ECOLE NATIONALE SUPERIEURE DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

UNION-DISCIPLINE-TRAVAIL

PROPOSITION D'UNE PROCEDURE D'ANONYMISATION DE DONNEES D'ENQUETES : CAS DE L'ENQUETE ACADEMIQUE SUR LE BIEN ETRE DES JEUNES ET LES OPPORTUNITES ECONOMIQUES A KORHOGO (2024)

Année académique 2023- 2024

REALISE PAR :

COULIBALY Al Aziz N'golo

DJIBRILLA Issa

Elèves ingénieurs Statisticiens Economistes

SOUS LA SUPERVISION DE :

M. STEPHANE N'ZI

Enseignant-Chercheur à l'ENSEA

Août 2024

REMERCIEMENTS

Nous tenons à exprimer notre profonde gratitude envers toutes les personnes qui ont contribué à la réalisation de ce document.

Nous remercions tout d'abord notre superviseur, M. Stéphane N'Zi, pour son soutien constant, ses conseils avisés et sa disponibilité tout au long de ce projet. Surtout pour sa patience à notre égard malgré les difficultés rencontrées par rapport au suivi du programme.

Nos sincères remerciements vont également à nos familles et amis pour leur soutien inconditionnel et leur encouragement tout au long de cette aventure académique.

Nous adressons nos remerciements à l'École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSEA) pour nous avoir fourni un environnement propice à l'apprentissage et à la recherche.

Enfin, nous exprimons notre reconnaissance envers toutes les personnes dont les travaux ont nourri notre réflexion et ont contribué à l'avancement des connaissances dans le domaine de l'anonymisation des données.

SOMMAIRE

REMERCIEMENTS i

AVANT-PROPOS iv

LISTE DES SIGLES ET TABLEAUX, ... v

Liste des Sigles v

Liste des Tableaux vi

ABSTRACT ET RESUME vii

INTRODUCTION GENERALE 10

I. Contexte et Justification 1

II. Problématique 3

III. Objectif général 4

IV. Objectifs spécifiques 4

V. Hypothèses 4

VI. Plan 4

CHAPITRE 1 : CADRE CONCEPTUEL ET REVUE DE LITTERATURE 10

I. INTRODUCTION 6

II. CADRE CONCEPTUEL 6

1. Concepts fondamentaux et objectifs de l'anonymisation 6

2. Risques et modèles d'attaques 7

3. Modèles de protection des données 10

4. Techniques d'anonymisation 16

5. Utilité des données anonymisées --Cadre légal et éthique 18

III. REVUE DE LITTERRATURE 20

1. Revue de littérature théorique 20

2. Revue Empirique 23

IV. CONCLUSION 25

APPROCHE METHODOLOGIQUE 10

I. INTRODUCTION 27

II. PHASE EXPLORATOIRE : AVANT L'APPLICATION DES METHODES 27

1. Évaluation de la nécessité de la confidentialité du point de vue légal 27

2. Présentation de la base de données et hypothèses 28

3. Identification et consolidation des variables clé 29

4. Choix du type de diffusion et élaboration des scénarios d'intrusion 30
5. Sélection des variables clé 31
III. PRESENTATION DES MODELES UTILISES 33
1. Procédure d'anonymisation par SDC classique 33
2. Modèles de génération de sonnées synthétiques 35
IV. APPLICATION DES MODELES 37
1. Processus SDC 37
2. Les méthodes génératives 38
3. Indicateurs d'évaluation du risque sur les bases de données 38
4. Indicateurs de mesure de l'utilité des données 39
V. CONCLUSION 41
RESULTATS ET DISCUSSION 10
I. INTRODUCTION 42
II. RESULTATS DE L'ANONYMISATION 42
***1.** Résultats du processus SDC 42*
2. Résultats comparés du SDC et des autres modèles génératives 43
III. TEST SUR DES MODELES ECONOMETRIQUES 46
IV. DISCUSSION DES RESULTATS 48
1. Anonymisation et Risques de Divulgation 49
2. Comparaison des Modèles de Génération de Données Synthétiques 49
V. Conclusion 50
CONCLUSION GENERALE 51
REFERENCES BIBLIOGRAPHIQUES viii
ANNEXES 10
TABLE DES MATIERES xv

AVANT-PROPOS

L'avènement de l'ère numérique a engendré une explosion de données sans précédent, transformant fondamentalement la façon dont nous collectons, stockons, et analysons l'information. Au cœur de cette révolution se trouve la question importante de la protection de la vie privée et de la confidentialité des données. En effet, alors que les données deviennent de plus en plus accessibles et exploitables, les préoccupations concernant la sécurité et la confidentialité des informations personnelles n'ont jamais été aussi pressantes.

Ce document est le fruit d'une exploration approfondie des défis et des opportunités liés à la protection des données, avec un accent particulier sur l'anonymisation des données d'enquêtes, en se basant sur l'exemple de l'Enquête Académique portant sur les bien-être des jeunes et les opportunités économiques à Korhogo. Notre démarche s'inscrit dans un contexte où les avancées technologiques, telles que l'intelligence artificielle et l'apprentissage automatique, offrent de nouvelles possibilités pour préserver la confidentialité tout en préservant l'utilité des données pour la recherche et l'analyse.

À travers ce travail, nous nous efforçons de proposer une procédure d'anonymisation des données d'enquêtes qui intègre des méthodes telles que le Contrôle de Divulgation Statistique (SDC) et les avancées contemporaines en matière de production de données synthétiques, tout en tenant compte des exigences légales et éthiques en matière de protection des données. Notre objectif est de contribuer à l'avancement des pratiques en matière de protection de la vie privée dans le domaine des enquêtes sociales, en proposant des solutions pratiques et efficaces pour concilier l'impératif de confidentialité avec la nécessité de partager des données pour la recherche et la prise de décision.

Nous tenons à exprimer notre gratitude envers toutes les personnes qui ont contribué à la réalisation de ce travail, en particulier notre superviseur, M. Stéphane N'Zi, dont la flexibilité nous a permis d'élargir notre champ de réflexion. Nous espérons que ce travail apportera une contribution significative à la communauté scientifique et aux professionnels engagés dans la protection des données et la promotion de la vie privée dans un monde de plus en plus connecté.

LISTE DES SIGLES ET TABLEAUX,

Liste des Sigles

Adj R^2 : Adjusted R^2

AOL : America Online

CAP : Correct Attribution Probability

CCPA : California Consumer Privacy Act

CSTest : Chi-Squared Test

DP: Differential Privacy

ENSEA : Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

GAN: Generative Adversarial Networks

HIPAA: Health Insurance Portability and Accountability Act

IL1: Information Loss 1

IPEDA: Information Protection and Electronic Documents Act

KL: Kullback Leibler

KLD: Kullback Leibler Divergence

KS: Kolmogorov Smirnov

LPD: Local Differential Privacy

MIMIC-III: Medical Information Mart for Intensive Care III

POPIA: Protection of Personal Information Act

PRAM: Post Randomization Method

PDPA: Personal Data Protection Act

PUF : Public Use Files

RGPD : Règlement Général sur la Protection des Données

RGPH : Recensement Général de la Population et de l'Habitat

QID: Quasi-Identifiant

SMC: Securised Multiparty Calcul

SSA: Statistic South Africa

SUF: Scientific Use Files

ZD : Zone de dénombrement

Liste des Tableaux

Tableau 1: Une base de données fictive des étudiants de l'ENSEA 11
Tableau 2: Exemple de k-anonymat -Table 2-anonyme de la base fictive 11
Tableau 3: Exemple de l-diversity ; Table 2-diverse de la base fictive 12
Tableau 4: Quelques législations sur la protection des données 19
Tableau 5: Résultats de l'anonymisation par SDC classique 42
Tableau 6: Résultats de l'anonymisation avec les modèles génératifs 43
Tableau 7: Résultats du modèle linéaire simple en fonction des bases anonymisés. 46

ABSTRACT ET RESUME

Résumé

Cette étude explore les défis et les opportunités liés à la protection des données, avec un accent particulier sur l'anonymisation des données d'enquêtes. L'objectif est de proposer une procédure d'anonymisation intégrant des méthodes telles que le Contrôle de Divulgation Statistique (SDC) et les avancées en matière de production de données synthétiques, tout en tenant compte des exigences légales et éthiques. Une procédure d'anonymisation est proposée en utilisant le SDC et les modèles de génération de données synthétiques (GAN et copules Gaussiens) en se basant sur une préparation contextuelle des données inspirée de la procédure SDC. L'efficacité des méthodes est comparée pour garantir la confidentialité des individus et l'utilité des données. Les résultats montrent que les modèles génératifs sont supérieurs au SDC en termes de confidentialité. Cependant, ils ont tendance à surestimer ou sous-estimer certaines dépendances entre les variables. Cela fait du SDC classique, une procédure adaptée aux anonymisations de type SUF et des modèles génératifs, notamment le *Copula GAN* et la *TVAE* adaptées aux PUF. Cette étude contribue à l'avancement des pratiques de protection de la vie privée dans les enquêtes sociales, en proposant des solutions pour concilier confidentialité et partage de données pour la recherche.

Abstract

This research explores the challenges and opportunities related to data protection, with a particular focus on survey data anonymization. The objective is to propose an anonymization procedure that integrates methods such as Statistical Disclosure Control (SDC) and advancements in synthetic data generation, while considering legal and ethical requirements. A proposed anonymization procedure utilizes SDC and synthetic data generation models (GANs and Gaussian copulas), based on a contextual data preparation inspired by the SDC procedure. The effectiveness of the methods is compared to ensure individual privacy and data utility. The results show that generative models are superior to SDC in terms of privacy. However, they tend to overestimate or underestimate certain dependencies between variables. This makes traditional SDC a suitable procedure for SUF-type anonymizations, while generative models, notably *Copula GAN* and *TVAE*, are well-suited for PUF. This study contributes to the advancement of privacy protection practices in social surveys by proposing solutions to reconcile privacy and data sharing for research.

INTRODUCTION GENERALE

I. Contexte et Justification

Le monde actuel est témoin de bouleversements majeurs qui façonneront les décennies à venir, parmi lesquels le développement rapide de l'intelligence artificielle (IA) occupe une place prépondérante. Bien que l'avènement d'internet eût posé les bases de cette révolution ; le développement des modèles d'IA basés sur la génération de données l'accélère. Les sciences sont aussi entrainées dans cette révolution. La Statistique n'en est pas non plus épargnée car elle en est la base et le cœur de ce phénomène. Si le carburant de la Statistique, c'est la donnée, le web quant à lui, représente une plateforme de diffusion, de partage de celle-ci, et l'homme ne s'en est pas privé. En effet, selon Alehmans (2017), la diffusion des données statistiques sur Internet a pris son essor dès les années 1978, mais c'est en décembre 2007, lors d'une réunion à Sébastopol, que des leaders d'Internet comme Tim O'Reilly et Lawrence Lessig ont plaidé en faveur de la libre diffusion des données publiques, inspirés par les principes de *l'open source*.

Cette ouverture des données officielles a cependant suscité des préoccupations croissantes quant à la protection de la vie privée des individus concernés. En réalité, un incident notable est survenu en 2006 lorsque l'entreprise américaine AOL a publié une base de données contenant les recherches de 600 000 utilisateurs sur une période de 3 mois, après avoir retiré les *identifiants directs (Noms, prénoms, Numéro matricule…)*. Bien que la base de données exclue les informations directement identifiables *(pseudonymisation)*, des journalistes ont réussi à identifier certains utilisateurs, mettant ainsi en lumière les lacunes en matière de protection de la vie privée. Pour identifier les utilisateurs, les journalistes ont utilisé une combinaison de techniques de recoupement d'informations. Ils ont analysé les termes de recherche contenus dans la base de données et les ont croisés avec des informations publiques disponibles en ligne. Par exemple, en examinant des recherches spécifiques et récurrentes liées à des adresses locales, des noms ou des intérêts personnels, ils ont pu associer certains schémas de recherche à des individus spécifiques. Cette procédure a démontré qu'il est possible de réidentifier des personnes même à partir de données pseudonymisés en utilisant des informations de contexte et des techniques de data mining. Donc, les techniques statistiques peuvent nuire à la confidentialité. Il y a donc nécessité *d'anonymiser* convenablement les bases de données (modifier la base de données de sorte que les informations sur des individus de la base ou non soient connues à partir de celle-ci)

Les défis liés à la diffusion des données ont conduit à la création de la Convention 108 du Conseil de l'Europe en 1981, établissant des principes pour la protection des données personnelles. Complétée par le Règlement Général sur la Protection des Données (RGPD) en 2018, cette convention a renforcé la réglementation en matière de protection des données pour toutes les organisations et pays membres de l'Europe (Arben, 2020). Cependant, les défis persistent avec l'expansion rapide de l'Internet. De plus, la Directive européenne 95/46/CE exige une protection adéquate des données personnelles transférées vers des pays tiers, ce qui a conduit à l'élaboration de textes législatifs dans de nombreux pays africains, soulignant l'importance croissante de la protection des données à l'échelle mondiale (Alex, 2012).

Si des lois sont émises, la recherche scientifique quant à elles n'est pas muette. Depuis longtemps, les chercheurs ont tenté de trouver des solutions en allant au-delà des identifiants directs. Une des méthodes les plus répandues est la procédure du *Statistical Disclosure Control* (*SDC*), ou *Contrôle de Divulgation Statistique en français*. Le SDC repose sur des scénarios (imitation des attaques de pirates) simples basés sur les connaissances déterministes des attaquants, et se concentre sur un ensemble limité de variables supposées connues des attaquants (Hundepool et al., 2012). Par exemple, lors d'une attaque, les connaissances de l'attaquant sur des variables comme l'âge et le sexe peuvent être utilisées pour ré-identifier des individus dans une base de données anonymisée. Cependant, il existe des possibilités pour un pirate de mener des inférences sur les données, ce qui est facilité aujourd'hui par la masse critique de données disponibles sur internet. Une illustration frappante de cette vulnérabilité est l'étude publiée en 2019 par Rocher et al. Ils ont démontré que 99,98% des Américains peuvent être correctement ré-identifiés dans n'importe quelle base de données anonymisé utilisant 15 attributs *(variables)* démographiques. Leur travail utilise des techniques de machine learning avancées pour montrer les limites des méthodes de SDC classiques. Encore une fois, l'avancée des modèles statistiques peut compromettre la confidentialité des données.

Pour contrer ces risques, des modèles théoriques probabilistes comme la Differential Privacy (DP), ou Confidentialité Différentielle en français, ont été développés. La DP est considérée comme l'une des plus grandes avancées en protection des données statistiques en raison de sa solidité mathématique (Dwork et al., 2006) que, selon nous, pour son éloquence. Par exemple, la DP ajoute du bruit aléatoire aux résultats des requêtes sur une base de données, de manière que les

informations spécifiques sur un individu ne puissent pas être déterminées avec une haute probabilité. Elle rend ainsi les tentatives de ré-identification pratiquement impossibles. Les avancées dans les réseaux de neurones, notamment avec les Generative Adversarial Networks (GAN), ont permis de générer des données synthétiques qui imitent les propriétés statistiques des données réelles sans révéler d'informations sensibles spécifiques. Ces évolutions montrent comment les techniques modernes s'adaptent et tirent parti des capacités technologiques pour améliorer la protection des données.

II. Problématique

Dans le contexte actuel de protection des données personnelles, il est nécessaire que les producteurs de données respectent les lois et réglementations lors de la collecte, du stockage et surtout de la publication des données. Bien que les pratiques actuelles soient souvent conformes aux normes statistiques, elles ne suffisent pas pour répondre aux exigences du mouvement Open Data. La diffusion en ligne des données nécessite une protection renforcée de la vie privée des individus concernés. Les méthodes classiques d'anonymisation, comme le SDC, se révèlent vulnérables face aux techniques avancées de ré-identification, notamment celles utilisant des techniques d'inférence sophistiquées.

De leur côté, les données issues d'enquêtes sociales, illustrent cette vulnérabilité. Même après anonymisation par des méthodes classiques comme le SDC, des attaquants malveillants peuvent encore ré-identifier des individus en utilisant ces informations publiquement accessibles. Cela compromet la confidentialité des données anonymisées. D'un autre côté, nous nous demandons laquelle de ces techniques est la meilleure pour anonymiser une base de données comportant presqu'autant de variables que d'individus.

