


M1 Science Cognitives
Composante Sciences numériques, en Sciences Cognitives, traitement automatique des langues et innovation

Projet Tutoré Partie II : Réalisation

Étude de la qualité des données synthétiques pour les séries temporelles

Oscar Bidou, Anais Mignot–Charvillat, Youcef Namoun, Enzo Verbanaz, Rania Ait Chabane, Ayoub Maanni, Othmane Taoussi

Encadrant de projet : Azim Roussanaly

Laboratoire Ouvert en Learning Analytics
Laboratoire Lorrain de Recherche en Informatique et ses Applications
Institut des sciences du Digital, Management et Cognition
Université de Lorraine

November 3, 2025

Contents

1 Introduction

La donnée est la denrée la plus précieuse dans le monde digital et économique actuel. L'utilisation de masses de données devient indispensable pour réaliser des avancements dans le domaine scientifique de la recherche ou dans le domaine de l'industrie. Premièrement, bien que généralement abondante, il y a des cas de figure ou l'on peut manquer de données. Les processus de récupération de données peuvent aussi s'avérer être compliqués, surtout lorsqu'il s'agit de données confidentielles. De plus, le coût lié à l'accès et à l'utilisation de ces données peut parfois être élevé. C'est dans ce contexte que les algorithmes de génération de données interviennent ; ils fournissent une solution intéressante pour pallier aux problèmes de disponibilité et de confidentialité des données réelles. Il est toutefois crucial de considérer la qualité de ces données synthétiques générées : ont-elles les mêmes propriétés statistiques que les données originales ? Permettent-elles de faire des analyses correctes et précises ? Et finalement, est-ce que les défis de confidentialité sont respectés ? L'article ci-présent permet de répondre à ces questions et d'étudier ce domaine émergent de l'intelligence artificielle.

1.1 Cadre du projet

1.1.1 Structure

Le Laboratoire Lorrain de Recherche en Informatique et ses Applications, appelé LORIA, a pour mission la recherche fondamentale et appliquée en sciences informatiques. Dans le cadre de ses recherches, le LORIA a développé la plateforme LOLA, Laboratoire Ouvert en Learning Analytics. C'est un outil de valorisation des données éducatives permettant le développement de nombreuses technologies pour l'éducation. L'application permet de centraliser des données afin d'en faciliter l'accès aux organisations partenaires du LORIA et leur garantir un dépot sécurité. Par exemple, il est possible pour le Centre national d'enseignement à distance (CNED) et toute la communauté éducative utilisant LOLA de tester et d'évaluer leurs algorithmes sur une base de données commune, permettant ainsi la comparaison de performance de chacun. La plateforme fournit aussi des modèles en open source afin que d'autres organisations puissent analyser leurs propres données. De plus, on retrouve des applications de visualisations, des indicateurs d'évaluations ainsi que diverses ressources d'accompagnement des utilisateurs.

1.1.2 Les données dans l'IA

L'abondance croissante de données est aujourd'hui incontestable. Cependant, la majeure partie des données récoltées ne sont pas exploitables. En effet, en plus des préoccupations sur la qualité des données, la protection de la vie privée des individus est primordiale et impose des contraintes significatives sur la gestion de ces données. Les réglementations, notamment le Règlement Général sur la Protection des Données (RGPD), érigent des barrières strictes quant à la manière dont les informations personnelles doivent être traitées (Commission Nationale de l'Informatique et des Libertés (CNIL) 2022). Elle instaure notamment des différences claires à propos du statut des données en fonctions qu'elles soient pseudonymisées ou anonymisées. Des données strictement anonymes ne sont en effet plus considérés comme des données personnelles et sont beaucoup plus facilement exploitables au contraires des données anonymisées. Dans ce cadre, l'anonymisation des données devient une étape cruciale. Le changement d'identifiants et de nombreuses méthodes de limitation de divulgation existent, tant au niveau individuel en ajoutant du bruit par exemple (Hirsch 2005) qu'au niveau du jeu de données global avec de l'échange d'attributs ou de la microaggregation (Defays and Anwar 1998). Mais l'anonymisation ne se limite plus simplement à des modifications superficielles telles que celles énoncées. En effet, il est aujourd'hui possible en détectant des schémas entre

plusieurs jeux de données de désanonymiser des données (Narayanan and Shmatikov 2008). Par ailleurs, des algorithmes modernes d'apprentissage automatique et d'analyse de données contournent aussi ces techniques primitives et retrouvent facilement des informations sensibles (Kurtukova, Romanov, and Fedotova 2019). Ces méthodes traditionnelles d'anonymisation s'avèrent alors insuffisantes face à des techniques d'analyse de plus en plus sophistiquées et ne satisfont donc pas les exigences de la CNIL pour anonymiser des données. De plus, les données ne sont plus des entités isolées, mais font souvent partie d'un réseau complexe d'informations interconnectées. Ainsi, les méthodes d'anonymisation doivent désormais prendre en compte les relations et les interactions entre les données pour éviter toute possibilité de réidentification.

1.1.3 Synthétisation des données (définition et enjeux)

L'enjeu de la collecte de donnée réside donc désormais dans la capacité à effectuer une anonymisation qualitative, allant au-delà des approches conventionnelles comme l'inversion de certaines caractéristiques (sexes, salaires etc.) ou le bruitage. L'objectif est de garantir une protection robuste des données tout en préservant leur utilité et leur pertinence pour les analyses et les applications. Dans cette perspective, la synthétisation des données émerge comme une stratégie prometteuse. Cette approche va au-delà de la simple modification d'éléments identifiables, en générant des ensembles de données synthétiques qui on pour but de préserver les distributions statistiques et caractéristiques structurelles des données d'origine, tout en tentant de garantir l'impossibilité de relier ces données à des individus spécifiques.

