L'ANONYMISATION DES DONNÉES STRATÉGIQUES AVEC AVATAR

Livre blanc coordonné par :

Confiance ai

AIRBUS · Air Liquide · Confiance ai · CIVITEO · Ethik-IA · GICAT · MyData-TRUST · Nantes Université · OCTOPIZE · pwc · Renault Group · sopra steria · SystemX · THALES · UNSW SYDNEY · VENDÉE LE DÉPARTEMENT

OCTOPIZE REMERCIE

Nos contributeurs

AIRBUS

Air Liquide

Confiance ai

CIVITEO

Ethik - IA
Garantie Humaine de l'

GICAT

MyData-TRUST

Nantes Université

pwc

Renault Group

sopra steria

SystemX

THALES

UNSW SYDNEY

VENDÉE
LE DÉPARTEMENT

et les partenaires de Confiance.ai

Air Liquide

AIRBUS

Atos

cea

Inria

NAVAL GROUP

Renault Group

SAFRAN

SAINT EXUPÉRY

sopra steria

SystemX

THALES

Valeo

FRANCE 2030

Ce travail a été soutenu par le gouvernement français dans le cadre du programme "France 2030", au sein de l'Institut de Recherche Technologique SystemX dans le cadre du programme Confiance.ai.

Edito

Au nom de toute l'équipe d'Octopize, nous dédions ce livre blanc à notre cher collègue Rémy, disparu ce 19 janvier. ***Rémy, tu vas nous manquer !***

Nos pensées les plus sincères vont à sa famille et ses proches.

C'est avec beaucoup de fierté que nous vous partageons ces travaux. Je vous souhaite une excellente lecture.

Olivier Breillacq

Dirigeant & Fondateur
Octopize

L'executive order du 28/02/2024 de Joe Biden constitue la mesure la plus importante jamais prise par un président pour protéger la sécurité des données personnelles des Américains. Ce décret autorise le procureur général à en empêcher le transfert à grande échelle, vers des pays qui suscitent des inquiétudes, et prévoit des garanties pour d'autres activités susceptibles de donner lieu à des transferts de données vers d'autres pays. Dans un contexte étatsunien peu propice à de telles entraves, c'est dire l'importance d'un tel enjeu. Ce choix politique nous oblige à la réciproque et appelle des solutions techniques, si possible nationales, qui en permettent l'effectivité.
L'avatarisation des données proposée par Octopize répond à cette problématique, parce qu'elle répond aux cahiers des charges de l'IA (ORIA, par exemple, essentielle pour notre recherche), assure la sécurité des données personnelles des Français et contribue à réduire notre handicap face aux grands groupes américains.

Philippe Latombe

Député de la **Vendée**,
Secrétaire de la **Commission des Lois**,
Commissaire à la **CNIL**

Sommaire

INTRODUCTION

Confiance.ai[1] est le pilier technologique du Grand Défi « Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle » lancé par le **Conseil de l'innovation**. Il s'agit du plus grand programme de recherche technologique du plan **#AIforHumanity**, qui vise à faire de la France l'un des pays leaders en matière d'intelligence artificielle (IA). Il a pour ambition de développer un **environnement méthodologique outillé** de confiance au service de la conception et de l'intégration d'une **IA sûre, fiable et sécurisée** dans les systèmes critiques (automobile, aéronautique, défense et sécurité, énergie, industrie…). Les avancées réalisées depuis 2021 sont consultables au travers d'un Livre Blanc[2] et de publications scientifiques disponibles dans la collection HAL.[3]

Dans un **environnement industriel** toujours plus connecté, les données jouent un rôle clé dans l'**optimisation des processus** et des **performances**. Toutefois, le partage et l'utilisation de ces données expose de nouveaux enjeux, notamment celui de leur **protection**. La nécessité de préserver la **confidentialité** tout en permettant leur **exploitation** pousse à explorer des techniques de désensibilisation. **L'anonymisation** regroupe un ensemble de méthodes visant à altérer les données afin de rendre impossible l'identification des individus, tout en conservant un haut niveau d'utilité.

Dans le cadre des **données industrielles**, elle peut être envisagée comme une stratégie de protection, adaptée aux défis spécifiques de ce secteur. Ce livre blanc présente les **résultats de travaux** explorant les perspectives de désensibilisation des données industrielles. Ces recherches se concentrent sur des **données collectées par des capteurs**, mises à disposition par un membre du programme Confiance.ai pour l'entraînement de modèles de détection d'anomalies. En plus de cette approche de désensibilisation, une attention particulière est portée à l'**anonymisation des séries temporelles**, un type de données largement présent dans les environnements industriels.

Par sa compatibilité avec plusieurs types de données dont les séries temporelles, la **méthode avatar développée par Octopize** est particulièrement adaptée à cette étude.
Par cette approche, nous démontrons comment l'anonymisation peut non seulement **sécuriser les données** mais aussi maintenir leur **utilité pour l'analyse** et la modélisation. Ainsi, ce livre blanc fournit des perspectives précieuses pour allier protection des données et innovation industrielle, ouvrant la voie à une **gestion plus sûre et efficace des informations sensibles.**

1 https://www.confiance.ai/
2 https://www.confiance.ai/contenus-media/
3 https://hal.science/CONFIANCEAI/search/index/?q=*&rows=30&sort=producedDate_tdate+desc

L'apport des méthodes d'anonymisation des données stratégiques basées sur l'avatarisation ouvre des perspectives inédites : amélioration de la protection de la vie privée, meilleure équité et une utilité avérée pour le développement de systèmes d'IA (machine learning).

Appliquée à la santé, cette approche facilite le partage de données sensibles. Elle permet notamment de prédire les résultats d'un essai clinique et valider un projet de recherche, accélérer les phases de contractualisation entre les acteurs, ou encore offrir des possibilités en matière de remplacement des personnes humaines dans les pratiques interventionnelles de recherche.

Octopize fait la démarche de qualification et de preuve de leur processus d'anonymisation pour obtenir une donnée de synthèse anonyme et statistiquement pertinente. Associé à un modèle d'IA encadré par une supervision humaine, l'avatarisation garantit une analyse robuste et éthique, catalysant innovation et protection.

La mobilisation de ces outils d'avatarisation et de Garantie Humaine pourrait offrir des garanties de sécurité et de conformité AI Act pour les patients et les usagers, tant au niveau de la protection des données que sur les garanties de confiance pour les produits de santé développés à l'aide de données artificielles.

David Gruson

Directeur de Programme Santé à Domicile,
La Poste Santé & Autonomie
Fondateur, **ETHIK-IA**

DONNÉES STRATÉGIQUES

B.

B.1 De quoi parle-t-on ?

Une **donnée stratégique** est une information spécifique considérée comme essentielle pour une organisation. Souvent confidentielles, les données stratégiques représentent une valeur certaine pour l'organisation. Elles favorisent la compréhension du contexte concurrentiel, des tendances du marché, des opportunités ou des risques. L'identification, la collecte, l'analyse et l'utilisation efficace des données stratégiques sont essentielles pour élaborer et mettre en œuvre des plans stratégiques et maintenir un avantage concurrentiel sur le marché.

B.2 Pourquoi les protéger ?

Les données stratégiques peuvent contenir des **informations sensibles** sur les objectifs commerciaux, les stratégies de marché, les innovations en cours de développement, les données financières etc. La divulgation de ces informations à des concurrents ou à des parties non autorisées peut compromettre la **position concurrentielle** de l'entreprise.

De nombreuses données stratégiques sont par ailleurs soumises à des **réglementations strictes** en matière de confidentialité et de protection des données, telles que le Règlement Général sur la Protection des Données (**RGPD**), le **secret des affaires**, ou l'**IG 13100** s'agissant des informations classifiées. Les entreprises sont tenues de se conformer à ces réglementations pour éviter d'être sanctionnées.

Au-delà du cadre juridique, la protection des données stratégiques renforce la **confiance** des clients, des partenaires commerciaux ou des investisseurs dans l'entreprise, démontrant son engagement envers la sécurité et la confidentialité des informations.

Protéger les données stratégiques est donc essentiel pour garantir la **compétitivité**, la **conformité** réglementaire, la **confiance** des parties prenantes, et la **pérennité** de l'entreprise, en particulier dans les environnements commerciaux concurrentiels ou s'agissant de domaines intrinsèquement sensibles (énergie, défense, etc.).

B.3 Quelles données pour quel usage ?

L'un des membres du programme Confiance.ai possède et exploite des **unités de production** équipées de capteurs permettant de surveiller leur bon fonctionnement. Dans le cadre du programme, cet industriel nous a transmis les données issues de ces capteurs pour traiter un **cas d'usage lié à la détection d'anomalies**. Étant donné la nature des secteurs concernés, ces données présentent une certaine sensibilité, notamment en ce qui concerne les procédés de production. Leur protection représente ainsi un enjeu stratégique majeur. Par conséquent, ces données ont été préalablement rendues **non-identifiantes** avant leur partage.

L'anonymisation de la donnée est un sujet central pour l'avenir du Machine Learning. Il est indispensable de se doter de technologies qui permettent de protéger les données sensibles tout en offrant une exploitation large de ces données qui contiennent des informations riches. L'accessibilité à ces données va nécessairement renforcer la fiabilité des système d'Intelligence Artificielle.

Les travaux menés par Octopize et Sopra Steria décrits dans ce livre blanc permet de lever des freins à l'utilisation du Machine Learning dans des domaines pour lequels la confidentialité de la donnée est essentielle, en particulier la santé, la défense, ou l'énergie mais aussi plus largement dans tous les cas où le secret de la donnée brute est important. L'approche garantit leur sécurité avec des métriques concrètes qui sont absolument nécessaires à une démarche de confiance. Ce dernier point est une des grandes forces de ce travail, car toutes les approches ne sont pas aujourd'hui assorties de telles métriques.

Yves Nicolas

AI group Program Director
Deputy Group CTO
Sopra Steria

L'ANONYMISATION, VECTEUR DE PROTECTION DES DONNÉES STRATÉGIQUES

C.

C.1 Qu'est-ce que l'anonymisation ?

Selon la définition de la **CNIL**, l'anonymisation est un traitement réalisé sur les données personnelles, qui consiste "à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible".

Le Comité européen de la protection des données (**CEPD**) a défini trois critères qui permettent de s'assurer qu'une donnée est réellement anonyme :

- **L'individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
- **La corrélation** : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
- **L'inférence** : il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu.

C.2 Dans quel cas parler d'anonymisation ?

Selon sa définition, le périmètre d'application de **l'anonymisation** est celui des données personnelles. La notion d'individu est présente dans chacun des critères que sont l'individualisation, la corrélation et l'inférence.

Au-delà, l'anonymisation consiste à s'assurer qu'il n'est pas possible de remonter à l'individu à l'origine des données. En d'autres termes, il s'agit de **rendre impossible la réidentification** des informations ayant engendré lesdites données.

En ce sens, si le terme réfère littéralement aux données personnelles, l'anonymisation permet de protéger plus largement les données sensibles, confidentielles, ou stratégiques. Anonymiser revient alors à abaisser le seuil de sensibilité des données traitées.

C.3 L'anonymisation adaptée aux données stratégiques

L'anonymisation des données stratégiques est essentielle pour protéger les **informations confidentielles** d'une organisation, tout en rendant possible leur utilisation à des fins **analytiques ou de recherche.**

L'anonymisation consiste à modifier ces données de manière à ce qu'elles ne puissent plus être directement associées à des individus ou à des entités spécifiques, tout en préservant leur utilité pour l'analyse. Cela peut inclure **l'agrégation** ou la **généralisation** des données pour empêcher toute identification indirecte. La vérification de la suppression des risques d'inférence, de corrélation et d'individualisation permet de s'assurer de la réduction des risques de compromission de la confidentialité.

Octopize a rejoint Cyber@StationF, l'accélérateur de startups de Thales dédié à la cybersécurité, en 2024. Cette collaboration représente un enjeu majeur pour le traitement des données stratégiques et confidentielles en Défense. Ensemble, nous explorons l'application de leur méthode d'anonymisation avatar aux données de Défense, en particulier pour l'entraînement d'algorithmes de Machine Learning. Forts des avancées réalisées dans le cadre du programme d'accélération, Octopize bénéficie de l'accompagnement de coachs techniques et business de Thales, leur permettant de tirer parti de leur expertise. Ils rencontrent également des clients de Thales pour identifier des cas d'usage concrets. Un proof of concept (POC) est actuellement en cours, et nous prévoyons de partager les résultats dans un futur livre blanc consacré à l'IA et à la Défense.