Il est donc impératif de développer des procédures d'anonymisation des données d'enquêtes qui tiennent compte de l'ensemble des risques. Les approches innovantes, telles que la génération de données synthétiques, offrent des solutions robustes pour protéger les données sensibles tout en permettant leur utilisation pour des analyses et des recherches essentielles. Ces techniques avancées permettent de garantir que les informations restent confidentielles, répondant ainsi aux besoins croissants en matière de protection des données à l'ère de *l'Open Data.*

III. Objectif général

L'objectif de cette étude est *de proposer une procédure d'anonymisation des données intégrant les méthodes du SDC et les modèles génératifs tout en maintenant l'utilité des données.*

IV. Objectifs spécifiques

Cette étude vise spécifiquement à **:**

1. Évaluer les exigences légales et éthiques en matière de protection des données personnelles, en se concentrant sur les normes et législations nationales pertinentes.
2. Proposer une procédure d'anonymisation des données spécifique en utilisant la procédure SDC et la méthode GAN.
3. Comparer l'efficacité des procédures d'anonymisation SDC et GAN dans le cadre des enquêtes sociales.
4. Déterminer la meilleure approche d'anonymisation des données pour garantir à la fois la confidentialité des individus et l'utilité des données pour les analyses statistiques.

V. Hypothèses

Dans la conduite de cette recherche, nous faisons l'hypothèse que :

1. Les modèles génératifs sont meilleurs que le SDC du point de vue de la confidentialité ;
2. Les modèles génératifs sont meilleurs que le SDC du point de vue de l'utilité des données.

VI. Plan

Cet écrit est structuré autour quatre (04) chapitres principaux hormis l'introduction. Le premier chapitre, *Cadre Conceptuel et Revue de Littérature*, explore les concepts clés liés à l'anonymisation des données, les risques de ré-identification et une revue détaillée des méthodes classiques et modernes de protection des données. Le deuxième chapitre, *Approche éthodologique*, décrit la méthodologie adoptée pour l'étude, y compris l'analyse contextuelle de la base de données ; les techniques d'anonymisation utilisées, les critères de performance évalués, et la procédure d'application sur les données réelles. Le suivant, *Résultats et Analyse*, présente les résultats de l'anonymisation des données en utilisant différentes méthodes, compare leur efficacité et discute des implications des résultats obtenus. Enfin, le cinquième chapitre, *Discussion et Conclusion*,

synthétise les principaux résultats, discute les limites méthodologiques, propose des recommandations pratiques pour la diffusion des données anonymisées.

CHAPITRE 1 : CADRE CONCEPTUEL ET REVUE DE LITTERATURE

I. INTRODUCTION

Dans le contexte actuel de la prolifération des données et de l'importance croissante de la protection de la vie privée, l'anonymisation des données revêt une importance capitale. Ce chapitre présente le cadre conceptuel et une revue de la littérature portant sur les différentes techniques d'anonymisation et de génération de données synthétiques. L'objectif est de fournir une compréhension approfondie des concepts clés, des méthodes et des défis associés à l'anonymisation des données, particulièrement dans le cadre des enquêtes sociales. Nous explorons les principes fondamentaux de la confidentialité et de l'anonymisation, les risques de réidentification, ainsi que les modèles de protection des données et les techniques d'anonymisation classiques et avancées. Cette revue de littérature est essentielle pour situer notre recherche dans le paysage actuel et pour identifier les lacunes et les opportunités d'amélioration dans les méthodes existantes.

II. CADRE CONCEPTUEL

Nous présentons ici une structure théorique qui guide la recherche en définissant les concepts clés, les variables et les relations entre eux. Il sert de plan pour comprendre comment les différents éléments de l'étude interagissent et s'influencent mutuellement.

1. Concepts fondamentaux et objectifs de l'anonymisation

a) Confidentialité et Anonymisation

La confidentialité des données fait référence à la protection des informations contre l'accès non autorisé et la divulgation non intentionnelle. Elle vise à garantir que les données sensibles ne sont accessibles qu'aux personnes autorisées.

Différence entre anonymisation et pseudonymisation :

- ***Anonymisation*** : Modification des données pour qu'elles ne puissent plus être associées à un individu identifiable de manière irréversible.
- ***Pseudonymisation*** : Remplacement des identifiants directs par des pseudonymes, permettant de retrouver l'identité originale sous certaines conditions sécurisées.

''L'objectif de ce travail est d'anonymiser les données et non d'appliquer la pseudonymisation.''

Les objectifs de l'anonymisation sont

- **La protection de la vie privée** : Empêcher que les données ne soient retracées jusqu'à des individus spécifiques ;
- **La conformité légale** : Se conformer aux réglementations en matière de protection des données personnelles.
- **Facilitation du partage de données** : Permettre l'utilisation et le partage des données pour la recherche et l'analyse sans compromettre la confidentialité.

La confidentialité n'a de sens que parce qu'il y a une sensibilité dans les données.

b) *Données Sensibles et Quasi-Identifiants*

Identifiants directs : Informations permettant d'identifier directement une personne *(ex. Nom et prénoms).*

Quasi-identifiants : Données qui, combinées avec d'autres informations, peuvent identifier une personne *(ex. : date de naissance, code postal).*

Données sensibles : Informations qui, si divulguées, peuvent compromettre la vie privée ou la sécurité d'un individu *(ex : données médicales, financières, et de localisation.)*

Ces variables n'ont pas le même rôle dans les procédures d'anonymisation. Il y a un risque associé à chacun en fonction de son type et de sa structure.

2. *Risques et modèles d'attaques*

a) *Réidentification et risques de réidentification*

Ré-identification : C'est le processus par lequel des données anonymisées ou pseudonymisées sont combinées avec d'autres sources d'information pour identifier à nouveau les individus. Le risque associé à la réalisation d'un tel évènement est appelé *risque de réidentification*.

NB : La ré-identification des individus peut avoir des impacts significatifs sur la vie privée, notamment la stigmatisation et discrimination (Dwork et al., 2006), la violation de la vie privée en exposant des informations personnelles qui devraient rester confidentielles, telles que leur état de santé, leurs croyances religieuses, leurs préférences sexuelles, etc. (Machanavajjhala et al., 2007). Il y a même des conséquences socio-économiques en affectant la capacité des individus à obtenir un emploi, à accéder à des services sociaux ou à bénéficier de certaines opportunités économiques

(Li et al., 2007).

Exemples de situation faisant prévaloir les risques de ré-identification :

- *Croisement de données médicales anonymisées avec des bases de données publiques pour identifier des patients.*
- *Utilisation de métadonnées de localisation pour identifier les mouvements et comportements des individus.*
- *Combinaison d'informations démographiques anonymisées avec des enregistrements électoraux pour retrouver les identités.*

b) Mesures du risque de réidentification

Ces mesures permettent d'évaluer le risque que des données anonymisées puissent être réidentifiées et associées à des individus spécifiques, et d'évaluer l'efficacité des techniques d'anonymisation. Voici quelques-unes des types de mesures couramment utilisées :

- **Probabilité de réidentification** : Estimation de la probabilité qu'un individu puisse être identifié à partir des données anonymisées.
- **Lien entre les enregistrements** : Examine la corrélation entre les enregistrements anonymisés et d'autres ensembles de données accessibles au public ou privées.
- **Mesures de similarité** : Évaluent la similitude entre les enregistrements anonymisés et les données non anonymisées.
- **Entropie de l'information** : Mesure du degré d'incertitude quant à l'identification des individus à partir des données anonymisées. *Plus l'entropie est faible, plus le risque de réidentification est élevé.*
- **Mesures de distance** : Évaluent la distance entre les enregistrements anonymisés et les données réelles, *par exemple, la distance euclidienne ou de Hamming.*
- **Mesures de sensibilité** : Examinent comment les modifications apportées aux données peuvent affecter la probabilité de réidentification.

c) Scénarios et modèles d'attaques

Les scénarios d'attaques décrivent comment des quasi-identifiants peuvent être utilisés pour compromettre la confidentialité des données anonymisées. Comprendre ces scénarios est essentiel pour renforcer les mesures de protection des données. Voici les principales techniques utilisées par les attaquants pour ré-identifier des individus à partir de données anonymisées :

i. Attaques par correspondance des enregistrements

Ces attaques visent à relier des enregistrements similaires ou identiques provenant de différentes bases de données. Même si les données sont anonymisées, des recoupements minutieux peuvent permettre la ré-identification des individus. Samarati et Sweeney (1998) ont montré que les attaquants exploitent souvent des attributs quasi-identifiants tels que la date de naissance, le genre et le code postal pour faire correspondre des enregistrements de différentes sources. Par exemple, croiser des informations médicales anonymisées avec une base de données électorale peut permettre de retrouver l'identité des patients.

ii. Attaques par correspondance des tables

Elles consistent à rechercher et associer des enregistrements provenant de différentes tables pour révéler des informations sensibles. Li et al. (2007) ont montré que cette attaque repose sur l'exploitation des chevauchements d'attributs entre différentes tables partiellement anonymisées. En croisant des informations de plusieurs sources, les attaquants peuvent reconstituer des profils complets des individus. *Par exemple, associer des données de santé anonymisées avec des informations d'assurance sociale peut permettre de reconstruire l'historique médical complet d'une personne.*

iii. Attaques probabilistes

Il s'agit d'attaques qui utilisent des modèles statistiques et probabilistes pour associer des enregistrements et révéler des informations sensibles. Ces méthodes sophistiquées permettent aux attaquants de déduire des informations avec un haut degré de certitude, même en présence de bruit ou d'autres techniques d'anonymisation (Machanavajjhala et al., 2006). Les attaquants exploitent les distributions statistiques des données et utilisent des algorithmes d'apprentissage pour estimer les correspondances probables entre différents ensembles de données. *Par exemple, même si les données de santé sont anonymisées, les schémas de traitement et les fréquences de visite peuvent permettre de déduire l'identité des patients avec une certaine probabilité.*

iv. Autres attaques

Outre les attaques par correspondance des enregistrements, des tables et les attaques probabilistes, d'autres méthodes peuvent être utilisées pour compromettre l'anonymisation des données :

Attaques par inférence : Elles utilisent des connaissances a priori et des règles logiques pour déduire des informations sensibles à partir des données disponibles (Dwork et al., 2006). *Par exemple, en sachant qu'un petit nombre de personnes souffrent d'une maladie rare dans une base de données, il est possible de ré-identifier ces individus même si les données sont anonymisées*.

Attaques basées sur les réseaux sociaux : Elles servent des informations provenant de réseaux sociaux pour compléter et vérifier des données anonymisées. Monreale et al. (2011) ont montré que les schémas d'interaction sociale et les réseaux de contacts peuvent révéler des informations personnelles lorsqu'ils sont croisés avec des bases de données anonymisées.

Attaques d'homogénéité : Ces attaques se produisent lorsqu'un ensemble de données *k-anonymisé (voir section suivante : 3-a-i)* a des enregistrements identiques dans un groupe pour un attribut sensible spécifique. *Par exemple, si tous les individus d'un groupe k-anonyme ont la même maladie, un attaquant qui sait qu'un individu est dans ce groupe peut déduire l'attribut sensible de cette personne.*

Attaques de fond de panier : Se produisent lorsqu'un attaquant utilise des connaissances externes pour déduire des informations sensibles à partir de données anonymisées.

3. *Modèles de protection des données*

a) *Modèles de protection syntaxique*

Les modèles de protection syntaxique utilisent la structure des données pour garantir la confidentialité.

i. *k-anonymat*

Une table est dite k-anonyme lorsque tout enregistrement dans la table avec une valeur de QID a au moins k-1 autres enregistrements qui ont la même valeur de QID. Cela implique que la probabilité d'attribuer une victime à un enregistrement à travers les QID est au maximum $\mathbf{1/k}$.

Exemple : *Dans ce tableau contenant des informations fictives sur les étudiants de l'ENSEA, les quasi-identifiants sont Classe, Age et Sexe et l'information sensible est Revenue/mois. Si l'attaquant sait que le seul étudiant de la classe AS3 susceptible d'avoir 45 ans est un professionnel qu'il connait, il est plus facile pour lui de l'attribuer à l'enregistrement 2 et connaitre son revenu mensuel.*

Tableau 1: Une base de données fictive des étudiants de l'ENSEA

Classe	*Age*	*Sexe*	*Revenue/mois (FCFA)*
AS3	*21*	*H*	*50 000*
AS3	*45*	*H*	*600 000*
AS1	*18*	*F*	*100 000*
AS1	*18*	*F*	*1000*

Source : Auteurs

En regroupant les âges par classe moins de 20 ans et plus de 20 ans, on obtient une table 2-anonyme.

Tableau 2: Exemple de k-anonymat -Table 2-anonyme de la base fictive

Classe	*Age*	*Sexe*	*Revenue/mois*
AS3	*20+*	*H*	*50 000*
AS3	*20+*	*H*	*600 000*
AS1	*20-*	*F*	*100 000*
AS1	*20-*	*F*	*100 000*

Source : Auteurs

Avantages et limites : Le k-anonymat empêche les attaques par correspondances des enregistrements en formant une partition des enregistrements de sorte que chaque groupe ait au moins k enregistrements identiques sur les QID. Cependant, il ne tient pas compte du fait que plusieurs enregistrements peuvent concerner le même individu, ce qui le rend moins que k-anonyme dans certains cas.

ii. L-Diversité

Le principe de l-diversité exige que chaque groupe de QID contienne au moins l valeurs distinctes et bien représentées pour l'attribut sensible. Cela empêche les attaques par correspondances des attributs en garantissant une diversité suffisante des valeurs sensibles dans chaque groupe.

Exemple : *Dans notre exemple précédent, nous avons utilisé le k-anonymat pour regrouper les âges par classe afin d'assurer un niveau minimal d'anonymat. Maintenant, appliquons la l-diversité pour garantir une diversité suffisante des valeurs sensibles (revenu mensuel) dans chaque groupe de quasi-identifiants. Supposons que nous voulions assurer une l-diversité de 2. Cela signifie que*

chaque groupe de quasi-identifiants doit contenir au moins 2 valeurs distinctes pour l'attribut sensible (revenu mensuel). Pour assurer une l-diversité de 2, nous devons nous assurer qu'il y a au moins 2 valeurs distinctes de revenu mensuel dans chaque groupe de quasi-identifiants.
Si nous regroupons les âges par classe comme précédemment, nous obtenons :

Tableau 3: Exemple de l-diversity ; Table 2-diverse de la base fictive

Classe	**Age**	**Sexe**	**Revenue/mois**
AS3	20+	H	50 000
AS3	20+	H	600 000
AS1	20-	F	100 000
AS1	20-	F	100 000

Source : Auteurs

Dans ce tableau, chaque groupe de quasi-identifiants contient au moins deux valeurs distinctes pour le revenu mensuel : 50 000 et 600 000 pour AS3, et 100 000 pour AS1. Ainsi, nous respectons le critère de l-diversité de 2. Cela complète le processus d'anonymisation en utilisant à la fois le k-anonymat pour l'anonymisation générale et la l-diversité pour garantir une diversité suffisante des valeurs sensibles dans chaque groupe de quasi-identifiants.

Comparaison avec k-anonymat : Tandis que le k-anonymat se concentre sur les QID pour empêcher les attaques par correspondances des enregistrements, la l-diversité vise à prévenir la corrélation entre les QID et les attributs sensibles. Elle complète le k-anonymat en ajoutant une contrainte sur la distribution des valeurs sensibles.

iii. T-closeness

La t-closeness repose sur le principe selon lequel *la distribution des données sensibles* dans chaque groupe de quasi-identifiants doit être *proche* de la distribution globale des données sensibles, selon une mesure de similarité spécifiée. L'objectif est de garantir que les informations sensibles ne sont pas trop spécifiques à un sous-groupe particulier d'individus, tout en préservant l'utilité des données pour l'analyse statistique.

Exemple : *Reprenons le cas des données des étudiants de l'ENSEA et appliquons la t-closeness pour anonymiser ces données. Nous voulons garantir que la distribution marginale des revenus*

mensuels dans chaque groupe de quasi-identifiants (classe, âge, sexe) est proche de la distribution globale des revenus mensuels dans l'ensemble des données. ***Calculons d'abord la distribution globale des revenus mensuels dans l'ensemble des données :***

- *Revenu mensuel : {50 000, 600 000, 100 000}*

Maintenant, regroupons les données par classe, comme nous l'avons fait précédemment (Tableau 3). Dans chaque groupe de quasi-identifiants (classe), examinons la distribution marginale du revenu mensuel :

- *Pour AS3 (20+), la distribution marginale est {50 000, 600 000}.*
- *Pour AS1 (20-), la distribution marginale est {100 000, 100 000}.*

Maintenant, ***comparons ces distributions marginales à la distribution globale :***

- *Distribution globale du revenu mensuel : {50 000, 600 000, 100 000}*

Pour ***vérifier la t-closeness*** *dans nos données d'étudiants de l'ENSEA, nous devons avoir la répartition des revenus mensuels dans chaque groupe de quasi-identifiants (par exemple, chaque classe) avec la répartition globale des revenus dans l'ensemble des données. On peut utiliser une mesure de divergence (exemple : la Distance de Kullback-Leibler (KLD) ou la distance de Earth Mover.) pour quantifier cette différence. Si la divergence est inférieure à un seuil* ***t,*** *la t-closeness est respectée pour ce groupe. Sinon, nous devons ajuster notre anonymisation pour atteindre la t-closeness pour tous les groupes de quasi-identifiants.*

Avantages *:* la t-closeness prend en compte la distribution des données sensibles, ce qui permet de préserver la représentativité des données pour l'analyse statistique.