Ces données peuvent ainsi servir de substitut aux ensembles réels dans des scénarios de test, de développement d'algorithmes ou d'autres applications nécessitant l'utilisation de données sensibles tout en respectant scrupuleusement les normes de protection des données. Cette approche innovante s'inscrit dans une démarche visant à concilier les impératifs de la recherche et du développement avec les exigences éthiques et réglementaires en matière de confidentialité des données.

La synthétisation peut se faire à deux niveaux :

– **Données synthétiques partielles :**

Les données partiellement synthétiques ne synthétisent qu'une partie des variables, concentrant l'effort sur un sous-ensemble spécifique. L'objectif est similaire à celui des données entièrement synthétiques, mais en se concentrant généralement sur les variables les plus sensibles, tout en laissant les autres inchangées (Reiter and Mitra 2009). Cette approche permet de maintenir un équilibre entre la valeur analytique et les risques de divulgation.

– **Données synthétiques complètes :**

Les données totalement synthétiques conservent les relations, les distributions graphiques et les propriétés statistiques similaires aux données d'origine, tout en étant dépourvues d'informations provenant de données réelles (Raghunathan 2021). Malgré leur origine artificielle, ces données permettent de parvenir aux mêmes conclusions que si elles étaient dérivées de données enregistrées.

1.2 Modélisation du problème

Les données que nous utiliserons dans le cadre de la création de données synthétiques sont des traces d'apprentissage de la norme xApi qui ont la particularité d'être des données sous forme de série temporelle. C'est-à-dire qu'elles ont la contrainte de devoir rester pertinentes aussi dans l'axe de la temporalité. En effet, l'étude du comportement des utilisateurs d'une plateforme d'Edtech

implique forcément de pouvoir suivre des évolutions et des tendances qui se dessinent au fil de l'apprentissage.

1.2.1 Time series

Les séries temporelles, ou "time series" en anglais, sont des séquences de données discrètes recueillies ou mesurées au fil du temps. L'utilisation de séries temporelles remontent de manière informelle jusqu'à l'antiquité, où elles étaient utilisées en astronomie pour étudier les mouvements célestes. Au 17e siècle on commence à les utiliser pour enregistrer et dater des prélèvements météorologiques ou des prix de marchandises et d'actions. Au 20e siècle, avec l'essor des statistiques et de l'informatique, elles ont pris une importance accrue dans l'économie, la finance, la santé, l'ingénierie ou tout autre domaine nécessitant l'étude de données dans le temps. Plus récemment, le big data et l'IA ont largement accru leur utilisation. Désormais, la machine peut apprendre d'elle même de ces données ce qui rend la manipulation d'un grand nombre d'entre elles plus efficace et donc précieuses.

Les séries temporelles diffèrent donc des dataset plus classiques en cela qu'elles sont structurées dans un ordre chronologique avec ce que l'on appelle des timestamps qui renvoient à un moment temporel précis. Dans notre cas, les données recueillies seront très nombreuses à cause du nombre d'utilisateurs à prendre en compte et du type de données que l'on recueille, à savoir toutes les actions qu'ils effectuent sur une plateforme. Elles sont aussi irrégulières dans le temps et variées dans leur qualité contrairement à des données de relevé de température ou de prix d'action par exemple. Les données sont donc dépendantes les unes des autres et ne sont souvent pertinentes que prises en ensemble pour étudier une variation au cours du temps ou une suite d'action contrairement au datasets plus classique. Le traitement de données de séries temporelles exige parfois des méthodes spécifiques pour prendre en compte cet aspect de temps. Il est possible d'analyser des time series avec plusieurs méthodes en fonction de notre but, de nos moyens ainsi que du type de données. Dans le but de garantir une bonne fiabilité des données synthétiques que nous devrons évaluer, il est important de comprendre dans quel contexte elles seront exploitées. Les données recueillies sont des time series décrivant les comportements d'utilisateurs interagissant avec un système numérique. Nos données synthétiques devront donc simuler efficacement les comportement d'utilisateurs pris indépendamment ainsi que des tendances communes à tous. Le but étant de déceler des tendances qui permettront de prédire des caractéristiques communes entre les utilisateurs grâce à leur comportement comparés. De simples statistiques descriptives ne suffiront donc pas pour capter la complexité des interactions entre tant de variables. Nous verrons plus tard des méthodes adaptées pour notre cas particulier.

1.2.2 Dataset

Ainsi, pour effectuer les différents tests de données nous utiliserons un jeu de données qui proviennent d'un learning analytics. Les données sont collectées sur un site qui enregistre les intéractions des utilisateurs avec la plateforme. Chaque action effectuée est ajoutée à la base de données chronologiquement, indicé par un timecode et caractérisée notamment par un nom, un verbe et un objet. Par exemple, un champ de la base pourrait correspondre à :

Timecode	Acteur	Verbe	Objet
"2015-02-04 17:54:00"	Gaëtan Martin	Connexion	Cours de Statistique

Table 1: Exemple d'une donnée temporelle

1.3 Objectif et motivation

L'objectif de la présente étude bibliographique est de citer brièvement les méthodes de génération de données synthétiques des plus utilisés, puis d'analyser de manière approfondie les techniques et méthodes permettant d'évaluer la qualité des données générées. Il est à noter que dans ce rapport, nous nous intéressons aux données synthétiques complètes.