Marine Martinez
Program Lead Cyber@StationF
Thales

Dans le secteur de la Défense et de la Sécurité, la donnée reste un élément central, tant au sein de la planification et la conduite des opérations terrestre comme des opérations d'armement, qu'au sein de la conception et la fabrication des équipements des forces armées par l'industrie de défense.
Au-delà, ce sont aussi des enjeux quant à l'explosion des données à traiter sur les théâtres d'opération, tout autant que des considérations croissantes d'utilisation des données issues de l'entraînement des forces armées, toujours plus digitalisées. Octopize est ainsi un exemple de start-up ayant un rôle à jouer dans la capacité des forces et des entreprises à échanger de la donnée, parmi les autres pépites technologiques françaises accompagnée par le label innovation et accélérateur de start-up GENERATE du GICAT.

Hubert Raymond

Responsable de l'Innovation et du programme GENERATE,
Référent Contrats Publics,
GICAT (Groupement des industries françaises de défense et de sécurité terrestres et aéroterrestres)

LES TECHNIQUES D'ANONYMISATION DES DONNÉES

D.

D.1 Panorama des familles d'anonymisation existantes

Le CEPD définit deux grandes familles de techniques d'anonymisation : la **randomisation** et la **généralisation**.

◆ (i) La **randomisation** consiste à modifier les attributs dans un jeu de données de telle sorte qu'ils soient moins précis, tout en conservant la répartition globale. Cette technique permet de protéger le jeu de données du risque d'inférence. Dans les techniques de randomisation, on peut par exemple citer l'ajout de bruit, la permutation et la confidentialité différentielle.

◆ (ii) La **généralisation** consiste à modifier l'échelle des attributs des jeux de données, ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes. Cette technique permet d'éviter l'individualisation d'un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d'autres. Dans les techniques de généralisation, on peut par exemple citer l'agrégation, le k-anonymat, le l-diversité ou encore le t-proximité.

Chacune des techniques d'anonymisation peut être appropriée, selon les circonstances et le contexte, pour atteindre la finalité souhaitée sans compromettre le droit des personnes concernées au respect de leur **vie privée**.

D.2 Méthode avatar d'Octopize

D.2.1 Principes

La **méthode avatar** est une approche unique de génération de données synthétiques anonymes, qui préserve la structure et la pertinence statistique du jeu de données original tout en respectant la confidentialité liée auxdites données. Cette technique utilise une approche **centrée sur l'individu** en créant des simulations locales basées sur ce dernier, ce qui rend la simulation d'un avatar unique. La méthode avatar est conçue pour répondre aux trois critères énoncés par le **CEPD** pour évaluer la robustesse d'un processus d'anonymisation.

Comparé à d'autres techniques telles que les arbres de décision et les Generative Adversarial Networks (**GAN**), le logiciel avatar démontre une utilité similaire dans la préservation de la structure et de la pertinence statistique du jeu de données d'origine. En outre, le logiciel avatar inclut des **mesures de confidentialité** qui permettent d'évaluer la protection apportée aux données anonymisées au regard des trois critères définis par le CEPD.

La méthode avatar prend en entrée des données originales et produit des **données synthétiques et anonymes** de même taille et de même nature. Les données numériques restent numériques, les données catégorielles restent catégorielles, etc. Le cœur de la méthode est illustré dans la Figure 1 et décrit ci-après.

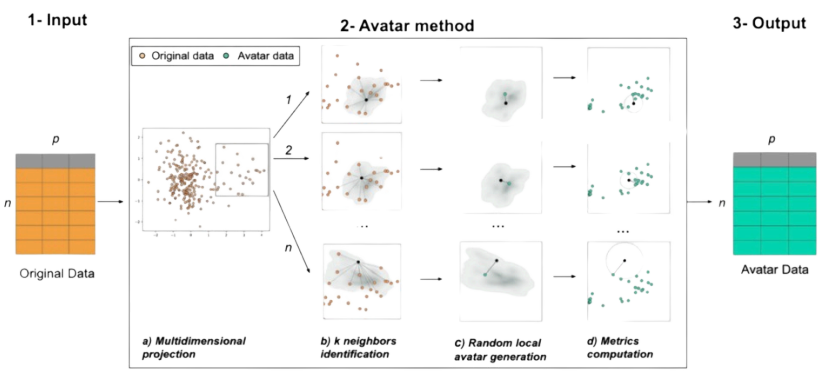

Figure 1 : Principes de la méthode avatar.

[1] Article sur la méthode dans Nature Digital Medicine : https://www.nature.com/articles/s41746-023-00771-5

D.2.2 Projection multidimensionnelle

Les données originales sont projetées dans un **espace multidimensionnel** approprié à l'aide de techniques de réduction de dimensions telles que l'analyse factorielle des données mixtes (FAMD), l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM). Les transformations utilisées doivent être **réversibles**, c'est-à-dire qu'il existe une transformation inverse qui permet de revenir à l'espace de représentation d'origine.

Cette étape transforme les individus, qui sont initialement décrits par plusieurs caractéristiques numériques et catégorielles, en **coordonnées numériques structurées** qui facilitent le calcul des distances entre les individus. Elle réduit également la dimensionnalité du jeu de données afin de mettre en évidence les informations les plus pertinentes.

D.2.3 Calcul des k-voisins

Les distances entre voisins sont ensuite calculées entre tous les points de cet espace afin d'appliquer un **algorithme de k-voisins les plus proches (KNN)**. Celui-ci définit une zone locale autour de chaque coordonnée - chacune étant la projection d'un individu à partir des données originales - définie par ses plus proches voisins.

D.2.4 Génération aléatoire de coordonnées avatars

Pour chacune de ces zones locales, une simulation unique est tirée de manière pseudo-aléatoire, créant une nouvelle coordonnée à l'intérieur de la zone, que nous appelons l'avatar de la coordonnée d'origine. Cette simulation est influencée par la distance entre le point d'origine et chacun de ses voisins, par un **poids aléatoire** suivant une **distribution exponentielle** et par un facteur de **contribution aléatoire** pour chaque voisin.

Cela permet aux simulations non-déterministes d'être considérées comme un processus irréversible, ce qui est une condition nécessaire à la conservation de la confidentialité.

D.2.5 Inversion de la transformation pour revenir à l'encodage d'origine

Une fois qu'une donnée synthétique a été générée pour chaque individu, les coordonnées de l'avatar sont inversées pour revenir à l'encodage original, en conservant le type des attributs originaux (catégoriques, numériques, etc.). Bien qu'il soit impossible de récupérer les données originales à partir des données avatar, la **structure de l'ensemble de données est préservée** comme illustré dans la Figure 2.

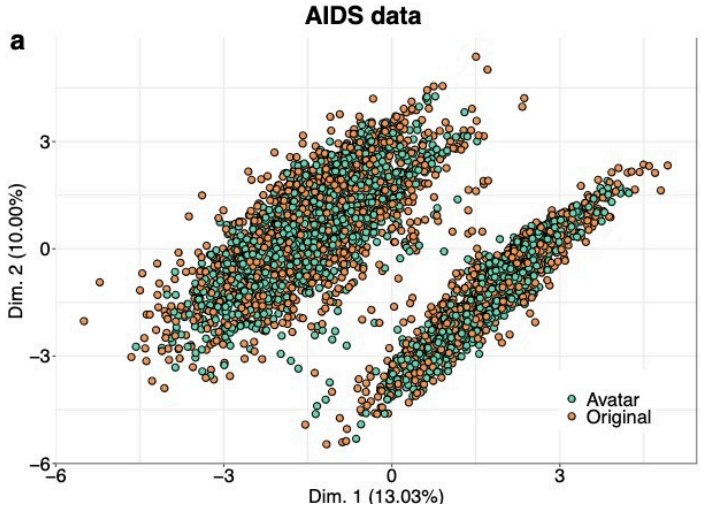

Figure 2 : Conservation de la structure du jeu de données après anonymisation : les points avatars couvrent globalement l'ensemble des points originaux, exception faite des points isolés représentant des individus extrêmes. Les deux dimensions de la FAMD expliquant le plus de variances sont représentées.

D.2.6 Calcul des paramètres de la confidentialité des données

Le logiciel avatar comprend des **métriques de mesure de la confidentialité** des données anonymisées qui sont essentielles pour prouver la protection apportée aux données. Le **rapport d'anonymisation** détaillant ces métriques, et automatiquement généré par le logiciel, constitue une réelle analyse de risque. Les métriques visent à couvrir différents scénarios d'attaque et répondent aux trois critères du CEPD : individualisation, corrélation et inférence.
La première famille de métriques vise à évaluer la protection d'un jeu de données contre les attaques **d'individualisation**.

Ces attaques peuvent prendre différentes formes, ce qui nécessite différentes mesures complémentaires. Lors du calcul de ces métriques, les liens entre individus originaux et traités sont conservés temporairement. Ces liens permettent de valider et de quantifier le succès des attaques simulées.

Ci-dessous trois exemples de métriques d'individualisation automatiquement calculées par le **logiciel avatar** :

◆ **Distance To Closest**. Pour calculer la DTC, on mesure la distance entre chaque individu synthétique et son original le plus proche. La valeur médiane est conservée afin d'avoir une seule valeur représentative associée à cette mesure. Le raisonnement qui sous-tend la DTC est que si chaque individu synthétique est proche d'un original, l'ensemble de données pourrait présenter un risque d'individualisation. Toutefois, une DTC faible ne signifie pas nécessairement qu'il y a un risque et, par conséquent, le Closest Distances Ratio doit être mesuré pour le compléter.

◆ **Closest Distances Ratio**. De même que pour le DTC, le CDR est calculé en mesurant d'abord la distance entre un avatar et son individu original le plus proche, divisée par la distance avec son second individu original le plus proche. En d'autres termes, on mesure la distance entre les deux individus originaux les plus proches. Si le rapport est élevé, les deux originaux les plus proches sont à la même distance et il est donc impossible de les distinguer avec certitude en pratique. A partir des ratios calculés pour chaque individu traité, la médiane est conservée pour fournir une valeur unique de CDR. Il existe un risque d'individualisation lorsque le DTC et le CDR sont tous deux faibles.

◆ **Hidden Rate**. La Hidden Rate est la probabilité qu'un attaquant commette une erreur en reliant un individu à son avatar (individu synthétique) le plus similaire. C'est là que le lien entre l'original et l'avatar qui a été conservé temporairement devient utile.

◆ **Local Cloaking**. Pour obtenir le Local Cloaking, le nombre d'avatars entre un individu et l'avatar qu'il a généré est calculé pour chacun des individus. Le Local Cloaking est la valeur médiane obtenue. Notons que la Hidden Rate et le Local Cloaking sont liés puisque la Hidden Rate représente le nombre d'individus pour lesquels l'avatar d'un individu est le plus proche avatar de cet individu.

La deuxième famille de métriques répond au critère de **corrélation**. Ces métriques répondent à un scénario d'attaque courant et probable.

L'attaquant dispose d'un jeu de données traitées et d'une base de données d'identification externe (par exemple, un registre des électeurs) contenant des informations communes avec les données traitées (par exemple, l'âge, le sexe, le code postal). Plus il y a d'informations en commun entre les deux bases de données, plus l'attaque sera efficace.

Le taux de protection contre les corrélations (**Correlation Protection Rate**) est une métrique qui permet d'évaluer le pourcentage d'individus qui ne seraient pas reliés avec succès à leur homologue synthétique si l'attaquant utilisait une source de données externe. Les variables sélectionnées comme étant communes aux deux bases de données doivent être susceptibles d'être trouvées dans une source de données externe.

Pour couvrir le pire des scénarios, nous supposons que les mêmes individus sont présents dans les deux bases de données.
En pratique, certains individus de la base de données anonymisée ne sont pas présents dans la source de données externe et vice versa. Cette métrique repose également sur le fait que le lien entre l'original et le synthétique est conservé temporairement. Ce lien est utilisé pour mesurer combien d'appariements sont incorrects.

Les métriques qui répondent au critère **d'inférence** correspondent à un autre type d'attaque ; l'attaquant cherche à déduire des informations supplémentaires sur un individu à partir des données anonymisées disponibles.