Limites *:* La principale limite de la t-closeness réside dans sa complexité de mise en œuvre, en particulier pour *déterminer la mesure de similarité à utiliser pour évaluer la proximité entre les distributions.*

Comparaison aux autres méthodes : Comparée à d'autres méthodes d'anonymisation telles que le k-anonymat et la l-diversité, la t-closeness offre un niveau de protection de la vie privée plus élevé en prenant en compte la distribution des données sensibles.

iv. δ-Diversity / Presence

Pour contrer la possibilité de liaisons de table, ce modèle a été introduit et avec l'idée de limiter la probabilité d'inférer la présence de tout enregistrement potentiel de victime dans une plage

spécifiée $\delta = (\delta min, \delta max)$ (Nergiz et al., 2007 ; Wong et al., 2006*)*. Cette notion de présence δ peut indirectement empêcher les liaisons d'enregistrements et d'attributs, car si l'attaquant a une confiance maximale de δ% que l'enregistrement de la victime ciblée est présent dans la table publiée, alors la probabilité d'une liaison réussie vers son enregistrement et son attribut sensible est également limitée à δ%.

b) Modèles de protection statistique

Les modèles de protection statistique utilisent des techniques basées sur les propriétés statistiques des données pour assurer la confidentialité.

i. *k-map*

Proposé par Reiter (2005), ce modèle garantit qu'un enregistrement anonymisé est indistinguable d'au moins k autres enregistrements *dans une population plus large.* Formellement, chaque enregistrement dans la base de données anonymisée correspond à *au moins k enregistrements dans la population de référence.*

ii. *Seuils de risque moyen*

Ces méthodes définissent des seuils pour le risque moyen d'identification, comme proposé par Skinner et Holmes (1998). Formellement, le risque moyen d'identification est le risque attendu qu'un individu soit correctement identifié dans une base de données anonymisée.

iii. *Modèles de super-population*

Ces modèles, tels que présentés par Fienberg et McIntyre (2004), utilisent des hypothèses de super-population pour estimer et limiter le risque d'identification. Les modèles de super-population considèrent la base de données comme un échantillon d'une population ***hypothétique plus large***, permettant ainsi d'estimer le risque d'identification.

c) Modèles de Protection Sémantique

Ils visent à fournir des garanties de confidentialité en utilisant des techniques basées sur des ***interprétations mathématiques*** ou ***théoriques de la confidentialité.***

i. (ε, δ)-differential privacy

Introduit par Dwork et al. (2006), ce modèle assure que la *probabilité d'obtenir* ***un certain résultat*** *d'analyse est presque la même, que* ***les données d'un individu spécifique*** *soient* ***incluses ou non*** *dans la base de données*. Formellement, pour tout ensemble de résultats ***S*** et pour toutes les bases de données ***D1*** et ***D2*** qui diffèrent ***d'au plus un*** enregistrement, la probabilité que l'algorithme de confidentialité différentielle produise ***S*** satisfait

$$Pr[M(D1) \in S] \leq \epsilon \cdot Pr[M(D2) \in S] + \delta$$

Dans l'exemple *des étudiants de l'ENSEA, avec la confidentialité différentielle, lorsqu'on interroge la base de données pour obtenir des statistiques sur les revenus moyens des étudiants, les résultats ne doivent pas permettre à un attaquant de déterminer si un étudiant spécifique est inclus dans ces statistiques ou non.*

Avantages :

- ✓ La confidentialité différentielle offre ***une garantie mathématique*** solide de la confidentialité des données, ***indépendamment des connaissances ou des capacités de l'attaquant***.
- ✓ Elle permet de protéger la vie privée des individus tout en permettant l'utilisation des données pour des analyses statistiques et des requêtes.

Limites et comparaison aux autres méthodes :

- ✓ Malgré une forte garantie de confidentialité, elle peut nécessiter des mécanismes de perturbation des données qui peuvent réduire leur utilité pour certaines analyses.
- ✓ Comparée à d'autres méthodes d'anonymisation telles que le k-anonymat, la l-diversité et la t-closeness, la confidentialité différentielle offre une protection plus forte.

ii. Approche de désidentification basée sur la théorie des jeux

Cette approche, présentée par Ghosh et Roth (2011), modélise la désidentification comme *un jeu stratégique* entre un *attaquant* et un *défenseur*. L'attaquant cherche à identifier les individus dans une base de données anonymisée en utilisant des informations accessibles publiquement, tandis que le défenseur cherche à protéger la vie privée des individus en modifiant la base de données pour rendre l'identification difficile ou impossible. L'attaquant utilise des techniques d'appariement d'enregistrements, d'attaques de recoupement et d'autres méthodes statistiques et algorithmiques pour maximiser le nombre de réidentifications correctes. Les stratégies du défenseur reposent sur des techniques de suppression, généralisation, ajout de bruit et échange de données (data swapping)

pour anonymiser la base de données, avec pour objectif de minimiser le risque de réidentification tout en préservant l'utilité des données. La solution recherchée est un équilibre de Nash.

4. *Techniques d'anonymisation*

a) *Techniques d'anonymisation classiques : Le Contrôle de Divulgation Statistique (SDC)*

L'anonymisation de base de données requiert de prendre en compte dans la démarche l'ensemble des failles exploitables présentées ci-dessus et mettre en place une combinaison de contres attaques permettant d'assurer la protection des informations. Le SDC est un ensemble de techniques visant à protéger la confidentialité des données tout en les rendant utilisables pour des analyses statistiques. Son objectif principal est de réduire le risque de divulgation des données sensibles tout en préservant leur utilité analytique. Pour atteindre cet objectif, le SDC utilise diverses stratégies d'anonymisation des données visant à minimiser les risques de divulgation à travers *le k-anonymat, l-diversité, ou* bien d'autres propriétés souhaitées. Parmi les techniques utilisées en SDC, on a par exemple :

- ✓ ***Suppression :*** *Éliminer les identifiants directs de la base de données. Cela réduit immédiatement le risque de ré-identification directe.*
- ✓ ***Généralisation :*** *Il s'agit de remplacer les valeurs spécifiques par des catégories plus générales pour réduire la granularité des données.* ***Par exemple :*** *20-30 ans au lieu de 25 ans et les trois premiers chiffres du code postal au lieu du code complet.*
- ✓ ***Perturbation :*** *On ajoute du bruit aux données pour masquer les valeurs réelles tout en conservant les tendances globales.* ***Exemple :*** *Ajout de bruit gaussien aux valeurs numériques ou l'utilisation de techniques de micro-agrégation pour brouiller légèrement les données.*

b) *Anonymisation Basée sur la Génération de Données Synthétiques*

L'anonymisation basée sur la génération de données synthétiques est une approche innovante pour protéger la confidentialité des données tout en préservant leur utilité pour l'analyse. Cette méthode consiste à créer des données artificielles qui imitent les propriétés statistiques des données réelles, rendant ainsi les informations sensibles non identifiables. Les Réseaux Génératifs Antagonistes (GAN) sont une des techniques les plus prometteuses dans ce domaine.

i. Réseaux Génératifs Antagonistes (GAN)

Les GAN sont une classe de modèles d'apprentissage profond utilisés pour générer des données synthétiques réalistes à partir d'un ensemble de données d'entraînement. Proposés par Goodfellow et al. en 2014, les GAN se composent de deux réseaux neuronaux antagonistes : le générateur et le discriminateur.

- ✓ **Générateur** : Le générateur crée de nouvelles données en les échantillonnant à partir d'une distribution de probabilité latente. L'objectif est de produire des données indiscernables des données réelles.
- ✓ **Discriminateur** : Le discriminateur tente de distinguer entre les données réelles et les données générées par le générateur. Il s'améliore à chaque interaction en identifiant les données synthétiques.

Les deux réseaux sont entraînés de manière concurrente, s'améliorant mutuellement au fil du temps. Le générateur apprend à créer des données de plus en plus réalistes, tandis que le discriminateur devient plus habile à détecter les données synthétiques. Ce processus d'apprentissage antagoniste permet d'atteindre un équilibre où les données générées sont pratiquement indiscernables des données réelles.

ii. Autres Techniques de génération de données synthétiques

En plus des GAN, plusieurs autres techniques sont utilisées pour générer des données synthétiques.

- ✓ **Conditional Tabular GAN (CTGAN)** : Spécifiquement conçu pour les données tabulaires, CTGAN gère efficacement les dépendances entre colonnes ;
- ✓ **Copula Gaussian** : Utilise des fonctions de copules pour conserver les structures de dépendance complexes des données réelles.
- ✓ **Autoencodeurs Variationnels (VAE)** : Apprennent à encoder et décoder les données dans une distribution latente, produisant des données synthétiques réalistes.

Les données synthétiques offrent plusieurs avantages, notamment la protection de la confidentialité en éliminant les liens directs avec les données réelles, la conservation des propriétés statistiques des données originales pour des analyses utiles, et la facilitation du partage de données entre organisations tout en respectant les régulations de protection des données. Cependant, elles présentent aussi des défis, tels que l'instabilité de l'entraînement des modèles nécessitant des ajustements précis, une diversité limitée des données générées réduisant leur représentativité, et

une efficacité variable du discriminateur pouvant compromettre la qualité des données synthétiques produites.

5. *Utilité des données anonymisées --Cadre légal et éthique*

a) *Utilité et mesures d'utilité*

Dans le processus d'anonymisation, on évalue à la fin l'intérêt des données obtenues. Cela passe par la notion d'utilité. ***L'utilité*** *des données fait référence à leur valeur et à leur pertinence pour les analyses et les applications prévues.* ***Les mesures d'utilité*** sont des ***métriques*** utilisées pour évaluer cela. Quelques indicateurs sont la vérification de ***la préservation des caractéristiques statistiques*** (la moyenne, la médiane, l'écart-type, etc) après ***anonymisation***, ***des tendances et de la qualité des prévisions.*** Toutefois, il est aussi possible de tester les données sur des cas réels comme on l'aurait fait avec les données réelles. Il s'agit d'appliquer les mêmes analyses aux deux bases pour évaluer la pertinence des données obtenues *(ex : modèles économétriques)*

Etant donné la nécessité de la préservation de l'utilité des données ; il faut trouver un équilibre entre la protection de la confidentialité des données et la préservation de leur utilité pour garantir qu'elles restent utiles tout en minimisant le risque de divulgation d'informations sensibles. Cela passe également par la prise en compte des exigences légales et éthiques.

b) *Législations et réglementations internationales*

Nous présentons ici quelques lois et principes éthiques justifiant l'anonymisation des données et surtout dans les enquêtes sociales. Les législations sur la protection des données varient dans le monde entier, mais partagent un objectif commun de garantir la confidentialité et la sécurité des informations personnelles. Comme on peut le constater aussi, les pays africains ne sont pas en marge de cette série de législations bien que leur implication soit tardive par rapport aux autres *(voir tableau ci-dessous)*. Les approches vont de sanctions sévères pour les violations, telles que celles observées aux États-Unis, à des normes rigoureuses de consentement explicite et de transparence, comme le prévoit le *Règlement Général sur la Protection des Données (RGPD)* de l'Union Européenne. Chaque cadre réglementaire reflète les valeurs et les préoccupations de sa région respective, mais converge vers une protection renforcée des données personnelles.

Tableau 4: Quelques législations sur la protection des données

Année	Législation	Pays/Région	Particularités
1996	*HIPAA (Health Insurance Portability and Accountability Act)*	États-Unis	✓ Normes nationales pour la protection des informations médicales ✓ Sanctions pour les violations, ✓ ***Exigences de sécurité des données électroniques.***
2000	*Data Protection Act*	Royaume-Uni	✓ ***Obligations pour protéger les données,*** ✓ Droits des individus sur leurs informations, ✓ Pénalités pour le non-respect.
2013	*Loi n° 2013-450 relative à la protection des données à caractère personnel*	*Côte d'Ivoire*	✓ Autorité de régulation, ✓ Sanctions pour les violations, ✓ ***Protection contre l'utilisation non autorisée et le traitement abusif.***
2013	*Protection of Personal Information Act (POPIA)*	*Afrique du Sud*	✓ Consentement explicite, ✓ Audits réguliers et rapports de conformité.
2018	*Règlement Général sur la Protection des Données (RGPD)*	Union Européenne	✓ Harmonisation des lois, ✓ ***Transparence,*** ✓ Consentement clair, ✓ Sanctions sévères.
2019	*Personal Data Protection Act (PDPA)*	*Nigeria*	✓ Normes strictes, ✓ Sanctions, ✓ Mesures de sécurité techniques et organisationnelles.
2019	*Data Protection Act*	*Kenya*	✓ Consentement, ✓ ***Commission de protection des données,*** ✓ Sanctions.
2020	*California Consumer Privacy Act (CCPA)*	Californie, États-Unis	✓ Droit de connaitre les données collectées, ✓ Possibilité de refus de vente de ses données, ✓ ***Obligation de divulgation des données collectées.***
2020	*Personal Information Protection and Electronic Documents Act (PIPEDA)*	Canada	✓ ***Protection des données dans les activités commerciales,*** ✓ Consentement, ✓ Évaluations régulières de la conformité.

Source : Auteurs.

Comme on le constate, en Afrique, les pays comme l'Afrique du Sud, le Nigeria et le Kenya ont

adopté des lois de protection des données pour répondre aux défis croissants liés à la confidentialité numérique. En Côte d'Ivoire, la *Loi n° 2013-450* relative à la protection des données à caractère personnel témoigne de l'engagement du pays à garantir la confidentialité des informations dans un contexte numérique évolutif, avec des autorités de régulation et des sanctions pour assurer le respect des normes de protection des données.

III. REVUE DE LITTERRATURE

1. Revue de littérature théorique

a) Évolution des méthodes de protection contre les attaques de ré-identification

L'évolution des modèles d'anonymisation des données témoigne d'une dynamique constante entre la protection des informations privées et les techniques d'attaque sophistiquées. Le k-anonymat, introduit par Samarati et Sweeney (1998), a marqué un tournant en rendant chaque enregistrement indistinguable au sein d'un groupe de k-1 autres enregistrements, ce qui diminue le risque de ré-identification. Toutefois, ce modèle a montré ses limites face aux attaques d'homogénéité et par fond de connaissance. Pour pallier ces faiblesses, Wang et Fung (2006) ont proposé le (X, Y)-anonymat, une extension du k-anonymat qui impose des contraintes supplémentaires sur les relations entre les ensembles d'attributs. Ce modèle améliore la résistance aux attaques par correspondance en tenant compte des structures et interdépendances des données, mais sa mise en œuvre devient plus complexe.

La l-diversité, introduite également en 2006 par Machanavajjhala et al., va plus loin en garantissant une diversité suffisante des valeurs sensibles dans chaque groupe d'équivalence, rendant plus difficile la déduction d'informations sensibles. Cependant, la réalisation d'une diversité adéquate peut parfois être un défi. En réponse aux besoins de flexibilité, Wang et al. ont élaboré en 2005 le Confidence Bounding, qui fixe un seuil de confiance sur la probabilité de deviner correctement une information sensible. Bien qu'offrant une certaine flexibilité, cette méthode reste complexe à mettre en œuvre et à ajuster selon les spécificités des données. Le modèle de t-closeness de Li et al. (2007), a été introduit pour minimiser les risques d'inférence en exigeant que la distribution des attributs sensibles dans chaque groupe soit proche de la distribution globale. Cependant, la mesure

de cette similarité peut être subjective, introduisant une certaine variabilité dans l'efficacité de la protection.