2 Génération de données synthétiques pour les time-series

Les modèles de génération de données synthétiques pour les séries temporelles consiste à créer des ensembles de données temporelles fictifs qui reproduisent certaines caractéristiques statistiques et temporelles des données réelles. Ils devraient préserver les dynamiques temporelles, de sorte que les nouvelles séquences respectent les relations originales entre les variables à travers le temps. Cela peut être utile pour augmenter la taille de l'ensemble de données, tester des modèles, et bien entendu, préserver la confidentialité des données réelles. Pour des raisons de concision, nous allons expliquer brièvement les deux grandes catégories de méthodes de génération de données synthétiques pour les time-series à savoir les réseaux de neurones et les méthodes supervisées.

2.1 Les réseaux de neurones

Les GANs sont des modèles de réseaux de neurones largement utilisés pour la génération de données pour différentes raisons. Ils permettent de générer des données de haute qualité dans des domaines très variés, aussi bien généralistes que très spécifiques. Ils sont particulièrement utiles dans les domaines où les données sont sensibles ou limitées comme le secteur de la santé (Hazra and Byun 2020) ou de l'éducation (Bethencourt-Aguilar et al. 2023) dans notre cas. Ils comprennent deux parties principales : un générateur et un discriminateur. Le générateur crée de nouvelles données synthétiques, tandis que le discriminateur apprend à différencier les données synthétiques des données réelles. Les deux parties sont donc entraînées de manière adversative, le générateur cherchant lui à générer des données les plus réalistes possibles. A la fin de ce processus itératif, si tout s'est bien déroulé comme prévu, le discriminateur n'est plus capable de distinguer les vraies données des synthétiques. Ceci permet d'obtenir un générateur capable de produire des données hautement réalistes.

Pour commencer l'entraînement du générateur et du discriminateur, il est possible d'utiliser des structures de neurones pré-entraînés (Grigoryev, Voynov, and Babenko 2022) ou de partir de zéro. On entraîne ensuite le générateur sur les données que l'on veut imiter. Celui-ci génère alors un espace latent ou chaque dimension représente une caractéristiques qu'une donnée synthétique possède. Ces données sont ensuite évaluées par le discriminateur. Si le discriminateur se trompe, alors il ajuste les paramètres de son réseau de neurones pour augmenter les chances qu'il ne se trompe pas la prochaine fois que l'on lui présente la donnée en question. A noter que s'il il ajuste trop ces paramètres, il pourrait ne plus discriminer d'autres données, c'est ce que l'on appelle du surentraînement. Il doit donc trouver un perpétuel équilibre pour être le plus efficace possible en moyenne. Au contraire, si le discriminateur ne se trompe pas, alors c'est le générateur qui doit ajuster ses paramètres pour produire des données plus réalistes.

2.1.1 La variante timeGan

TimeGAN est une variante de GAN qui permet d'intégrer une composante temporelle à la génération de données (Yoon, Jarrett, and Schaar 2019). Elle se compose de quatre composants de réseau : une fonction d'incorporation, une fonction de récupération, un générateur de séquences, et un discriminateur de séquences. À la différence des GANs classiques, TimeGAN intègre des composants d'autoencodage, permettant d'encoder des caractéristiques, de générer des représentations, et d'évoluer dans le temps simultanément. La fonction d'incorporation transforme les données temporelles en un espace latent, créant une représentation abstraite des séquences temporelles. La fonction de récupération permet de retrouver les données originales à partir de l'espace latent, contribuant ainsi à une reconstruction précise des séquences.

Le générateur de séquences (Sequence Generator) et le discriminateur de séquences (Sequence Discriminator) de TimeGAN sont des composants similaires à ceux du GAN classique. Cependant, TimeGAN introduit des éléments spécifiques pour la modélisation temporelle, synchronisant les dynamiques des données réelles et synthétiques grâce à une perte supervisée. Cette approche vise à capturer les motifs temporels complexes des séries temporelles pour générer des données synthétiques préservant les caractéristiques temporelles importantes (Dash et al. 2020).

2.2 Les méthodes supervisées

Les méthodes supervisées pour générer des données synthétiques de séries temporelles utilisent des modèles classiques tels que des modèles de régression (Cano and Torra 2009) ou des méthodes basées sur des règles. Contrairement aux GANs qui adoptent une approche non supervisée, ces méthodes nécessitent des ensembles de données étiquetés pour l'entraînement. Le modèle, formé sur des caractéristiques temporelles et des relations entre les variables à partir de données réelles, est ensuite capable de générer de nouvelles séquences temporelles en respectant les schémas et les relations apprises.

Les modèles supervisés pour la prédiction de séquences assurent un contrôle précis sur les dynamiques du réseau, étant basés sur des données étiquetées où les séquences d'entrée sont associées à des étiquettes de sortie. Cette approche explicite permet aux modèles d'apprendre des motifs et des relations au sein des séquences, facilitant ainsi la prédiction de nouvelles séquences non vues. Le terme "dynamiques du réseau" fait référence à la manière dont l'état interne et les paramètres du modèle évoluent au fil du temps lors du traitement des séquences. En étant "déterministes", ces modèles produisent toujours la même sortie pour une séquence d'entrée donnée, contrairement aux modèles probabilistes ou stochastiques qui peuvent générer des résultats différents pour une même entrée en raison de l'incertitude inhérente.