La métrique d'inférence calcule la possibilité de déduire, avec une probabilité significative, la valeur originale d'une variable cible à partir des valeurs d'autres variables traitées. La métrique d'inférence peut être utilisée sur des **cibles numériques et catégorielles**. Lorsque la cible est numérique, on parle de métrique d'inférence de régression et on évalue la protection comme la différence absolue moyenne entre la valeur prédite par l'attaquant et la valeur numérique originale.

Pour Air Liquide Healthcare, la protection des données personnelles est une responsabilité majeure pour une utilisation raisonnée et conforme de l'information. Permettre d'élargir le potentiel d'investigation en générant des jeux de données anonymisées avec avatar ouvre des perspectives immenses, en particulier dans le domaine de la Santé. La définition d'algorithmes de prédiction d'observance grâce à l'IA permet d'adapter les plans d'accompagnement des patients dans le traitement de leur maladie chronique ou encore d'aider au diagnostic précoce de l'évolution de la maladie. Ce sont des exemples concrets pour lesquels l'anonymisation des données rend possible ces innovations au bénéfice du patient.

Olivier Gruet

Programs Director & Chief Data officer
Air Liquide Healthcare

D'autre part, nous parlons de métrique d'inférence de classification lorsque la cible est catégorielle et le niveau de protection est représenté par la précision de la prédiction.

Les métriques détaillées ci-dessus ne sont qu'un aperçu de l'ensemble des métriques mises à disposition dans le **rapport d'anonymisation automatiquement généré par le logiciel avatar**. Une telle méthodologie permet de générer des jeux de données anonymes avec un modèle entièrement explicable et des mesures de confidentialité concrètes qui permettent à l'utilisateur de mesurer le degré de protection.

D.3 Application de la méthode avatar pour ce cas d'usage

La méthode avatar, telle que décrite ci-dessus, est une **méthode d'anonymisation**. Elle vise donc à **protéger** l'individu à l'origine des données, puisque l'anonymisation s'applique aux données personnelles. Néanmoins et comme indiqué ci-dessus, la méthode peut tout à fait s'entendre au sens de la **désensibilisation** de données confidentielles ou stratégiques. Dans ce cas-là, l'individu, basiquement représenté par une ligne dans un jeu de donnée, peut être assimilé à une machine, un capteur, ou même, un espace-temps selon la nature du jeu de données que l'on souhaite anonymiser.

Dans un contexte général, anonymiser des données requiert d'avoir une notion d'individu ou d'entité dans les données puisque l'anonymisation modifie les données dans le but de protéger ces individus en les cachant parmi leurs voisins respectifs. Pour ce cas d'usage industriel, la notion d'individu n'est pas définie puisque les données d'une variable se rapportent à une seule machine. L'approche retenue pour permettre l'anonymisation du jeu de données consiste à **découper le jeu de données en plages temporelles**, chacune étant assimilée à un individu. Ce choix de segmentation répond à un besoin métier. Dans d'autres contextes, cette segmentation peut être différente voire bien souvent naturelle en fonction de ce que l'on souhaite protéger.

Par exemple, les données d'une machine partagées par plusieurs utilisateurs peuvent être naturellement divisées en segments représentant des sessions d'utilisations distinctes. Une fois anonymisés, ces segments représentent toujours des sessions d'utilisation mais il est impossible de les réidentifier ainsi que l'utilisateur à qui elles sont rattachées. Un autre cas d'usage peut nous amener à anonymiser des données provenant de plusieurs machines du même type (par exemple, des respirateurs). Dans ce cas, l'entité à protéger est l'utilisateur de chacune des machines (que ce soit une entreprise ou un individu).

La segmentation peut donc naturellement être faite par les identifiants des machines ou par des identifiants de sessions.

Les données exploitées pour les analyses de ce document proviennent de capteurs enregistrant des relevés durant des opérations continues dans un environnement industriel. Ces relevés forment des suites de mesures, autrement dit des **séries temporelles**.

Les **séries temporelles** diffèrent des données dites **tabulaires** ou **statiques** dans le fait que la relation entre des relevés successifs fait partie intégrante des données. C'est elle qui permet de définir des **tendances** ou des **évolutions** dans le temps (par exemple hausse de pression ou de température...). Les grands principes de la méthode avatar décrits jusqu'à présent peuvent être utilisés sur des données de types séries temporelles. En effet, les étapes de projections, de calcul de voisins et de génération de coordonnées synthétiques restent cohérentes.

Cependant, le **type de projection** est différent entre données tabulaires et séries temporelles. Les approches de **projections tabulaires** utilisées dans la méthode avatar sont l'analyse en composantes principales (ACP) et ses dérivées. Leur utilisation sur des séries temporelles auraient pour effet de perdre toute information relative au séquencement des points. Il est donc nécessaire d'utiliser une **méthode de projection ou de transformation** propre aux séries temporelles. Il existe plusieurs techniques de ce type comme les transformées de Fourier, les transformées en cosinus discrètes ou encore la décomposition en ondelettes. Ces approches ont déjà été utilisées dans le contexte de la génération d'avatars dans des contextes médicaux [2].

Il existe également une version adaptée de **l'ACP** pour le domaine fonctionnel, permettant de modéliser une variable en fonction d'une autre. Cette ACP fonctionnelle (FPCA) peut être appliquée aux séries temporelles, car ces données représentent une fonction traduisant l'évolution d'une variable dans le temps. Cette méthode est également idéale pour la réduction de dimensions, ce qui en a fait notre choix privilégié. De plus, tout comme l'ACP classique, la FPCA permet d'effectuer une **transformation inverse**, permettant ainsi de revenir des coordonnées aux variables initiales, comme dans le jeu de données d'origine. Pour en savoir davantage sur la FPCA, nous recommandons l'**article de Wang et al.** [3].

En pratique, les données peuvent inclure plusieurs variables temporelles, qui peuvent avoir des **fréquences d'échantillonnage** différentes, être périodiques ou non. Par ailleurs, ces variables sont souvent associées à des **données fixes**. La méthode avatar s'adapte à ce type de contexte. Les différentes étapes qui rendent cela possible sont illustrées et détaillées dans la Figure 3.

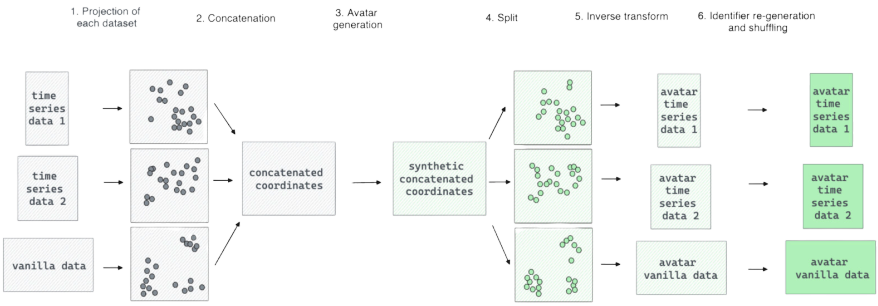

Figure 3 : Différentes étapes dans l'anonymisation de données mixtes comportant des séries temporelles et des données statiques.

Tout d'abord, chacun des jeux de données est projeté dans un **espace numérique** (étape 1). Comme mentionné précédemment, les données temporelles sont projetées avec la FPCA, tandis que les variables statiques (dénotées vanilla dans la Figure 3) sont projetées avec l'ACP ou ses dérivées. Ensuite, pour anonymiser l'ensemble des données en une seule étape, les coordonnées obtenues à partir des différentes projections sont **concaténées** (étape 2). Le processus de **génération des coordonnées synthétiques** est alors appliqué (étape 3), puis ces coordonnées synthétiques sont redivisées par **jeu de données** (étape 4) pour permettre la **transformation inverse** (étape 5). Au final, cet enchaînement d'étapes produit autant de jeux de données en sortie qu'il y en avait en entrée.

L'exemple présenté dans la Figure 4 permet de se rendre compte visuellement du résultat d'une anonymisation de séries temporelles avec la méthode avatar. On voit en particulier que les **tendances et les caractéristiques globales sont conservées** mais que certaines séquences de valeurs propres à une seule entité ne le sont pas. Cela met en évidence une bonne conservation du **signal** ainsi qu'un apport de **privacy**.

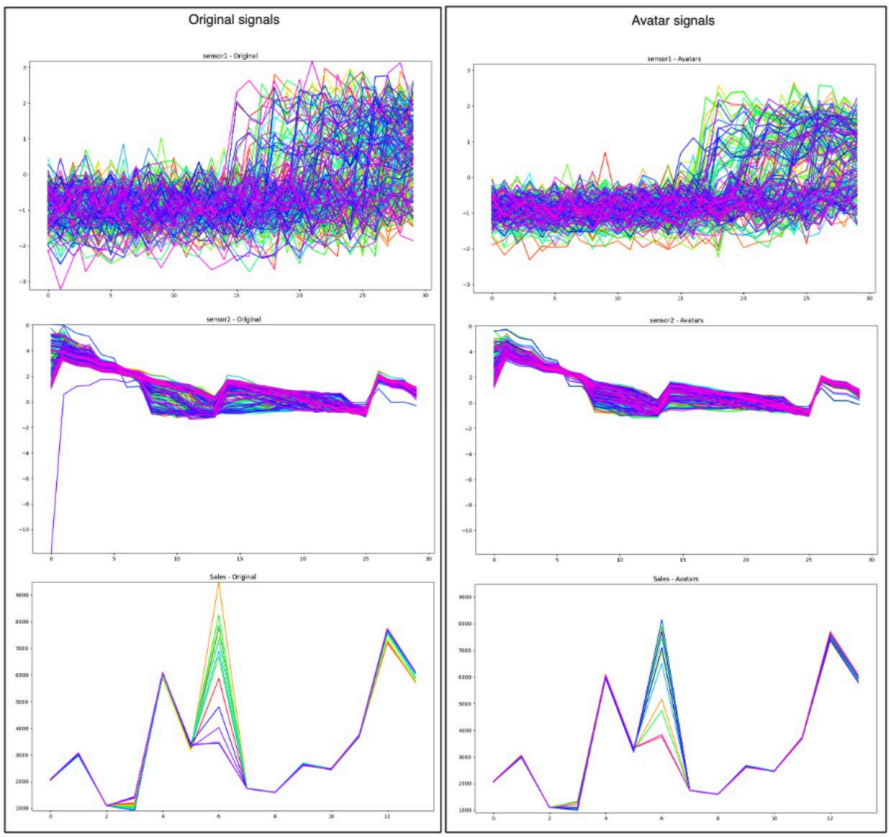

Figure 4: Exemples de 3 variables séries temporelles originales (à gauche) et leurs avatars (à droite).

La simple suppression des champs de données personnelles directement identifiantes (nom, prénom…) d'un jeu de données ne permet pas toujours une anonymisation du jeu de données. En effet, plus celui-ci contient des champs renseignés à la maille individuelle, plus les personnes seront identifiables en croisant les différentes caractéristiques des différentes colonnes, comme dans le jeu Qui est-ce ?

Dans ce cas, l'avatarisation du jeu de données prend tout son sens. Celle-ci consiste à générer un nouveau jeu de données avatar dans lequel aucun point de données issu du jeu de données initial ne se retrouvera. Le nuage de points avatar est conçu pour préserver les propriétés statistiques du nuage de points initial, tout en s'adaptant aux exigences spécifiques du cas d'usage. En générant un jeu de données synthétique, elle ouvre la voie à une exploitation statistique tout en minimisant le risque d'exposition de données sensibles. À ce titre, elle peut permettre un changement de paradigme dans la gestion des données.

Alexis Rouet

Chief Data Officer HR
Renault Group

PRODUCTION DE JEUX DE DONNÉES ANONYMISÉES

E.1 Choix du périmètre

Bien que la méthode avatar puisse s'appliquer à plusieurs capteurs, ce livre blanc se concentre sur les **résultats obtenus avec l'algorithme de détection d'anomalies 1D-CNN**, appliqué à un capteur déjà analysé dans des travaux précédents de Confiance.ai. Les données de ce capteur, rendues préalablement non-identifiantes, ont été fournies par l'industriel.

Les données des capteurs couvrent une **période d'un an**. Pour l'entraînement du modèle 1D-CNN, elles sont divisées en training et testing sets et au sein de chaque set, les données sont divisées en segments. Chacun de ces segments représente alors un échantillon pour l'entraînement du modèle.