En 2011, Monreale et al. ont présenté le modèle de m-invariance, qui assure qu'un enregistrement soit indiscernable parmi m autres enregistrements, ajoutant une couche de protection supplémentaire. Ce modèle, bien que renforçant la sécurité, nécessite une sélection subjective de la valeur de m et peut rester vulnérable aux attaques d'homogénéité. Enfin, la confidentialité différentielle, développée par Dwork et al. en 2006, se distingue par son approche de protection robuste, ajoutant un bruit aléatoire contrôlé aux réponses des bases de données pour préserver la confidentialité des informations individuelles. Les variantes telles que ε-differential privacy et (ε, δ) -differential privacy offrent des ajustements flexibles en fonction des besoins spécifiques, tandis que la local differential privacy (LDP), en appliquant le bruit directement sur les données des utilisateurs avant leur partage, renforce la confidentialité en éliminant la nécessité de faire confiance à une entité centralisée.

b) Techniques d'anonymisation classiques

Les premières méthodes, telles que la suppression, ont consisté à éliminer complètement les identifiants directs pour réduire les risques de réidentification. Toutefois, cette approche, comme le soulignait Sweeney (2002), ne suffit pas toujours à prévenir la réidentification par le biais d'informations quasi-identifiantes. La généralisation a été développée pour améliorer la protection en remplaçant des valeurs spécifiques par des valeurs plus larges, telles que les tranches d'âge, ce qui permet d'atteindre le k-anonymat proposé par Samarati et Sweeney en 1998. Cependant, cette technique peut entraîner une perte d'information utile pour l'analyse. La perturbation est quant à elle essentielle pour atteindre la confidentialité différentielle (Dwork et al., 2006). D'autres techniques comme l'agrégation combinent les données de plusieurs individus en une seule statistique, ce qui réduit les risques de réidentification tout en préservant les tendances globales. Enfin, le masquage, qui remplace les données sensibles par des valeurs fictives ou générées aléatoirement, a été développé pour offrir une protection supplémentaire, bien que Narayanan et Shmatikov (2008) aient montré que même les données masquées peuvent rester vulnérables à des attaques sophistiquées.

c) *Techniques d'anonymisation basées sur les données distribuées*

Les techniques de sécurisation des données distribuées jouent un rôle essentiel dans l'anonymisation des données en facilitant l'analyse de données réparties entre plusieurs parties. Elles offrent des approches complémentaires pour l'anonymisation des données dans des contextes distribués. Tout d'abord, *le calcul multipartie sécurisé (SMC),* développé par Andrew Yao en 1982, permet à plusieurs parties de collaborer pour effectuer des calculs sur des données privées sans révéler ces données entre elles. Malgré les défis liés aux ressources de communication et de calcul. Des améliorations telles les *circuits cryptés* (Garbled Circuits) ont été introduites pour accroître l'efficacité et la scalabilité de SMC. Plus tard, on verra apparaitre une nouvelle technique : le *chiffrement homomorphique*, popularisé par Craig Gentry (2009), qui permet de réaliser des calculs sur des données chiffrées sans avoir besoin de les déchiffrer. Cela garanti ainsi la confidentialité tout au long du processus de traitement et d'analyses sur des données sensibles. Bien que coûteux en ressources computationnelles, les récents efforts pour optimiser les algorithmes de chiffrement homomorphique ont rendu cette approche plus viable à grande échelle. Enfin, *l'apprentissage fédéré,* introduit par McMahan et al. (2017), entraîne des modèles sur des données réparties sur plusieurs dispositifs ou serveurs tout en gardant les données localement. Dans son principe, l'anonymisation est conservée à travers la construction de modèles prédictifs sans centraliser les données sensibles, réduisant ainsi les risques de divulgation d'informations personnelles. Bien que l'apprentissage fédéré fasse face à des défis liés à l'hétérogénéité des données distribuées, des solutions telles que *le cryptage différentiel* et les mises à jour sécurisées des modèles ont été développées pour renforcer sa sécurité et sa robustesse.

d) *Anonymisation basée sur les données synthétiques*

Depuis l'introduction des *Réseaux Antagonistes Génératifs (GAN)* par Ian Goodfellow (2014), qui utilisent un générateur et un discriminateur pour créer des données synthétiques, l'anonymisation à partir de données synthétiques a évolué pour intégrer des techniques de confidentialité différentielle afin de surmonter les risques de ré-identification. En 2019, le *Conditional Tabular GAN (CTGAN)* a été développé pour améliorer la gestion des dépendances entre colonnes dans les données tabulaires, offrant une meilleure préservation des relations statistiques. Parallèlement, la méthode de *Copula Gaussian*, utilise des fonctions de copules pour conserver les structures de dépendance complexes des données réelles (Liu et al. ;2020). *Les autoencodeurs variationnels*

(VAE), introduits par Kingma et Welling (2013), ont également contribué à cette avancée en apprenant à encoder et décoder les données dans une distribution latente, avec le *Tabular Variational AutoEncoder (TVAE)* comme une version optimisée pour les données tabulaires. Enfin, des outils tels que *Synthpop* et *DataSynthesizer* exploitent des méthodes statistiques et algorithmiques variées pour générer des données synthétiques, offrant des alternatives robustes pour l'anonymisation tout en préservant les propriétés originales des ensembles de données (Nowok et al., 2016).

2. *Revue Empirique*

a) *Anonymisation des données de santé*

Diverses études ont exploré des méthodes d'anonymisation pour garantir la confidentialité des données de santé tout en permettant des analyses utiles. Par exemple, Gkoulalas-Divanis et al. (2014) ont étudié l'application de modèles telles que le k-anonymat, l-diversité et t-closeness sur les dossiers de santé électroniques. Ces méthodes réduisent significativement les risques de ré-identification tout en maintenant une utilité élevée des données pour les chercheurs. Cependant, elles peuvent entraîner une perte d'information notable, affectant la précision des analyses, et présentent des défis en termes de complexité et de coût computationnel. De plus, Johnson et al. (2016) ont travaillé sur la base de données *Medical Information Mart for Intensive Care III* (MIMIC-III), contenant des informations de santé anonymisées sur des patients en soins intensifs. Ils ont utilisé des techniques de confidentialité différentielle pour anonymiser ces données. Toutefois, l'ajout de bruit pour garantir la confidentialité affecte la précision des modèles prédictifs, et la mise en œuvre de cette méthode dans des ensembles de données vastes et hétérogènes comme MIMIC-III reste complexe.

b) *L'Anonymisation des données d'enquête sociale en Afrique*

En Afrique, l'évolution de l'anonymisation des données d'enquête sociale révèle une progression constante dans l'adaptation et l'application de techniques pour protéger la confidentialité des données sensibles tout en maximisant leur utilité pour la recherche. Par exemple, El Emam et al. (2013) ont mené une étude en Afrique de l'Ouest, utilisant le k-anonymat pour anonymiser les données démographiques et de santé. La méthode s'est révélée efficace, mais il a été nécessaire de

l'ajuster pour tenir compte des spécificités des populations locales. Cinq (05) ans plus tard, Kinyanjui et ses collègues au Kenya ont appliqué les techniques d'anonymisation en utilisant la suppression de données, la généralisation et la permutation pour en réduire le risque de ré-identification. Cependant, cette étude a révélé des défis, notamment la perte de précision dans les analyses en raison de la suppression ou de la généralisation excessive des données. En 2017, *Statistics South Africa* a utilisé la confidentialité différentielle pour anonymiser les données du recensement. Cela consistait à ajouter du bruit aux données pour masquer les informations sensibles tout en permettant des analyses précises. Bien qu'elle offre une protection robuste contre la ré-identification, même lorsque plusieurs sources de données sont combinées, elle nécessite des ressources computationnelles importantes et une expertise technique avancée, posant des défis dans les contextes à ressources limitées. Enfin, Mwangi et al. (2020) ont appliqué des techniques de perturbation des données pour anonymiser les informations des agriculteurs en Afrique de l'Est. Cela a été efficace pour protéger les informations des agriculteurs, mais elle nécessite une calibration précise pour éviter de compromettre la qualité des données. Par ailleurs, elle a mis en évidence l'importance de l'ajustement des méthodes d'anonymisation aux spécificités locales ou communautaires pour garantir que les données restent utiles pour les analyses tout en assurant une protection adéquate de la confidentialité.

c) *Utilisation des données synthétiques : une revue empirique*

L'utilisation des données synthétiques est devenue une méthode de plus en plus adoptée pour pallier les limitations de partage de données sensibles tout en conservant la possibilité de réaliser des analyses utiles et précises.

i. *Dans le domaine de la santé*

Choi et al. (2017) ont utilisé des GAN pour créer des données synthétiques de dossiers médicaux électroniques. Ils ont démontré que les modèles prédictifs formés sur des données synthétiques présentaient une performance similaire à ceux formés sur des données réelles, tout en garantissant la confidentialité des patients. Cependant, ils ont noté que les GAN peuvent introduire des complexités computationnelles et nécessitent une grande quantité de données pour un entraînement efficace. De même, Beaulieu-Jones et al. (2019) ont utilisé des auto-encodeurs variationnels (VAE) pour générer des données de soins intensifs. Ils ont constaté que les données synthétiques maintenaient une grande fidélité statistique et étaient suffisamment réalistes pour être utilisées dans

des recherches ultérieures sans risque de ré-identification.

ii. Dans le secteur financier

Patki et al. (2016) ont démontré l'efficacité des techniques de génération de données synthétiques, incluant les GAN et les modèles basés sur les copules, pour les transactions financières. Leur approche a permis de générer des données synthétiques tout en conservant les propriétés statistiques des données originales. Cela est important pour les banques et les institutions financières qui doivent partager des données avec des chercheurs ou des partenaires tout en respectant les régulations strictes de confidentialité. Cependant, l'étude a révélé que la complexité de ces méthodes peut entraîner des défis en termes de mise en œuvre et de coûts computationnels. Xu et al. (2019) ont utilisé les GAN pour les données de transactions financières et ont montré que ces modèles peuvent créer des ensembles de données réalistes qui conservent la distribution et les corrélations des données originales, facilitant ainsi la détection de fraudes et d'autres analyses financières. Ils ont également noté que l'utilisation des GAN peut être limitée par la nécessité de données de haute qualité pour l'entraînement.

iii. Dans le contexte des enquêtes sociales

El Emam et al. (2015) ont souligné que les données synthétiques pouvaient réduire les risques de divulgation tout en fournissant des résultats analytiques comparables à ceux obtenus avec des données réelles. Cependant, ils ont mentionné que les techniques d'anonymisation doivent être soigneusement calibrées pour éviter une perte excessive de précision. Snoke et al. (2018) ont étudié l'impact des données synthétiques sur l'analyse des enquêtes de santé et ont conclu que les données synthétiques permettaient de préserver la confidentialité tout en conservant des structures de données essentielles pour les analyses statistiques. Cependant, ils ont souligné que l'introduction de bruit pour garantir la confidentialité peut parfois compromettre la qualité des données synthétiques.

IV. CONCLUSION

La revue de la littérature et l'élaboration du cadre conceptuel ont permis de mettre en lumière les diverses approches et techniques d'anonymisation des données, ainsi que leurs avantages et inconvénients. Les méthodes traditionnelles, telles que le k-anonymat et la l-diversité, restent des

outils essentiels pour protéger la confidentialité des données, mais elles présentent des limites face aux attaques de plus en plus sophistiquées. Les techniques de génération de données synthétiques, notamment celles basées sur les GAN et les copules, offrent des solutions prometteuses pour préserver à la fois la confidentialité et l'utilité des données, bien qu'elles nécessitent des ajustements et une calibration précise pour éviter les biais.

La compréhension des risques de réidentification et des modèles de protection est fondamentale pour concevoir des stratégies d'anonymisation efficaces. Par ailleurs, les avancées dans les méthodes d'apprentissage profond et les techniques de cryptographie apportent de nouvelles perspectives pour renforcer la protection des données sensibles. En conclusion, le choix des techniques d'anonymisation doit être guidé par les objectifs spécifiques de protection et d'utilisation des données, en tenant compte des exigences légales et éthiques.

APPROCHE METHODOLOGIQUE

I. INTRODUCTION

Dans cette étude, notre objectif principal est d'évaluer l'efficacité des méthodes d'anonymisation et de génération de données synthétiques dans la préservation de la confidentialité et de l'utilité des données. Pour ce faire, nous avons adopté une méthodologie qui combine des techniques d'anonymisation traditionnelles et des modèles de génération de données synthétiques. Ce chapitre détaille les étapes méthodologiques suivies, les choix de paramétrisation effectués, ainsi que les outils et techniques utilisés pour atteindre nos objectifs de recherche.

Nous commençons par l'analyse du contexte d'anonymisation de notre base de données. Cette analyse est inspirée de la procédure SDC classique. Ensuite, nous décrivons les modèles utilisés dans cette étude, notamment la procédure SDC classique, les GAN et les copules gaussiennes. Nous présentons également les critères de performance utilisés pour évaluer l'efficacité des méthodes d'anonymisation, y compris la perte d'information et le risque de divulgation. Enfin, nous détaillons la procédure d'application des techniques d'anonymisation et des modèles génératifs sur les données réelles, en expliquant les choix méthodologiques et les justifications.

II. PHASE EXPLORATOIRE : AVANT L'APPLICATION DES METHODES

1. Évaluation de la nécessité de la confidentialité du point de vue légal

Les règlements sur la protection des données personnelles en Côte d'Ivoire, tels que définis par l'ARTCI, énoncent des directives strictes pour la collecte, le traitement et l'utilisation des données à caractère personnel. Dans le cadre de notre analyse pour évaluer la nécessité de la confidentialité des données, il est impératif de prendre en compte ces réglementations.

Premièrement, l'objectif et le champ d'application de la loi ARTCI définissent que toute entité, qu'elle soit publique ou privée, est tenue de respecter les normes de protection des données lors de la collecte, du traitement, de la transmission et de l'utilisation de données personnelles. Cela souligne l'importance de garantir la confidentialité des données tout au long de leur cycle de vie.

Deuxièmement, les formalités nécessaires aux traitements des données à caractère personnel, telles que la déclaration préalable à l'autorité de protection des données, démontrent que tout traitement de données doit être conforme aux exigences légales en matière de protection des données. Cette procédure garantit que les responsabilités en matière de protection des données sont maintenues et que les traitements, en particulier ceux impliquant des données sensibles, sont soumis à un examen attentif par les autorités compétentes.

Enfin, les principes directeurs du traitement des données à caractère personnel énoncés par ARTCI mettent en évidence l'importance d'une anonymisation totale et irréversible des données pour prévenir la réidentification. Cela signifie que les mesures de confidentialité doivent être mises en place de manière robuste pour garantir que les données personnelles restent protégées, même en cas de tentative de dé-anonymisation.

Nous retenons que l'analyse des réglementations ARTCI confirme que le traitement des données pour protéger leur confidentialité est non seulement requis mais également une exigence légale en Côte d'Ivoire. En conséquence, l'utilisation de méthodes d'anonymisation est essentielle pour garantir la conformité aux réglementations en vigueur tout en préservant l'utilité des données en cas de publication de celles-ci.

2. Présentation de la base de données et hypothèses

La base de données que nous utilisons provient de l'Enquête Académique 2024 de l'ENSEA, réalisée à Korhogo sur le bien-être des jeunes et les potentialités économiques. La base de données est sous forme brute et n'a pas encore suivi le processus de traitement des données. Les unités statistiques sont donc des jeunes issus des ménages. Elle est issue d'un échantillonnage de type probabiliste, mais nous n'avons pas les poids de sondage. Elle contient 951 variables et plus de 1000 lignes. De façon générale, les techniques d'anonymisation sont appliquées à une base qui peut dans son état actuel faire l'objet de publication. *Nous émettons donc les hypothèses de travail suivantes :*

- ***Hypothèse 1 :*** *La base de données est prête pour la publication et nécessite uniquement l'application de techniques d'anonymisation pour protéger les individus ;*
- ***Hypothèse 2 :*** *on suppose que la base de données n'est pas issue d'un échantillonnage aléatoire, donc qu'il n'y a pas de poids de sondage.*

L'hypothèse 2 a une implication importante dans notre méthodologie. Cela nous permet de faire abstraction du niveau ménage de la base de données et de considérer que nous avons uniquement des jeunes gens comme individus. *Sinon, dans la pratique, nous aurions procédé à l'anonymisation des variables au niveau du ménage avant de passer au niveau individuel, conformément à la procédure SDC classique.*

3. Identification et consolidation des variables clé

Nous commençons par une classification des variables de la base de données. Les variables sensibles incluent des informations sur la santé, la situation financière, et d'autres aspects très personnels, dont la divulgation pourrait mener à des discriminations, des stigmatisations ou d'autres impacts négatifs.

1. **Retrait des identifiants directs** : Nous retirons les identifiants directs de la base de données *(ex. : nom et prénoms).*
2. **Consolidation des variables** : Nous modifions la base de données pour réduire le nombre de variables liées possibles. Cette étape est importante pour minimiser les incohérences dans les données anonymisées.

*Exemple de consolidation : La section B du questionnaire de l'enquête retrace les activités d'une journée entière des jeunes avec 456 variables, résultant de la répartition des 24 heures entre 16 activités, plus 1 activité quelconque à signaler, 1 variable pour préciser l'activité, et 1 variable pour les réponses à une question à choix multiple. Cela crée 24 * (16 + 1 + 1) = 456 variables liées. Pour éviter les incohérences, nous avons consolidé ces variables en calculant le temps affecté à chaque activité, réduisant ainsi les variables à 16+2 (résumé des activités pour le travail et le loisir), soit une réduction de 99,60%.*

Scores de bien-être : En se référant au questionnaire, nous avons réduit 12 variables en un score de bien-être mental compris entre 0 et 12. Nous avons également construit le score de relation avec les proches et le score de relation avec les autres à partir des variables des sous-sections de la section D. Nous avons ensuite supprimé les variables non informatives *(ex. : variables à modalité unique, variables avec plus de 90% de valeurs manquantes, variables redondantes).*

NB : *À cette étape, si nous avions des poids de sondages ainsi que la méthodologie de l'enquête, nous devrions les décrire et ajuster les poids si nécessaire.*

Nous avons constaté que certaines activités ne sont quasiment pas pratiquées, nous pouvons encore réduire en transformant les 18 variables en une variable **temps de travail (ou un autre temps de loisir)** permettant d'évaluer le temps allouer aux activités qui ont pour conséquence directe de réduire l'utilité des individus. *Pour ce qui est du temps allouer au loisir, il peut être déduit de l'autre. Nous utilisons le questionnaire dans lequel il est spécifié des différentes distinctions.*
Nous avons opté pour la variable temps de loisir.
Ce genre d'ajustements ont été effectués maintes fois pour obtenir une base avec suffisamment de données informatives.