3 Étude de la qualité des données synthétiques

Une fois que les données synthétiques sont générées, il est crucial d'étudier la qualité de ces données. Lorsqu'on parle de qualité, aussi appelé utilité ou valeur, plusieurs critères sont importants à prendre en compte. Tout d'abord, les enregistrements générés doivent être indiscernables des enregistrements réels, ce qui correspond au critère de fidélité. Ensuite, les données synthétiques doivent être aussi utiles que les données réelles lorsqu'elles sont utilisées à des fins prédictives similaires. Ceci est d'autant plus important dans le contexte de notre projet, qui vise à mettre à disposition d'importantes quantités de données synthétiques aussi utiles que des données réelles pour des objectifs de recherche dans le domaine de l'éducation. Les résultats doivent donc être cohérents et aussi proches que possible des résultats que l'on obtiendrait en utilisant les données

réelles. Enfin, la génération des données synthétiques doit prendre en compte l'aspect confidentiel des enregistrements. Le problème de confidentialité peut être particulièrement critique dans le cas où un seul enregistrement contient un critère particulier. Dans les données synthétiques, si nous générons le même enregistrement avec un seul changement, il est facile de discerner de quel individu il s'agit. Un exemple de ce type de données est le suivant :

Timecode	Acteur	Verbe	Objet
"2023-12-21 00:09:00"	Oliver Vert	Consultation	Cours d'Art Plastique

Table 2: Enregistrement réel

Supposons qu'il n'y a qu'un seul élève qui a consulté le cours d'art plastique, un cours facultatif. Dans ce cas, l'omission de cet enregistrement serait un manquement au niveau de la fidélité et la création d'un enregistrement similaire avec un nom d'élève différent permettrait quand même de faire le lien avec l'enregistrement initial, et d'identifier l'individu concerné, nous devons donc faire un compromis entre la conservation des caractéristiques utiles et celles permettant de réidentifier les personnes.

Dans cette partie, nous allons nous concentrer sur les méthodes d'étude de la qualité des données générées, ainsi que leurs limites. Nous séparons ces méthodes en deux catégories, à savoir les méthodes qui assurent la fidélité des données synthétiques aux données réelles et leur utilité et celles qui répondent à l'exigence de confidentialité.

3.1 Confidentialité des données synthétiques

Une des raisons principales de l'utilisation des données synthétiques est le respect de contraintes de confidentialité dans les données réelles notamment lorsque les données réelles impliquent des informations personnelles ou sensibles. Il est donc crucial de s'assurer qu'il est impossible de retrouver une information confidentielle dans un enregistrement synthétique. À cette fin, nous employons des mesures visant à évaluer à quel point ce critère est respecté.

3.1.1 Étude des corrélations inter-tables

L'étude de la corrélations entre les données réelles et les données synthétiques est une méthode qui permet de s'assurer que les liens entre ces deux data-sets sont minimales et qu'aucune inférence d'information sensible ne peut être faite à partir des données générées. Deux méthodes peuvent être employées pour cette mesure :

– **Le coefficient de corrélation de Spearman :**
Le coefficient de corrélation de Spearman mesure la corrélation linéaire entre deux variables X et Y, en se basant sur les rangs des valeurs plutôt que sur les valeurs brutes (Vega-Márquez et al. 2020). Dans notre cas, il sera utilisé pour comparer les corrélations entre les variables des données original D_{org} et les variables des données synthétique D_{syn}.
Pour ce faire, les variables dans D_{org} et D_{syn} son classé par ordre croissant, et pour chaque valeur un rang correspondant est attribué. Par la suite le coefficient de corrélation de Spearman $\rho_{R(X),R(Y)}$ entre chaque paire de variables est calculé selon la formule suivante :

$$\rho_{R(X),R(Y)} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

où $\sum d_i^2$ est la somme des carrés des différences entre les rangs des paires de valeurs liées et n est le nombre total de paires de valeurs.

Un coefficient de Spearman proche de 1 indique une concordance parfaite dans l'ordre des valeurs entre D_{org} et D_{syn}, tandis qu'un coefficient proche de -1 indique une discordance parfaite. Un coefficient proche de zéro suggère l'absence de corrélation.

– **Le V de Cramer :**
Le V de Cramer est une méthode statistique utilisé pour évaluer l'association entre deux variables catégorielles en se basant sur le test du χ^2 chi deux (Bergsma 2013).

$$V = \sqrt{\frac{\chi^2}{n \times \min(k-1, r-1)}}$$

où χ^2 est la statistique du chi deux, n est le nombre total d'observations, k est le nombre de lignes dans la table de contingence, r est le nombre de colonnes dans la table de contingence.

Score de contingence :

$$C_{score} = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Son application spécifique dans l'évaluation des données synthétiques vise à quantifier la corrélation entre les données originales et les données synthétique. En évaluant la similitude des structures catégorielles entre les deux ensembles de données, le V de Cramer contribue à assurer que les données synthétiques préservent les caractéristiques essentielles des données réelles sans compromettre la confidentialité.

3.1.2 Méthode empirique de holdout en utilisant la distance par rapport au plus proche enregistrement

Cette approche repose sur la division des données originales en deux ensembles égaux : un ensemble d'entraînement (Train) et un ensemble de validation (Holdout). À partir de l'ensemble d'entraînement (Train), une fusion est réalisée pour créer un ensemble synthétique (Synthetic).

La méthode évalue la proximité entre l'ensemble Train et l'ensemble synthétique (D(s, t)), ainsi que la proximité entre l'ensemble Holdout et l'ensemble synthétique (D(s, h))en utilisant la distance to closest record DCR. Ces calcul permettront d'évaluer la qualité de l'ensemble synthétique en faisant une comparaison entre les deux distances calculées.