Il existe plusieurs manières de segmenter des séries temporelles. Dans le cadre de ces travaux, 3 approches ont été considérées :

◆ **Segmentation par fenêtre de temps fixe et exclusive**. Avec cette segmentation, la série temporelle est divisée en plusieurs séries de taille prédéfinie (par exemple 7h). Chaque pas de temps n'est représenté que dans un seul segment (exclusivité).

◆ **Segmentation par détection de pic**. Cette méthode consiste à appliquer une étape de détection de pics pour identifier les pas de temps où le signal atteint ses valeurs les plus élevées. Le découpage en segments s'effectue autour de ces pics, de manière à ce qu'ils se situent au centre du segment, tout en respectant une taille prédéfinie (par exemple 7 heures). La segmentation par détection de pic est particulièrement recommandée pour des données cycliques. Elle permet de préserver les corrélations potentielles entre les pas de temps et les valeurs du signal, tout en restant compatible avec une réduction de dimensions.

◆ **Segmentation par fenêtre glissante**. Enfin, la troisième méthode de segmentation utilise une fenêtre glissante de taille prédéfinie qui est décalée d'un certain nombre de pas de temps entre 2 segments. Ce décalage est contrôlé par un paramètre de chevauchement *stride*. Selon le stride utilisé, cette segmentation peut potentiellement générer un grand nombre de segments (à stride=1, le nombre de segments est maximum).

Étant donné que le modèle cible de détection d'anomalies est un **réseau de neurones**, un large volume de données est nécessaire pour garantir des résultats optimaux. Pour cette raison, la segmentation par **fenêtre glissante** avec un stride de 1 a été retenue pour cette analyse. Toutefois, la segmentation par **détection de pic** a été utilisée pour certaines illustrations.

Les segments étant modifiés au cours du processus d'anonymisation, il devient impossible de les regrouper pour reconstituer des données représentant l'évolution des variables sur l'ensemble de l'année. Néanmoins, ces segments restent parfaitement adaptés pour l'**entraînement de modèles.**

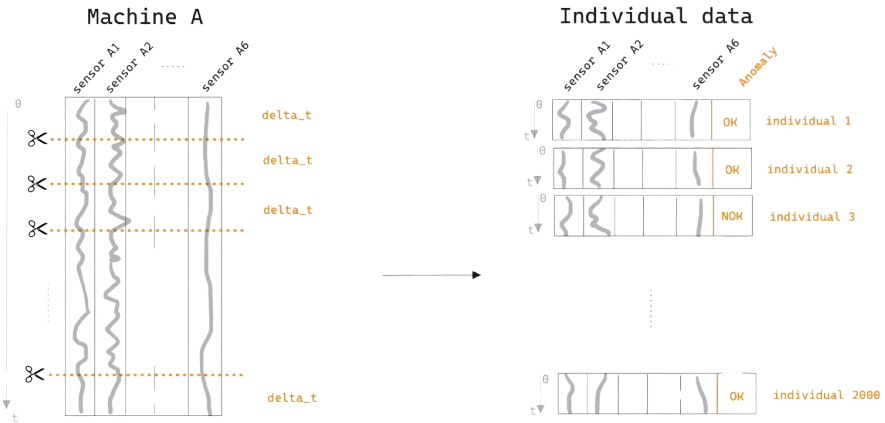

Figure 5: Segmentation des données annuelles d'une machine ou d'un système en entités individuelles (segments). Dans cet exemple, une année de données génère 2000 individus qui peuvent être anonymisés.

E.2 Contraintes rencontrées

L'anonymisation de données est un processus se basant sur la **modélisation de données**, tout comme la détection d'anomalies. De manière générale, l'anonymisation de données doit être effectuée sur des **populations d'individus** qui font partie d'un même contexte et qui peuvent donc être comparées entre eux. Ce qui est valable pour des individus l'est aussi pour des données issues de machines.

Pour s'inscrire dans un même **contexte**, des données temporelles doivent être définies sur une **même plage de temps**. Une étape de **normalisation du temps** est donc appliquée en amont de l'anonymisation. Cette étape redéfinit chaque segment sur une plage de temps allant de t=0 (début du segment) à t=1 (fin du segment). Ainsi, une fois normalisées, tous les segments peuvent **être comparés** entre eux comme illustrés ci-dessous.

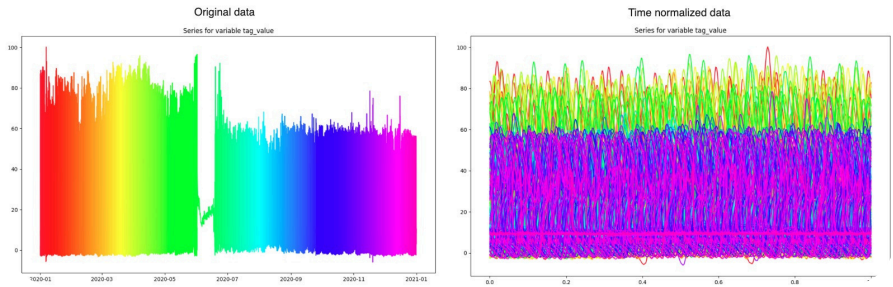

Figure 6 : Normalisation des pas de temps pour les données du cas d'usage. Chaque couleur représente une entité distincte. La plage temporelle originale couvre une année complète. Suite à la normalisation, tous les points sont dans l'intervalle de temps [0, 1].

Une deuxième contrainte doit être respectée. Celle-ci est propre à la **méthode FPCA** utilisée pour projeter les séries temporelles. La FPCA nécessite que les séries temporelles soient alignées, c'est-à-dire qu'elles aient des **mesures effectuées aux mêmes pas de temps**. Les courbes d'une variable devront donc être rééchantillonnées si les mesures ne sont pas effectuées à la même fréquence ou en général aux mêmes pas de temps. Pour cela, un certain nombre de pas de temps périodiques est défini entre t=0 et t=1, et les valeurs des capteurs sont inférées pour chacun de ces pas à l'aide d'une interpolation linéaire.

Il est important de noter que le nombre de pas de temps ne peut dépasser le nombre d'entités distinctes présentes dans les données à anonymiser, ce qui constitue une contrainte de la FPCA. Par conséquent, une **perte de précision** peut survenir dans le cas de données comportant peu d'entités. Cependant, cette limitation ne s'applique pas aux données segmentées par fenêtre glissante avec un stride égal à 1.

E.3 Évaluation de la protection apportée aux données

La méthode avatar modifie les données de manière à transformer chaque entité de façon distincte, dans le but de **maximiser l'utilité** tout en garantissant la protection de la **vie privée**. Cette méthode est **paramétrable**, permettant aux utilisateurs d'ajuster le niveau de confidentialité en fonction de leur cas d'usage. Il est donc primordial pour l'utilisateur de valider les jeux de données générés en évaluant les **métriques de confidentialité.**

Il est important de noter que le calcul de **métriques de confidentialité** est recommandé pour valider tout jeu de données anonymisé, quelle que soit la méthode utilisée. Cela inclut même les approches présentées comme **private by design**, celles-ci étant également paramétrables et pouvant, par conséquent, produire des données non anonymes.

La nécessité d'implanter des solutions d'IA de confiance est un enjeu majeur pour les PME et, plus largement, pour toutes les entreprises souhaitant maintenir ou rehausser leur productivité et leur croissance. Aujourd'hui, l'IA n'est plus un atout : c'est désormais une nécessité.

L'IA de confiance repose avant tout sur des données de confiance. Développer des solutions personnalisées ou s'appuyer sur des solutions existantes nécessite de comprendre les défis industriels, mais aussi de garantir la qualité, la fiabilité et la confidentialité des données partagées. C'est précisément ce que permet la méthode avatar.

C'est la clé pour :

- *rassurer les partenaires commerciaux dans leurs projets d'opérationnalisation,*
- *protéger les citoyens en se conformant aux cadres réglementaires et aux régulations internationales,*
- *et accélérer le déploiement de solutions d'IA de confiance au sein des entreprises.*

Seules des données de confiance permettront d'assurer la réussite des projets d'IA tout en répondant aux exigences de sécurité, confidentialité et de transparence.

Marie-Pierre Habas-Gerard

Directrice Générale - Consortium industriel en Intelligence Artificielle industrielle
Confiance IA Canada

Le **cas d'usage** se compose exclusivement de données temporelles, et seules les métriques répondant au critère **d'individualisation** du RGPD seront évaluées. L'état de l'art concernant les métriques de confidentialité pour les séries temporelles ne permet pas d'identifier des métriques spécifiques. Toutefois, il est possible d'utiliser les métriques d'individualisation présentées en D.2.6, car elles reposent sur des coordonnées dans un **espace projeté**. Étant donné que nous appliquons une projection (FPCA) pour traiter les séries temporelles, il devient possible de calculer des métriques telles que la **Hidden Rate**, le **Local Cloaking**, ainsi que toute autre métrique calculées sur des coordonnées.

Les métriques de confidentialité obtenues lors d'un run d'anonymisation avec **k=20** sont présentées dans le Tableau 1. Pour rappel, k est le paramètre qui ajuste le **compromis entre la préservation du signal et la confidentialité**. Les objectifs (cibles Octopize) indiquées pour chaque métrique sont fournis à titre indicatif et représentent un **niveau d'anonymisation fiable** pour la plupart des cas d'usage. En pratique, les objectifs de confidentialité doivent être définis **en fonction du cas d'usage** spécifique. Par exemple, les objectifs de confidentialité doivent être plus élevés lorsque l'objectif de l'anonymisation est de rendre les données accessibles en open data, par rapport à un partage de données entre services d'une même entité. En effet, le risque et l'impact d'une fuite potentielle de données sont plus importants dans le premier cas.

Métriques de privacy	Valeur mesurée	Objectif moyen
Hidden rate	93.9%	>90%
Local cloaking	10	>=5
Distance to closest	10.0	>0.2
Closest distance ratio	0.83	>0.3
Row direct match protection	99.9%	>90%

Tableau 1 : Métriques de privacy mesurées sur 1 run d'anonymisation avec la solution avatar du jeu de donnée avec k=20.

Les **métriques** présentées ici répondent donc favorablement au critère **d'individualisation** du RGPD.

Pour aller plus loin dans la compréhension de l'apport de la méthode avatar, il est possible de **comparer et d'interpréter les données avant et après anonymisation**. Afin de faciliter cette comparaison, nous pouvons utiliser les mêmes données de capteurs mais segmentées avec une **méthode de détection de pic**. Cela a pour effet d'aligner les courbes avec le pic le plus proéminent à t=0.5. Il est ainsi plus simple d'**identifier les tendances** générales des signaux mais aussi d'**identifier les signaux rares** et potentiellement ré-identifiables.

Les Figures 7 à 10 représentent **quatre anonymisations distinctes** (pour 4 mois de données différents). Elles illustrent en particulier le fait que les **signaux rares sont corrigés** afin de ne plus laisser paraître les phénomènes rares. Dans certains cas (Figures 9) où plusieurs modes sont présents, nous observons que ces différents modes sont **conservés**. Cela n'est possible que si suffisamment de signaux suivent ces modes de fonctionnement, les rendant ainsi suffisamment fréquents et **enlevant tout risque de réidentification**.

Nantes s'engage depuis plusieurs années au service d'une gestion éthique des données que ce soit dans sa charte métropolitaine de la donnée ou plus récemment à travers sa doctrine pour poser un cadre de régulation des usages de l'IA. La confiance, la transparence et la maîtrise des données publiques sont des valeurs que la collectivité s'efforce ainsi de traduire au quotidien dans ses nombreux projets. Dans ce contexte, pouvoir concilier qualité statistique des données et impératifs de protection des données personnelles des usagers est une question centrale. Les travaux menés par Octopize sur l'anonymisation des données sensibles avec avatar permettent de faire avancer concrètement cet enjeu en contribuant à ouvrir des perspectives nouvelles pour développer des projets d'intérêt général au service par exemple de la santé globale ou encore de la réduction des consommations énergétiques, à la fois respectueux des personnes et participant à la confiance vis-à-vis de ces dispositifs.

