opérations de désidentification et les critères d'utilisabilité et de confidentialité de façon précise et sans ambiguïté.

Les techniques de désidentification proposées ici s'appuient sur l'ISO/IEC 20889:2018 [1] ainsi que sur les avancées de la littérature scientifique. Ce document complète et raffine ces techniques selon les implémentations existantes et disponibles dans l'Environnement de Traitement Intermédiaire.

Chaque technique est décrite par leur contexte d'utilisation, des cas d'exemple concret ainsi que qu'un guide des paramètres à configurer pour chacune d'elles.

Les techniques de désidentifications sont également complétées par une normalisation des mesures de confidentialité et d'utilisabilité des données. Ces mesures permettront d'évaluer la qualité et la pertinence des données après une ou plusieurs transformations. L'évaluation de ces critères pourront être partagé avec les demandeurs, après analyse de la confidentialité de ces critères par le Data Privacy Engineer.

Enfin, la Section 4 offre une synthèse globale des techniques et mesures présentées, en soulignant leur complémentarité et leur contribution à un traitement responsable des données.

2. Techniques de désidentification

Les techniques de désidentification visent à réduire la capacité d'identifier ou de réidentifier une personne à partir d'un jeu de données, tout en préservant l'utilité analytique des données. La norme ISO/IEC 20889:2018 [1] établit un cadre de référence international en définissant une terminologie commune et en classifiant les méthodes appropriées pour diminuer efficacement le risque de réidentification. Parallèlement, le RGPD (notamment dans ses considérants et aux articles 4 et 25) encourage l'usage de la pseudonymisation et de l'anonymisation comme leviers de minimisation des données, permettant de tirer parti des données à caractère personnel tout en renforçant la protection des droits des personnes concernées [1].

La norme ISO/IEC 20889:2018, relative aux techniques de désidentification, mentionne également la nécessité de réduire l'exposition des données en appliquant des transformations telles que la suppression, la généralisation ou l'agrégation. Bien que le ciblage ou la suppression verticale ne soient pas des techniques de désidentification dans la norme ISO/IEC 20889 :2018, elles constituent des *mesures organisationnelles* et *techniques de réduction* ou de *transformation des données* essentielles pour garantir la conformité et la sécurité des traitements relatifs au données à caractère personnelle mais également des données autrement protégées.

En effet, le principe de minimisation des données est l'un des fondements de la protection des données personnelles, tel que défini par l'article 5(1)(c) du Règlement général sur la protection des données (RGPD[1]). Ce principe impose aux responsables de traitement et aux sous-traitants de réduire la quantité et la granularité des données collectées et utilisées, afin de limiter les risques pour les personnes concernées. Les mesures de minimisation regroupent l'ensemble des techniques et pratiques permettant d'atteindre cet objectif, avant ou en complément des mécanismes de dé-identification.

2.1 Outils statistiques

2.1.1 Echantillonnage aléatoire stratifié

Un échantillonnage stratifié est une méthode d'échantillonnage à partir d'une population. Cette technique consiste à sélectionner des données de manière aléatoire tout en garantissant une répartition homogène selon des variables spécifiques. Contrairement à la génération de données

[1] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32016R0679

fictives, les données obtenues restent réelles. L'objectif est de respecter les critères définis par le demandeur tout en assurant une représentativité des différentes strates[2].

Contexte d'utilisation

Par exemple, un demandeur peut souhaiter réaliser un échantillonnage comprenant X individus par tranche d'âge et par catégorie socio-professionnelle. Pour cela, il doit :

- Identifier les variables à prendre en compte (ex. tranche d'âge, catégorie socio-professionnelle).
- Définir le nombre d'individus souhaité dans chaque échantillon (X individus).

Le demandeur peut également indiquer qu'il souhaite obtenir le nombre maximum possible d'enregistrements par échantillon. Dans ce cas, le système analysera la distribution des données et retiendra le plus petit nombre disponible parmi les combinaisons de valeurs.

Il est important de noter que cette technique peut échouer si les conditions initiales ne sont pas respectées (par exemple, si une tranche d'âge ou une catégorie professionnelle contient moins d'individus que X). Dans ce cas, il est nécessaire de définir une stratégie pour gérer ces situations :

- Conserver la valeur avec un nombre d'individus réduit ;
- Supprimer la valeur de l'échantillon ;
- Arrêter le traitement.

Paramètres

Nom de la technique	**Echantillonnage aléatoire stratifié**
Paramètres	
ID	**Description**
VARIABLE_LIST	Liste des variables qui doivent être pris en compte pour l'échantillonnage.
STRATEGIE	Une valeur parmi : COMPUTE_MAX : le plus grand nombre possible d'individus KEEP_UNSATISFACTORY : conserve les valeurs non satisfaites

[2] Échantillonnage stratifié — Wikipédia

	REMOVE_UNSATISFACTORY : supprime les valeurs non satisfaites STOP_PROCESS : Arrête le traitement en cas d'échec
COHORTE_COUNT	Le nombre d'individu par échantillon. Ce paramètre n'est pas pris en considération dans le cadre d'une stratégie basée sur le plus grand nombre d'individu possible par échantillon.

Exemple d'utilisation

Pour obtenir un jeu de données comprenant 1000 individus par tranche d'âge et par catégorie socio-professionnelle, insérez et adaptez l'opération suivante :

Tableau 1: Exemple d'utilisation de l'échantillonnage aléatoire stratifié

Process ID	TRT-EXEMPLE-1_1
Description	Echantillonnage sur les tranches d'âge et catégorie socio-professionnelles
Technique Name	Echantillonnage aléatoire stratifié
Paramètres	
ID	**Name**
VARIABLE_LIST	AGE, PROFESSION
STRATEGIE	STOP_PROCESS
COHORTE_COUNT	1000

2.1.2 Macro-agrégation d'enregistrements

La macro-agrégation agit sur un jeu de données et regroupement des enregistrements entre eux en fonction des valeurs d'un ou plusieurs champs. Cette technique vise à réduire la granularité des données d'un jeu de données afin de limiter les risques de réidentification tout en préservant leur utilité statistique.

Dans le cadre des pratiques de dé-identification, deux méthodes d'agrégation sont distinguées dans la norme ISO/IEC 20889:2018 [1] :

- Macro-agrégation, décrite dans cette section, qui opère à un niveau plus global en regroupant les données selon des variables de regroupement et en calculant des agrégats pour les autres variables, produisant ainsi des macro-données adaptées pour les rapports ou à l'analyse consolidée.

- Micro-agrégation (Section 2.4), qui consiste à regrouper des enregistrements en petits ensembles homogènes et à remplacer les valeurs individuelles par une valeur représentative (moyenne, médiane, etc.), afin de conserver la structure statistique tout en masquant les données exactes.

Le jeu de données résultant d'une macro-agrégation contient deux types de variables :

- Les variables dont les valeurs permettent de regrouper les enregistrements entre eux (par exemple, localité, secteur).
- Les variables agrégées (moyenne, somme, etc.) pour un même ensemble de valeurs de variables groupées.

Contexte d'utilisation

Les données agrégées (macro-niveau) résultantes de la macro-agrégation sont souvent utilisées pour les rapports ou analyses globales. Example : calculer l'âge moyen par localité ou le revenu total par commune.

L'objectif de la macro-agrégation est de réduire la granularité des données en passant d'un niveau individuel à un niveau agrégé (ex. par région ou catégorie).

Dans cette opération, dans le jeu de donnée résultant :

- Les variables de regroupement seront préservées.
- La transformation des variables agrégées seront détaillées dans la description du traitement.
- Les autres variables ne seront pas conservées dans le jeu de données résultant.

Paramètres

<table>
<tr><td>Nom de la technique</td><td colspan="3">Macro-agrégation</td></tr>
<tr><td>Paramètres</td><td colspan="3"></td></tr>
<tr><td>ID</td><td colspan="3">Description</td></tr>
<tr><td>AGGREGATION_VARIABLES</td><td colspan="3">Liste des variables sur lesquels l'agrégation doit être faite. Ces variables seront présentes dans le jeu de données résultants.</td></tr>
<tr><td rowspan="2">AGGREGATED_VARIABLES</td><td colspan="3">Lister toutes les variables qui doivent être conservées dans le jeu de données en plus des variables d'agrégation. Pour chaque variable à conserver, précisez la méthode d'agrégation à employer.

Pour une variable à agréger, plusieurs OUTPUTs peuvent être spécifiées. Par exemple, pour une variable numérique, on peut définir un output avec la valeur moyenne, un output avec la valeur maximale...

