Génération de signaux anonymes à partir de données non anonymes par modèle de mélange linéaire local

Alban-Félix BARRETEAU[1,2] Olivier REGNIER-COUDERT[1] Eric LE CARPENTIER[2] Saïd MOUSSAOUI[2]

[1]Octopize, Nantes, France

[2]Nantes Université, Ecole Centrale de Nantes, LS2N UMR CNRS 6004, France

Résumé – L'objet de cette communication est de proposer une technique de génération de signaux synthétiques anonymes à partir d'une base de données non-anonyme. La méthode adoptée consiste à construire un modèle statistique dans lequel chaque signal de la base de données initiale est représenté comme un mélange linéaire de ses voisins. Ce modèle génératif est ensuite utilisé pour simuler un ensemble de signaux synthétiques différents de ceux de la base de données initiale mais contenant toutes les informations nécessaires pour son exploitation. La qualité de la base de données simulée est quantifiée par des mesures et d'utilité et de vie privée. Cette technique est appliquée à une base de signaux ECG réels.

Abstract – The aim of this work is to propose a technique of anonymization through signal generation from a non-anonymous database. The adopted method consist of a statistical mixing model where each signal is considered as a mixture of its neighbors in the database. This generative model is then used to simulate a set of synthetic signals containing all the information present in the initial database. The quality of the simulated database is assessed in terms of privacy and utility criteria. The proposed approach is applied to a database of real ECG signals.

1 Introduction

Une donnée personnelle est définie comme étant « toute information se rapportant à une personne physique » [5]. L'anonymisation est un processus irréversible qui rend impossible la ré-identification de l'individu à l'origine de la donnée. Une donnée anonymisée n'est alors plus considérée comme une donnée personnelle, et n'est donc plus soumise aux règles du Règlement Général pour la Protection des Données (RGPD). Un des secteurs où les contraintes RGPD sont strictement encadrées est celui de la santé. Par exemple, les signaux biomédicaux tels que les signaux électrocardiographiques (ECG) sont des données personnelles au sens du RGPD et leur exploitation est très réglementée. Entre autres, sans le consentement éclairé de l'individu, ils ne peuvent pas être utilisés pour une autre finalité que celle pour laquelle ils ont été enregistrés. Pourtant, ces données peuvent être utiles à des fins de recherche clinique, notamment pour le prototypage d'algorithmes de traitement du signal et d'aide à la décision. Plus particulièrement, l'entraînement d'algorithmes de traitement par apprentissage statistique nécessite l'exploitation de larges bases de données. Les bases de données réelles étant difficilement accessibles de part les restrictions juridiques et contenant des informations personnelles sensibles qui ne sont pas nécessaires au traitement souhaité, l'anonymisation des données permet de répondre au mieux à ces différentes problématiques.

Les techniques classiques d'anonymisation, par généralisation des données (k-anonymat [10]) ou par ajout de bruit (*differential privacy* [6]), sont utilisées pour certaines analyses statistiques sur des données tabulaires, mais ces techniques ne sont pas adaptées à la spécificité de données plus complexes tels que des signaux ou des images. Dans ce travail, on se focalise sur les techniques par génération de données synthétiques car celles-ci sont plus adaptées au contexte de ces bases de données contenant beaucoup d'information. En particulier, on s'intéresse à la méthode avatar présentée dans [7]. Cette méthode propose de générer un avatar pour chaque signal d'une base de données comme un mélange linéaire des ses voisins avec des coefficients de mélange choisis aléatoirement. La première contribution de ce travail est de proposer un modèle statistique sur la distribution des coefficients de mélange adapté à la base de données. La seconde contribution consiste en la proposition d'une stratégie permettant d'appliquer cette méthode à des signaux ECG.

D'autres méthodes basés sur des GAN (Generative Adversarial Networks) proposent également la génération de données ECG. Dans [1], l'objectif est de faire de l'augmentation de données : les données générées se veulent proches des données originales et ne sont pas anonymes. [11] compare deux GAN pour la génération d'ECG anonymes mais, bien qu'ils arrivent à de bons résultats en termes d'utilité, les ECG synthétiques n'apportent aucune preuve en termes de privacy. Ces méthodes s'apparentent donc plus à des méthodes de synthèse d'ECG, plutôt qu'à des méthodes d'anonymisation.

Nous allons d'abord présenter succinctement le principe de la méthode avatar, sur laquelle ce travail s'appuie. Ensuite, nous proposerons une technique alternative à cette méthode, par mélange linéaire suivant une loi de Dirichlet. Enfin, nous décrirons un processus de génération d'ECG anonymes utilisé pour évaluer la méthode proposée.