Francky Trichet
Vice-Président, **Nantes Université**
(Numérique Responsable et Nouveaux Usages)
@franckytrichet

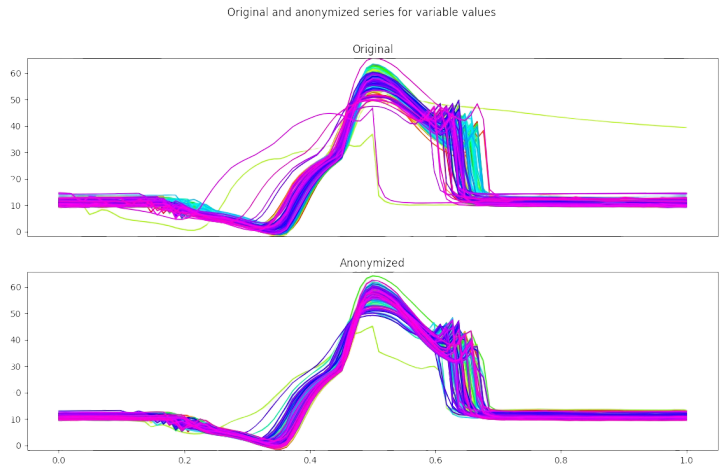

Figure 7 : Anonymisation de signaux du mois de mai alignés avec détection de pic: les signaux rares ne sont plus discernables dans les avatars.

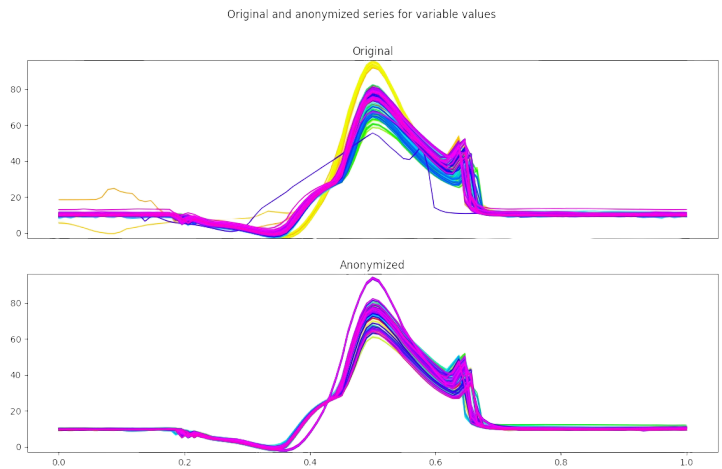

Figure 8 : Anonymisation de signaux du mois de juillet alignés avec détection de pic: les signaux rares ne sont plus discernables dans les avatars.

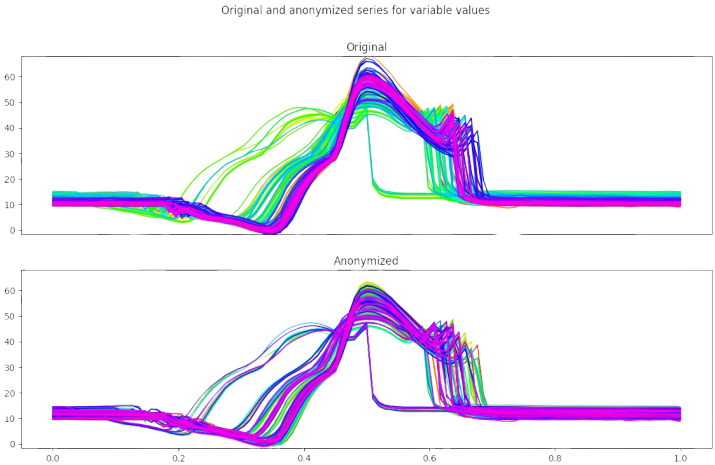

Figure 9 : Anonymisation de signaux du mois d'août alignés avec détection de pic: deux modes de fonctionnement sont conservés dans les avatars.

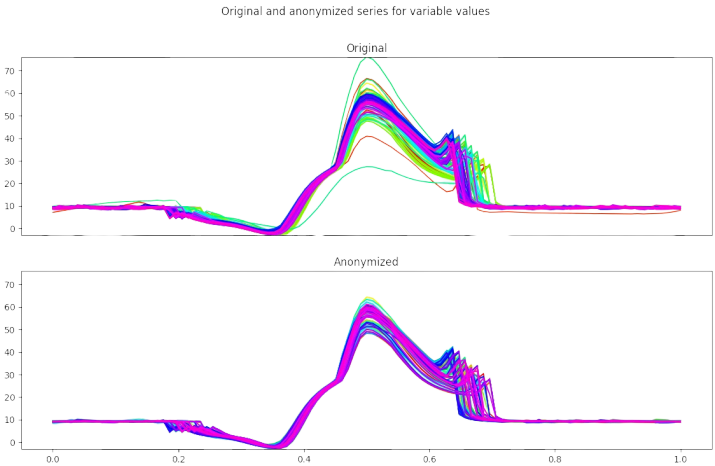

Figure 10 : Anonymisation de signaux du mois de décembre alignés avec détection de pic: un seul mode de fonctionnement est conservé dans les avatars.

L'anonymisation est un sujet clé dans la manipulation à grande échelle de données initialement personnelles. En effet, en tant que processus rendant quasi impossible toute ré-identification des personnes concernées, l'anonymisation permet de s'affranchir de nombreuses contraintes posées par le RGPD tout en assurant un certain secret du set de données initial. En matière de Machine Learning notamment cela permet de lever de nombreux freins dans la constitution des sets de données d'apprentissage des modèles d'IA. En effet, les contraintes peuvent être nombreuses pour pouvoir utiliser de façon licite des sets de données personnelles dites sensibles ou non. Or, le non-respect de ces contraintes peut avoir des conséquences dommageables multiples pouvant aller jusqu'à rendre illicite le modèle d'IA illégalement entrainé.

Les travaux menés par Octopize et Sopra Steria décrits dans ce livre blanc ont pour objet de démontrer comment l'anonymisation au travers de la méthode avatar développée par Octopize «peut non seulement sécuriser les données mais aussi maintenir leur utilité pour l'analyse et la modélisation ». Ces paramètres sont évidemment essentiels au développement d'une IA de confiance conforme aux exigences réglementaires.

Valérie Aumage

Head of IP/IT/Data Privacy
Avocat à la Cour
PwC

E.4 Évaluation générique du maintien de la valeur statistique des données

En complément de l'évaluation de la **confidentialité** du jeu de données anonymisé, des métriques génériques d'utilité peuvent être calculées. Étant donné leur nature générique, ces métriques ne nécessitent pas de connaître ni de simuler l'utilisation future des données, ce qui pourrait être coûteux en termes de temps de calcul (par exemple, pour une utilisation en machine learning). Ainsi, elles permettent d'effectuer des **itérations rapides** lors du paramétrage de la solution avatar.

Le tableau ci-dessous résume les métriques d'utilité calculées sur les données transmises. Les métriques sont divisées en 2 familles : les **métriques globales**, calculées sur l'ensemble des points, toutes séries confondues; et les **métriques calculées au niveau individuel**, c'est-à-dire sur chaque série.

Métriques d'utilité	Valeur mesurée	Objectif
Métriques globales		
Pointwise Hellinger distance	0.01	<0.2

Métriques individuelles (% de différence)		
Series mean	0.00%	<5%
Series minimum	0.01%	<10%
Series maximum	0.00%	<10%
Series sum of values	0.00%	<5%
Series entropy (20 bins)	0.00%	<5%
Autocorrelation (10)	0.00%	<5%

Tableau 2 : Métriques d'utilité mesurées sur 1 run d'anonymisation avec la solution avatar du jeu de donnée avec k=20.

La **distance d'Hellinger** est une distance entre 2 distributions. Plus sa valeur est grande, plus les distributions comparées sont différentes. Pour une distance d'Hellinger de type **pointwise**, la distribution des valeurs relevées est comparée avec son équivalente construite avec les valeurs anonymisées. Bien que donnée à titre indicatif seulement, une distance d'Hellinger inférieure à 0.2 est considérée comme indicateur d'une **bonne conservation du signal.**

Les métriques individuelles sont calculées à partir **d'indicateurs** (features) générés sur chaque série. Les valeurs sont **moyennées** et la **différence relative** entre les valeurs originales et avatar est exprimée en pourcentage. Les indicateurs extrait des données sont : la **moyenne** d'une série (series mean), son **minimum** (series minimum) et son **maximum** (series maximum), la **somme** de ses valeurs (series sum of values), son **entropie** (series entropy - qui peut être interprétée comme représentant la complexité de la séries) et son **autocorrelation** (corrélation entre points d'une même série).

Les **objectifs** donnés à titre indicatif pour les différences observées sur ces indicateurs sont **atteints**. Enfin, une **interprétation visuelle** des données générées vient confirmer la **conservation des tendances** présentes dans les données d'origine comme le montrent les Figures 11 et 12.

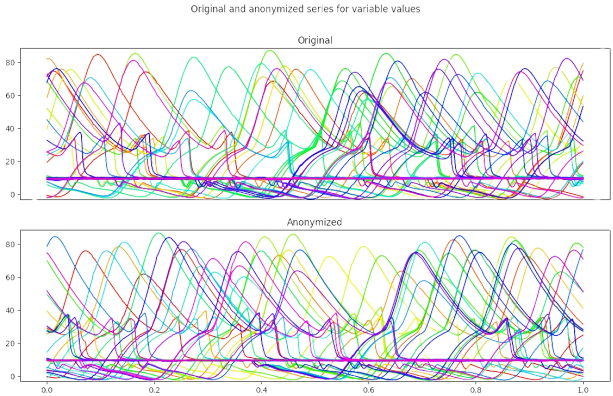

Figure 11 : Ensemble des courbes originales (en haut) et avatars (en bas) pour les données du cas d'usage. 50 courbes sont représentées pour des raisons de lisibilité.

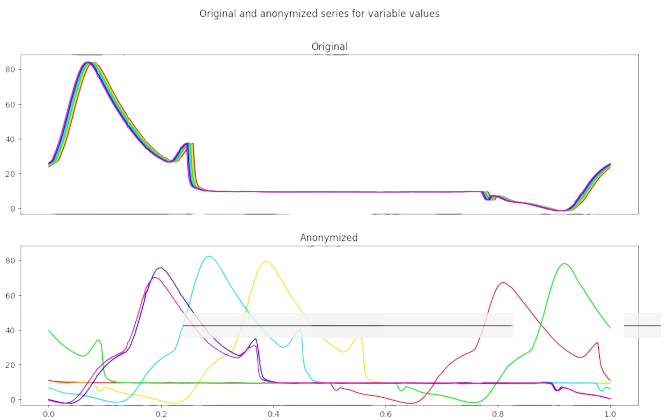

Figure 12 : Sélection des 6 premières courbes originales (en haut) et avatars (en bas) pour les données du cas d'usage.

Les métriques génériques permettent d'**identifier de potentielles pertes d'utilité** dans la phase de paramétrage de la solution avatar et donne une première estimation de la conservation du signal. Il est cependant recommandé de poursuivre l'**évaluation de l'utilité** des données anonymisées en s'assurant qu'elles sont utilisables pour le cas d'usage ciblé. La section qui suit se focalise donc sur l'**apprentissage de modèles de détection d'anomalies** à partir de ces données.

Le développement de nouveaux médicaments et traitements implique très souvent la réutilisation de données initialement collectées dans le cadre des soins de santé ou de précédents projets de recherche. L'accès à ces données est soumis à un cadre règlementaire stricte dont la conformité justifie le recours à des technologies renforçant la protection des données personnelles.

Les solutions de génération de données synthétiques dont la méthode « Avatar » développée par Octopize vise à protéger l'intégrité des données tout en garantissant la confidentialité de celle-ci.

Cette méthode représente un progrès significatif permettant notamment de sortir du champ d'application du Règlement Général de Protection des Données (RGPD), de lever les frontières du transfert des données en dehors de l'Union Européenne tout en réduisant les coûts opérationnels et les délais associés à l'obtention de consentements répétés.