L'ensemble est spécifié sous forme d'une table :</td></tr>
<tr><td>INPUT</td><td>OUTPUT</td><td>Method</td></tr>
</table>

	Identifiant de la variable à agréger	Identifiant de la variable contentant les valeurs agrégées.	Méthode parmi la liste suivante : SUM MEAN MAX MIN MEDIAN MODE PERCENTILE STANDARD_DEVIATION COUNT

Exemple d'utilisation

Sur un jeu de données contenant les âges des personnes par localité et que l'on veut obtenir une agrégation des âges sur les localités, la variable de regroupement sera la localité et la variable agrégée sera la variable âge. On peut spécifier plusieurs méthodes d'agrégation pour une variable (dans l'exemple : moyenne, minimum et maximum). Le Plan de mesures appropriées contiendra alors la description suivante :

Tableau 2: Exemple d'utilisation de la macro-agrégation

Process ID	TRT-EXEMPLE-1_2		
Description	Agrégation de l'âge sur les localités		
Technique Name	Macro-agrégation		
Paramètres			
ID	**Name**		
AGGREGATION_VARIABLES	LOCALITE		
AGGREGATED_VARIABLES	**INPUT**	**OUTPUT**	**Method**
	AGE	AGE_MOYEN	MEAN
	AGE	AGE_MAX	MAX
	AGE	AGE_MIN	MIN

2.2 Suppression et masquage

La technique de suppression est une méthode de désidentification consistant à retirer complètement certaines variables d'un jeu de données afin de réduire le risque de réidentification. Son objectif est d'éliminer les identifiants directs (nom, adresse, numéro de sécurité sociale) ou des quasi-identifiants (date de naissance, code postal) pour empêcher qu'une personne puisse être identifiée à partir des données restantes. Cette approche est classée parmi les techniques « destructives » car elle supprime

l'information plutôt que de la transformer. Le cas d'utilisation de la technique de la suppression est particulièrement adapté lorsque les variables ou les enregistrements supprimés ne sont pas indispensables à l'analyse [1]. Cependant, il y a plusieurs types de suppressions.

2.2.1 Suppression verticale

La suppression verticale permet de supprimer totalement une variable (c'est-à-dire une colonne dans les données tabulaires) d'un jeu de données.

Contexte d'utilisation

Cette mesure vise à la fois à respecter le principe de minimisation des données prévu par le RGPD (article 5(1)(c)) mais aussi de de réduire au maximum la quantité d'information disponible dans l'ETS afin de :

- Limiter les risques de réidentification.
- Réduire l'exposition des données sensibles.
- Optimiser la performance des traitements.

La suppression verticale devra donc être employée systématiquement pour toutes les variables non utiles à l'analyse ultérieur par le Data Scientist dans l'Environnement de traitement sécurisé.

Pour effectuer une suppression de variable, le demandeur devra considérer deux cas de figures :

- Soit la variable n'est pas utilisée dans les traitements de désidentification, alors il demande au détenteur d'effectuer cette suppression afin la mise à disposition des données.
- Soit la variable est utilisée dans les traitements de désidentification ou de croisement, alors le demandeur devra spécifier explicitement un traitement de suppression dans la liste des traitements de désidentification (c.à.d. après utilisation dans l'ETI).

Paramètres

Nom de la technique	**Suppression verticale**
Paramètres	
ID	**Description**
VARIABLE_LIST	Liste des variables à supprimer.

Exemple d'utilisation

Une étude veut une cibler la population de plus de 50 ans mais n'a pas besoin de l'âge dans ce cadre. La variable « AGE » doit être considérée dans le plan de mesures appropriées pour effectuer le ciblage mais cette variable doit être supprimée dans le jeu de données cible. Dans ce cadre, la variable sera marquée comme « à utiliser » dans la section « description des sources » ; le traitement de suppression sera alors à spécifier dans la liste des traitements de la manière suivante :

Tableau 3: Exemple d'utilisation de la suppression verticale

Process ID	TRT-EXEMPLE-2_1
Description	Suppression des variables inutiles
Technique Name	Suppression verticale
Paramètres	
ID	**Value**
VARIABLE_LIST	AGE

2.2.2 Ciblage

Le ciblage consiste à ne conserver que les enregistrements spécifiques. Les enregistrements sont conservés selon une règle de rétention des enregistrements.

Contexte d'utilisation

Le ciblage est un processus fonctionnel qui consiste à sélectionner un sous-ensemble d'individus ou d'éléments à partir d'un jeu de données, en fonction de critères définis, afin de leur appliquer une action et un but spécifique. Le ciblage est nécessaire pour faire correspondre les données mises à disposition à la finalité métier et à la stratégie de traitement. Cette technique permet de ne conserver dans un jeu de données uniquement les enregistrements respectant une règle spécifiée. Le but est de :

- Identifier une population pertinente pour une finalité déterminée.
- Optimiser l'efficacité des actions en réduisant la portée à ce qui est nécessaire.

Le ciblage sera donc employé pour cibler les éléments à faire apparaître dans le jeu de données cibles. Cette technique est à utiliser pour modifier la portée géographique, temporelle ou en catégories d'individus du jeu de données cible par rapport à celle des jeux de données sources. Il est à noter que la portée peut également être modifiée lors des croisements de plusieurs jeux de données sources.

Le contexte d'utilisation du ciblage est basé sur :

- Un critère de sélection de domaine tels que démographiques, comportementaux, géographiques, temporels, etc.
- Une finalité explicite où chaque ciblage doit être justifié par une finalité légitime et documentée.
- L'impact sur les droits des personnes qui peut entraîner des risques de discrimination ou de profilage et qui nécessite une analyse du Data Protection Impact Assessment (DPIA).
- Les risques relatifs aux données autrement protégées.

Paramètres

Nom de la technique	**Ciblage**
Paramètres	

ID	Description
RETENTION_RULE	Règle de conservation des enregistrements La règle peut s'appliquer sur plusieurs variables du jeu de données. Elle est écrite sous la forme d'une formule booléenne avec les opérateurs suivants : ▪ Opérateurs booléens : OR, AND, NOT ▪ Opérateurs mathématiques : = ≠ < ≤ > ≥ ≈ ▪ Opérateurs ensemblistes : IN (xx,yy,zz) ▪ Type de base : nombres, booleen (True, False), String.

Exemple d'utilisation

Une étude veut une cibler la population de plus de 50 ans. La variable « date de naissance » doit être considérée dans le plan de mesures appropriées pour effectuer le ciblage.

Tableau 4: Exemple d'utilisation du ciblage

Process ID	TRT-EXEMPLE-2_2
Description	Ciblage des personnes de plus de 50 ans
Technique Name	Ciblage
Paramètres	
ID	**Value**
RETENTION_RULE	AGE > 50

2.2.3 Suppression horizontale

La suppression horizontale ou filtrage concerne la suppression totale d'enregistrement dans un jeu de données. La suppression des enregistrements est effectuée selon une règle de suppression.

Contexte d'utilisation

A l'instar du ciblage, la suppression horizontale vise à limiter :

- Limiter le volume des données à ce qui est strictement nécessaire pour la finalité du traitement.
- Réduire les risques liés à la conservation ou à l'utilisation de données non pertinentes.
- Préparer un jeu de données cible cohérent avec les besoins métier.

Cette technique pourra également être employée pour supprimer les données aberrantes ou « outliers » qui sont, en statistique, des observations distantes des autres observations présentes dans

les jeux de données. Ces données présente des risques d'individualisation importants dans un jeu de données.

Cette technique peut être utilisée dans le cadre de outliers connus à priori dans les jeux de données sources soit parce qu'ils ont été détectés par les détenteurs des données, soit parce qu'ils ont été identifiés comme tel lors de l'analyse de risque effectuée sur les données à mettre à disposition.

Paramètres

Nom de la technique	**Suppression horizontale par règle**
Paramètres	
ID	**Description**
DELETION_RULE	Règle de suppression des enregistrements La règle peut s'appliquer sur plusieurs variables du jeu de données. Elle est écrite sous la forme d'une formule booléenne avec les opérateurs suivants : ▪ Opérateurs booléens : OR, AND, NOT ▪ Opérateurs mathématiques : = ≠ < ≤ > ≥ ≈ ▪ Opérateurs ensemblistes : IN (xx,yy,zz) ▪ Type de base : Nombres, Booleen (True, False), String

Exemple d'utilisation

L'analyse de risque a montré que la population dont l'âge est supérieur à 90 ans présente des risques de réidentification. L'ensemble des données sur les personnes de plus de 90 ans sont supprimées du jeu de données. Le plan de mesures appropriés contient donc l'opération suivante :

Tableau 5: Exemple d'utilisation de la suppression horizontale

Process ID	TRT-EXEMPLE-2_2
Description	Suppression des données relatives aux personnes de plus de 50 ans
Technique Name	Suppression horizontale
Paramètres	
ID	**Value**
DELETION_RULE	AGE > 50

2.2.4 Substitution des identifiants personnels

Cette technique consiste à remplacer les valeurs d'une variable, correspondant souvent à un identifiant personnel, par un pseudonyme. La valeur du PIDs est remplacé par sa valeur encryptée. La technique est souvent appelée pseudonymisation des identifiants personnels.

Contexte d'utilisation

La désidentification des identifiants personnels est obligatoire dans le cadre de la mise à disposition des données dans l'ETS. La technique privilégiée doit être la suppression verticale. Toutefois, il peut être envisagé que ces identifiants personnels soient traités par substitution.

Dans cette opération, l'ensemble des variables du jeu de données sont conservées. Seules les valeurs dans les variables PID sont remplacées par des pseudonyme.

Un remplacement peut être déterministe dans un contexte particulier : dans un contexte, le pseudonyme sera toujours le même pour un PID donné. Dans le cadre du DGA, seul les contextes suivants sont autorisés :

- Aucun : pas de contexte défini, le pseudonyme sera toujours différent.
- Dataset : le pseudonyme sera le même pour un même PID du même data set. Il sera différent pour des datasets différents.
- Projet : le pseudonyme sera le même pour un même PID de n'importe quel data set du projet.