La DCR ou distance au plus proche enregistrement se calcule comme suit :

$$d(s, r) = \sum_{j=1}^{P} d_j$$

où d_j représente la distance entre les valeurs de la j-ième colonne pour les individus r et s. La nature de la colonne j influence l'expression de d_j :

- Pour une colonne catégorielle,

$$d_j = \begin{cases} 0, & \text{si } s_j = r_j \\ 1, & \text{si } s_j \neq r_j \end{cases}$$

- Pour une colonne numérique,

$$d_j = |s_j - r_j|$$

Ainsi, la DCR (Distance au Plus Proche Enregistrement) est définie pour chaque individu $s \in S$ comme la distance minimale avec tous les individus $r \in X$:

$$\text{DCR}(s) = \min_{r \in X} d(s, r)$$

En cas d'interchangeabilité idéale des ensembles d'entraînement et de validation par rapport aux données synthétisées, la proportion η_T des observations plus proches de la base d'entraînement que de la base de validation est égale à 0.5, dans ce cas, il n'est pas possible de distinguer un enregistrement synthétique d'un enregistrement réel.

3.2 Fidélité des données synthétiques

La fidélité des données est aussi cruciale que leur confidentialité. Toutefois, il convient de souligner que ces deux aspects peuvent souvent être en désaccord.

En effet, renforcer la confidentialité d'un ensemble de données synthétiques peut compromettre sa représentativité statistique. Ainsi, évaluer la fidélité des données générées par des synthétiseurs devient essentiel pour assurer la qualité et l'adéquation des échantillons produits. Cette évaluation englobe des mesures telles que l'analyse des corrélations au sein des tables de données et l'examen des distributions des variables.

La fidélité d'un dataset synthétique par rapport à un dataset réel se réfère à la capacité du dataset synthétique à reproduire fidèlement les caractéristiques statistiques, les distributions, les corrélations, et autres propriétés essentielles du dataset réel à partir duquel il a été généré. En d'autres termes, un dataset synthétique est considéré comme fidèle s'il maintient la structure de dépendance entre les variables, les distributions marginales des variables individuelles, les relations entre les variables, et d'autres propriétés statistiques essentielles du dataset réel, tout en préservant l'anonymat et la confidentialité des données originales. Ainsi, la fidélité garantit que le dataset synthétique peut être utilisé de manière fiable pour effectuer des analyses et des études, tout en minimisant le risque de divulgation d'informations sensibles ou identifiables.

3.2.1 Étude des corrélations intra-tables

Ici, il s'agira de minimiser la corrélation entre deux variables de la même table. Cela signifie que, pour chaque paire de colonnes dans la table réelle, on examine comment elle se compare à la même paire de colonnes dans la table synthétique en termes de corrélation et de distribution. On s'intéressera à deux tests statistiques en fonction du type de variable :

La distance en variation totale (TVC) : pour les données catégorielles

Ce test calcule la distance de variation totale entre les colonnes réelles et synthétiques. Pour ce faire, il calcule d'abord la fréquence de chaque valeur de catégorie et l'exprime sous forme de probabilité. Cette statistique compare les différences de probabilités, comme le montre la formule ci-dessous :

$$\delta(O_j, S_j) = \frac{1}{2} \sum_{\omega \in \Omega_i} |P(O_j = \omega) - P(S_j = \omega)|$$

Avec $(O_j)_{1\leq i\leq p}$ les variables originales et $(S_j)_{1\leq i\leq p}$ les variables synthétiques, pour chaque couple (O_j, S_j), les densités de probabilité de toutes les modalités possibles $\omega \in \Omega_j$ d'une colonne j sont calculées.

Le score final s'exprime par :

$$\text{TVC}(O_j, S_j) = 1 - \delta(O_j, S_j)$$

un score plus élevé signifie une meilleure qualité

Distance entre données réelle et données synthétique Toujours dans le but d'étudier la similitude des modèles, il est pertinent d'examiner les distances les données réelles et les données synthétiques. Si cette différence est importante, les données synthétiques générées pourraient ne pas être considérées comme fidèle. En revanche, si l'écart est faible, cela indique que les deux ensembles de données produisent des prédictions similaires, attestant ainsi de la qualité des données synthétiques.

L'une des méthodes les plus utilisées est l'erreur quadratique moyenne (MSE). C'est une mesure globale de la fidélité des données. Elle est obtenue en moyennant la distance euclidienne carrée entre les individus de $\mathcal{D}_{\text{syn}}$ et $\mathcal{D}_{\text{orig}}$.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}(x_i^{\text{syn}} - x_i^{\text{orig}})^2$$

Ici, N représente le nombre de prédictions, et x_i^{syn} et x_i^{orig} représentent respectivement le i-ième individus de $\mathcal{D}_{\text{syn}}$ et $\mathcal{D}_{\text{orig}}$. Plus la MSE est proche de 0, plus les données synthétique sont similaires.

D'autres méthodes proche de la MSE existe, notamment la dynamic time warping (DTW) qui est une méthode très similaire à une MSE mais qui est plus robuste aux différent algorithmes de génération de donnée synthétique. La dynamic time warping propose une évaluation globale et non linéaire des séries temporelles (Matthieu Herrmann 2023). La DTW ne repose pas sur une distance euclidienne, mais sur une recherche de motifs de séries temporelles, n'étant pas forcément de même taille. De cette façon, la DTW ne souffre pas du trop grand nombre de dimensions, ni des irrégularités entre $\mathcal{D}_{syn}$ et $\mathcal{D}_{orig}$. Une application de la DTW permet une robustesse et une adaptabilité bien plus importantes qu'une MSE.
Enfin, la divergence de Kullback-Liebler est une méthode d'évaluation du critère spatial des données. Elle représente le degré de dissimilarité entre les probabilités de voisinages de $\mathcal{D}_{syn}$ et $\mathcal{D}_{orig}$. La méthode est détaillée plus en profondeur lors de l'explication de l'intégration de voisins stochastiques distribués en temps.