4. Choix du type de diffusion et élaboration des scénarios d'intrusion

a) Type de diffusion

Le type de diffusion dépend des besoins des producteurs de données, des responsables du traitement pour la confidentialité et des utilisateurs potentiels. Il faut commencer par prioriser les variables à publier. Nous allons toutefois appliquer nos méthodes à toute la base restante à chaque étape. Les choix de variables s'opéreront de façon endogène au processus d'anonymisation mis en place.

Nous décidons de produire deux types de fichiers à publier en fonction du public cible.

- ✓ Nous établissons une première anonymisation de type *Fichiers à Usage Scientifique* (*Scientific Use Files* SUF) nécessitant un faible niveau de protection et un niveau élevé d'utilité.
- ✓ Nous anonymisons ensuite les données initiales sur la base d'une publication de type *Fichiers à Usage Public (Public Usage Files : PUF). Ce type de diffusion nécessite une forte anonymisation des données.*

b) Elaboration des Scénarios d'Intrusion

Au niveau des scénarios potentiels de divulgation et les variables clés utilisées pour identifier les individus, nous en dégageons *Trois (03)* **:**

i. *Scénario 1 : Intrusion basique*

- ✓ *Hypothèse* : L'attaquant dispose uniquement des données publiquement disponibles.
- ✓ *Méthode :* Tentative de ré-identification à partir des données générales disponibles sur Internet sans accès à des informations supplémentaires spécifiques ou d'autres sources comme le Recensement Général de la Population et de l'Habitat (RGPH)

ii. *Scénario 2 : Intrusion avec données Quasi-Identifiantes*

- ✓ *Hypothèse* : L'attaquant dispose de données quasi-identifiantes telles que l'âge, le sexe et la localité.
- ✓ *Méthode* : Il utilise des techniques de correspondance de modèles pour relier les enregistrements anonymisés à des individus spécifiques en utilisant des informations quasi-identifiantes.

iii. *Scénario 3 : Intrusion avancée*

- ✓ *Hypothèse* : L'attaquant dispose de sources de données multiples et de capacités de calcul avancées.
- ✓ *Méthode* : Combinaison de données de plusieurs sources et application de techniques de machine learning pour tenter de ré-identifier les enregistrements anonymisés.

5. *Sélection des variables clé*

a) Quasi-identifiants

Les quasi-identifiants qualitatifs choisis sont :
id04 (Localité), **id09** (Milieu de résidence), **cj2** (Sexe), **cj3** (Lieu de naissance), **cj4** (Résidence continue), **cj7** (Lieu de provenance), **cj10** (Actuellement à l'école), **cj13** (Statut d'occupation), **cj15**

(Emploi précédent), **cj17** (Secteur d'activité), **bn20** (Tranche de revenu), **bn41** (Ordinateur personnel), **cas1** (Permis de conduire)

Les quasi-identifiants quantitatifs sont des données numériques :
cj_age (Âge), **cj5** (Résidence actuelle), **bn7** (Poids), **bn6** (Taille)

b) Variables sensibles

Ces variables, qu'elles soient quantitatives ou qualitatives, leur divulgation pourrait compromettre la vie privée et la sécurité des individus en exposant des informations personnelles critiques touchant à la santé mentale, aux finances, aux relations sociales, aux expériences personnelles et à d'autres aspects délicats
Les variables sensibles quantitatives sont des mesures numériques dans notre base incluent des informations sur la santé mentale, les économies mensuelles, et les relations sociales.

NH (Temps de loisir total), **bn2345** (Relation avec les autres), **bn450** (Dépression ou anxiété), **bn70** (Bien-être mental), **bn26** (Économies mensuelles)

Les variables sensibles qualitatives couvrent divers aspects comme les raisons de migration, le statut économique et familial, la consommation de substances, et les expériences personnelles.

cj9 (Raison venue à Korhogo), **cj11** (Raison arrêt des études), **cj19** (Obtention emploi principal), **cj20** (Autre emploi), **cj22** (Démarches emploi semaine écoulée), **cj24** (Statut vital père avant 15 ans), **cj25** (Statut vital mère avant 15 ans), **cj26** (Résidence à 15 ans), **cj27** (Statut des parents biologiques), **cj28** (Niveau d'instruction du tuteur), **cj29** (Niveau d'instruction du père), **cj30** (Niveau d'instruction de la mère), **cj31** (Tuteur travaillait-il?), **cj32** (Difficultés financières de la famille), **cj33** (Statut économique du soutien financier), **cj34** (Type d'entreprise du soutien financier), **cj35** (Secteur d'activité du soutien financier), **cas3_0** (Relation avec la famille), **cas4** (Proche consommant une substance

avant 15 ans), **cas6** (Présence de maquis, bars, fumoirs), **cas7** (Consommation d'alcool), **cas21** (Consommation de tabac), **bn1** (Aide pour aller de l'avant), **bn10** (Souffrance d'une maladie chronique), **bn12** (Existence de maladies héréditaires), **bn14** (Satisfaction de l'état de santé), **bn15** (Victime d'un crime ou d'une agression), **bn18** (Sentiment d'être rejeté), **mt1** (Situation socio-professionnelle).

III. PRESENTATION DES MODELES UTILISES

Dans cette partie, nous présenter l'ensemble de modèles convoqués dans notre démonstration. On commence par les procédures SDC classiques appliquées et ensuite, on présente les méthodes de génération de données synthétiques utilisés. Les modèles utilisés sont proposés en raison des divers scénarii présentés. Le SDC classique servira à nous protéger contre le scénario 1 et les méthodes génératives contre les scénarii 1, 2 et 3.

1. Procédure d'anonymisation par SDC classique

Nous avons utilisé des méthodes comme la suppression locale pour atteindre le k-anonymat et la l-diversité. De plus, les méthodes perturbatrices ont été appliquées.

a) Ajout de bruit

Ces méthodes introduisent de l'incertitude sans altérer significativement la structure des données. Elles sont fondées pour la plupart sur le principe de masquage matriciel en appliquant à la base de données *(X)* par une transformation du type : $\boldsymbol{Z = AXB + C}$, où *A* modifie les enregistrements (lignes) et *B* les variables (colonnes) et *C* est un bruit. Les techniques de perturbations consistent tout simplement à calculer Z.

Perturbation par un bruit additif : On applique $Z = X + \epsilon$. Tout dépend de la distribution de ϵ.

Bruits indépendants : $X \sim N(\mu, \Sigma); \epsilon \sim N(0, \Sigma_\epsilon)$ avec $\Sigma_\epsilon = \alpha \ diag({\sigma_1}^2, \dots, {\sigma_p}^2)$

Où ${\sigma_i}^2 = Var(X_i)$. Dans cette spécification, la moyenne et la covariance sont préservées et les coefficients de corrélation linéaires ne le sont pas. *(En effet,* $Var(\boldsymbol{Z_j}) = Var(\boldsymbol{X_j}) + \alpha Var(\boldsymbol{X_j}) =$

$(1+\alpha)$ *Donc,* $\rho_{Z_j,Z_i} = \frac{1}{1+\alpha}\,\rho_{X_j,X_i} \neq \rho_{X_j,X_i}$).

Bruits corrélés : *On choisit donc* ϵ *telle que* $\boldsymbol{\Sigma_\epsilon} = \alpha\,\boldsymbol{\Sigma}$. Cela nous permet d'avoir $\boldsymbol{\Sigma_Z} = \boldsymbol{\Sigma} + \alpha\,\boldsymbol{\Sigma} = (1+\alpha)\boldsymbol{\Sigma}$. Donc $\rho_{Z_j,Z_i} = \frac{1+\alpha}{1+\alpha}\,\rho_{X_j,X_i} = \rho_{X_j,X_i}$, la corrélation linéaire est est conservée. La matrice de variance variance-covariance des variables perturbées est toutefois biaisée. Nous préférons toutefois, cette spécification à l'autre**.** Mais les deux ne sont pas beaucoup utilisées en raison de leur faible apport en termes de protection de données, en particulier contre les individus extrêmes pour la variable concernée *(Hundepool et al., 2012).* Ainsi, elle peut protéger contre la divulgation d'attributs mais pas contre la divulgation d'identité. *Par exemple, même si le chiffre d'affaires des entreprises d'une base de données est modifié par ces méthodes, l'identité de l'entreprise ayant le plus grand chiffre d'affaires dans un secteur reste connue.*

Bruit et transformation linéaire : On peut corriger le biais introduit dans la matrice lors de l'ajout de bruits corrélés. En effet, on appliqe une transformation linéaire à la suite du bruit.

$Z = c(X+\epsilon) + D$ où X et ϵ sont spécifiées comme ci-dessus et les paramètres c et d_j *sont tels que :* $E(\boldsymbol{Z_j}) = E(\boldsymbol{X_j})$ et $Var(\boldsymbol{G_j}) = Var(\boldsymbol{X_j})$. *C'est-à-dire* $d_j = (1-c)E(X_j)$

NB : les paramètre *c doit rester secret, sinon un utilisateur pourrait revenir aux données initiales. C'est comme une clé de cryptage en cryptographie.*

Comme on peut le constater, l'ajout de bruits additifs n'a quasiment pas d'effet sur les très grandes valeurs mais modifie significativement les plus faibles. D'autres méthodes dont été développées pour améliorer ce défaut, notamment à travers l'introduction de bruits multiplicatifs.

Perturbation par un bruit multiplicatif : On applique $Z = W \bullet X$ *où* $\bullet$ est le produit matriciel de Hadamard, (produits composants par composante) où $\boldsymbol{W}$ est une matrice dont les colonnes sont des variables d'espérance $\mathbf{1}$ et de variance $\sigma_W^2 > 0$. Dans *sdcMicro*, un package présent dans le logiciel R, ; on a la méthode de perturbation aléatoire par une matrice orthogonale (*Random Orthogonal Matrix Maskin*g (ROMM) en anglais)**.** Elle consiste à générer une matrice orthogonale aléatoire $\boldsymbol{A}$, issue d'une distribution G définie sur le groupe des matrices orthogonales qui laissent le vecteur colonne identité $\mathbf{1}_n$ invariant.

b) *La méthode PRAM (Post-RAndomisation Method)*

La méthode PRAM (Post-RAndomisation Method) applique une perturbation aléatoire aux données

catégorielles en définissant des probabilités de transition des modalités d'une variable *X* à une variable perturbée *Z.* La probabilité de passage de la modalité *k* de *X* à la modalité ***l*** de *Z* est notée $p_{kl} = P(Z = l|X = k)$. Chaque enregistrement du fichier de données est soumis à cette procédure. *Il est impératif de ne pas diffuser la matrice de transition, car cela permettrait aux utilisateurs de retrouver les données initiales sans biais.* La matrice PRAM peut parfois produire des modifications *incohérentes*, comme changer une femme enceinte en homme enceinte. Pour éviter cela, on utilise une matrice PRAM invariante qui maintient en moyenne la distribution de la variable perturbée.

2. *Modèles de génération de sonnées synthétiques*

a) *Structure de Base du GAN*

Le GAN se compose de deux réseaux de neurones : un générateur G et un discriminateur D.

Le générateur G prend un vecteur aléatoire z comme entrée et produit des données synthétiques $G(z)$. *Le discriminateur D* tente de distinguer entre les données synthétiques $G(z)$ et les données réelles x .

Fonction de perte :

Les fonctions de perte pour le GAN sont essentielles pour son apprentissage. Elles guident les ajustements des poids des réseaux du générateur et du discriminateur.

$$min_G max_D E_{x\sim p_{data}}[logD(x)] + E_{z\sim p_z}\left[log\left(1 - D\big(G(z)\big)\right)\right]$$

$z \sim p_z$ *Signifie que z provient de la distribution des vecteurs aléatoires*

$x \sim p_{data}$ *Signifie que les données* x *proviennent de la distribution des données réelles*

- ✓ $E_{x\sim p_{data}}[logD(x)]$*:* Cette partie de la fonction de perte maximise la probabilité que le discriminateur D classe correctement les données réelles x comme étant réelles.
- ✓ $E_{z\sim p_z}\left[log\left(1 - D\big(G(z)\big)\right)\right]$ *:* Cette partie minimise la probabilité que le discriminateur D classe incorrectement les données synthétiques $G(z)$ comme étant réelles.
- ✓ Le générateur G est entraîné pour minimiser $log\left(1 - D\big(G(z)\big)\right)$, ce qui signifie qu'il apprend à produire des données $G(z)$ que le discriminateur ne peut pas distinguer des données réelles.

Pour le discriminateur D : $min_D \frac{1}{2}E_{x\sim pdata}[(D(x)-b)^2] + \frac{1}{2}E_{z\sim pz}\left[\left(D\left(G(z)\right)-a\right)^2\right]$ *b et a sont des constantes qui représentent les labels pour les vraies et fausses données respectivement.*

Pour le générateur G : $min_G \frac{1}{2}E_{z\sim pz}\left[\left(D\left(G(z)\right)-\delta\right)^2\right]$ *où δ est une constante représentant le label idéal que le générateur souhaite atteindre pour tromper le discriminateur.*

b) Le CTGAN (Conditional Tabular GAN)

Il s'agit d'une version des GAN dans laquelle le générateur produit des données synthétiques à partir d'un vecteur de bruit et de variables conditionnelles (c) , tandis que le discriminateur essaie de distinguer les données synthétiques des données réelles. La fonction de perte est donnée par :

$$min_G max_D E_{x\sim p_{data}}\left[logD(x)\right] + E_{z\sim p_z}\left[log\left(1-D\left(G(z|c)\right)\right)\right]$$

c) Le modèle de génération basé sur les copules gaussiennes (Copula Gauss)

Dans ce modèle, on utilise les copules pour modéliser les dépendances entre les variables et générer des données synthétiques. *Une copule est une fonction qui lie les marges univariées d'une distribution multivariée, permettant de modéliser des dépendances complexes.* Le processus de génération implique deux étapes.

- *D'abord, modéliser la dépendance entre les variables avec une copule ;*
- *ensuite, générer des échantillons à partir de cette copule.*

d) Le Tabular Variational Autoencoder (TVAE)

Le VAE fonctionne en tirant un échantillon z de la distribution de codes $p_{model}(z)$, introduit dans un réseau générateur $g(z)$, et en échantillonnant x à partir de $p_{model}(x|z)$.L'architecture du VAE comprend un encodeur et un décodeur, ce dernier étant utilisé à des fins génératives en encodant une entrée comme une distribution sur l'espace latent plutôt qu'un point unique.

La fonction de perte du VAE combine le coût de reconstruction et la divergence de Kullback-Leibler (DKL), assurant que la distribution latente se rapproche de la loi normale *N (0,1).*

$$\mathcal{L} = E_{q(z|x)}[log p(x \mid z)] - DKL(q(z \mid x)||p(z))$$

L'astuce de reparamétrisation permet la rétropropagation à travers un nœud d'échantillonnage stochastique, en décomposant le vecteur latent comme suit : $z = \mu + \sigma\epsilon$, où μ et σ sont des paramètres appris, et ϵ est $N(0,1)$. Cette décomposition permet de séparer les parties entraînables et stochastiques du modèle, facilitant l'entraînement efficace du VAE.

e) Le copule-GAN

Le Copula-GAN combine les copules et les réseaux adversariaux génératifs (GAN) pour générer des données synthétiques en conservant les dépendances complexes entre les variables. Il les copules pour modéliser les relations entre les variables et les GAN pour produire des données réalistes. Le processus de génération de données avec le Copula-GAN se déroule en trois étapes principales.

- *Tout d'abord, les copules sont utilisées pour capturer les dépendances entre les variables d'origine;*
- *ensuite, un GAN est entraîné sur ces variables transformées pour générer des données synthétiques.*
- *enfin, les échantillons générés par le GAN sont transformés en utilisant les copules pour obtenir des données synthétiques qui conservent les structures de dépendance des données d'origine.*

IV. APPLICATION DES MODELES

On a deux catégories de modèles. Les modèles de génération de données synthétiques et d'un autre côté, les modèles classiques de protection.