Si un nom de variable cible est spécifié, la variable source est supprimée et la variable cible est ajoutée. Il est interdit de conserver à la fois la variable source et la variable cible dans le jeu de données résultant de cette transformation.

Paramètres

Nom de la technique	**Substitution de variables**
Paramètres	
ID	**Description**
VARIABLE	Variable sur laquelle la substitution doit être faite.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs substituées (pseudonymes).
CONTEXT	AUCUN : la substitution sera aléatoire DATASET : le pseudonyme sera le même au sein du jeu de donnée mais sera différent pour deux data sets différents. PROJECT : le pseudonyme sera le même pour tous les jeux de données du projet.

Exemple d'utilisation

Tableau 6: Exemple d'utilisation de la substitution d'identifiant personnel

Process ID	TRT-EXEMPLE-2_4
Description	Cryptage des matricules S2_RESID_MAT
Technique Name	Substitution des identifiants personnels
Paramètres	
ID	Name
VARIABLE	S2_RESID_MAT
TARGET_VARIABLE	S2_RESID_PSEU
CONTEXT	PROJECT

2.2.5 Masquage

La technique de suppression par masquage consiste à remplacer tout ou partie des valeurs d'une variable par des caractères neutres (par exemple « *** » ou « XXXX ») ou des symboles, afin de rendre l'information illisible tout en conservant la structure des valeurs. L'objectif de suppression par masquage est d'empêcher la divulgation d'informations sensibles ou identifiantes tout en maintenant la cohérence du format pour des traitements ultérieurs. Cette méthode est classée parmi les techniques non destructives, car elle ne supprime pas la variable mais en altère le contenu.

Contexte d'utilisation

Le masquage est particulièrement utile dans les environnements où les données doivent être partiellement visibles pour des raisons opérationnelles, comme l'affichage des quatre derniers chiffres d'une carte bancaire. Cette approche équilibre confidentialité et utilité, mais doit être appliquée avec des règles claires pour éviter toute réidentification indirecte.

Paramètres

Nom de la technique	**Substitution par masquage**
Paramètres	
ID	**Description**
VARIABLE	Variable sur laquelle la substitution doit être faite.
SELECTION_RULE (OPTION)	Règle de sélection des enregistrements à modifier La règle peut s'appliquer sur plusieurs variables du jeu de données. Elle est écrite sous la forme d'une formule booléenne avec les opérateurs suivants : ▪ Opérateurs booléens : OR, AND, NOT ▪ Opérateurs mathématiques : = ≠ < ≤ > ≥ ≈

	▪ Opérateurs ensemblistes : IN (xx,yy,zz) ▪ Type de base : Nombres, Booleen (True, False), String Si le paramètre n'est pas précisé, la subsera appliqué à l'ensemble des enregistrements.
SUBSTITUTION_RULE	Règle de substitution des valeurs Par exemple : - Remplacer les 5 premiers caractères par XXXXX - Affecter la valeur '*'

Exemple d'utilisation

Ne conserver dans l'IBAN que les 4 premiers caractères correspondant à la Banque.

Tableau 7: Exemple d'utilisation du masquage

Process ID	TRT-EXEMPLE-2_5
Description	Traitement des code IBAN
Technic Name	**Substitution par masquage**
Paramètres	
ID	Name
VARIABLE	IBAN
SELECTION_RULE	
SUBSTITUTION_RULE	Conserver uniquement les 4 premiers caractères

2.3 Généralisation

La généralisation est une technique de dé-identification qui consiste à remplacer des valeurs précises par des valeurs plus globales ou moins spécifiques (ex. convertir une donnée continue en intervalle, regrouper une valeur détaillée en catégorie). Cette méthode est reconnue par ISO/IEC 20889:2018 [1] comme une forme de réduction de la précision des données pour diminuer le risque de ré-identification. L'objectifs de la généralisation :

- Réduire la granularité des données afin de protéger la vie privée.
- Maintenir un équilibre entre confidentialité et utilité, en conservant la structure statistique (distribution, proportions).

Les cas d'utilisation de la généralisation permettent :

- Réduction de granularité pour la confidentialité qui consiste à transformer des valeurs précises tels que l'âge, revenu, localisation en catégories ou intervalles pour limiter le risque de ré-identification.
- Préparation de données pour analyses statistiques afin de créer des classes homogènes (ex. tranches d'âge, zones géographiques) pour faciliter les calculs et visualisations.
- Rapport : Publier des données agrégées (exemple, par région ou tranche).
- Segmentation métier qui consiste à regrouper des professions, catégories socio-économiques ou zones pour simplifier les analyses et les décisions.

2.3.1 Rounding absolu

Arrondi des valeurs numériques à l'arrondi inférieur d'un multiple.

Contexte d'utilisation

Toutes variables numériques à caractère personnel ou protégées.

Paramètres

Nom de la technique	**Rounding absolu**
Paramètres	
ID	**Description**
VARIABLE	Identifiant de la variable à transformer.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
ROUDING_INCREMENT	Incrément à considérer soit m dans la formule suivante : $Z = \text{round}(x,\mathbf{m}) = \text{round}(x/\mathbf{m})\cdot\mathbf{m}$ **Où** - m est l'incrément à considérer - x : la valeur initiale - Z : la valeur après transformation

Exemple

La technique suivante permettra d'arrondir l'âge aux multiples de 5 inférieurs.

Nom de la technique	**Rounding absolu**
Paramètres	
ID	**Description**
VARIABLE	AGE
TARGET_VARIABLE	AGE_ARRONDI
ROUDING_INCREMENT	5

La transformation opérée donnera alors :

Valeur initiale de l'Age	Valeur transformée de l'âge
7	5
13	10
46	45
27	25
72	70

2.3.2 Rounding relatif

Une technique de généralisation consiste également à prendre en compte un arrondi basé sur le centième ou millième des valeurs maximum et minimum.

Correspond à la formule suivante : Round (10^1 * (V-Vmin) / (Vmax – Vmin))

Contexte d'utilisation

Toutes variables numériques à caractère personnel ou protégée.

Paramètres

Nom de la technique	**Rounding relatif**
Paramètres	
ID	**Description**
VARIABLE	Identifiant de la variable à transformer.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
ROUDING_UNIT	Percentile à considérer soit p dans la formule suivante : Z = Round (10^p * (x-Vmin) / (Vmax – Vmin)) Où: • P : le percentile • Z : la valeur transformée • x : la valeur initiale • Vmin = le minimum de toutes les valeurs de la variable • Vmax = le maximum de toutes les valeurs de la variable

	Le percentile est exprimé en en puissance de 10. Par exemple, un rounding sur la centaine sera exprimé par le chiffre 2.

Exemple

La technique suivante permettra d'arrondir la consommation électrique sur une année en percentile de l'écart entre les valeurs max et min :

Nom de la technique	**Rounding absolu**
Paramètres	
ID	**Description**
VARIABLE	CONSOMMATION
TARGET_VARIABLE	CONSOMMATION_ARRONDIE
ROUDING_UNIT	2

La transformation opérée donnera alors :

Valeur initiale de CONSOMMATION	**Valeur transformée**
20	3
5	0
7	0
24	3
27	4
320	63
300	60
505	100

2.3.3 Généralisation des codes postaux Luxembourgeois

Les code postaux Luxembourgeois ont une granularité fine parfois correspondant à une rue. Ils sont donc généralement considérés comme des quasi-identifiants. Il est donc conseillé de les généraliser :

- Au premier chiffre pour un découpage en zone plus importante
- Aux deux premiers chiffres du code postal pour une granularité fine.

Contexte d'utilisation

Toute variable de code postal Luxembourgeois.

Paramètres

Nom de la technique	**Généralisation de codes postaux**
Paramètres	

ID	Description
VARIABLE	Identifiant de la variable à transformer.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
ROUDING_UNIT	1 ou 2 en fonction de la granularité désirée.

2.3.4 Généralisation des dates/instant (datetime)

La généralisation de date/instant se fera selon plusieurs niveaux de généralisation : années, mois, jours, heure, minutes ou secondes.

Contexte d'utilisation

Toute variable de date.

Paramètres

Nom de la technique	Généralisation de date
Paramètres	
ID	**Description**
VARIABLE	Identifiant de la variable à transformer.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
ROUDING_UNIT	YEAR ou MONTH ou DAY ou HOUR ou MIN ou SEC en fonction de la granularité désirée.

2.3.5 Généralisation de Période

Une généralisation dans permettant de transformer une période sous forme de deux dates « From » et « to » en nombre d'année, de mois, etc...

Contexte d'utilisation

Tout couple de dates ou d'instant qui définissent une période.

Paramètres

Nom de la technique	Généralisation de période
Paramètres	
ID	**Description**
FROM_VARIABLE	Variable correspondant au début de la période.
TO_ VARIABLE	Variable correspondant à la fin de la période.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.

ROUDING_UNIT	YEAR ou MONTH ou DAY ou HOUR ou MIN ou SEC en fonction de la granularité désirée.