La statistique de Kolmogorov-Smirnov (KSC) : pour les données numériques

Dans notre cas d'étude, nous ne disposons pas de données numériques, mais il est intéressant de considérer des mesures adaptées si l'on venait à travailler avec ce type de données à l'avenir. Les mesures pour les données numériques jouent un rôle crucial dans diverses analyses, notamment pour évaluer la tendance centrale, la dispersion, et d'autres caractéristiques statistiques essentielles. En présence de données numériques, des outils comme la moyenne, la médiane, l'écart type, et les coefficients de corrélation deviennent pertinents pour déduire des informations significatives.

La statistique de Kolmogorov-Smirnov mesure la distance entre les fonctions de distribution des variables $(O_j)_{1\leq i\leq p}$ et $(S_j)_{1\leq i\leq p}$ pour déterminer à quel point ils sont similaires ou différents, tel que :

$$KS(O_j, S_j) = 1 - \max_{\omega \in \Omega_j} |P(O_j \leq \omega) - P(S_j \leq \omega)|$$

Ainsi, un score KS plus élevé serait celui qui présente la meilleure concordance avec les données originales en termes de distribution.

Entropie croisée

L'entropie croisée est une mesure utilisée pour évaluer la qualité des modèles de génération de données en comparant la distribution des données synthétiques à celle des données réelles (Jamin and Humeau-Heurtier 2020). Elle est particulièrement pertinente pour les tâches de classification et de génération de données, où il est essentiel de quantifier la divergence entre les distributions des données générées et des données originales. Elle mesure la différence entre deux distributions de probabilité P et Q, où P est la distribution des données réelles et Q est la distribution des données synthétiques. La formule de l'entropie croisée est la suivante :

$$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

Ici, $P(x)$ représente la probabilité réelle de l'événement x et $Q(x)$ représente la probabilité prédite par le modèle pour le même événement.

L'entropie croisée est utilisée pour quantifier à quel point les données synthétiques capturent les propriétés statistiques des données réelles. Une entropie croisée plus faible indique une meilleure correspondance entre les deux distributions, ce qui signifie que les données synthétiques imitent fidèlement les données réelles.

3.2.2 Visualisation

La visualisation pour appréhender et évaluer la fidélité de ces données offre une perspective complémentaire à l'étude de corrélation, en rendant les relations complexes entre les données synthétiques et réelles plus accessibles et compréhensibles. Elle permet de déceler visuellement les similitudes et les divergences, et d'apprécier la capacité des données synthétiques à refléter fidèlement la structure et la dynamique des données originales.

Afin de pouvoir visualiser les données, il est nécessaire de ne sélectionner qu'une partie des informations à représenter. De nombreuses méthodes de réduction dimensionnelle existent aujourd'hui, permettant de réduire les dimensions du jeu de données. Lors de l'utilisation de méthodes de réduction dimensionnelle pour les séries temporelles, il est crucial de considérer un pré-traitement ou un encodage adapté de la dimension temporelle. Ces étapes supplémentaires permettent de s'assurer que les caractéristiques temporelles importantes sont préservées et correctement représentées dans le processus de réduction. Cela aide à maintenir l'intégrité des dynamiques temporelles, garantissant que les visualisations et analyses qui en découlent sont à la fois significatives et fiables.

La réduction dimensionnelle est une méthode mathématique et informatique qui consiste à transposer les données dans un espace de dimension plus petite.
L'objectif est de trouver un sous-espace vectoriel X' de dimension p' de notre jeu de données X de dimension p. La mise en avant de certaines propriétés de X permettrait une évaluation des similitudes entre $\mathcal{D}_{syn}$ et $\mathcal{D}_{orig}$.

Il est cependant impératif de garder à l'esprit qu'il n'existe pas de méthode de réduction dimensionnelle parfaite. Par défaut, toute réduction de dimensionnalité entraîne une perte d'information

des données. Chaque algorithme de réduction de dimensionnalité a des limites importantes et performe au mieux pour des types spécifiques de données. Un algorithme peut regrouper efficacement un jeu de données tout en étant inefficace sur un autre. Une bonne compréhension des données, ainsi que des méthodes de projection, est nécessaire afin d'obtenir des résultats les plus significatifs possibles.

Une visualisation des données réduites permet d'offrir une meilleure interprétation de celles-ci ainsi que certains tests statistiques.

Les algorithmes étudiés sont :

1. ACP ainsi que TD-ACP
2. MDS ainsi que TMDS

Méthode de la fenêtre temporelle

Dans notre cas, où les données consistent en des séries temporelles, les algorithmes de réduction de dimensionnalité classiques tels que l'ACP, MDS, et t-SNE ne permettent pas de capturer la dynamique temporelle des données. L'adaptabilité aux séquences temporelles est nécessaire pour obtenir une représentation des tendances évolutives des séries temporelles.

Afin de répondre au besoin d'une analyse plus fine et spatiotemporelle, la méthode de la fenêtre temporelle a été développée (Morishita 2021, Jäckle et al. 2015). Cette méthode consiste à diviser les données en m fenêtres, découpant ainsi la séquence temporelle. La taille de la fenêtre dépendra des données. Pour obtenir une observation fine des variations des données, un plus grand nombre de fenêtres est nécessaire, tandis qu'un nombre plus restreint de fenêtres permettrait d'analyser plus globalement les tendances.