1. Processus SDC

Tout d'abord, *le k-Anonymat* est appliqué avec *k=3* pour s'assurer que chaque enregistrement partage les valeurs de quasi-identifiants avec au moins deux autres enregistrements. Cela réduit le risque de réidentification en formant des groupes de taille minimale. Une évaluation du risque de divulgation est effectuée pour chaque variable sensible. Les cinq variables sensibles présentant le plus haut risque de divulgation sont identifiées pour des mesures supplémentaires de protection.

Ensuite la l-diversity est appliquée avec *l=3* sur les cinq (05) variables sensibles les plus risquées. Cela garantit une diversité suffisante des valeurs sensibles au sein de chaque groupe de quasi-identifiants, rendant plus difficile la réidentification des individus. Pour les autres variables sensibles catégorielles, la méthode PRAM est appliquée, offrant ainsi une protection supplémentaire contre la réidentification. Quant aux variables numériques, un bruit corrélé de niveau *0,5* leur est ajouté.

2. *Les méthodes génératives*

On utilise les **modèles CTGAN** (Conditional Tabular GAN), **Copula Gauss** (Gaussian Copula), **TVAE** (Tabular Variational Autoencoder), **CopulaGAN** (Copula GAN).

Après Apprentissage de la structure des données. On peut générer autant de données qu'on veut. On peut même le faire selon des distributions calées à l'avance. *Ici, nous appliquons ensuite une génération conditionnellement à la variable* ***id05*** *représentant la Zone de Dénombrement (ZD). Cela permet de générer les individus sachant que la ZD est donnée. Cette façon de faire permet de construire des bases quasiment similaires aux données initiales. Le choix des variables de génération conditionnelles impacte sur la similitude des données générées*

3. *Indicateurs d'évaluation du risque sur les bases de données*

a) *Risque de divulgation*

Le risque de divulgation est généralement évalué à partir de la probabilité qu'une observation individuelle puisse être ré-identifiée dans les données anonymisées. Une méthode courante pour calculer ce risque est la suivante **:**

Fréquence relative : Calculer la fréquence relative de chaque combinaison de quasi-identifiants dans les données anonymisées.

En effet, étant donné F_c le nombre d'individus dans la population de provenance de la base (population totale) ayant la clé ***c*** *(combinaison de QID), la* probabilité de réidentification pour la clé *c*, est $\frac{1}{F_c}$. Il s'agit du risque, et si F_c est inconnue, on l'estime à travers un modèle statistique. Et dans ce cas, le risque associé à la clé *c* est $r_c = E\left(\frac{1}{F_c}\middle|\ f_c\right)$. *(Avec* f_c *la fréquence de la clé dans*

l'échantillon)

Risque global : Le risque global de divulgation peut être calculé comme la somme des inverses des fréquences relatives des combinaisons uniques de quasi-identifiants.

b) Categorical CAP (Classification Accuracy for Privacy)

La métrique CategoricalCAP évalue le risque de divulgation d'informations sensibles en mesurant la difficulté pour un attaquant de deviner correctement ces informations à partir de données synthétiques. Elle utilise l'algorithme Correct Attribution Probability (CAP), qui identifie les lignes correspondantes dans les données synthétiques pour tenter de deviner les valeurs sensibles des données réelles. Le score final varie de 0 à 1, où 1 indique une protection totale et 0 une vulnérabilité totale. Cette métrique est conçue pour des données catégorielles et booléennes, mais ne supporte pas encore les valeurs manquantes.

4. Indicateurs de mesure de l'utilité des données

a) Perte d'information (IL1)

La IL1 quantifie la perte d'information en comparant les distributions des variables avant et après l'anonymisation. On calcule la somme des déviations observées chez tous les individus pour les variables continues.

b) Différence des valeurs propres (Eigenvalues)

La différence des valeurs propres est utilisée pour évaluer la préservation de la structure statistique des données. Elle compare les matrices de covariance des données originales et anonymisées.

NB : Celui-là et la IL1 sont valables dans l'application de la procédure SDC classique

c) CSTest (Column Similarity Test)

Le CSTest utilise le test du chi-carré pour comparer les distributions des données synthétiques et réelles, en mesurant la similarité entre les colonnes originales et synthétiques. Des valeurs plus élevées (proches de 1) indiquent une plus grande similarité. Il est compatible avec les données catégorielles. Nous l'utiliserons pour évaluer l'utilité des données catégorielles.

d) *KS Complement (Kolmogorov-Smirnov Test Complement) :*

Compatible avec les données numériques continues et les valeurs datetime converties en valeurs numériques, elle ignore les valeurs manquantes. Utilisant la statistique de *Kolmogorov-Smirnov*, elle transforme une distribution numérique en fonction de distribution cumulative (CDF) et mesure la différence maximale entre les CDFs des données réelles et synthétiques. Un score de 1 indique une similarité exacte, tandis qu'un score de 0 indique une dissimilarité totale.

e) *Continuous KL Divergence et la Discrete KL Divergence*

Ces deux (02) métriquent utilisent la divergence de Kullback-Leibler (KLD) pour comparer la similitude entre les données réelles et synthétiques. La **ContinuousKLDivergence** s'applique aux colonnes de données continues. Elle convertit ces valeurs en catégories par *binning*, puis calcule *l'entropie relative* entre les fréquences observées et attendues des histogrammes 2D des données réelles et synthétiques. Le score final est normalisé avec la formule $1/(1+KLD)$, allant de 0 *(aucune similitude)* à 1 (similitude maximale). Cela permet de vérifier que les données synthétiques reflètent bien les distributions des données réelles.

La **DiscreteKLDivergence** fonctionne de manière similaire mais s'applique aux données discrètes ou booléennes. Ces métriques assurent *que les relations et les dépendances présentes dans les données originales sont bien préservées dans les données synthétiques*, ce qui est important pour leur utilité.

f) *Test sur des modèles économétriques :*

En se mettant à la place des utilisateurs des données, nous effectuer des tests sur des modèles économétriques. Pour des raisons de simplicité, on choisit des modèles linéaires. Le but est d'évaluer la capacité des bases de données à reproduire les mêmes résultats que la base de données originale. *Nous voulons vérifier si l'essentiel de l'information contenue dans l'ensemble de la base de données est aussi contenu dans les versions anonymisées. On regresse à cet effet la variable 'bn70', le score de bien être mental, sur d'autres prédicteurs de l'état mental et des caractéristiques sociodémographiques… continus et qualitatifs…*

Le modèle utilisé est spécifié comme suit :

$$bn70 = \beta_0 + \beta 1 \cdot cj\_age + \beta_2\, cj2 + \beta_3 \cdot cj13 + \beta_4 \cdot cj10 + \beta_5 \cdot cas7 + \beta_6 \cdot bn2345 + \beta_7 \cdot mt1 + \epsilon$$

Avec : ***bn70****: Bien-être mental ,* ***cj_age****: Âge de l'enquêté ;* ***cj2*** *: Sexe de l'enquêté ;* ***cj13:*** *CJ13. Quel est votre statut d'occupation actuellement ?* ***CJ10****: Partez-vous actuellement à l'école ?;* ***cas7****: Avez-vous déjà consommé de l'alcool ?;* ***bn2345****: score de relation avec les autres ;* ***mt1****:Comment jugez-vous votre situation socio-professionnelle ?*

V. CONCLUSION

En conclusion, ce chapitre a présenté une méthodologie structurée et détaillée pour évaluer l'efficacité des méthodes d'anonymisation et de génération de données synthétiques. Les choix méthodologiques effectués, tels que la sélection des quasi-identifiants et des variables sensibles, ainsi que les paramètres de modélisation, ont été présentés. Cette méthodologie, combinant dans un même écrit une procédure d'application des méthodes basés sur la SDC classique et des méthodes avancées de génération de données synthétiques, offre un cadre intéressant pour les chercheurs et les praticiens souhaitant appliquer des techniques d'anonymisation et de génération de données synthétiques dans divers contextes. Cela contribue ainsi à l'avancement des pratiques de protection des données tout en facilitant les analyses statistiques fiables.

RESULTATS ET DISCUSSION

I. INTRODUCTION

Cette section présente les résultats obtenus suite à l'application des différentes techniques d'anonymisation sur notre jeu de données. *L'objectif est d'évaluer comment ces méthodes parviennent à protéger la confidentialité des individus tout en préservant l'utilité des données pour des analyses statistiques. Cela est fait relativement à la base de données originales comme référence.* Nous débutons par l'évaluation des résultats des méthodes traditionnelles d'anonymisation. Nous analysons leur capacité à réduire les risques de ré-identification et à maintenir les caractéristiques essentielles des données. Ensuite, nous comparons ces résultats avec ceux obtenus par les modèles de génération de données synthétiques, tels que les GAN (Generative Adversarial Networks), les copules gaussiennes et les autoencodeurs variationnels. Chaque modèle est évalué en fonction de critères spécifiques tels que le risque de divulgation, la perte d'information et la fidélité des distributions statistiques. Les résultats obtenus serviront de guide pour choisir les techniques les plus appropriées en fonction des besoins spécifiques de protection de la confidentialité et d'analyse des données.

II. RESULTATS DE L'ANONYMISATION

1. Résultats du processus SDC

Les résultats de l'application des méthodes classiques sont consignés dans le tableau suivant.

Tableau 5: Résultats de l'anonymisation par SDC classique

Critères	***Avant Anonymisation***	***Après Anonymisation***
2-anonymat	*92 violations (10,099%)*	*0 violations* ***(0,000%)***
3-anonymat	*311 violations (34,138%)*	*0 violations* ***(0,000%)***
5-anonymat	*448 violations (49,177%)*	*139 violations* ***(15,258%)***
Risque de divulgation	*Entre 0,00% et 100,0%*	*Entre* ***0,00%*** *et* ***14,16%***
Perte d'information (IL1)	*0*	***8460,46***
Différence des valeurs propres	*0%*	***4,580%***

Source : Auteurs

L'anonymisation des données a considérablement réduit les risques de ré-identification. Les violations de 2-anonymat et 3-anonymat ont été entièrement éliminées, indiquant une forte amélioration de la confidentialité.

Le risque de divulgation a également été réduit, passant d'une fourchette maximale de 100% à 14.16%. En contrepartie, il y a eu une perte d'information relativement faible et des modifications dans la structure des données, comme le montre l'augmentation de l'IL1 et de la différence des valeurs propres.

En résumé, l'anonymisation des données a été efficace pour atteindre le 3-anonymat et maintenir la qualité des données.

2. Résultats comparés du SDC et des autres modèles génératives

Tableau 6: Résultats de l'anonymisation avec les modèles génératifs

Modèle	Categorical CAP	CS Test	KS Complement	Continuous KL Divergence	Discrete KL Divergence	Moyenne
Base réelle	0,929	**1,000**	1,000	1,000	1,000	0,985
SDC classique	0,995	0,222	***0,984***	***0,993***	0,498	0,524
CTGAN	***1,000***	0,806	0,656	0,760	0,582	0,735
CopulaGauss	***1,000***	***0,992***	***0,772***	***0,833***	**0,897**	**0,890**
TVAE	***1,000***	**0,950**	0,610	0,676	**0,786**	**0,776**
CopulaGAN	***1,000***	0,853	0,658	0,702	0,617	0,742

Source : Auteurs

a) Analyse des résultats de chaque modèle

i. Base Réelle

La base réelle sert de référence pour évaluer les autres modèles. Les métriques de cette base montrent des valeurs maximales en termes de préservation des distributions, car il s'agit des données d'origine non altérées. Les scores pour CS Test, KS Complement, Continuous KL Divergence, et Discrete KL Divergence sont tous à 1, indiquant une parfaite fidélité à l'original. Le score moyen de 0.985 montre que les données réelles conservent leur utilité statistique presque entièrement intacte.

Cependant, cela signifie également qu'il n'y a aucune protection contre le risque de divulgation, avec un Categorical CAP de 0.929 indiquant un risque élevé. En résumé, bien que les données réelles soient idéales pour l'analyse en termes d'exactitude et d'intégrité, elles ne protègent pas du tout la confidentialité des individus.

ii. SDC Classique

Le SDC classique (Statistical Disclosure Control) applique des techniques d'anonymisation pour protéger la confidentialité des données. Avec un Categorical CAP de 0.995, le risque de divulgation est *fortement réduit*. Toutefois, les scores des tests CS (0.222) et Discrete KL Divergence (0.498) *révèlent une altération significative des distributions des données catégorielles et discrètes* en raison des suppressions et transformations appliquées pour atteindre le 3-anonymat, la 3-diversité. Le SDC classique réussit donc à protéger les données sensibles, mais au prix d'une utilité réduite. La similarité entre les distributions des colonnes réelles et anonymisées est faible, rendant les données moins fiables pour certaines analyses. Bien que les distributions continues soient relativement bien conservées (KS Complement = 0.984, Continuous KL Divergence = 0.993), l'impact global sur l'utilité des données est négatif, comme le montre le score moyen de 0.524.

iii. CTGAN

Ce modèle utilise des réseaux génératifs antagonistes pour créer des données synthétiques. Ce modèle atteint un score parfait de 1 pour le Categorical CAP, éliminant le risque de divulgation. Il a réussi à créer des données synthétiques avec une protection robuste contre la divulgatioN. Les scores pour les tests CS (0.806), KS Complement (0.656), et Continuous KL Divergence (0.760) indiquent que les distributions catégorielles et continues sont bien préservées, bien que perfectibles.

iv. Copula Gaussian

Le Copula Gaussian est un modèle basé sur des fonctions de copules pour générer des données synthétiques. Avec un Categorical CAP de 1.000, il assure une protection complète contre la divulgation. Les scores pour CS Test (0.992), KS Complement (0.772), Continuous KL Divergence (0.833), et Discrete KL Divergence (0.897) sont élevés, montrant une excellente préservation des distributions.

Le Copula Gaussian se distingue donc par sa capacité à préserver l'utilité des données tout en offrant une protection robuste contre la divulgation. Les données synthétiques générées par ce

modèle sont très similaires aux données réelles en termes de distributions, ce qui les rend idéales pour des analyses statistiques. Le modèle parvient à équilibrer efficacement confidentialité et utilité, faisant de *Copulan Gaussian* un choix de premier plan pour des applications nécessitant à la fois une forte protection des données et une haute qualité analytique.

v. TVAE

Il utilise des *autoencodeurs variationnels* pour générer des données synthétiques. Il offre une protection complète avec un Categorical CAP de 1.000. Les scores des tests CS (0.950), KS Complement (0.610), et Continuous KL Divergence (0.676) montrent une bonne, mais perfectible, conservation des distributions. Bien que les scores pour les distributions continues soient légèrement inférieurs à ceux des données catégorielles, Il réussit à générer des données synthétiques avec une protection robuste contre la divulgation et une utilité raisonnable. Les données synthétiques produites par TVAE sont particulièrement utiles pour des applications où la confidentialité est primordiale, mais où une certaine déviation des distributions originales est acceptable.

vi. Copula GAN

En combinant des techniques de réseaux génératifs antagonistes et de copules pour générer des données synthétiques. Avec un Categorical CAP de 1.000, il assure une protection complète contre la divulgation. Les scores pour CS Test (0.853), KS Complement (0.658), Continuous KL Divergence (0.702), et Discrete KL Divergence (0.617) montrent une conservation modérée des distributions. Les distributions catégorielles sont mieux conservées que les distributions continues, suggérant des améliorations possibles pour les données numériques. Ce modèle est adapté pour des situations où la confidentialité est essentielle, mais où l'utilité des données doit être maintenue à un niveau raisonnable pour des analyses fiables.

b) Synthèse

Les résultats montrent que les différents modèles d'anonymisation et de génération de données synthétiques présentent des performances variées en fonction des mesures spécifiques. La base réelle sert de référence avec des valeurs proches de 1 dans la plupart des catégories.

SDC classique montre une bonne performance en Categorical CAP et KS Test, mais une très faible performance en CS Test et Continuous KL Divergence. Des problèmes dans la préservation de certaines caractéristiques statistiques des données, ce qui est logique puisqu'il s'agit d'une version altérée de base initiale.

Bien que tous les modèles présentent des avantages et des inconvénients, ***Copula Gaussian*** *se distingue comme le modèle le plus équilibré et performant pour générer des données synthétiques tout en préservant les propriétés des données originales. Cependant, le modèle* ***TVAE*** *présente aussi les germes d'un bon conservateur-protecteur bien que présentant quelques lacunes notamment dans le CS Test et le KS Test.*

III. TEST SUR DES MODELES ECONOMETRIQUES

a) Présentation des résultats

Les résultats des coefficients de régression pour les différentes bases anonymisées par les modèles ci-dessus présentés (Base réelle, SDC classique, CTGAN, CopulaGauss, TVAE, CopulaGAN) sont comparés pour évaluer comment ces données anonymisées reproduisent les mêmes résultats que les données réelles.

Les coefficients de régression sont présentés pour les variables suivantes : cj_age, cj2_Masculin, cj10_Oui, cas7_Oui, mt1_Neutre, mt1_Pas du tout satisfaisante, mt1_Pas satisfaisante, et mt1_Trés satisfaisante. L'ajustement du modèle est évalué par l'Adj R^2 pour chaque ensemble de données.