Gregory Collet

Early Stage Success Manager, DPD certifié
My Data-Trust

En tant que responsable de l'Initiative Climate Links, un consortium mondial en développement dédié à la mise en relation des municipalités locales avec les technologies pertinentes aux ODD, je reconnais l'immense valeur des techniques d'anonymisation présentées dans ce livre blanc. Notre projet repose sur la capacité d'analyser de vastes ensembles de données sur les besoins municipaux, les offres de technologies durables et les contextes politiques. La méthode « avatar », avec son approche rigoureuse de génération de données synthétiques tout en préservant la pertinence statistique, offre une solution puissante. L'anonymisation permettra à Climate Links de construire et d'analyser son graphe de connaissances sans compromettre la confidentialité des parties prenantes ni révéler des renseignements commerciaux sensibles. En adoptant ces techniques, nous pouvons instaurer la confiance entre nos partenaires, accélérer l'adéquation des solutions durables aux besoins locaux et, finalement, contribuer à la mise en œuvre effective des Objectifs de Développement Durable des Nations Unies. Le consortium Climate Links est impatient d'intégrer cette technologie pour connecter les acteurs locaux et les fournisseurs mondiaux de manière sécurisée, éthique et efficace.

Newton H. Campbell Jr.

Maître de Conférences Adjoint (Senior Adjunct Lecturer),
UNSW (University of New South Wales), Sydney

VALIDATION DE MODÈLES D'APPRENTISSAGE SUR DONNÉES ANONYMISÉES

F.1 Modèle de détection d'anomalie choisi

La méthode retenue pour évaluer la pertinence des données anonymisées est une approche de **détection d'anomalies non supervisée** provenant du programme Confiance.ai. Elle opère sur des **séries temporelles univariées** régulièrement échantillonnées. Elle est fondée sur une méthode en deux étapes utilisant des **architectures 1D-CNN** profondes :

◆ **L'étape d'apprentissage de la représentation** : utilisation de tâches prétextes pour apprendre une représentation d'échantillons de données dites "normales", c'est-à-dire sans anomalies, de manière auto supervisée. Dans cette étape, le modèle apprend à reconstruire les données.

◆ **Détection des anomalies** : une anomalie est détectée à chaque fois que le score d'anomalie de l'échantillon de données testé est supérieur à un seuil, c'est-à-dire lorsque la reconstruction du signal est d'une qualité insuffisante. Les scores d'anomalie sont calculés pour chaque point d'un échantillon de données. Afin de disposer du score le plus robuste possible, pour un point donné, le score d'anomalie est calculé comme la moyenne de tous les scores de reconstruction associés aux échantillons de données fenêtrés contenant ce point. Cette manière de calculer le score permet de surcroît de proposer une localisation temporelle de l'anomalie dans la fenêtre considérée.

Ce mode opératoire a été choisi car il est considéré comme l'une des **approches les plus matures** du programme Confiance.ai en termes de détection d'anomalie non supervisée à la date de l'expérimentation. Les travaux [4], [5] et [6] mentionnent des approches similaires.

Il faut noter quelques points importants dans l'application de la méthode :

- ◆ L'entraînement doit être réalisé sur des **données supposées ne pas contenir d'anomalies,**
- ◆ Le procédé est particulièrement efficace dans le cas de signaux périodiques, lorsque la taille de la fenêtre choisie est égale à la **période du signal**,
- ◆ Comme pour beaucoup de méthodes d'apprentissage en traitement du signal, il est recommandé de procéder à un **découpage en séquence avec recouvrement**, afin d'augmenter le dataset d'apprentissage en taille mais aussi en diversité,
- ◆ Les données d'entrée sont **standardisées**.

L'objectif est d'obtenir 2 modèles : un modèle entraîné sur un **jeu de données original** (que nous nommerons modèle Original) et un modèle entraîné sur des **données avatars** construites à partir du même jeu de données original (que nous nommerons modèle avatar). Ce dernier point est important puisqu'il faut alors **anticiper l'étape de segmentation du signal** avec recouvrement qui doit nécessairement être réalisée en amont de l'anonymisation.

Les équipes de CIVITEO accompagnent des administrations locales nombreuses dans leur gestion des données. Toutes (communes, intercommunalités, départements ou régions) utilisent des données de plus en plus massives pour remplir au quotidien leurs missions de service public. Mais toutes sont attentives à la protection de la vie privée comme à l'empreinte carbone de leurs outils numériques. L'utilisation des avatars, données synthétiques anonymes, représente à l'évidence une formidable perspective, notamment pour entraîner les systèmes qui auront recours au machine learning dans des domaines aussi variés que la gestion de l'énergie ou de l'eau, la gestion des déchets, les déplacements mais aussi l'action sociale, l'éducation ou encore la prévention en santé.

Jacques Priol
Président, expert data & IA,
Auteur et conférencier,
Cabinet CIVITEO

F.2 Résultat des tests comparatifs d'entraînement du modèle

L'évaluation des données avatars se fait en entraînant le **modèle 1D-CNN** une première fois sur des données d'origine et une seconde sur des données anonymisées. La Figure 13 résume la procédure.

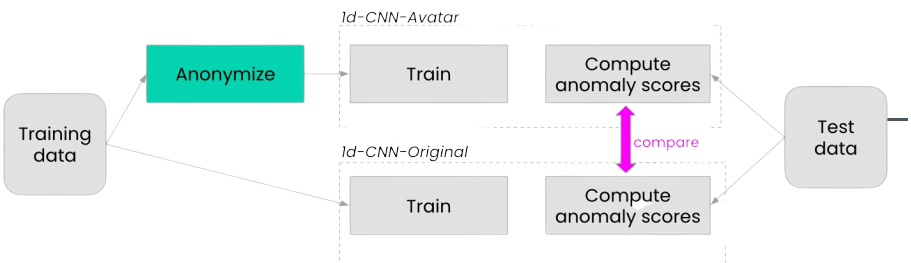

Figure 13: Entraînement et comparaison des scores d'anomalie produits par 1D-CNN entraîné respectivement sur données d'origine et données anonymisées.

Un jeu de données **d'entraînement** (training data) est choisi parmi l'ensemble des données disponibles. Ce jeu de données est anonymisé et utilisé pour l'entraînement du modèle, tout comme sa version non anonymisée. Le **paramétrage** du 1D-CNN est le même pour les deux entraînements.

Des données distinctes des données d'entraînement sont choisies en tant que **données de test** (test data) pour évaluer la qualité des modèles. Notons que ces données de tests sont rendues **non-identifiantes** (ou pseudonymisées). Ainsi, le résultat de l'évaluation permettra de conclure sur la capacité d'**utiliser des données anonymes** pour entraîner des modèles appliqués par la suite à des données réelles.

Cela est rendu possible par le fait que la **méthode avatar conserve les propriétés statistiques** du jeu de données, chose primordiale pour un réseau neuronal. Chacun des modèles est utilisé pour calculer des **scores d'anomalies** pour les données de test. L'entraînement du 1D-CNN étant non-déterministe, 10 entraînements ont été réalisés afin de produire des **intervalles de confiance**.

La mesure utilisée pour évaluer l'efficacité de la reconstruction par le modèle 1D-CNN est la **Mean-Square Error** (MAE). Les entraînements ont été réalisés sur **un mois** de données et lancés 10 fois. Les scores présentés dans la Table 3 sont moyennés sur les 10 lancements.

	Modèle entraîné sur données originales	Modèle entraîné sur données avatars
MAE entre les données d'origine et les données d'origine reconstruites (scores finaux moyennés sur l'ensemble des fenêtres)	0.04	0.03
MAE entre les données avatars et les données avatars reconstruites (scores des séquences avatarisées)		0.02

Tableau 3: Erreur de reconstruction (MAE) du 1D-CNN entrainé sur données d'origine et avatars.

On peut alors noter que :

- Les deux modèles ont **convergé**,
- La reconstruction des avatars par le modèle avatar est **un peu meilleure** que la reconstruction des données d'origine par le modèle original,
- La reconstruction des données d'origine par le modèle avatar est **un peu meilleure** que la reconstruction des données d'origine par le modèle original.

Ces deux dernières observations peuvent être interprétées à travers le **processus d'anonymisation** lui-même. Comme expliqué précédemment, l'anonymisation vise à **modéliser les modes de la distribution** d'origine et à générer des individus correspondant à ces modes. Par conséquent, le processus tend à **atténuer les comportements** marginaux, réalisant ainsi, d'une certaine manière, une forme de **modélisation de la normalité**. Cela pourrait faciliter le travail d'apprentissage.

La Figure 14 présente les **scores d'anomalie** pour les deux modèles sur l'ensemble des données couvrant la période étudiée. Les résultats montrent les **intervalles de confiance** à 95 % pour ces scores, tandis que la Figure 15 montre ces mêmes résultats, mais uniquement pour le **mois de mai**, ce mois ayant enregistré **les points les plus anormaux.**

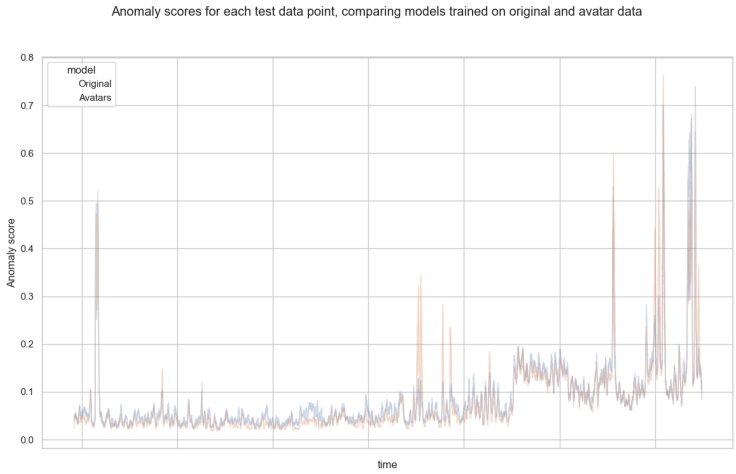

Figure 14: Comparaison des scores d'anomalies obtenus sur 10 runs pour l'ensemble des données.

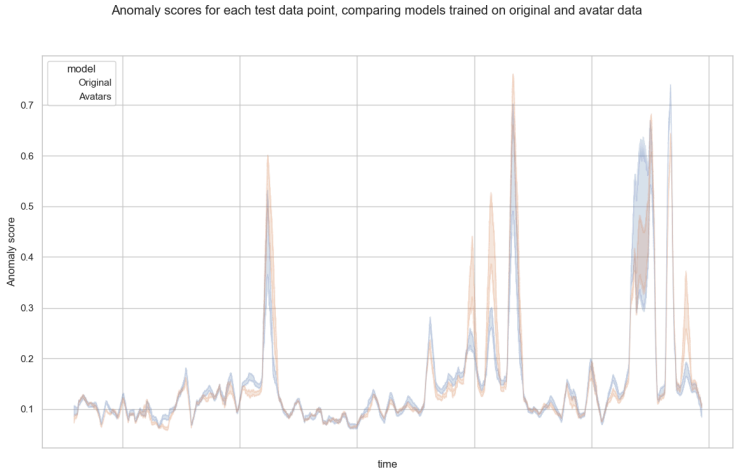

Figure 15: Comparaison des scores d'anomalies pour le mois de mai obtenus sur 10 runs.

Enfin, les Figures 16 et 17 montrent les **scores** obtenus lors d'une itération, ainsi que le **signal** autour de deux anomalies confirmées, visibles respectivement en **janvier et en mai**. Ces deux anomalies, validées par des experts métier, sont identifiées de manière relativement **similaire**.

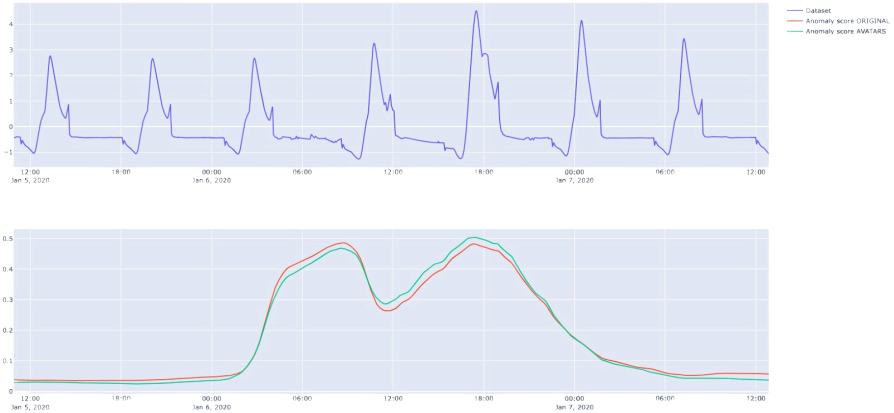

Figure 16 : Scores d'anomalie autour d'une anomalie connue du mois de janvier: l'anomalie résulte en des scores originaux et avatars élevés.