Pour chaque fenêtre temporelle, il est nécessaire de créer une matrice de données regroupant toutes les données survenant pendant cette période. Ainsi, les données seront divisées en m matrices de données temporelles. Sur chacune d'entre elles, notre méthode de réduction de la dimensionnalité sera appliquée. Soit m le nombre de fenêtres temporelles et n_i le nombre de données de la matrice correspondant à la i-ème fenêtre m_i. $\sum_{i=0}^{m} n_i = N$

Chaque matrice associée à une fenêtre aura donc son propre projecteur. De cette manière, chaque période temporelle, définie par la taille de la fenêtre, aura ses propres coefficients reflétant les tendances des données au cours de cette période.

ACP – Analyse en Composante Principales L'analyse en composantes principales est la méthode de réduction de dimension la plus simple et la plus connue. Cette méthode linéaire réduit la dimension des données de manière à expliquer le maximum de variance des données dans p' dimensions. Concrètement, une ACP tourne l'objet dans l'espace afin d'avoir une représentation en p' dimensions qui maximise la variance perçus. Il est nécessaire de trouver la matrice de projection W telle que :

$$X' = WX$$

Cette matrice W est constituée des p' vecteurs propres de la matrice de covariance des données normalisées X_{norm} qui expliquent le plus de variance. La matrice de covariance de X_{norm} est donnée par :

$$Cov(X_{norm}) = \frac{1}{N-1} X_{norm}^T X_{norm}$$

Le ratio de variance expliquée d'un vecteur propre est donné par la valeur propre associé à celui-ci, divisé par la somme de tous les valeurs propres. Soit λ_i une valeur propre, alors la variance

expliquée du i-ème vecteur propre v_i est donnée par :

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_i + ... + \lambda_p}$$

La simplicité de cette méthode se traduit non seulement par un coût computationnel significatif de l'ordre de $O(N.p^2)$, mais également par une perte d'information considérable. En effet, lors de l'application d'une Analyse en Composantes Principales (ACP), il y as une perte de la variance expliquée par les vecteurs propres non représentés (la variance expliquée des $p - p'$ vecteurs propres restant). Si la relation entre les vecteurs propres et la variance expliquée est linéaire, l'ACP présente peu, voire aucun intérêt. Dans des espaces de grandes dimensions, la projection de notre matrice X dans un espace bidimensionnel est pratiquement synonyme d'une importante perte d'information. Cette perte risque d'empêcher de capturer correctement la structure sous-jacente des données.

TD-ACP – Analyse en Composantes Principales dépendante du temps:

La création d'un nouvel algorithme, appelé td-ACP (Analyse en Composantes Principales dépendante du temps), en utilisant la méthode des fenêtres temporelles dans le cas de l'ACP s'est avérée performante dans le traitement des séries temporelles (Morishita 2021). En termes simples, Morishitaa applique une ACP sur chaque sous-ensemble de dimensions $n_i \times p$ des données X. Ainsi, la matrice de projection, associée à chaque fenêtre, équivaut aux deux vecteurs propres de la matrice de covariance du sous-ensemble équivalent normalisé qui explique le plus de variance.

Cette méthode a une complexité plus élevée qu'une ACP en raison de la multiplicité des opérations, elle est estimée à $O(m.N.p^2)$. Il est crucial de garder à l'esprit les limites de cet algorithme. En plus de la perte de variance inhérente aux ACP, les résultats sont d'autant plus difficiles à interpréter en raison de la complexité de l'opération. En effet, un mauvais choix de la taille des fenêtres temporelles peut entraîner des discontinuités dans les représentations. De plus, cette méthode reste une projection linéaire des données. Une relation plus complexe entre les composantes temporelles rendra difficilement interprétables des résultats. Il est donc nécessaire d'avoir une bonne connaissance du domaine des données afin de sélectionner au mieux la taille des fenêtres.

Cependant, dans le cas ou la perte de variance expliquée n'est pas trop importante, cette méthode devrais permettre une bonne visualisation des tendances différentes de $\mathcal{D}_{syn}$ et $\mathcal{D}_{orig}$. De plus, couplée a une mesure d'IO, la TD-ACP permettrait une approche multidimensionnel d'évaluation de l'utilité des données. Si les relations ne sont pas linéaire, appliquer une méthode du noyau (projection dans une nouvelle dimension avec une distance cosinus) devrais permettre une meilleur analyse d'une IO couplé a une TD-ACP.

MDS – Scaling Multi-dimensionnel Il existe deux types de scaling multidimensionnel, le métrique et le non-métrique. Celui utilisé par Jäckle, la MDS métrique, aussi appelée analyse de coordonnées principales (à ne pas confondre avec une Analyse de Composantes Principales), revient à effectuer une ACP, non pas sur une matrice de corrélation, mais sur une matrice de distance des données (Jäckle et al. 2015). L'article de Jäckle définit la distance entre deux individus de cette façon :

$$\text{distance}(A, B) = \frac{1}{p} \sum_{i=1}^{p} [A_i \neq B_i] \cdot w_i$$

Avec $[A_i \neq B_i]$ la distance euclidienne entre deux individus A et B distincts et $W = (w_1, w_2, ..., w_i, ..., w_p)$ le vecteurs poids des dimensions. Cependant, il est libre à l'auteur de définir lui-même la fonction de distance qu'il préfère.