2 Méthodologie

2.1 Principe de la méthode Avatar

Considérons un jeu de donnée $\mathcal{X}$, composé de N individus et de V variables les décrivant. La méthode avatar crée un nouveau jeu de données de même taille $N \times V$, qui conserve l'information utile mais qui préserve les individus originaux

de la divulgation d'informations personnelles et sensibles. Un avatar est généré pour chacun des individus dans le jeu de données d'origine.

La première étape est de projeter les données d'origine dans un espace de représentation adapté qui doit conserver l'essentiel de l'information ainsi que permettre le calcul de distances entre les individus sous forme de matrice de dimension réduite $N \times D$. Cette transformation doit également être inversible. Dans le contexte de données tabulaires, cette étape peut être le résultat d'une d'une analyse factorielle de donnée mixtes (AFDM) [9].

La deuxième étape est d'identifier les K plus proches voisins dans l'espace projeté, qui servent de base à la génération de chaque avatar. Un poids $\omega_{n,k}$ est attribué pour chaque voisin $v_{n,k}$, $k \in \{1, \ldots, K\}$ de chaque individu $n \in \{1, \ldots, N\}$. La méthode empirique de génération des poids basée sur la distance aux voisins, une loi exponentielle et des permutations aléatoires expliquée dans [7] est une méthode empirique qui a montré des résultats satisfaisants dans le contexte de données tabulaires. Nous proposons ici une alternative, décrite plus en détail dans la section suivante, qui prend en compte la distribution de l'ensemble du jeu de données original. Les coordonnées des individus synthétiques sont ensuite générées comme combinaison linéaire de ses K plus proches voisins :

$$y_n = \sum_{k=1}^{K} \omega_{n,k} v_{n,k} \tag{1}$$

Enfin, la transformation inverse est appliquée sur ces coordonnées pour ré-exprimer les avatars dans l'espace de représentation d'origine.

2.2 Proposition de méthode par mélange linéaire suivant une loi de Dirichlet

Nous proposons ici une alternative à la méthode de calcul des poids présentée dans [7]. Dans cette variante, chaque individu est considéré comme mélange linéaire de ses plus proches voisins. Les coefficients de mélange sont obtenus par moindres carrés sous contrainte de positivité et de somme à un [4]. On obtient une matrice des coefficients de mélange $\mathbf{A}$. Les coefficients de mélange sont supposés être tirés suivant une loi de Dirichlet dont les paramètres $\boldsymbol{\alpha}$ sont estimés par maximum de la vraisemblance $p(\mathbf{A}|\alpha)$. Cette alternative permet une meilleure appréhension de la méthode ainsi qu'un calibrage des poids de génération des avatars plus fin. L'algorithme de génération des coordonnées avatars à partir des coordonnées originales est décrit dans l'algorithme **??**.

2.3 Processus de génération d'ECG anonymisé

Dans la littérature, l'algorithme Chronos [3] propose une adaptation de la méthode avatar [7] pour l'anonymisation de signaux, par génération de données synthétiques. Celui-ci a été utilisé pour générer des cycles uniques d'électrocardiogrammes (ECG). Cette nouvelle proposition permet quant à elle la génération d'ECG complet, avec plusieurs cycles. De par la spécificité des signaux ECG, plusieurs étapes préalable à la procédure d'anonymisation sont nécessaires.

Algorithme 1 : Génération de coordonnées Avatars par mélange linéaire suivant une loi de Dirichlet

Calcul de la matrice des distances entre les individus

pour *chaque individu n dans N* **faire**

  Sélection des K plus proches voisins au n-ième individu et construction de sa matrice des voisins :

$$\mathbf{V}_n^{K\times D} = \begin{bmatrix} v_{n,1} & \cdots & v_{n,K} \end{bmatrix}^\top$$

  avec $v_{n,k}$ les coordonnées du k-ième voisin à x_n

  Estimation du vecteur d'abondance $\mathbf{a}_n$ tel que : $x_n^{1\times D} = \mathbf{a}_n^{1\times K}\ \mathbf{V}_n^{K\times D}$ sous contrainte de positivité et de somme à un.

fin

Estimation des K paramètres de la loi de Dirichlet $\boldsymbol{\alpha}$=$[\alpha_1, \cdots, \alpha_K]$ à partir de la matrice des abondances $\mathbf{A}^{N\times K} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{bmatrix}^\top$ avec l'estimateur du maximum de vraisemblance $p(\mathbf{A}|\boldsymbol{\alpha})$