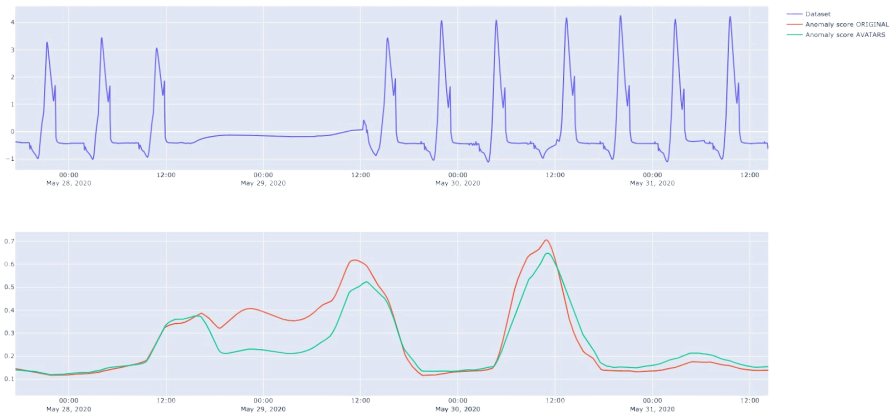

Figure 17 : Scores d'anomalie autour d'une anomalie connue du mois de mai : l'anomalie résulte en des scores originaux et avatars élevés

Nous observons :

◆ Indépendamment du modèle, des **modes de fonctionnement différents** peuvent apparaître au fil du temps. Ainsi, les scores obtenus en avril et mai sont nettement plus élevés que ceux de janvier ou février. Cela doit être pris en compte lors de l'interprétation des résultats.

◆ **La corrélation entre les scores originaux et les scores avatars** est confirmée par le calcul des coefficients de Pearson (voir tableau et figure ci-dessous). Plus la corrélation est proche de 1, plus les scores évoluent de manière similaire. Dans ce cas, toutes les corrélations sont supérieures à 0,7, seuil au-delà duquel une corrélation est généralement considérée comme forte. À l'exception du mois de mars, le coefficient de corrélation dépasse 0,9, indiquant une très forte corrélation.

Periode de test	Pearson correlations
Janvier	0.99
Mars	0.74
Avril	0.95
Mai	0.94
Période complète	0.96

Tableau 4 : Corrélations entre scores d'anomalie originaux et avatars sur différentes périodes : les coefficients sont au-delà de 0.7, ce qui représente une forte corrélation.

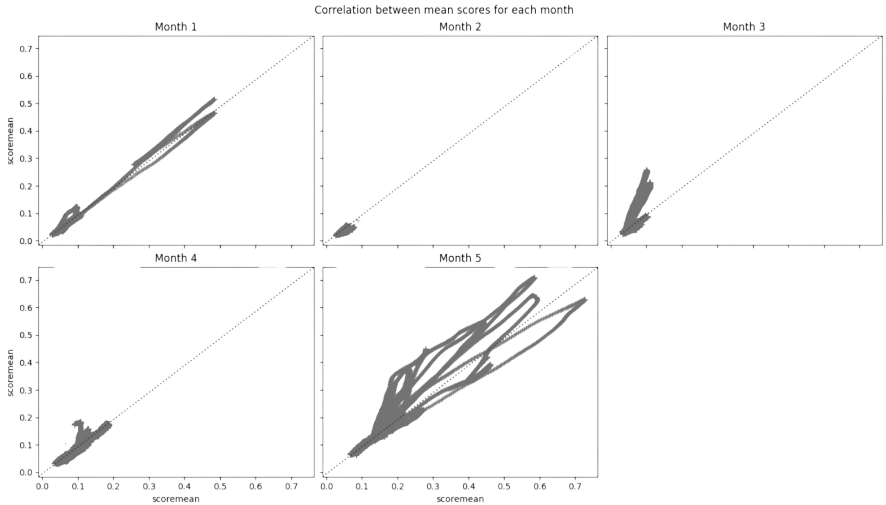

Figure 18 : Corrélations entre scores originaux et scores avatars: les points sont concentrés autour de l'axe y=x, illustrant une forte corrélation.

Pour approfondir la comparaison des scores, concentrons-nous sur les **pas de temps** où les scores produits par le modèle avatar dépassent l'**intervalle de confiance** associé au modèle original. Pour ces pas de temps, nous mesurons la **différence maximale**, que nous classons comme **positive** si le score avatar est supérieur au score original, ou **négative** dans le cas inverse. Afin de tenir compte de l'amplitude du score dans le calcul des différences, celles-ci sont normalisées en les divisant par le score. Cela permet de **comparer** les différences entre les pas de temps de manière cohérente.

Ces valeurs peuvent être réparties en trois catégories :

◆ **Différences proches de 0** (par exemple différence dans l'intervalle [-0.5,0.5]). Ces différences peuvent s'expliquer par plusieurs facteurs, tels que l'incertitude du modèle. En pratique, les différences observées sur ces pas de temps ont un impact très limité sur la détection des anomalies.

◆ **Différences largement positives** (par exemple différence>0.5). Ces différences ont tendance à produire des Faux Positifs, c'est-à-dire de créer des alertes non valides ou à attribuer des scores très larges à des alertes existantes. Dans le cadre des données de cet industriel, le but de la détection d'anomalie est de faire remonter auprès d'experts des pas de temps suspects.

Les faux positifs ont donc pour effet d'augmenter la mobilisation d'expertise humaine mais ne représente pas un danger comme pourrait l'être des faux négatifs. Pour ce qui est des scores d'anomalies largement plus élevés sur des anomalies réelles, cela ne représente pas de risque particulier. La gestion de ces faux positifs peut se faire par l'adaptation du seuil utilisé pour remonter une alerte.

◆ **Différences largement négatives** (par exemple différence <0.5). Ces différences ont tendance à générer des faux négatifs, c'est-à-dire de ne pas créer d'alertes lors d'événements suspects. Dans un contexte de détection d'anomalie, les faux négatifs sont plus impactant que les faux positifs.

La Figure 19 illustre les **différences observées** sur l'ensemble de la période. Des différences largement positives sont présentes. En comparant avec les scores d'anomalies (Figure 20), on constate que ces différences n'apparaissent que sur des **scores déjà élevés**, correspondant donc à des anomalies potentielles. Par ailleurs, aucune différence largement négative n'est observée, ce qui indique que l'entraînement du 1D-CNN sur des données avatars **n'accroît pas le risque de faux négatifs.**

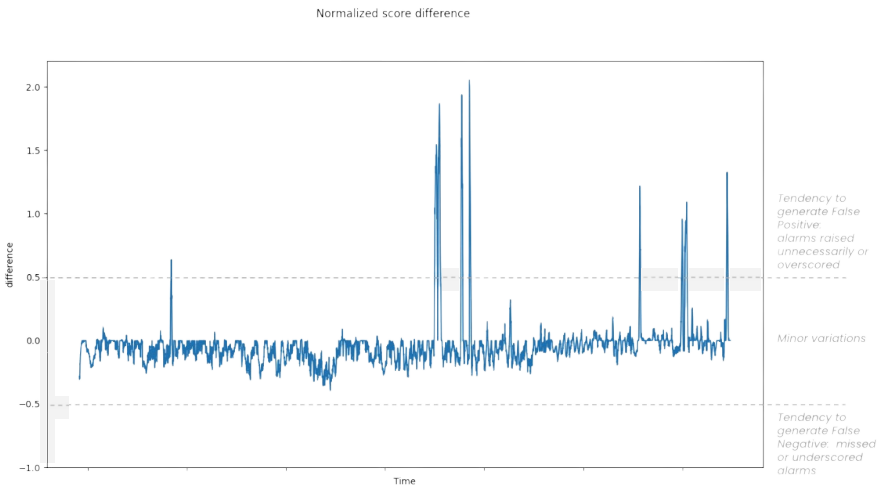

Figure 19: Classification des différences normalisées obtenues sur le cas d'usage.

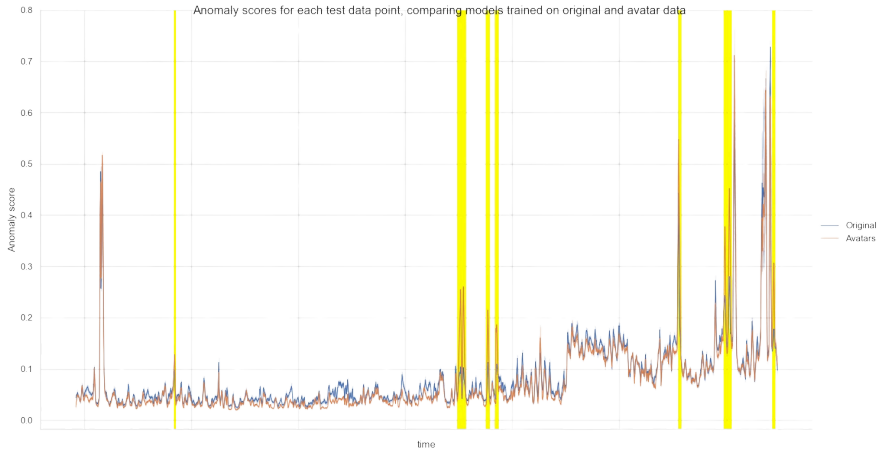

Figure 20: Score d'anomalies obtenus par les deux modèles et différences normalisées supérieures à 0.5: les différences les plus larges sont observées sur des pas de temps avec des scores originaux déjà élevés.

En pratique, un modèle de détection d'anomalie est souvent associé à **un seuil au-delà duquel une alarme peut être déclenchée**. Pour ce cas d'usage, cette alarme a pour but de faire remonter à des experts un **comportement suspect** pour une analyse approfondie. Dans les travaux précédents sur 1D-CNN, le **seuil recommandé** est déterminé en prenant le 99ème percentile des scores d'anomalies obtenus sur les données d'entraînement. Ainsi, les seuils de **0.085** et **0.061** sont respectivement obtenus pour les modèles originaux et avatars.

En appliquant ces seuils aux scores obtenus sur les mois de test, il est possible de calculer le **nombre** et le **pourcentage d'alertes** qui seraient remontées par ces modèles. Nous présentons ces résultats ci-dessous.

Mois	Alertes, modèle original	Alertes, modèle avatars
Janvier	3011 (7%)	5548 (12%)
Février (train)	418 (1%)	418 (1%)
Mars	7124 (16%)	9328 (21%)
Avril	30069 (70%)	32746 (76%)
Mai	42533 (95%)	44640 (100%)

Tableau 5 : Nombre d'alertes créées par chacun des modèles.

Parce que le fonctionnement nominal du système étudié **évolue dans le temps**, il est important de considérer ces résultats mois par mois. La différence en nombre d'alertes entre les deux modèles est relativement **constante**. Le modèle entraîné sur les avatars génère 5% d'alertes en plus que son équivalent entraîné sur les données originales.

Ce phénomène peut principalement s'expliquer par le fait que l'anonymisation a tendance à **recentrer les individus** et en particulier les plus extrêmes. Bien que cela soit bénéfique d'un point de vue **privacy** et **réduction de la sensibilité** de la donnée, des différences dans les queues de distributions sont probables.

Un **calcul de seuil** se basant sur les mêmes hypothèses (en prenant par exemple le 99ème décile) peut produire des différences telles que celles observées sur l'analyse des données de l'industriel. La forte croissance du nombre d'alertes en fonction du mois quelque soit le modèle s'explique par le **profil des données** avec un régime nominal en fin de période qui diffère de celui de février utilisé pour l'entraînement comme illustré précédemment.

Les données dans le champ particulier de la santé au travail sont devenues un outil à la fois de recherche sur les risques professionnels ou de désinsertion. Plus récemment, des indicateurs se sont développés de suivi des actions menées par les organismes chargés du suivi des salariés et des agents que sont les services de prévention et de santé au travail interentreprises ou autonome, et des services assignés aux fonctions publiques. Ces données sont également un moyen de dialogue social entre les acteurs de ce domaine particulier. Cependant, du fait de la sensibilité extrême de ces données, seuls des rapports avec données agrégées ne sont envisageable en pratique jusqu'à lors.