2.3.6 Généralisation par table de correspondance

La généralisation par table de correspondance fait bien partie des techniques de généralisation classées dans la norme ISO/IEC 20889:2018 [1]. Une généralisation de valeur dans laquelle une table définit pour chaque valeur de la variable à généraliser la valeur plus générale. La généralisation par table de correspondance consiste à utiliser une table prédéfinie qui associe chaque valeur possible à une valeur agrégée (exemple. Correspondance de codes postaux vers des communes, tranches d'âge vers catégories).

Contexte d'utilisation

La généralisation par table de correspondance est utilisée lorsque l'on souhaite remplacer des valeurs précises par des catégories prédéfinies afin de réduire la granularité des données et limiter le risque de réidentification. Cette technique est particulièrement adaptée pour :

- Variables catégorielles ou numériques sensibles (exemple, codes postaux, âges, professions).
- Création de classes cohérentes selon des règles métier ou réglementaires (exemple, regrouper des codes postaux en zones géographiques, des âges en tranches).
- Maintien de la cohérence analytique : les tables de correspondance garantissent une transformation uniforme et reproductible.

Les cas d'usage de la généralisation par table de correspondance seraient potentiellement dans le cadre :

- Segmentation (exemple, regrouper des professions en catégories socio-professionnelles).
- Analyses statistiques (exemple, transformer des âges en tranches pour histogrammes).
- Rapport réglementaire (exemple, regrouper des zones pour anonymiser des données géographiques).

Paramètres

Nom de la technique	**Généralisation de période**
Paramètres	
ID	**Description**
VARIABLE	Variable qui doit être généralisée.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
File	Un fichier CSV permettant d'effectuer le mapping entre la valeur et la généralisation à obtenir. Le fichier CSV sera attaché à la demande. Le fichier CSV peut disposer de plusieurs colonnes. La première colonne correspondra toujours à la valeur à généraliser.

	La deuxième colonne comprendra la valeur généralisée de la granularité la plus fine. Les colonnes suivantes pourront définir un niveau plus important.

2.3.7 Regroupement en intervalle

Le regroupement d'intervalle, également nommé « data bucketing » ou « data bining » agrège des données numériques en intervalles de valeurs.

Contexte d'utilisation

Toutes variables numériques à caractère personnel ou protégée.

Le data **bucketing** sera utilisée pour :

- Réduire la complexité des données numériques continues,
- Créer des catégories ou des intervalles (appelés "bins" ou "classes"),
- Faciliter l'analyse ou la visualisation (par exemple dans les histogrammes).

Paramètres

Nom de la technique	**Data bucketing**
Paramètres	
ID	**Description**
VARIABLE	Variable dont les valeurs doivent être transformées.
TARGET_VARIABLE	Nom de variable qui contiendra les valeurs transformées.
METHOD	FIXED_EDGES : défini l'ensemble des valeurs encadrant les bins sous forme d'un tableau. FIXED_SIZE : Défini la taille des bins à créer FIXED_NUMBER : Le nombre de bins est fixe et les bins auront tous la même taille. Le demandeur doit spécifier ce nombre.

2.4 Micro-agrégation

La micro-agrégation est une technique de désidentification de données numériques. Elle consiste à regrouper les enregistrements en ensembles homogènes (strates) d'un nombre minimum k, puis à remplacer les valeurs individuelles par une méthode représentative du groupe tels que la moyenne, médiane, centroïde ou intervalle. Ainsi, chaque groupe sera agrégé en un unique enregistrement. L'objectif est de réduire le risque de réidentification tout en conservant des propriétés statistiques de l'ensemble de données [1] [2].

Contexte d'utilisation

Utiliser la micro-agrégation est sur données numériques individuelles (exemple, revenus, âges) permet de :

- Préserver la structure et la distribution des données : la moyenne de chaque groupe permet de maintenir des propriétés statistiques importantes (moyenne globale, variances, corrélations).
- Permettre d'utiliser les données pour analyses statistiques et modèles prédictifs, car l'information essentielle reste intacte.
- Assurer un bon équilibre entre confidentialité (via la réduction du risque de réidentification) et fidélité des données.

A noter que la différence entre la micro-agrégation et la macro-agrégation réside principalement dans le niveau d'agrégation et la finalité. La micro-agrégation consiste à regrouper des enregistrements individuels en petits groupes homogènes d'une taille minimale (k) et à remplacer les valeurs originales par une valeur représentative, généralement la moyenne ou la médiane du groupe. Cette technique vise à réduire le risque de réidentification tout en conservant la structure statistique fine des données individuelles comme définie dans le ISO 20889 [1]. À l'inverse, la macro-agrégation opère à un niveau plus global : elle regroupe les données selon des variables de regroupement (par exemple, localité ou secteur) et calcule des agrégats pour les autres variables, produisant ainsi des macro-données. Cette approche supprime la granularité individuelle. En résumé, la micro-agrégation protège en partie la confidentialité tout en maintenant une certaine finesse analytique, tandis que la macro-agrégation transforme les données en informations consolidées à un niveau agrégé.

Paramètres

<table>
<tr><td>Nom de la technique</td><td colspan="3">Micro-agrégation</td></tr>
<tr><td>Paramètres</td><td colspan="3"></td></tr>
<tr><td>ID</td><td colspan="3">Description</td></tr>
<tr><td>K</td><td colspan="3">Fixe la taille minimale du groupe.</td></tr>
<tr><td rowspan="3">AGGREGATED_VARIABLES</td><td colspan="3">Lister toutes les variables qui doivent être conservées dans le jeu de données. Pour chaque variable à conserver, précisez la méthode d'agrégation à employer.

Pour une variable à agréger, plusieurs OUTPUTs peuvent être spécifiées. Par exemple, pour une variable numérique, on peut définir un output avec la valeur moyenne, un output avec la valeur maximale…

L'ensemble est spécifié sous forme d'une table :</td></tr>
<tr><td>INPUT</td><td>OUTPUT</td><td>Méthode</td></tr>
<tr><td>Identifiant de la variable à agréger</td><td>Identifiant de la variable agrégée.</td><td>Méthode parmi la liste suivante :

SUM</td></tr>
</table>

			MEAN MEDIAN
DISTANCE_DEFINITION	Pour chaque variable non numérique, une définition des distances peut être effectuées. Cette définition se fera sous la forme d'une matrice de distance. Cette matrice de distance sera attachée en format CSVs au plan de mesures appropriées.		
	VARIABLE	Matrice	
	Identifiant de la variable non numérique	Nom du fichier contenant la matrice de distance entre les valeurs de la variable.	

2.5 Modèles formels d'anonymisation et de confidentialité

L'anonymisation consiste à appliquer des techniques de désidentification afin de réduire le risque de réidentification des individus dans un jeu de données. Un modèle formel de mesure de la confidentialité appliqué à l'anonymisation permet d'évaluer ce risque et, dans certains cas, de fournir des garanties mathématiques. Ce type de modèle tient compte du contexte du cas d'usage et reposent sur des paramètres définis à partir d'une analyse empirique, incluant les risques liés à la réidentification par isolement, liaison ou inférence. Trois modèles connus dans la littérature scientifique et ISO 20889:2018 : K-anonymity, L-diversity, et T-closeness.

2.5.1 Modèle K-anonymity

Le *K-anonymity* consiste à regrouper les informations de manière que chaque identifiant (personne) ne puisse pas être distingué des autres dans un groupe d'au moins *K* individus [3] [4]. Cela réduit les risques de retrouver une personne précise (la probabilité est limitée à 1/K). Cependant, cette méthode ne suffit pas à empêcher certaines déductions ou inférences basées sur les données restantes.

Contexte d'utilisation

La K-anonymity s'applique particulièrement dans les cas où :

- Les données contiennent des quasi-identifiants (ex. âge, code postal, genre) qui, combinés, peuvent permettre d'identifier une personne.
- L'objectif est de conserver la structure des données pour des analyses statistiques simples (ex. moyennes, fréquences) sans divulguer d'informations individuelles.
- Les organisations veulent un mécanisme simple et interprétable pour garantir qu'aucun individu ne soit isolé dans le jeu de données : chaque enregistrement doit être indiscernable parmi au moins *K* individus.

K-anonymity est adaptée aux jeux de données tabulaires où l'on peut regrouper ou généraliser des variables pour former des classes d'équivalence, tout en maintenant une utilité minimale pour des analyses descriptives.

En pratique, pour appliquer la méthode K-anonymity, il est indispensable de :

1. Identifier les quasi-identifiants : Ce sont les variables qui, combinées, peuvent permettre la réidentification.
2. Définir le seuil K : Spécifier la valeur minimale de *K* (taille des classes d'équivalence) à partir de laquelle l'anonymisation est considérée comme satisfaisante.
3. Préciser la transformation pour chaque variable : chaque quasi-identifiant doit être traité par une technique appropriée, telle que :
    a. Généralisation (ex. regrouper les âges en tranches)
    b. Masquage (ex. remplacer une partie du code postal par des astérisques)
    c. Suppression (ex. retirer le genre si non nécessaire)

Les techniques comme la généralisation (ex. regrouper les âges en tranches) ou le masquage (ex. remplacer une partie du code postal par des astérisques) sont utilisés itérativement pour diminuer la granularité des valeurs lorsque la condition de K-anonymity n'est pas respectée.