Calcul de la matrice de pondération $\boldsymbol{\Omega}^{N\times K}$ comme N tirages de la distribution de Dirichlet $D(\alpha)$

pour *chaque individu n in N* **faire**

  Calcul des coordonnées des avatars : $y_n = \boldsymbol{\Omega}_n^{N\times K}\ \mathbf{V}_n^{K\times D}$

fin

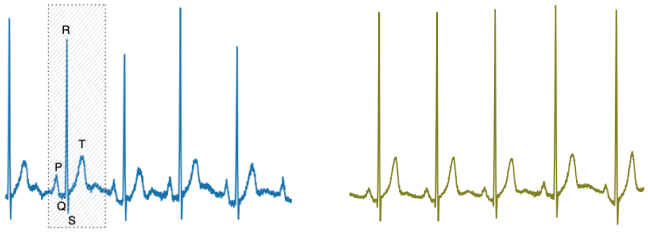

FIGURE 1 : Exemple de signal ECG. A gauche, un cycle complet est identifié sur un signal original ainsi que les ondes PQRST . A droite, son avatar.

1) Segmentation des cycles. Chaque ECG est décomposé par cycles. Le cycle a été choisie comme étant la durée entre deux ondes R. Les signaux R sont détectés comme étant les maxima locaux du signal. On introduit ensuite un décalage de telle sorte que le complexe QRS se trouve au centre du cycle.

2) Normalisation. Pour être comparable, les cycles sélectionnés sont centrés-réduits. Chacun des cycles est ré-échantillonné en un même nombre, 600, de points (normalisation de la fréquence).

3) Calcul des voisins. Les plus proches voisins d'un cycle normalisé sont calculés grâce au coefficient de corrélation entre ce cycle et la base de données des cycles normalisés.

4) Transformation. On applique une transformée en cosinus discret (DCT) [2] aux cycles normalisés. Cette transformation sans perte permet de condenser une majorité de l'énergie du

signal dans quelques coefficients et permet ainsi une meilleure conservation de l'information lors du processus de génération des données par mélange.

5) Generation des avatars. Les cycles normalisés sont anonymisés par la méthode d'avatarization suivant la loi de Dirichlet telle que définie précédemment. Les paramètres α de la loi de Dirichlet sont estimées comme le maximum de vraisemblance $p(\mathbf{A}|\alpha)$ avec $\boldsymbol{\alpha}$ la matrice des abondances des cycles normalisés en fonction de leur plus proches voisins. On traite également les paramètres de normalisation, la moyenne $\boldsymbol{\mu}$ et l'écart type $\boldsymbol{\sigma}$, pour chacun des cycles avec ses $\boldsymbol{K}$ plus proches voisins, afin d'assurer la dé-normalisation des avatars :

- $\tilde{\boldsymbol{\mu}} = \sum_{k=1}^{K} \boldsymbol{\omega_k} \boldsymbol{\mu_k}$
- $\tilde{\boldsymbol{\sigma}} = \sum_{k=1}^{K} \boldsymbol{\omega_k} \boldsymbol{\sigma_k}$

avec $\boldsymbol{\omega_k}$ issu d'un tirage de la distribution de Dirichlet avec les mêmes paramètres appris.

6) Concaténation des cycles. Pour reconstituer des ECG complets, les cycles avatarisés de chacun des signaux sont assemblés comme suit. 1) Chaque cycle est ré-échantillonné en nombre $\tilde{\boldsymbol{T}}$ de points tirés d'une loi normale dont les paramètres sont la moyenne et l'écart type de la longueur des cycles du signal original. 2) Concaténation de chacun des cycles avatars. Les amplitudes des cycles sont ajustées de telle sorte que l'amplitude initiale du cycle suivant soit identique à l'amplitude finale du cycle précédent. 3) En amont du signal, on ajoute une seconde fois le premier cycle avatarisé. De même, on ajoute autant de fois que nécessaire le dernier cycle à la fin du signal, afin d'obtenir un signal avatar au moins aussi long que celui d'origine, considérant les cycles incomplets qui se trouvent en début et en fin d'un enregistrement ECG. 4) On ajoute au signal avatar reconstitué la moyenne des amplitudes $\bar{\tilde{\boldsymbol{\mu}}}$ des cycles constituant le signal et on le multiplie par la moyenne des écarts-types $\bar{\tilde{\boldsymbol{\sigma}}}$. 5) Fenêtrage du signal avatar afin qu'il soit de même longueur que les signaux originaux.