C'est dans ce contexte que la solution de type Avatar présentées par Octopize sans réidentification possible ouvre une voie de réflexion et de collaboration entre l'ensemble des acteurs, y compris les salariés/agent eux-mêmes, sur ces questions de recherche, d'indicateurs et de dialogue bien entendu à un niveau macro (secteur, service de prévention ...). Il est possible d'imaginer des partages réels de données, sans risque pour le salarié ou agent à destinée des acteurs du monde du travail et de la prévention.

Pr Alexis Descatha

- Clinical Professor of Occupational Medicine, Epidemiology and Prevention**, Donald and Barbara Zucker School of Medicine, Hofstra/Northwell, USA**

- Inserm, Irset UMR1085 Equipe Ester, Université d'Angers, France

- Centre Antipoison et de Toxicovigilance Grand Ouest, Prévention, CHU d'Angers, France

F.3 Validation des résultats sur une deuxième analyse

Afin de valider les conclusions, l'analyse est **répliquée** sur une **autre période temporelle** pour le même capteur. Dans cette deuxième analyse, le mois d'août est utilisé pour entraîner le modèle 1D-CNN et les données des mois de juin à novembre utilisées comme test. Notons que sur cette période, les données du mois de juin reflètent un mode de fonctionnement **très différent** des autres mois, résultant ainsi sur des **scores d'anomalie élevés**. Ce phénomène est visible sur la Figure 6. Les résultats sont cohérents avec ceux obtenus sur la première analyse. En particulier :

- Des scores originaux et avatars fortement **corrélés** (Figures 21 et 22)
- Des différences de score n'engendrant pas de **faux négatifs** (Figure 23)

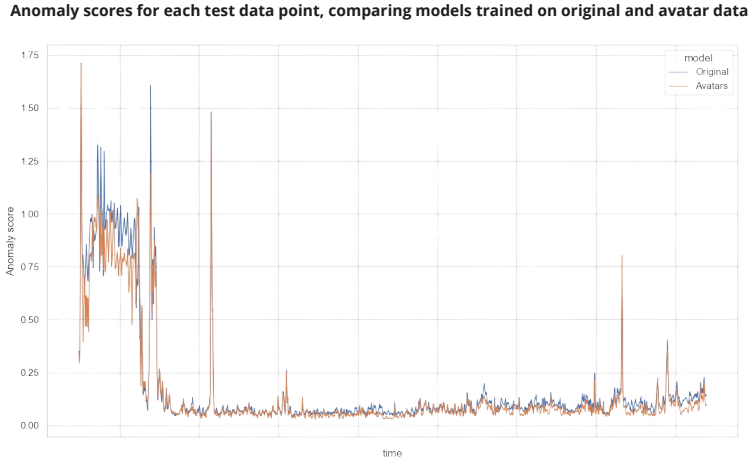

Figure 21: Scores d'anomalies pour les 2 modèles obtenus sur la deuxième analyse.

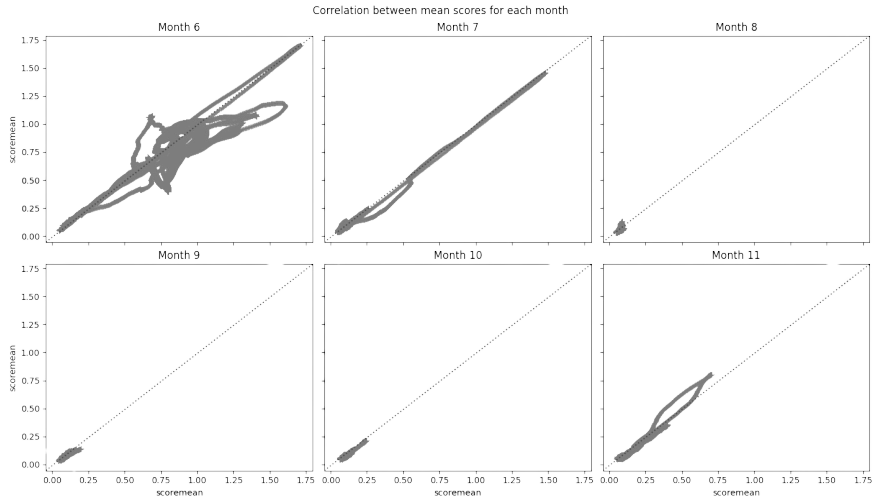

Figure 22 : Corrélation entre scores originaux et scores avatars sur la deuxième analyse.

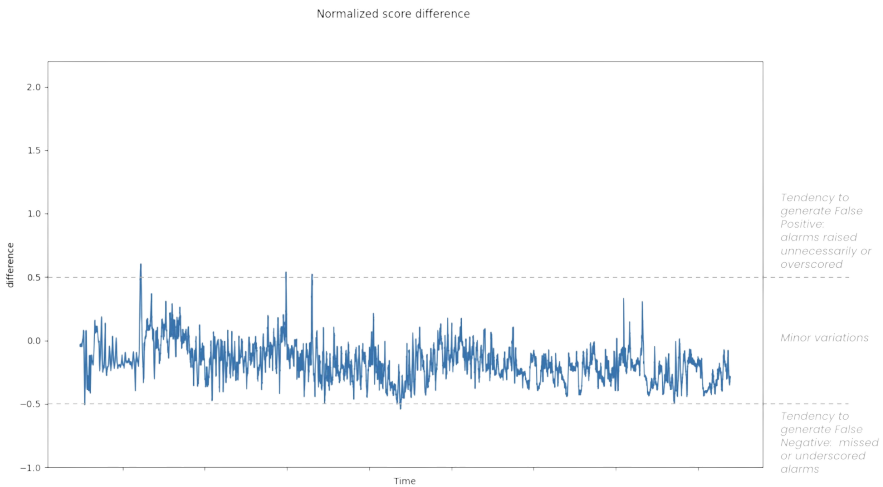

Figure 23 : Classification des différences normalisées obtenues sur la deuxième analyse.

Réduire le "time to market" des systèmes intégrant l'Intelligence Artificielle (IA) est un défi stratégique pour les organisations. Alors que l'essor de l'IA Générative a accéléré le développement des preuves de concept, le passage en production reste un frein majeur, avec de nombreux projets encore figés au stade expérimental. La stratégie de protection des données est, entre autres, un verrou important à lever dans un contexte industriel.

La communauté de l'IA de Confiance pour l'industrie s'est structurée autour du programme Confiance.ai afin de relever les nombreux défis associés à l'ingénierie de l'IA de confiance pour les systèmes critiques et réunit plus de 50 partenaires : des industriels issus de secteurs variés, des centres de recherche de premier plan et des startups innovantes comme « Octopize ».

Reconnue à l'échelle internationale, cette communauté s'ouvre et s'organise aujourd'hui en une fondation à l'échelle Européenne, « The European Trustworthy AI foundation ».

Paul Labrogère

Directeur Général / CEO

Institut de Recherche Technologique SystemX

CONCLUSION

Ce projet exploratoire, mené avec le soutien des équipes de **Confiance.ai**, de l'un de ses membres, et d'Octopize, a démontré la faisabilité de **désensibiliser des données industrielles** grâce à leur **anonymisation**. Il répond à un enjeu crucial : permettre le partage sécurisé de données sensibles entre partenaires. Notre objectif était de **partager des informations statistiques** sur des données industrielles sans divulguer leur valeur stratégique, comme les procédés de production.

Cette étude a validé l'**efficacité de la méthode avatar** pour ce cas d'usage, en permettant l'anonymisation de données issues de capteurs industriels. Elle démontre également la capacité de la méthode à traiter des **données complexes** telles que les séries temporelles, omniprésentes dans l'industrie. Pour qu'une donnée soit pleinement exploitable, elle doit trouver un **juste équilibre entre utilité et confidentialité**. Dans cette étude, nous voyons que les données avatar offrent ces deux garanties : des scores d'anomalies fortement **corrélés** et une preuve de la **confidentialité** fournie par les métriques de privacy.

De plus, nous avons illustré l'intérêt des données synthétiques anonymes pour l'entraînement de **modèles de Machine Learning** en contexte non supervisé. Nos résultats montrent que les modèles entraînés sur des données avatar affichent des **performances équivalentes** à ceux entraînés avec des données originales.

L'avatarisation pourrait même s'avérer bénéfique pour l'**apprentissage non supervisé**, en supprimant des **caractéristiques non représentatives** des jeux de données. Cela contribue à la modélisation de la normalité en amont de l'apprentissage des modèles de détection d'anomalies.

En permettant de combiner **confidentialité** et **performance**, l'avatarisation ouvre de nouvelles perspectives **d'innovation** dans l'industrie.

BIBLIOGRAPHIE

H.

1 Guillaudeux, M., Rousseau, O., Petot, J. *et al. Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymized data.* Nature Digital Medicine. 2023

2 Barreteau A-F, Regnier-Coudert, Le Carpentier E, Moussaoui S. *Génération de signaux anonymes à partir de données non anonymes par modèle de mélange linéaire local*. 2023. Colloque Francophone de Traitement du Signal et des Images GRETSI

3 Wang J-L, Chiou J-M, Muller H-G. *Functional data analysis*. Annual Review of Statistics and its application. (3)-1 2016

4 Bailly, R., Malfante, M., Allier, C., Ghenim, L, and Mars, J. (2021). *Self-supervised learning for anomaly detection on time series: application to cellular data*. Conférence sur l'Apprentissage automatique (CAp2021)

5 Bailly, R., Malfante, M., Allier, C., Ghenim, L, and Mars, J. (2021). *Deep anomaly detection using self-supervised learning: application to time series of cellular data*. In ASPAI 2021 - 3rd International Conference on Advances in Signal Processing and Artificial Intelligence, Porto, Portugal.

6 Li, D., Zhang, J., Zhang, Q., and Wei, X. (2017). *Classification of ecg signals based on 1d convolution neural network*. 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pages 1–6.

ABSTRACT

L'anonymisation des données stratégiques avec avatar

Ce livre blanc, rédigé avec le soutien du programme Confiance.ai, est un projet visant à développer des technologies de confiance pour l'intelligence artificielle, en particulier dans le domaine de la protection des données sensibles.

Dans le cadre d'un environnement industriel de plus en plus connecté, les données jouent un rôle primordial dans l'optimisation des processus. Cependant, leur partage et leur utilisation exposent à des risques accrus en matière de confidentialité. Ce livre blanc explore l'anonymisation comme solution pour protéger des données stratégiques tout en maintenant leur utilité.

La méthode avatar développée par Octopize permet la génération de données synthétiques préservant à la fois la structure statistique des données originales et leur confidentialité. Elle repose sur des techniques de projection multidimensionnelle pour créer des avatars de données et répond aux critères stricts d'anonymisation définis par la CNIL et le Comité Européen de Protection des Données, garantissant la non-réidentification des individus ou entités concernées.

Les travaux présentés ici portent sur un cas d'usage fourni par un membre du consortium, où des données de séries temporelles collectées par des capteurs ont été anonymisées pour servir à l'entraînement de modèles de détection d'anomalies basés sur des réseaux de neurones.

Les résultats montrent que l'anonymisation permet de protéger les données sans compromettre leur utilité pour des applications analytiques. Les modèles de détection d'anomalies entraînés sur des données anonymisées par avatar affichent des performances similaires à ceux entraînés sur des données originales. Ce livre blanc démontre ainsi la faisabilité de la protection des données sensibles dans un cadre industriel, tout en permettant l'innovation technologique par l'analyse de données sécurisées.

Auteurs

Olivier Regnier-Coudert

PhD, Data Scientist
Octopize

Amélie Bosca

Data Scientist
Sopra Steria & IRT System X

Mathieu Bleunven

Business Manager, Expert Défense & Mobilité
Octopize

Marie Berthon

Business Manager, Expert Défense
Octopize

Gabrielle Crolard

Responsable Communication & Marketing
Octopize

A propos

Ce livre blanc, réalisé dans le cadre du programme **Confiance.ai**, présente les travaux d'Octopize et d'un membre du consortium sur l'anonymisation des données industrielles. Il explore la méthode "avatar", une technique innovante de génération de données synthétiques garantissant à la fois confidentialité et utilité pour des applications comme la détection d'anomalies.

Le document met en lumière les enjeux de protection des données sensibles dans l'industrie tout en respectant les exigences des réglementations européennes.

Pour en savoir plus, contactez-nous à **contact@octopize.io** ou visitez www.octopize.io

OCTOPIZE
MIMETHIK DATA