La difficulté dans cette technique est d'identifier le paramètre K qui soit adapté au besoin de l'analyse. En effet, une valeur de K très grande peut rendre les données trop modifiées pour pouvoir être utilisable par le Data Scientist dans l'ETS. Dans ce cadre, il est recommandé d'utiliser une valeur de K tel que recommandé dans la littérature mais de compléter le plan de mesures appropriées avec des mesures d'utilité des jeux de données après transformation pour s'assurer de l'utilisabilité des données.

Paramètres

Nom de la technique	**K-anonymity**
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	Liste des variables constituant les identifiants
VARIABLE_LIST_QUASI_IDENT	Liste des variables constituant les quasi-identifiants
THRESHOLD_K	Le seuil correspondant à k dans le modèle k-anonymité en dessous duquel la solution n'est pas acceptable. Dans les meilleurs pratiques **K ≥ 5** pour une protection basique. K = de 10 à 20 pour une meilleure protection dans les grands ensembles de données.
Transformation 1	
VARIABLE QUASI_IDENT	Variable (quasi identifiant) sur laquelle la technique peut être employée.

TRANSFORMATION	Technique à utiliser.
Transformation 2	
VARIABLE QUASI_IDENT	Variable (quasi identifiant) sur laquelle la technique peut être employée.
TRANSFORMATION	Technique à utiliser.
….	

Exemple

Tableau 8 représente un exemple de jeu de données contenant des identifiants directs (ex. PatientID, Nom, Email), des quasi-identifiants (ex. Âge, Code postal, Genre) et une variable sensible (Maladie).

Tableau 8 Exemple de données sources à protéger

PatientID	Nom	Email	Âge	Code postal	Genre	Maladie
P001	A M	a.m@example.com	26	12345	F	Flu
P002	B L	b.l@example.com	27	12346	M	Cancer
P003	C B	c.b@example.com	33	12347	F	Diabetes
P004	D P	d.p@example.com	35	12345	M	Asthma
P005	E D	e.d@example.com	38	12346	F	Flu
P006	F R	f.r@example.com	39	12347	M	Diabetes
P007	G M	g.m@example.com	42	45610	F	Flu
P008	H L	h.l@example.com	44	45611	M	Asthma
P009	I L	i.l@example.com	45	45612	F	Cancer
P010	J M	j.m@example.com	48	45610	M	Diabetes
P011	K H	k.h@example.com	51	45611	M	Flu
P012	L F	l.f@example.com	58	45612	F	Cancer
P013	M G	m.g@example.com	61	78920	F	Flu
P014	N R	n.r@example.com	63	78921	M	Cancer
P015	O C	o.c@example.com	66	78922	F	Asthma
P016	P S	p.s@example.com	68	78920	M	Diabetes
P017	Q R	q.r@example.com	70	78921	M	Flu
P018	S C	s.c@example.com	72	78922	F	Asthma

Tableau 9 classifie les colonnes du Tableau 8 en d'identifiant direct, quasi-identifiant, variable sensible utile pour la K-anonymity.

Tableau 9 Exemple d'identifiant direct, quasi-identifiant, variable sensible.

Type	Colonnes	Définition	Traitement
Identifiant direct	PatientID, Name, Email	Permet d'identifier une personne à lui seul (ex.: nom, email).	Suppression. La suppression des identifiants direct est effectués par défauts.
Quasi-identifiant	Age, Code postale, Genre	Peut réidentifier en combinaison (ex.: âge + code postal).	Généralisation/masquage/suppression
Variable sensible	Disease	Information dont la divulgation crée un risque (santé).	Conservé pour analyse. Les variables sensibles ne sont pas pris en considération par K-anonymity mais contrôlé via L-diversity/T-closeness.

En se basant sur l'exemple du Tableau 8 une instanciation des paramètres à fournir pour spécifier l'opération de k-anonymity est monté dans le Tableau 10 :

Tableau 10: Exemple d'utilisation de la k-anonymity.

Nom de la technique	**K-anonymity**
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	PatientID, Nom et Prénom, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Code postale, Genre
THRESHOLD_K	5
Transformation 1	
VARIABLE	Code postale
TRANSFORMATION	Généralisation des codes postaux
ROUDING_UNIT	2 puis 1
Transformation 2	
VARIABLE	AGE
TRANSFORMATION	Regroupement en intervalle
METHOD	FIXED_SIZE=10

Transformation 3	
VARIABLE	Genre
TRANSFORMATION	MASQUAGE
SUBSTITUTION_RULE	Remplacer par une '*'

Tableau 11 montre la transformation des quasi-identifiants (tranches d'âge, masquage du code postal, suppression du genre) pour créer des classes d'équivalence (groupes) contenant au moins *K* individus (ici *K=5*). Cette approche réduit le risque de réidentification en « noyant » chaque individu dans un groupe suffisamment grand.

Tableau 11 Exemple de résultat de l'application K-anonymity.

Tranches d'âges	Code postal_Masqué	Genre	Maladie
25–39	123**	*	Flu
25–39	123**	*	Cancer
25–39	123**	*	Diabetes
25–39	123**	*	Asthma
25–39	123**	*	Flu
25–39	123**	*	Diabetes
40–59	456**	*	Flu
40–59	456**	*	Asthma
40–59	456**	*	Cancer
40–59	456**	*	Diabetes
40–59	456**	*	Flu
40–59	456**	*	Cancer
60–74	789**	*	Flu
60–74	789**	*	Cancer
60–74	789**	*	Asthma
60–74	789**	*	Diabetes
60–74	789**	*	Flu
60–74	789**	*	Asthma

2.5.2 Modèle L-Diversity

L-diversity est une amélioration de la méthode K-anonymity, conçue pour les jeux de données où la variabilité des variables sensibles est faible [5] [6]. Elle vise à empêcher les déductions trop faciles en garantissant que chaque groupe (classe d'équivalence) contient au moins *L* valeurs différentes pour la variable sensible (par exemple, la maladie). En pratique, cela signifie que même si deux personnes partagent les mêmes caractéristiques générales (âge, code postal), il sera difficile de deviner leur information sensible, car le groupe contient plusieurs valeurs possibles.

Contexte d'utilisation

La L-diversity est utilisée pour renforcer la protection offerte par la K-anonymity dans les jeux de données tabulaires. Elle intervient lorsque :

- Les données contiennent des variables sensibles (ex. maladie, revenu, opinion) et que l'on veut éviter les attaques par inférence.
- Chaque groupe d'équivalence formé par K-anonymity doit présenter une diversité suffisante pour les valeurs sensibles, afin qu'un attaquant ne puisse pas déduire l'information exacte d'un individu même s'il connaît son groupe.
- On souhaite limiter les risques liés aux distributions homogènes (ex. un groupe où tous les individus ont la même maladie), qui rendent K-anonymity insuffisante.

L-diversity est adaptée aux jeux de données tabulaires avec variables sensibles, pour garantir qu'au sein de chaque groupe anonymisé, il existe au moins L valeurs distinctes pour ces variables, réduisant ainsi le risque d'inférence.

Pour mettre en œuvre L-diversity, il est recommandé de :

- Identifier les quasi-identifiants : Ce sont les variables qui, combinées, peuvent permettre la réidentification (ex. âge, code postal, genre).
- Identifier les variables sensibles : Déterminer les colonnes contenant des informations sensibles (ex. maladie) et mesurer leur variabilité par classe.
- Fixer le seuil K de K-anonymity avant L (optionnel) : En pratique, L-diversity s'applique en complément de K-anonymity, K-anonymity commence par garantir que chaque classe contient au moins K individus (critère K-anonymity), puis on impose que ces classes respectent la diversité minimale L pour la variable sensible.
- Définir le seuil L (optionnel) : Spécifier la valeur minimale de L (nombre de valeurs distinctes de la variable sensible par classe) à partir de laquelle l'anonymisation est considérée comme satisfaisante. Une application de L-diversity seul, ajoute de la diversité dans les variables sensibles, mais ne garantit pas la taille minimale des classes.
- Préciser les transformations des quasi-identifiants (optionnel) : Pour atteindre à la fois les seuils K et L, chaque quasi-identifiant peut être traité par une technique appropriée :
    - Généralisation : (ex. regrouper les âges en tranches)
    - Masquage : (ex. remplacer une partie du code postal par des astérisques)
    - Suppression : (ex. retirer le genre si non nécessaire)

Paramètres

Nom de la technique	L-diversity
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	Liste des variables constituant les identifiants
VARIABLE_LIST_QUASI_IDENT	Liste des variables constituant le quasi-identifiant
VARIABLE_LIST_SENSIBLE	Liste des variables sensibles à considérer
THRESHOLD_K	Le seuil correspondant à k dans le modèle k-anonymité en dessous duquel la solution n'est pas acceptable. Dans les meilleurs pratiques **K ≥ 5** pour une protection basique. K = de 10 à 20 pour une meilleure protection dans les grands ensembles de données.
THRESHOLD_L	Le seuil correspondant à L dans le modèle L-diversity en dessous duquel la solution n'est pas acceptable. L ≥ 2 ou 3 pour une diversité minimale. L = 5 ou plus pour des variables sensibles.
Transformation 1	
VARIABLE QUASI_IDENT	Variable (quasi identifiant) sur laquelle la technique peut être employée.
TRANSFORMATION	Technique à utiliser.
Transformation 2	
VARIABLE	Variable (quasi identifiant) sur laquelle la technique peut être employée.
TRANSFORMATION	Technique à utiliser.