3 Métrique d'évaluation

Privacy. Afin d'évaluer la protection des données contre des attaques de ré-identification et plus précisément d'individualisation, deux métriques sont utilisées : la Hidden Rate (HR) et le Local Cloaking (LC). Au cours de l'anonymisation, le lien entre l'individu original et son avatar est temporairement conservé. Avant d'être supprimé de manière définitive, ce lien est utilisé pour évaluer le succès de divers scénarios d'attaque. HR mesure la probabilité qu'un attaquant obtienne une mauvaise association lorsqu'il assume que l'avatar le plus proche d'un individu est son avatar. LC mesure, pour chaque individu, le nombre d'avatars entre un individu et l'avatar qu'il a produit. On utilise le LC médian pour quantifier le niveau de protection pour le jeu de données dans son ensemble.

Deux types d'attaques sont considérées. La première se concentre sur les signaux dans leur forme d'origine. Le second type d'attaque se base sur des descripteurs extraits des signaux. Par exemple, dans le cas des ECG, un descripteur peut être la fréquence cardiaque moyenne ou l'amplitude maximale du signal. On peut ensuite définir une distance entre les vecteurs formés par l'ensemble des descripteurs calculés pour chaque signal.

Au final, on ne retient que le pire scénario possible, c'est à dire que l'on considère les métriques pour lesquelles l'attaque de ré-identification a été la plus efficace.

Utilité. Pour mesurer la conservation de l'information utile des signaux, il est important de définir en amont l'objectif de l'anonymisation des données : ces métriques ont pour but de valider la conservation des caractéristiques utiles du signal. En effet, le traitement qui est appliqué au signaux altère leurs caractéristiques identifiantes et modifie leurs propriétés à l'échelle de l'individu. Néanmoins, le processus doit conserver les caractéristiques statistiques globales du jeu de données. Ces caractéristiques sont extraites des signaux, à la fois sur le jeu de donnée original et le jeu de donnée avatar. Nous comparons ensuite les distributions de ces caractéristiques entre le jeu de données original et celui avatar avec la distance de Hellinger, bornée entre 0 et 1. Les relations de dépendances entre ces variables sont aussi évaluées avec la variation d'information normalisée.

4 Expérimentations et résultats

Présentation des données. Afin de tester la méthode, nous l'avons appliqué sur des signaux ECG provenant de la base de données Chapman [12]. Cette base de données regroupe des ECG 12 dérivations de 10646 patients, d'une durée de 10 secondes échantillonnées à 500Hz. Nous avons retenus pour ces expérimentations un échantillons 1000 signaux de la dérivation II d'une longueur de 5 secondes.

L'étape de segmentation des ECG est essentielle à la procédure de génération proposée. Néanmoins, l'objet de ce travail n'est pas la segmentation des ECG. C'est pourquoi nous avons sélectionnés ces 1000 signaux pour leur bonne détection des cycles par l'algorithme de segmentation utilisé [8].

Résultats. Nous avons évalué ces données sur la base des descripteurs suivants : 1) fréquence cardiaque, 2) durées entre deux ondes P, Q, R, S et T successives, 3) amplitudes des ondes P, Q, R, S et T, 4) durée des ondes P, R et T. Pour chacun de ces descripteurs, par signal ECG, on retient la valeur moyenne, l'écart type, le minimum et le maximum. On obtient 56 descripteurs pour chacun des signaux. Le nombre de plus proches voisins considérés pour le processus de génération des avatars est $\boldsymbol{K = 10}$. Nous appliquons la procédure de génération d'ECG anonymes avec la méthode suivant une loi de Dirichlet (noté Dirichlet dans le tableau ci-dessous) présentée et avec la méthode de génération des poids proposées dans [7] (noté $\boldsymbol{Original}$ dans le tableau) afin de comparer les deux approches, avec un paramètre $\boldsymbol{K}$ égal à 10 et 20. Les résultats obtenu pour les différentes métriques pour chacune des combinaisons testées sont résumées dans le tableau ci-dessous :

	$\boldsymbol{K = 10}$		$\boldsymbol{K = 20}$	
	Dirichlet	Original	Dirichlet	Original
Hellinger	**0,139**	0,143	0,144	0,145
VI	**0.0131**	0.0146	0.0142	0.0155
HR	91,8	92,1	92,8	**93,5**
LC	44	47	44	**52**

Hellinger représente la distance de Hellinger moyenne entre les distributions des descripteurs des signaux originaux et avatars. VI est la différence moyenne entre les matrices des variations d'information des descripteurs originaux et avatars. HR et LC sont respectivement les hidden rate, en pourcentage, et les local cloaking médian.

La génération d'ECG avec la méthode suivant une loi de Dirichlet semble mieux préserver la distribution des principales caractéristiques des données, tandis que la méthode empirique obtient de meilleurs résultats en termes de privacy. Dans les deux cas, nous n'observons pas de grosses différences.