Tableau 7: Résultats du modèle linéaire simple en fonction des bases anonymisés.

Coeficients de regression	***Base réelle***	***Base SDC classique***	***Base CTGAN***	***Base CopulaGauss***	***Base TVAE***	***Base CopulaGAN***
Adj R²	*0,0678*	***0,0673***	***0,0038***	***0,0100***	***0,2012***	*0,0026*
const	*0.000*	***-0.000***	***-0.000***	***-0.000***	***-0.000***	*-0.000*
cj_age	*0.163***	***0.160****	*-0.055*	*0.004*	***0.107****	*-0.003*
cj2_Masculin	*0.153***	***0.159****	*-0.005*	***0.073****	***0.235****	*0.016*
cj10_Oui	*0.049*	***0.044***	*0.067**	*-0.046*	*-0.031*	*0.030*
cas7_Oui	*0.010*	***0.011***	***0.034***	***0.021***	*0.061**	*0.003*
mt1_Neutre	*-0.118****	***-0.122****	*0.032*	*-0.002*	***-0.188****	*0.081**
mt1_Pas du tout satisfaisante	*-0.104****	***-0.090****	*-0.053*	***-0.105****		*0.058*
mt1_Pas satisfaisante	*-0.091***	***-0.095****	*0.044*	***-0.033***	***-0.287****	*0.041*
mt1_Très satisfaisante	*0.004*	***0.007***	*0.018*	*0.000*		*-0.050*

Source : Auteurs *(* : p-value<10% / **p-value <5%, /***p-value <1%)*

b) Base SDC Classique

Sur la base SDC classique, le modèle a un Adj R^2 de 0.0673, *très proche de la base réelle.* Les coefficients comme cj_age (0.160***), cj2_Masculin (0.159***), et mt1_Neutre (-0.122***) sont également très similaires aux données réelles, *indiquant une bonne conservation des relations.*

Le modèle SDC classique conserve bien les relations statistiques des données réelles grâce à des techniques d'anonymisation comme la suppression, et l'ajout de bruit (Hundepool et al., 2012). Ces méthodes permettent de masquer les informations sensibles tout en préservant la structure des données. Par exemple, le bruit ajouté a été calibré sur une matrice conservant les corrélations dans les données, pour ne pas perturber excessivement les distributions statistiques, ce qui explique la similitude des coefficients et de l'Adj R^2 avec les données réelles.

c) CTGAN

La base obtenue par CTGAN présente un Adj R^2 de 0.0038, beaucoup plus faible que la base réelle. Les coefficients diffèrent significativement, par exemple, cj_age (-0.055) et cj2_Masculin (-0.005), indiquant *une mauvaise conservation des relations*.

Le modèle CTGAN produit des résultats médiocres parce que les GAN (Generative Adversarial Networks) sont souvent sujets à des problèmes de mode collapse, où le générateur ne parvient pas à couvrir toute la distribution des données réelles (Goodfellow et al., 2014). De plus, les GAN ont tendance à être instables pendant l'entraînement, ce qui peut conduire à des représentations biaisées des relations statistiques complexes. Par conséquent, les coefficients et l'Adj R^2 diffèrent considérablement des valeurs réelles.

d) CopulaGauss

Le modèle estimé sur les données du CopulaGauss a un Adj R^2 de 0.0100. Bien que certains coefficients soient proches des données réelles, comme cj2_Masculin (0.073**) et mt1_Pas du tout satisfaisante (-0.105***), d'autres comme cj_age (0.004) diffèrent, indiquant *une conservation partielle des relations. Cependant, il faut noter que c'est le seul qui permet de suivre les relations entre les variables à partir des signes qui sont respectés. Seulement, il a tendance à sous-estimer les effets obtenus.*

En réalité, le modèle à Copules Gaussiennes, en utilisant des copules, permet de conserver les relations structurelles à un certain degré (Sklar, 1959). Cependant, les copules peuvent sous-estimer ou sur-estimer les dépendances entre variables lorsque la forme de la dépendance est complexe ou non linéaire. Cela explique pourquoi certains coefficients sont proches des valeurs réelles, mais d'autres montrent des différences significatives, et l'Adj R^2 est faible.

e) TVAE

La base TVAE affiche un Adj R^2 de 0.2012, le plus élevé parmi tous les modèles et même plus que celui du modèle de base, il a tendance *à surestimer les effets observés*. Les coefficients pour cj_age (0.107***), cj2_Masculin (0.235***), et mt1_Neutre (-0.188***) sont proches des données réelles, montrant *une bonne conservation des relations pour la plupart des variables grâce à* capacité du TVAE à capturer des dépendances complexes dans les données (Kingma & Welling, 2014). Toutefois, cette complexité peut également entraîner une surestimation des effets, comme indiqué par les coefficients plus élevés que ceux des données réelles.

f) Copula GAN

Cette base présente un Adj R^2 de 0.0026. Les coefficients diffèrent largement des données réelles, par exemple, cj_age (-0.003) et cj2_Masculin (0.016), indiquant une *mauvaise reproduction des relations présentes dans les données réelles.* Cela peut être dû à une combinaison des problèmes de mode collapse des GAN et à des limitations des copules pour modéliser des dépendances complexes (Goodfellow et al., 2014; Sklar, 1959). La faible Adj R^2 et les coefficients divergents par rapport aux données réelles indiquent une mauvaise capacité à reproduire fidèlement les relations, probablement en raison de la complexité accrue de la méthode combinée.

IV. DISCUSSION DES RESULTATS

Les résultats de cette étude montrent l'efficacité des méthodes d'anonymisation et de génération de données synthétiques pour réduire les risques de ré-identification tout en conservant, dans une certaine mesure, l'utilité des données. Ces observations s'alignent et divergent par rapport à d'autres études menées dans le domaine, fournissant une perspective plus nuancée.

1. *Anonymisation et Risques de Divulgation*

L'anonymisation des données a considérablement réduit les risques de ré-identification, comme en témoigne l'élimination des violations de 2-anonymat et 3-anonymat, et une réduction du risque de divulgation à un maximum de 14,16%. Ces résultats sont cohérents avec les travaux de Sweeney (2002) et El Emam et al. (2013), qui ont démontré que les méthodes traditionnelles d'anonymisation peuvent efficacement masquer les informations sensibles. Cependant, comme ces études l'ont également noté, cette réduction des risques entraîne souvent une perte d'information, ce que nous avons observé avec une augmentation de l'IL1 et des différences de valeurs propres après anonymisation. Ce compromis entre confidentialité et utilité est une constante dans le domaine de la protection des données.

2. *Comparaison des Modèles de Génération de Données Synthétiques*

Les modèles de génération de données synthétiques ont montré des performances variées. En premier lieu, le modèle **Copula Gaussian** a particulièrement bien réussi à préserver les distributions des données réelles, suivant les conclusions de Li et al. (2014) et Patki et al. (2016) sur l'efficacité des copules pour maintenir les dépendances structurelles. En revanche, les modèles **CTGAN** et **CopulaGAN**, bien qu'assurant une protection complète contre la divulgation, ont obtenu des scores plus faibles en termes de préservation des distributions, ce qui reflète les problèmes de mode collapse des GAN mentionnés par Goodfellow et al. (2014) et les défis supplémentaires posés par la modélisation des dépendances complexes avec des copules (Sklar, 1959). Par ailleurs, le modèle **TVAE** a montré une bonne capacité à préserver les relations statistiques des données réelles, comme observé par Kingma et Welling (2014). En outre, l'analyse des coefficients de régression a révélé que les bases **SDC classique** et **TVAE** conservent relativement bien les relations statistiques des données réelles, en accord avec les observations de Hundepool et al. (2012) sur l'importance de maintenir la structure des données pour les analyses économétriques. Toutefois, les bases générées par **CTGAN**, **Copula Gaussian**, et **CopulaGAN** montrent des Adj R^2 significativement plus faibles, mettant ainsi en évidence les limites des GAN dans la modélisation des dépendances complexes, comme noté par Choi et al. (2017).

V. Conclusion

Nous constatons que les méthodes d'anonymisation traditionnelles et les modèles de génération de données synthétiques peuvent efficacement réduire les risques de divulgation tout en maintenant l'utilité des données. Nos résultats confirment l'efficacité de ces techniques pour atténuer les risques, bien qu'il subsiste des défis dans la préservation de l'utilité des données. Comparativement à d'autres études, nos observations renforcent la nécessité d'un équilibre entre confidentialité et utilité des données.

Copula Gaussian se distingue comme une option équilibrée, capable de fournir une forte protection des données tout en préservant leur qualité analytique. Ce modèle est recommandé pour des applications nécessitant une préservation des dépendances structurelles complexes, équilibrant ainsi efficacement la confidentialité et l'utilité.

Ces modèles génératifs nous donnent une protection presque totale contre les risques évoqués dans les 3 scénarii présentés dans la méthodologie.

Par ailleurs, le **TVAE** offre également des performances prometteuses mais nécessite des ajustements pour minimiser la surestimation des effets observés. Bien qu'adapté pour des applications où une certaine déviation des distributions originales est acceptable, il est essentiel de prêter attention à sa tendance à surestimer les effets afin d'éviter des biais dans les analyses.

Pour les chercheurs et praticiens, ces conclusions fournissent des orientations précieuses dans le choix des techniques d'anonymisation et de génération de données synthétiques adaptées à leurs besoins spécifiques.

Le SDC Classique quant à lui, semble approprié pour des analyses économétriques simples où la préservation de la structure des données est cruciale, avec une faible perte d'information. Cependant, il est important de calibrer soigneusement les niveaux de bruit ajoutés pour préserver les corrélations essentielles, et cette méthode est particulièrement sensible au choix des quasi-identifiants, ce qui constitue une limitation fondamentale. *Il protège donc contre le scénario 1, mais cette sensibilité aux choix des QID et des variables sensibles limitent son utilisation.*

CONCLUSION GENERALE

En conclusion, le choix de la technique d'anonymisation doit être guidé par l'objectif spécifique visé à travers la publication des données, en équilibrant les besoins de confidentialité et d'utilité des données. Les résultats de cette étude fournissent des orientations précieuses pour les chercheurs et les praticiens dans leur sélection des méthodes d'anonymisation et de génération de données synthétiques adaptées à leurs besoins spécifiques. Nous recommandons la procédure suivante pour la publication de deux types de fichiers en fonction du public cible :

1. **Scientific Use Files (SUF) :** Ces fichiers seront anonymisés à l'aide des méthodes classiques d'anonymisation (SDC classique) souples. Cette approche permet de conserver les relations statistiques essentielles pour les analyses, ce qui est important pour les chercheurs et les scientifiques.
2. **Public Use Files (PUF)** : Ces fichiers nécessitent une anonymisation plus forte, justifiant l'utilisation de modèles basés sur la génération de données synthétiques, tels que TVAE et Copula GAN. Ces modèles assurent une protection robuste contre la divulgation, même si cela peut se faire au détriment de certaines caractéristiques des données originales. Après anonymisation, il est crucial de transmettre des détails sur les précautions à prendre par les utilisateurs avant l'interprétation des résultats obtenus. Toutefois, ces modèles doivent être appliquées à la suite d'une préparation contextuelle de la base de données inspirée de la procédure SDC classique.

Les modèles SDC classiques sont très sensibles au choix des quasi-identifiants et des variables sensibles, mais offre un cadre rigoureux d'analyse contextuelle.

Sur une base de plus de 900 variables initialement, réduite à environ 180, bien que l'anonymisation soit réussie, les risques de divulgation restent élevés. Une limite de notre étude à ce niveau est de ne pas modéliser les relations de dépendance entre les variables, une procédure qui implique des coûts computationnels énormes.

De même, il aurait été préférable de modéliser certaines dépendances entre les colonnes en tant que contraintes pour assurer la cohérence des données dans notre conception des modèles génératifs. Nous n'avons pas pris en compte la correction des incohérences induites par les techniques avant de tester les bases sur le modèle économétrique. Cela donne un intérêt aux techniques de génération de données synthétiques, qui apprennent *« toutes »* les dépendances entre les colonnes. Cela assure

une base vérifiant la confidentialité différentielle et valable statistiquement, bien que nécessitant quelques corrections pour assurer la cohérence.

Bien que nos résultats soient principalement axés sur le dualisme confidentialité-utilité plutôt que sur la cohérence interne, nous proposons qu'en cas de publication, il soit essentiel d'informer les utilisateurs des limitations ou de traiter en amont les bases de données en effectuant des ajustements. Cela devrait entraîner une grande vraisemblance des données, en particulier pour le modèle Copula Gaussian, qui présente des résultats très prometteurs.

REFERENCES BIBLIOGRAPHIQUES

1. Alehmans, J. (2017). Data Sharing and Open Data: Trends and Future Perspectives. *Journal of Data Management, 10*(2), 45-60.
2. Alex, A. (2012). Data Protection and Privacy Legislation in Africa: Progress and Challenges. *African Law Journal, 14*(1), 20-35.
3. Arben, T. (2020). The Evolution of Data Protection Regulations: From Convention 108 to GDPR. *International Journal of Data Privacy, 5*(1), 75-90.
4. Beaulieu-Jones, B. K., & Greene, C. S. (2019). Learning a Latent Space for Synthetic Health Data. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 121-128.
5. Binns, R., & C. G. (2018). Mitigating Bias in Data with Explanations. *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency (FAT)*.
6. Calder, T., & Žliobaitė, S. (2013). Why Unbiased Computational Processes Can Lead to Discrimination: A Case Study in Data Mining. *Proceedings of the 2013 ACM Conference on Knowledge Discovery and Data Mining (KDD)*, doi:10.1145/2487575.2488188.
7. Choi, E., Schuetz, A., Stewart, W. F., & Schulte, P. J. (2017). Using Generative Adversarial Networks for Synthetic Electronic Health Records. *Proceedings of the 26th International Conference on World Wide Web*, 1095-1104.
8. Dwork, C., & Lei, J. (2009). Differential Privacy and the Algorithms for Privacy. *Proceedings of the 30th International Conference on Computer Science*, 493-502.
9. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Proceedings of the Third Theory of Cryptography Conference (TCC 2006)*, 265-284.
10. Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science, 9*(3–4), 211-407.
11. El Emam, K., & Arbuckle, L. (2013). Anonymizing Health Data: Case Studies and Methods to Get You Started. *O'Reilly Media*.
12. El Emam, K., & Jonker, E. (2013). Privacy-Preserving Data Mining in Africa. *Proceedings of the 12th International Conference on Data Mining*, 129-138.
13. Fienberg, S. E., & McIntyre, J. (2004). Confidentiality and Data Access: A Review of Recent Developments. *Journal of Privacy and Confidentiality, 1*(1), 1-20.
14. Gentry, C. (2009). A Fully Homomorphic Encryption Scheme. PhD Thesis, Stanford University.
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative Adversarial Nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, 2672-2680.