Exemple

Dans le cadre du jeu de données sensibles de la section précédente, la technique de L-diversity peut être employée de la manière suivante :

Tableau 12: Exemple d'opération L-diversity

Nom de la technique	L-diversity
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	PatientID, Nom et Prénom, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Code postale, Genre
VARIABLE_LIST_SENSIBLE	Maladie

THRESHOLD_K	5
THRESHOLD_L	3
Transformation 1	
VARIABLE	Code postale
TRANSFORMATION	Masquage
Transformation 2	
VARIABLE	AGE
TRANSFORMATION	Regrouper les âges en tranches
Transformation 3	
VARIABLE	Genre
TRANSFORMATION	Suppression. Retirer le genre si non nécessaire.

Tableau 13 montre la transformation des quasi-identifiants en fonction de la variable sensible. Pour cela, chaque groupe doit contenir plusieurs valeurs différentes pour la variable sensible, ce qui rend toute inférence déterministe impossible. L-diversity compte le nombre de catégories distinctes de la variable sensible dans chaque classe. Dans notre cas, toutes les classes présentent L = 4 valeurs distinctes pour la variable « Maladie », alors que le seuil fixé à L=3, ce qui garantit la conformité. La différence entre l'exemple Tableau 11 et Tableau 13 se trouve dans les tranches d'âges. En effet les tranches d'âges ont été entendus afin de respecter le seuil fixé à L=3.

Tableau 13: Exemple de L-diversity avec L=3.

Tranches d'âges	Code postal_Masqué	Genre	Maladie
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

2.5.3 Modèle T-Closeness

T-closeness est une technique avancée qui améliore L-diversity pour les jeux de données où les variables sensibles sont peu variés ou répartis de manière inégale. Son objectif est de limiter les risques d'inférence statistique : chaque groupe doit avoir une distribution des valeurs sensibles proche de celle observée dans l'ensemble du jeu de données [7]. Cette proximité est mesurée par une distance et ne doit pas dépasser un seuil T. En pratique, cela permet de conserver des données anonymisées qui restent représentatives du jeu original, tout en protégeant la vie privée. Cependant, cette méthode peut réduire l'utilité des données, car elle supprime certaines corrélations entre les quasi-identifiants et les variables sensibles.

Contexte d'utilisation

La T-closeness est utilisée pour renforcer la protection offerte par K-anonymity et L-diversity dans les jeux de données tabulaires. Elle intervient lorsque :

- Les données contiennent des variables sensibles (ex. maladie, revenu) et que l'on veut éviter les attaques par inférence basées sur la distribution des valeurs sensibles.
- Chaque groupe d'équivalence formé par K-anonymity doit avoir une distribution des variables sensibles proche de la distribution globale, afin qu'un attaquant ne puisse pas déduire une information avec une forte probabilité.
- On souhaite limiter les risques liés aux groupes homogènes ou biaisés, qui rendent K-anonymity et L-diversity insuffisantes.

T-closeness s'applique aux données tabulaires contenant des variables sensibles. Elle garantit que la distribution des valeurs sensibles dans chaque groupe (formé par K-anonymity) reste proche de la distribution globale, en ne dépassant pas un seuil T, ce qui limite les risques d'inférence.

Pour mettre en œuvre T-closeness, il est recommandé de :

- Identifier les quasi-identifiants : Ce sont les variables qui, combinées, peuvent permettre la réidentification (ex. âge, code postal, genre).

- Identifier les variables sensibles : Déterminer les colonnes contenant des informations sensibles (ex. maladie) et calculer leur distribution globale.
- Fixer le seuil K de K-anonymity avant T (optionnel) : En pratique, T-closeness s'applique souvent en complément de K-anonymity. On commence par garantir que chaque classe contient au moins K individus (critère K-anonymity), puis on impose que la distribution des variables sensibles dans chaque classe soit proche de la distribution globale.
- Fixer le seuil K de K-anonymity et L de L-diversity avant T (optionnel) : En pratique, T-closeness s'applique souvent en complément de K-anonymity et L-diversity. On commence par garantir que chaque classe contient au moins K individus (critère K-anonymity), puis que ces classes respectent la diversité minimale L pour la variable sensible, avant d'imposer la proximité de distribution (critère T-closeness).
- Définir le seuil T : Spécifier la valeur maximale de distance entre la distribution des variables sensibles dans une classe et la distribution globale. Plus T est faible, plus la protection contre l'inférence est forte.
- Préciser les transformations des quasi-identifiants (optionnel) : Pour atteindre à la fois les seuils K et T, chaque quasi-identifiant peut être traité par une technique appropriée :
    - Généralisation : (ex. regrouper les âges en tranches)
    - Masquage : (ex. remplacer une partie du code postal par des astérisques)
    - Suppression : (ex. retirer le genre si non nécessaire)

Paramètres

Nom de la technique	**T-closeness**
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	Liste des variables constituant les identifiants
VARIABLE_LIST_QUASI_IDENT	Liste des variables constituant le quasi-identifiant
VARIABLE_LIST_ SENSIBLE	Liste des variables sensibles à considérer
THRESHOLD_K	Le seuil correspondant à k dans le modèle k-anonymité en dessous duquel la solution n'est pas acceptable. Dans les meilleurs pratiques **K ≥ 5** pour une protection basique. K = de 10 à 20 pour une meilleure protection dans les grands ensembles de données.
THRESHOLD_L	Le seuil correspondant à L dans le modèle L-diversity en dessous duquel la solution n'est pas acceptable. L ≥ 2 ou 3 pour une diversité minimal. L = 5 ou plus pour des variables sensibles.
THRESHOLD_T	Le seuil correspondant à T dans le modèle T-closeness au-dessus duquel la solution n'est pas acceptable. t ≤ 0.2 (20%) pour une confidentialité modérée.

	t ≤ 0,05–0,1 pour une forte confidentialité dans les ensembles de données sensibles.
Transformation 1	
VARIABLE	Variable (quasi-identifiant) sur laquelle la technique peut être employée.
TRANSFORMATION	Technique à utiliser.
Transformation 2	
VARIABLE	Variable (quasi-identifiant) sur laquelle la technique peut être employée.
TRANSFORMATION	Technique à utiliser.
….	

Exemple

Tableau 14: Exemple d'utilisation de la T-closenesss

Nom de la technique	**T-closeness**
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	PatientID, Nom et Prénom, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Code postale, Genre
THRESHOLD_K	5
THRESHOLD_T	0.05
Transformation 1	
VARIABLE	Code postale
TRANSFORMATION	Masquage
Transformation 2	
VARIABLE	AGE
TRANSFORMATION	Regrouper les âges en tranches
Transformation 3	
VARIABLE	Genre

TRANSFORMATION	Suppression. Retirer le genre si non nécessaire.

Tableau 15 illustre la transformation des quasi-identifiants en fonction de la variable sensible. Pour cela, chaque groupe doit présenter une distribution des valeurs sensibles proche de la distribution globale, ce qui limite fortement les risques d'inférence. T-closeness mesure la distance entre ces distributions et impose qu'elle soit inférieure au seuil T fixé. Dans notre cas, toutes les classes respectent T ≤ 0.05, ce qui garantit la conformité.

Tableau 15: Exemple de T-closeness avec T=0.05.

Tranches d'âges	Code postal_Masqué	Genre	Maladie
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

La β-likeness, introduit dans [8], est une méthode qui ne fait pas partie de l'ISO 20889 :2018, qui a pour but de protéger la sensibilité des données publiées après anonymisation (par exemple, des données médicales ou des enquêtes). Son objectif est d'empêcher qu'un attaquant devienne trop sûr de connaître une information sensible sur une personne après avoir vu les données anonymisées. Ceci est de garantir la robustesse de la confidentialité, même en connaissant la distribution a priori des variables sensibles, l'attaquant ne peut améliorer significativement sa prédiction pour chaque valeur. Si, avant la publication, la probabilité qu'une personne ait une certaine caractéristique (ex. une maladie) était faible, la publication ne doit pas augmenter cette probabilité de manière excessive. Le paramètre β fixe cette limite. Ainsi, même si quelqu'un analyse les données, il ne pourra pas être beaucoup plus certain qu'avant.

β-likeness est plus proche de t-closeness, il vise à réduire le risque d'inférence en maintenant une similarité entre distributions.

Contexte d'utilisation

La technique de β-likeness peut être utilisée dans les cas suivants :

- Publication des données pseudonymisées sans révéler d'informations sensibles.
- Application des méthodes par *généralisation* ou *altération* des valeurs sensibles pour limiter les risques d'inférence et ceci afin d'atteindre un niveau de β-likeness défini, tout en préservant la qualité des données.
- Un contrôle fin des microdonnées sensibles divulguée (plus strict que t-closeness ou l-diversité).
- Les valeurs sensibles sont déséquilibrées, rares ou très informatives.
- Situations où la protection de données doit être granulaire, en évitant que l'exposition d'une valeur sensible devienne trop probable pour chaque valeur individuelle, même dans des sous-groupes.