5 Conclusion

Dans cette communication, nous présentons une méthode de générations de données synthétiques alternatives à celle présentée par Octopize dans [7]. Cette méthode utilise une distribution de Dirichlet pour le tirage des coefficients de mélange attribué à chacun des voisins, suivant des paramètres appris sur la base de donnée originale. Cette contribution permet une meilleure appréhension du processus et semble mieux préserver l'information utile du signal.

Les métriques de privacy, sur la base des descripteurs extraits, montrent que les avatars sont protégés en regard du risque de ré-identification. Les métriques d'utilité, qui comparent les distributions de descripteurs des données ainsi que les relations de dépendances entre ceux-ci, prouvent que l'information utile du signal est conservée à travers ce processus d'anonymisation. Ainsi, les données avatars peuvent remplacer les données originales pour des études statistiques, des tâches de prédictions ou de classification.

Ce travail ne porte que sur le traitement d'une seule dérivation. Il pourrait être applicable aux autres dérivations, mais des travaux supplémentaires sont nécessaires pour s'assurer de la cohérence entre les différents signaux issus d'un même individu, et les métriques de privacy doivent être adaptées.

Le niveau d'anonymat des données a ici été évalué principalement en regard du risque de ré-identification, mais ce n'est pas le seul risque qui pèse sur les données. Un travail dédié à l'évaluation du niveau d'anonymat et des risques persistants sur les données générées semble nécessaire.

Chacune des étapes constituant le processus d'anonymisation des ECG peuvent faire l'objet d'une étude approfondie et être optimisées. Il existe également d'autres méthodes pour l'anonymisation d'ECG. Cette communication s'appuie sur l'exemple du cas d'usage d'un ECG pour illustrer le fonctionnement de la méthode présentée, sans en faire son objectif principal.

Références

[1] Edmond ADIB, Fatemeh AFGHAH et John J. PREVOST : Synthetic ecg signal generation using generative neural networks, 2022.

[2] N. AHMED, T. NATARAJAN et K.R. RAO : Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

[3] Zineb BENNIS et Pierre-Antoine GOURRAUD : Application of a novel anonymization method for electrocardiogram data. *In The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research*, pages 1–5, 2021.

[4] Emilie CHOUZENOUX, Maxime LEGENDRE, Saïd MOUSSAOUI et Jérôme IDIER : Fast Constrained Least Squares Spectral Unmixing Using Primal-Dual Interior-Point Optimization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):59–69, janvier 2014. Conference Name : IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[5] CNIL : Cnil - l'anonymisation des données personnelles.

[6] Cynthia DWORK : Differential privacy. *In 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 de *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, July 2006.

[7] Morgan GUILLAUDEUX, Olivia ROUSSEAU, Julien PETOT, Zineb BENNIS, Charles-Axel DEIN, Thomas GORONFLOT, Nicolas VINCE, Sophie LIMOU, Matilde KARAKACHOFF, Matthieu WARGNY *et al.* : Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1):37, 2023.

[8] Dominique MAKOWSKI, Tam PHAM, Zen J. LAU, Jan C. BRAMMER, François LESPINASSE, Hung PHAM, Christopher SCHÖLZEL et S. H. Annabel CHEN : NeuroKit2 : A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021.

[9] Jérôme PAGÈS : Analyse factorielle de donnees mixtes : principe et exemple d'application. *Revue de statistique appliquée*, 52(4):93–111, 2004.

[10] Latanya SWEENEY : k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[11] Vajira THAMBAWITA, Jonas L. ISAKSEN, Steven A. HICKS, Jonas GHOUSE, Gustav AHLBERG, Allan LINNEBERG, Niels GRARUP, Christina ELLERVIK, Morten Salling OLESEN, Torben HANSEN, Claus GRAFF, Niels-Henrik HOLSTEIN-RATHLOU, Inga STRÜMKE, Hugo L. HAMMER, Molly MALECKAR, Pål HALVORSEN, Michael A. RIEGLER et Jørgen K. KANTERS : DeepFake electrocardiograms : the beginning of the end for privacy issues in medicine. *medRxiv*, 2021. Publisher : Cold Spring Harbor Laboratory Press _eprint : https ://www.medrxiv.org/content/early/2021/05/10/2021.04.27.212

[12] Jianwei ZHENG, Jianming ZHANG, Sidy DANIOKO, Hai YAO, Hangyuan GUO et Cyril RAKOVSKI : A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, février 2020. Number : 1 Publisher : Nature Publishing Group.