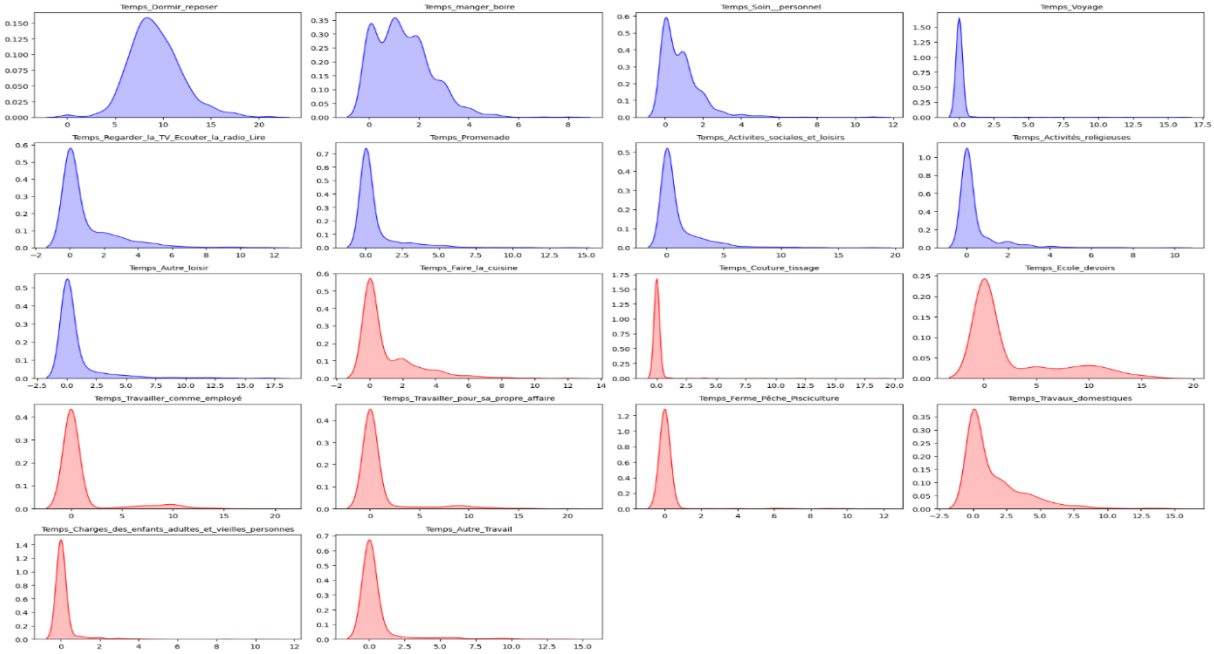
16. Gkoulalas-Divanis, A., & Loukides, G. (2014). Anonymizing Electronic Health Records for Clinical Research. *ACM Transactions on Knowledge Discovery from Data, 8*(2), 1-30.
17. Hundepool, A., Domingo-Ferrer, J., & Franconi, L. (2012). Statistical Disclosure Control. *Wiley Encyclopedia of Operations Research and Management Science*.
18. Hundepool, A., Roos, M., Schoutsen, E., & Verweij, M. (2012). Statistical Disclosure Control. *Wiley*.
19. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., & Feng, M. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data, 3*, 160035.
20. Kinyanjui, M., Mugo, S., & Kamau, E. (2018). Applying Data Anonymization Techniques in Social Surveys: Case Study of Kenya. *African Journal of Data Science, 10*(3), 45-56.
21. Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations*.
22. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.
23. Li, N., Li, T., & Venkatasubramanian, S. (2006). Closeness: A Privacy Model for the Era of Big Data. *IEEE Transactions on Knowledge and Data Engineering, 27*(4), 678-690.
24. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, 106-115.
25. Liu, L., & Wu, C. (2018). Survey on Privacy-Preserving Data Mining. *IEEE Transactions on Knowledge and Data Engineering, 30*(6), 1-16.
26. Liu, X., Li, C., & Han, J. (2020). Copula Gaussian: A Method for Preserving the Structure of Dependencies in Data Synthesis. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 678-685.
27. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, S. (2006). Privacy: Theory Meets Practice on the Map. *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 277-288.
28. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 3-37.
29. Monreale, A., Giannotti, F., & Pedreschi, D. (2011). M-Invariance: A New Privacy Model for Data Mining. *Proceedings of the 7th IEEE International Conference on Data Mining*, 290-299.
30. Monreale, A., Rinzivillo, S., & Trasarti, R. (2011). Mining Social Networks to Enhance

Privacy Protection in Anonymized Data. *IEEE Transactions on Knowledge and Data Engineering, 23*(6), 860-872.

31. Narayanan, A., & Shmatikov, V. (2008). How to Break Anonymity of the Netflix Prize Dataset. *Proceedings of the 30th IEEE Symposium on Security and Privacy*, 173-187.
32. Nergiz, M. E., & Clifton, C. (2007). How to Protect Privacy When Linking to External Databases. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109-118.
33. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1603-1612.
34. Reiter, J. P. (2005). Supply and Demand for Statistical Disclosure Control. *Journal of the American Statistical Association, 100*(470), 951-962.
35. Rocher, L., Hendrickx, J., & de Montjoye, Y.-A. (2019). Estimating the Success of Re-Identification Attacks. *Nature Communications, 10*(1), 1-8.
36. Samarati, P., & Sweeney, L. (1998). Generalizing Data to Provide Anonymity When Disclosing Information. *Proceedings of the 17th ACM SIGMOD International Conference on Management of Data*, 188-199.
37. Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557-570.
38. Wang, L., & Fung, B. C. M. (2006). Anonymizing Data with (X, Y)-Anonymity. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 207-216.
39. Wang, L., Li, N., & Yang, J. (2005). Confidence Bounding: Anonymizing Data by Fixing the Confidence of Sensitive Information. *Proceedings of the 7th International Conference on Data Mining*, 126-133.
40. Yang, Y., Zhang, Y., & Ma, L. (2019). CTGAN: Conditional Generative Adversarial Networks for Tabular Data. *arXiv preprint arXiv:1907.00503*.
41. Xu, J., Zhang, J., & Wang, X. (2020). Differential Privacy and Its Applications: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 32*(7), 1421-1437.
42. Yao, A. C. (1982). Protocols for Secure Computations. *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, 160-164.

ANNEXES

Graphique 1 : Exemple de consolidation : Transformation de 456 variables sur les habitudes de vie en 18 variables



Source : Auteurs

Tableau 8 : Une base de données fictive des étudiants de l'ENSEA

Classe	*Age*	*Sexe*	*Revenue/mois (FCFA)*
AS3	*21*	*H*	*50 000*
AS3	*45*	*H*	*600 000*
AS1	*18*	*F*	*100 000*
AS1	*18*	*F*	*1000*

Source : Auteurs

Tableau 9: Exemple de k-anonymat -Table 2-anonyme de la base fictive

Classe	*Age*	*Sexe*	*Revenue/mois*
AS3	*20+*	*H*	*50 000*
AS3	*20+*	*H*	*600 000*
AS1	*20-*	*F*	*100 000*
AS1	*20-*	*F*	*100 000*

Source : Auteurs

Tableau 10: Exemple de l-diversity ; Table 2-diverse de la base fictive

Classe	Age	Sexe	Revenue/mois
AS3	20+	H	50 000
AS3	20+	H	600 000
AS1	20-	F	100 000
AS1	20-	F	100 000

Source : Auteurs

Tableau 11: Quelques législations sur la protection des données

Année	***Législation***	***Pays/Région***	***Particularités***
1996	*HIPAA (Health Insurance Portability and Accountability Act)*	États-Unis	✓ Normes nationales pour la protection des informations médicales ✓ Sanctions pour les violations, ✓ ***Exigences de sécurité des données électroniques.***
2000	*Data Protection Act*	Royaume-Uni	✓ ***Obligations pour protéger les données,*** ✓ Droits des individus sur leurs informations, ✓ Pénalités pour le non-respect.
2013	*Loi n° 2013-450 relative à la protection des données à caractère personnel*	*Côte d'Ivoire*	✓ Autorité de régulation, ✓ Sanctions pour les violations, ✓ ***Protection contre l'utilisation non autorisée et le traitement abusif.***
2013	*Protection of Personal Information Act (POPIA)*	*Afrique du Sud*	✓ Consentement explicite, ✓ Audits réguliers et rapports de conformité.
2018	*Règlement Général sur la Protection des Données (RGPD)*	Union Européenne	✓ Harmonisation des lois, ✓ ***Transparence,*** ✓ Consentement clair, ✓ Sanctions sévères.
2019	*Personal Data Protection Act (PDPA)*	*Nigeria*	✓ Normes strictes, ✓ Sanctions, ✓ Mesures de sécurité techniques et organisationnelles.
2019	*Data Protection Act*	*Kenya*	✓ Consentement, ✓ ***Commission de protection des données,*** ✓ Sanctions.

2020	*California Consumer Privacy Act (CCPA)*	Californie, États-Unis	✓ Droit de connaitre les données collectées, ✓ Possibilité de refus de vente de ses données, ✓ ***Obligation de divulgation des données collectées.***
2020	*Personal Information Protection and Electronic Documents Act (PIPEDA)*	Canada	✓ ***Protection des données dans les activités commerciales,*** ✓ Consentement, ✓ Évaluations régulières de la conformité.

Source : Auteurs.

Tableau 12: Résultats de l'anonymisation par SDC classique

Critères	***Avant Anonymisation***	***Après Anonymisation***
2-anonymat	*92 violations (10,099%)*	*0 violations* ***(0,000%)***
3-anonymat	*311 violations (34,138%)*	*0 violations* ***(0,000%)***
5-anonymat	*448 violations (49,177%)*	*139 violations* ***(15,258%)***
Risque de divulgation	*Entre 0,00% et 100,0%*	*Entre* ***0,00%*** *et* ***14,16%***
Perte d'information (IL1)	*0*	***8460,46***
Différence des valeurs propres	*0%*	***4,580%***

Source : Auteurs

Tableau 13: Résultats de l'anonymisation avec les modèles génératifs

Modèle	Categorical CAP	CS Test	KS Complement	Continuous KL Divergence	Discrete KL Divergence	Moyenne
Base réelle	0,929	**1,000**	1,000	1,000	1,000	0,985
SDC classique	0,995	0,222	***0,984***	***0,993***	0,498	0,524
CTGAN	***1,000***	0,806	0,656	0,760	0,582	0,735
CopulaGauss	***1,000***	***0,992***	***0,772***	***0,833***	**0,897**	**0,890**
TVAE	***1,000***	**0,950**	0,610	0,676	**0,786**	**0,776**
CopulaGAN	***1,000***	0,853	0,658	0,702	0,617	0,742

Source : Auteurs

Tableau 14: Résultats du modèle linéaire simple en fonction des bases anonymisés.

Coeficients de regression	Base réelle	Base SDC classique	Base CTGAN	Base CopulaGauss	Base TVAE	Base CopulaGAN
Adj R^2	0,0678	**0,0673**	**0,0038**	**0,0100**	**0,2012**	0,0026
const	0.000	**-0.000**	**-0.000**	**-0.000**	**-0.000**	-0.000
cj_age	0.163***	**0.160*** **	-0.055	0.004	**0.107*** **	-0.003
cj2_Masculin	0.153***	**0.159*** **	-0.005	**0.073** **	**0.235*** **	0.016
cj10_Oui	0.049	**0.044**	0.067*	-0.046	-0.031	0.030
cas7_Oui	0.010	**0.011**	**0.034**	**0.021**	0.061*	0.003
mt1_Neutre	-0.118***	**-0.122*** **	0.032	-0.002	**-0.188*** **	0.081*
mt1_Pas du tout satisfaisante	-0.104***	**-0.090** **	-0.053	**-0.105*** **		0.058
mt1_Pas satisfaisante	-0.091**	**-0.095** **	0.044	**-0.033**	**-0.287*****	0.041
mt1_Très satisfaisante	0.004	**0.007**	0.018	0.000		-0.050

Source : Auteurs (* : p-value<10% / **p-value <5%, /***p-value <1%)

QUESTIONNAIRE DE L'ENQUETE *: https://enseaabidjan-my.sharepoint.com/:w:/g/personal/aziz_coulibaly_ensea_edu_ci/ES4yBHnpd_pJhcmD0v0sCjsBhxNoGJ63JH7aZ43LgIT-ag?e=sTdKrY&wdLOR=c5B8B716A-F144-4516-B485-44F7D1529863*

Le fichier de codes utilisés : https://github.com/alazizcoul/GT_Anonymization

TABLE DES MATIERES

REMERCIEMENTS i

AVANT-PROPOS iv

LISTE DES SIGLES ET TABLEAUX, ... v

Liste des Sigles v

Liste des Tableaux vi

ABSTRACT ET RESUME vii

INTRODUCTION GENERALE 10

I. Contexte et Justification 1

II. Problématique 3

III. Objectif général 4

IV. Objectifs spécifiques 4

V. Hypothèses 4

VI. Plan 4

CHAPITRE 1 : CADRE CONCEPTUEL ET REVUE DE LITTERATURE 10

I. INTRODUCTION 6

II. CADRE CONCEPTUEL 6

1. Concepts fondamentaux et objectifs de l'anonymisation 6

a) Confidentialité et Anonymisation 6

b) Données Sensibles et Quasi-Identifiants 7

2. Risques et modèles d'attaques 7

a) Réidentification et risques de réidentification 7

b) Mesures du risque de réidentification 8

c) Scénarios et modèles d'attaques 8

i. Attaques par correspondance des enregistrements 9

ii. Attaques par correspondance des tables 9

iii. Attaques probabilistes 9

iv. Autres attaques 9

3. Modèles de protection des données 10

a) Modèles de protection syntaxique ... 10
i. k-anonymat ... 10
ii. L-Diversité ... 11
iii. T-closeness ... 12
iv. δ-Diversity / Presence ... 13
b) Modèles de protection statistique ... 14
i. k-map ... 14
ii. Seuils de risque moyen ... 14
iii. Modèles de super-population ... 14
c) Modèles de Protection Sémantique ... 14
i. (ε, δ)-differential privacy ... 15
ii. Approche de désidentification basée sur la théorie des jeux ... 15
4. Techniques d'anonymisation ... 16
a) Techniques d'anonymisation classiques : Le Contrôle de Divulgation Statistique (SDC) ... 16
b) Anonymisation Basée sur la Génération de Données Synthétiques ... 16
i. Réseaux Génératifs Antagonistes (GAN) ... 17
ii. Autres Techniques de génération de données synthétiques ... 17
5. Utilité des données anonymisées --Cadre légal et éthique ... 18
a) Utilité et mesures d'utilité ... 18
b) Législations et réglementations internationales ... 18
III. REVUE DE LITTERRATURE ... 20
1. Revue de littérature théorique ... 20
a) Évolution des méthodes de protection contre les attaques de ré-identification ... 20
b) Techniques d'anonymisation classiques ... 21
c) Techniques d'anonymisation basées sur les données distribuées ... 22
d) Anonymisation basée sur les données synthétiques ... 22
2. Revue Empirique ... 23
a) Anonymisation des données de santé ... 23
b) L'Anonymisation des données d'enquête sociale en Afrique ... 23
c) Utilisation des données synthétiques : une revue empirique ... 24
i. Dans le domaine de la santé ... 24
ii. Dans le secteur financier ... 25
iii. Dans le contexte des enquêtes sociales ... 25

IV. CONCLUSION 25
APPROCHE METHODOLOGIQUE 10
I. INTRODUCTION 27
II. PHASE EXPLORATOIRE : AVANT L'APPLICATION DES METHODES 27
1. Évaluation de la nécessité de la confidentialité du point de vue légal 27
2. Présentation de la base de données et hypothèses 28
3. Identification et consolidation des variables clé 29
4. Choix du type de diffusion et élaboration des scénarios d'intrusion 30
a) Type de diffusion 30
b) Elaboration des Scénarios d'Intrusion 31
i. Scénario 1 : Intrusion basique 31
ii. Scénario 2 : Intrusion avec données Quasi-Identifiantes 31
iii. Scénario 3 : Intrusion avancée 31
5. Sélection des variables clé 31
a) Quasi-identifiants 31
b) Variables sensibles 32
III. PRESENTATION DES MODELES UTILISES 33
1. Procédure d'anonymisation par SDC classique 33
a) Ajout de bruit 33
b) La méthode PRAM (Post-RAndomisation Method) 34
2. Modèles de génération de sonnées synthétiques 35
a) Structure de Base du GAN 35
b) Le CTGAN (Conditional Tabular GAN) 36
c) Le modèle de génération basé sur les copules gaussiennes (Copula Gauss) 36
d) Le Tabular Variational Autoencoder (TVAE) 36
e) Le copule-GAN 37
IV. APPLICATION DES MODELES 37
1. Processus SDC 37
2. Les méthodes génératives 38
3. Indicateurs d'évaluation du risque sur les bases de données 38
a) Risque de divulgation 38
b) Categorical CAP (Classification Accuracy for Privacy) 39
4. Indicateurs de mesure de l'utilité des données 39

a) Perte d'information (IL1) ... *39*
b) Différence des valeurs propres (Eigenvalues) ... *39*
c) CSTest (Column Similarity Test) ... *39*
d) KS Complement (Kolmogorov-Smirnov Test Complement) : ... *40*
e) Continuous KL Divergence et la Discrete KL Divergence ... *40*
f) Test sur des modèles économétriques : ... *40*
V. CONCLUSION ... *41*
RESULTATS ET DISCUSSION ... *10*
I. INTRODUCTION ... *42*
II. RESULTATS DE L'ANONYMISATION ... *42*
***1.** Résultats du processus SDC* ... *42*
2. Résultats comparés du SDC et des autres modèles génératives ... *43*
a) Analyse des résultats de chaque modèle ... *43*
i. Base Réelle ... *43*
ii. SDC Classique ... *44*
iii. CTGAN ... *44*
iv. Copula Gaussian ... *44*
v. TVAE ... *45*
vi. Copula GAN ... *45*
b) Synthèse ... *45*
III. TEST SUR DES MODELES ECONOMETRIQUES ... *46*
a) Présentation des résultats ... *46*
b) Base SDC Classique ... *47*
c) CTGAN ... *47*
d) CopulaGauss ... *47*
e) TVAE ... *48*
f) Copula GAN ... *48*
IV. DISCUSSION DES RESULTATS ... *48*
1. Anonymisation et Risques de Divulgation ... *49*
2. Comparaison des Modèles de Génération de Données Synthétiques ... *49*
V. Conclusion ... *50*
CONCLUSION GENERALE ... *51*
REFERENCES BIBLIOGRAPHIQUES ... *viii*

ANNEXES.....*xi*
TABLE DES MATIERES.....*xv*