Pour mettre en œuvre la β-likeness, il est recommandé de :

- Définir ce qu'on veut protéger : Décider quelles analyses ou requêtes seront faites sur les données (par exemple, statistiques globales ou modèles prédictifs).
- Fixer le niveau de confidentialité (β) sera fixé de la même manière que les précédents seuils dans les autres métriques, après expérimentations empiriques. Les valeurs de β se trouve dans la fourchette [0.1,1]. Les petites valeurs sacrifient l'utilité mais offrent une forte protection, tandis que les plus grandes inversent la tendance.

Paramètres

Nom de la technique	β-likeness
Paramètres	
ID	**Description**

VARIABLE_LIST_IDENT	Liste des variables constituant les identifiants
VARIABLE_LIST_QUASI_IDENT	Liste des variables constituant le quasi-identifiant
VARIABLE_LIST_ SENSIBLE	Liste des variables sensibles à considérer
THRESHOLD_K (Option)	Le seuil correspondant à k dans le modèle k-anonymité en dessous duquel la solution n'est pas acceptable. Dans les meilleurs pratiques **K ≥ 5** pour une protection basique. K = de 10 à 20 pour une meilleure protection dans les grands ensembles de données. Le K est optionnel mais recommandé, il garantit k-anonymity pour les quasi-identifiants en plus de β-likeness.
β	Le seuil correspondant à β dans la β-likeness. Les petites valeurs ([0.1-0.2]) sacrifient l'utilité mais offrent une forte protection, tandis que les plus grandes inversent la tendance. La règle implicite du seuil β est pour chaque valeur sensible, sa probabilité dans un groupe ne doit pas augmenter de plus que le seuil β par rapport à sa probabilité globale.

Exemple

Tableau 16: Exemple d'utilisation de la β-likeness

Nom de la technique	**β-likeness**
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	PatientID, Nom et Prénom, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Code postale, Genre
VARIABLE_LIST_ SENSIBLE	Maladie
THRESHOLD_K (Option)	5
β	0.15
Transformation 1	
VARIABLE	Code postale
TRANSFORMATION	Masquage
Transformation 2	

VARIABLE	AGE
TRANSFORMATION	Regrouper les âges en tranches
Transformation 3	
VARIABLE	Genre
TRANSFORMATION	Suppression. Retirer le genre si non nécessaire.

2.5.5 Confidentialité différentielle

La confidentialité différentielle est un cadre mathématique garantissant la protection de la vie privée des individus dans un jeu de données. Contrairement aux approches classiques d'anonymisation, elle ne se limite pas à empêcher la réidentification : elle vise à réduire l'influence des données d'une personne sur le résultat des analyses. Pour cela, un bruit calibré est ajouté aux réponses aux requêtes, de sorte que la présence ou l'absence d'un individu modifie les résultats de manière minimale, contrôlée par un paramètre de confidentialité ε. Un ε faible offre une confidentialité forte mais réduit la précision des résultats, tandis qu'un ε élevé améliore la précision au détriment de la confidentialité. Cette approche fournit des garanties formelles contre de nombreuses attaques, même en présence d'informations externes. Cependant, trouver un équilibre entre confidentialité et utilité nécessite des techniques avancées et peut être coûteux en calcul [9].

Contexte d'utilisation

La confidentialité différentielle est particulièrement utile lorsque les requêtes portent sur des variables quasi-identifiants ou sensibles (par exemple : âge, revenu, état de santé) et que l'on souhaite publier des statistiques ou des agrégats sans révéler d'informations individuelles.

La confidentialité différentielle s'applique particulièrement dans les situations où :

- Des statistiques agrégées doivent être publiées (ex. recensements, indicateurs de santé publique) sans révéler d'informations individuelles.
- Des données sont exploitées pour l'apprentissage automatique ou la recherche tout en respectant des réglementations comme le RGPD.
- Des collaborations inter-clients (par exemple, santé, assurance) nécessitent des analyses croisées sans divulguer les données brutes.

- Des données numériques (continu ou discret). C'est le type principal utilisé en confidentialité différentielle. Pour les données catégoriques, un encodage est nécessaire (One-hot[3], ordinal encoding[4]).

Pour mettre en œuvre la confidentialité différentielle, il est recommandé de :

- Définir ce qu'on veut protéger : Décider quelles analyses ou requêtes seront faites sur les données (par exemple, statistiques globales ou modèles prédictifs).
- Fixer le niveau de confidentialité (ε) :
    - Petit ε (par exemple, 0.1) = forte confidentialité mais résultats moins précis.
    - Grand ε (par exemple, 10) = résultats plus précis mais confidentialité plus faible.

Paramètres

Nom de la technique	Confidentialité différentielle
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	Liste des variables constituant les identifiants
VARIABLE_LIST_QUASI_IDENT	Liste des variables constituant le quasi-identifiant
VARIABLE_LIST_ SENSIBLE	Liste des variables sensibles à considérer
ε	Le seuil correspondant à ε dans la confidentialité différentielle au-dessus duquel la solution n'est pas acceptable. • [0,1–0,5] : confidentialité élevée, distorsion plus importante. • [1–3] : pratique d'utilisation courante • >5 : confidentialité faible mais jeu de données précis.

[3] https://fr.wikipedia.org/wiki/Encodage_one-hot
[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html

Exemple

Tableau 17: Exemple d'utilisation de la Confidentialité différentielle

Nom de la technique	Confidentialité différentielle
Paramètres	
ID	**Description**
VARIABLE_LIST_IDENT	PatientID, Nom et Prénom, Email.
VARIABLE_LIST_QUASI_IDENT	Age, Code postale, Profession
VARIABLE_LIST_ SENSIBLE	Diagnostic médical, Revenu annuel
ε	1

2.5.6 Données synthétiques

Ce paragraphe regroupe un ensemble de technique qui consiste à créer des enregistrements entièrement nouveaux et artificiels tout en préservant les propriétés statistiques et les relations de l'ensemble de données d'origine sans contenir de données individuelles réelles.

Avatarisation

Le principe de la méthode d'avatarisation s'apparente aux techniques d'anonymisation complète, mais elle repose sur la création d'enregistrements synthétiques des données réelles. L'avatarisation s'appuie sur des méthodes de proximité tel que les *k* plus proches voisins (méthode d'apprentissage automatique) pour préserver la structure des données tout en diluant l'information individuelle. Ce procédé génère des avatars pour chaque enregistrement : des individus artificiels qui reproduisent les caractéristiques statistiques des données originales sans contenir d'informations personnelles.

Contexte d'utilisation

Le contexte d'utilisation des données avatarisés s'applique particulièrement dans les situations où :

- Entraînement de modèles IA sans risque de réidentification.
- Des statistiques agrégées doivent être publiées (ex. recensements, indicateurs de santé publique) sans révéler d'informations individuelles.

Des données sont exploitées pour l'apprentissage automatique (modèles IA) ou la recherche tout en respectant des réglementations comme le RGPD.

Dans le cadre de la génération de donnes synthétiques à travers l'avatarisation, certaines approches reposent sur des mécanismes qui ajustent le niveau de masquage des informations tout en appliquant des transformations ciblées sur les zones les plus exposées. Il est donc important pour le demandeur de spécifier des critères d'utilisabilité spécifique sur les données : le Hidden rate et le local cloacking.

L'objectif est de garantir une confidentialité robuste sans compromettre la qualité globale des données, en maintenant leur pertinence pour les analyses et les usages prédictifs.

Paramètres

Nom de la technique	**Avatarisation**
Paramètres	
ID	**Description**
K	Représente le nombre de voisins les plus proches utilisés pour garantir le niveau de confidentialité lors du processus d'avatarisation.
Seed (optionnelle)	Un entier aléatoire utilisé pour initialiser le générateur de nombres aléatoires dans l'algorithme d'avatarisation. En fixant une valeur pour *seed*, le même résultat est obtenu pour un même jeu de données et les mêmes paramètres.

3. Mesure d'utilisabilité et de confidentialité des données désidentifiées

La désidentification des données est une étape essentielle pour réduire les risques liés à la protection de la vie privée. Cependant, ces techniques entraînent souvent une diminution de la valeur analytique des données. Pour garantir que les données désidentifiées restent utiles, il est nécessaire de mesurer leur utilisabilité à travers des indicateurs de qualité. Ces indicateurs permettent d'évaluer dans quelle mesure le jeu de données conserve les caractéristiques nécessaires pour répondre aux objectifs spécifiques du demandeur, tout en respectant les exigences de confidentialité définies par la norme ISO/IEC 20889:2018.

Les indicateurs de qualité mesurent la similarité statistique entre les données réelles et les données désidentifiées. Si les données désidentifiées et réelles sont statistiquement similaires, on considère que les données désidentifiées sont de haute qualité. Ces indicateurs se veulent un objectif à atteindre, car il n'est pas toujours possible d'obtenir une qualité optimale (100%) pour tous les indicateurs.

3.1 Nombre d'enregistrement final

Calcul le nombre d'enregistrement. Ce KPIs de base sera utilisé comme KPI d'utilisabilité par défaut.

3.2 Taux de valeurs supprimées ou remplacées

Pour une variable donnée, calcul le nombre de valeurs supprimées ou remplacées par une transformation. Cette mesure peut être utilisée pour toute technique de transformation. Dans le

cadre de technique de synthétisation, cette mesure permet de voir la proportion de donnée réelle dans le jeu de données cible.

3.3 Taux d'enregistrement modifiés

Cette mesure représente la proportion d'enregistrements du jeu de données original qui ont été modifiés après application d'une technique de transformation (suppression, perturbation, agrégation, etc.). Elle s'applique à toutes les méthodes de désidentifications pour quantifier l'ampleur des changements apportés. Dans le cadre de technique de synthétisation, cette mesure permet de voir la proportion de donnée réelle dans le jeu de données cible.

3.4 Nombre de valeurs des variables

Le nombre des valeurs en pourcentage des combinaisons des valeurs distinctes pour un ensemble de variables avant et après désidentifications. Utilisée sur des quasi-identifiants, cette mesure est une mesure de confidentialités. Dans les techniques d'anonymisation, cette mesure est calculée par la moyenne des tailles des classes d'équivalence formées par les quasi-identifiants lors de l'anonymisation.

3.5 Taux de perte d'information

Taux de perte d'information mesure le nombre de valeurs d'une ou plusieurs variables qui ont disparu dont la taille à significativement diminuer lors des opérations de désidentification.

3.6 Taux de rétention

Le taux de rétention mesure le nombre d'enregistrements conservés après traitement de désidentification.

3.7 Taux de rétention de diversité

Le taux de rétention de diversité correspond au pourcentage de valeurs distinct d'une variable après désidentification par rapport aux valeurs uniques du même variable avant désidentification.

3.8 Classification Metric

Cette mesure attribue un score de pénalité à chaque ligne supprimée ou dont la valeur sensible diffère de la valeur majoritaire dans sa classe lors de l'utilisation des algorithmes L-diversity ou t-closeness, calculant ensuite la moyenne de ces pénalités.

3.9 Discernability Metric

La discernability metric combine le carré de la taille des classes d'équivalence et ajoute une pénalité pour la suppression des lignes [10] lors de l'utilisation des algorithmes K-anonymity, L-diversity ou t-closeness.

3.10 Distance de Kolmogorov-Smirnov

Le test de Kolmogorov–Smirnov (KS) est une méthode statistique non paramétrique qui mesure la différence entre deux distributions cumulatives : celle des données originales et celle des données désidentifiés. KS calcule la distance maximale entre les fonctions de distribution empirique des deux échantillons. Plus cette distance est faible, plus les distributions sont similaires[5]. Ce test s'applique principalement à des variables continues (par exemple : âge, revenu, température), car il repose sur des distributions ordonnées. Cependant, KS n'est pas adapté pour des variables catégorielles. Ainsi, des métriques comme le χ^2[6] ou la divergence de Jensen-Shannon seront privilégiés.

3.11 Distance de Jensen-Shannon

La distance de Jensen–Shannon (JSD) est une mesure basée sur la théorie de l'information qui évalue la similarité entre deux distributions de probabilité. Elle est dérivée de la divergence de Kullback–Leibler[7] mais présente deux avantages : elle est symétrique et bornée entre 0 (distributions identiques) et 1 (distributions totalement différentes). La JSD est particulièrement adaptée aux données catégorielles ou aux variables continues discrétisées en classes (par exemple : codes postaux, catégories de produits, ou revenus regroupés en tranches), car elle compare des distributions sous forme d'histogrammes ou de probabilités. Pour des données continues brutes, il est nécessaire de les regrouper en intervalles avant calcul [8].

3.12 Hellinger distance

La distance de Hillinger[9] mesure l'écart entre la distribution originale et la distribution de l'avatar. Cette métrique renvoie la moyenne des distances de Hellinger calculées pour chaque variable. La distance de Hellinger varie entre 0 et 1, où 0 représente l'absence de différence entre les deux distributions et 1 représente une grande différence, entre les deux distributions.

---

[5] Kolmogorov–Smirnov test - Wikipedia
[6] Test du χ^2 — Wikipédia
[7] Kullback–Leibler divergence - Wikipedia
[8] Jensen–Shannon divergence - Wikipedia
[9] Distance de Hellinger — Wikipédia

3.13 Similitude de corrélation

La similitude de corrélation calcule la différence de corrélation moyenne globale sur toutes les variables continues.

3.14 Hidden rate

Le Hidden Rate est utilisé pour évaluer la perte de lien entre une donnée originale et sa version synthétique, ce qui mesure la robustesse de l'anonymisation face au risque de réidentification. Il correspond à la proportion de données aléatoires appliqué pour réduire la visibilité des variables sensibles dans les jeux de données synthétiques. Ce chiffre est exprimé en pourcentage. Il mesure le pourcentage d'individus dont l'avatar n'est pas l'enregistrement synthétique le plus proche. Un taux élevé (≥ 95 %) signifie que presque personne n'a son avatar comme point le plus proche, garantissant une protection élevée. À l'inverse, un faible pourcentage indique que de nombreux individus sont trop proches de leur avatar, ce qui accroît le risque de réidentification.

3.15 Local Cloacking

Le Local Cloaking indique combien d'avatars sont plus proches d'un individu réel que son propre avatar. Plus ce nombre est élevé, plus il existe d'avatars intercalés entre l'original et son avatar, ce qui rend la réidentification difficile.

4. Conclusion

Tableau 18 présente un résumé des principales caractéristiques des techniques de minimisation décrites et indique les risques susceptibles d'être atténués par chacune d'elles, conformément aux recommandations de la norme ISO 20889 :2018 et la littérature scientifique [1] [2] [3].

Tableau 18: Propriétés des techniques de minimisation de données.

Nom de la technique	Fiabilité des données au niveau des enregistrements	Applicable aux types de valeurs	Applicable aux types de variables	Réduit le risque de		
				Isolement	Lien	Inférence
Techniques de minimisation						
Ciblage	Oui	Toutes	Toutes	Non	Non	Non
Echantillonnage aléatoire stratifié	Oui	Catégorielle (strates), continue (valeurs dans	Quasi identifiants et sensibles	Non	Non	Non

		chaque strate)				

Tableau 19 présente un résumé des principales caractéristiques des techniques de désidentification décrites dans ce document. Ce tableau indique aussi les risques par chacune d'elles, conformément aux recommandations de la norme ISO 20889 :2018 [5] et la littérature scientifique [6] [7]. Les définitions des indices du tableau sont :

- Indice 1 : Sauf si l'anonymat K est mis en œuvre à l'aide de la microagrégation.
- Les indices 2 et 3 sont décrites à partir de la littérature. 2: Isolement : Protégé (hérite de K-Anonymity). Lien : Amélioré (car diversité des valeurs sensibles réduit la probabilité de lien exact). Inférence : Réduit (car il y a plusieurs valeurs possibles pour la variable sensible). 3 : Isolement : Protégé (hérite de K-Anonymity). Lien : Protégé (car la distribution est contrôlée). Inférence : Très bien réduit.
- 4 : les métriques de désidentifications devraient être mesurées pour confirmer la réduction du risques. N/A : Non Applicable

Tableau 19: Propriétés des techniques de désidentifications de données.

Nom de la technique	Utilités des données au niveau des enregistrements	Applicable aux types de valeurs	Applicable aux types d'variables	Réduit le risque de		
				Isolement	Lien	Inférence
Outils statistiques						
Echantillonnage						
Agrégation	N/A	Continue, discret	Toutes	Oui	Oui	Oui
Suppression	Oui					
Suppression verticale/horizontale	Oui	N/A	N/A	Partiellement	Partiellement	Partiellement
Masquage	Oui	Catégorique	Identifiants directs	Oui	Partiellement	Non
Généralisation	Oui	Toutes, Tout est sujet à interprétation.	Identifiants directs			

Rounding	Oui	Continue	Identifiants directs	Non	Partiellement	Partiellement
Généralisation des codes postaux/dates/ période/table de correspondance	Oui	Continue, ordinale	Identifiants directs , Quasi identifiants et	Non	Partiellement	Partiellement
Pseudonimisation	Oui	Catégorique	Identifiants directs	Non	Partiellement	Non
Randomisation	No		Identification des variables.			
Micro aggregation	No	Continue	Identifiants indirects et tous les variables.	Non	Partiellement	Partiellement
Modèles formels						
K-anonymity	Oui[1]	Toutes	Quasi identifiants	Oui	Partiellement	No
L-diversity	Oui[2]	Toutes	Quasi identifiants et sensibles	Oui[2]	Oui[2]	Oui[2]
T-closeness	Oui[3]	Toutes	Quasi identifiants et sensibles	Oui[3]	Oui[3]	Oui[3]
Confidentialité différentielle	No	Toutes	Identification des variables	Oui	Oui	Partiellement
β-likeness	Oui	Toutes	Quasi identifiants et sensibles	Non	Non	Oui
Données synthétiques						

Avatarisation	Oui	Toutes	Identifiant, Quasi identifiants et sensibles	Oui[4]	Oui[4]	Oui[4]

Bibliographie

[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, "Microaggregation," *In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia 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 generalization 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. 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-Anonymization.," *ICDE,* pp. 217-228, 2005.