Similairement à une ACP, la MDS linéaire n'est pas forcément évidente à interpréter. En effet, là où une ACP représente les p' dimensions maximisant la variance de l'objet, une MDS représente

les p' dimensions maximisant les distances entre les individus. Cette méthode souffre donc des mêmes problèmes qu'une ACP avec en plus une notion de distance qui peut s'avérer problématique pour les données avec des grandes dimensions. La complexité d'une MDS métrique est relativement importante, elle est de l'ordre de $O(N^2.p)$ ce qui est supérieur à une ACP classique.

T-MDS – Scaling multi-dimensionnel temporel Afin d'intégrer le facteur temporel des données, Jäckle utilise la méthode de la fenêtre temporelle (Jäckle et al. 2015). Cependant, Jäckle a une approche différente que la méthode de la TD-ACP. Il effectue une MDS à chaque fenêtre afin de projeter les données dans un espace de une dimension (et non deux comme une TD-ACP). Il pose $p' = 1$. Il représente ensuite ce vecteur en fonction du timecode des individus. Il retrouve donc une représentation bi-dimensionnel. Cependant, une MDS n'est pas invariante face aux rotations. En effet, certains vecteurs peuvent se retrouver avec une rotation de 180°, ce qui empêcherait une identification correcte des clusters. Il faut donc effectuer un "slice flip". Pour ce faire, Jäckle multiplie par -1 la valeur des données.

La complexité d'une telle opération est de l'ordre de $O(N^3)$ dans le pire des cas. Le coût conditionnel est donc très important. Il est aussi bon de souligner la difficulté d'interprétation des résultats. Non seulement le résultat souffrira du trop grand nombre de dimensions de notre espace original, mais le résultat final dépendra grandement de la taille des fenêtres temporelles choisies. De plus, il ne faut pas oublier qu'une MDS ne représente uniquement qu'un certain pourcentage de la distance choisit entre les individus et rien d'autre. Une intégration de cette méthode avec d'autres graphes ou d'autres analyses statistiques peut être nécessaire.

La TMDS devrait fournir une évolution visuelle de la distance entre les individus. De manière similaire à la TD-PCA, l'utilisation de la TMDS couplée à une mesure d'IO devrait permettre une évaluation statistique du degré de recouvrement de la régression de l'évolution des distances inter-individus en fonction du temps.

3.3 Utilité des données synthétiques

Dans le contexte de l'évaluation des données synthétiques, l'utilité se définit par la capacité de ces données à répondre de manière efficace et pertinente à des besoins spécifiques. Elle est évaluée en fonction de la pertinence et de l'applicabilité des données dans des contextes réels. Cette évaluation englobe la manière dont les données synthétiques peuvent être utilisées pour l'entraînement de modèles, la simulation de scénarios, ou pour la réalisation d'analyses statistiques. L'utilité est donc un indicateur crucial de la valeur pratique des données synthétiques, déterminant si elles peuvent constituer un substitut fiable et efficace aux données réelles dans divers domaines d'application.

3.3.1 Étude des corrélations

Afin de mesurer l'utilité du jeu de données synthétique généré, une approche consiste à étudier les corrélations entre les deux jeux de données D_{syn} et D_{org}, elle vise à évaluer si les données synthétiques peuvent être utilisées de manière interchangeable avec les données réelles pour certaines analyses statistiques. L'étude de la covariance réalisable avec un test de Pearson peut indiquer si D_{syn} conserve les relations statistiques observées dans D_{org}, moyennant la normalité des données. En effet, le test de Pearson mesure le degré de relation linéaire entre deux variables. Le coefficient de corrélation ρ est calculé en faisant le rapport entre la covariance de deux variables et le produit de leur écart type :

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Ce coefficient est ainsi calculé pour toutes les variables souhaitées dans les jeux de données original et synthétique. Ensuite, les coefficients respectifs des deux ensembles sont comparés, que ce soit visuellement grâce à des heatmaps ou par le biais d'un test de Student, par exemple. Des coefficients similaires suggèrent que les relations linéaires sont bien conservées dans les données synthétiques. Il est à noter que si les données présentent des relations non linéaires, d'autres méthodes d'analyse pourraient être nécessaires. Il est notamment possible d'utiliser le coefficient de corrélation de Spearman. Pour ce faire, chaque valeur des variables X et Y est remplacée par son rang dans l'ensemble de données, donnant de nouvelles variables $R(X)$ et $R(Y)$. La formule devient alors :

$$\rho_{R(X),R(Y)} = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

L'étude des corrélations dans l'évaluation des données synthétiques joue un rôle clé à la fois dans la mesure de l'utilité, de la fidélité et de la confidentialité. Ainsi, les analyses de corrélations sont des éléments importants de l'étude globale de la qualité des données synthétiques, soulignant son impact sur plusieurs facettes de la qualité des données.

3.3.2 Étude des chevauchements

Définis par Karr, il existe deux types différents de mesures statistiques permettant l'évaluation de la distribution des données générées (Karr et al. 2006) : les mesures étroites et les mesures larges. Les mesures étroites, comprenant le chevauchement d'intervalle de confiance (IO - Intervals Overlap) et le chevauchement d'ellipsoïdes (EO - Ellipsoid Overlap), permettent une analyse fine et spécifique à un certain type de jeux de données. En opposition, les mesures larges, comme l'erreur quadratique moyenne (MSE), sont plus grossières et proposent une analyse plus générale des distributions.

- *Chevauchement d'intervalle de confiance*

  La première mesure étroite nommée par Karr est le chevauchement des intervalles de confiance (Karr et al. 2006). Cette mesure unidimensionnelle permet une évaluation de la similarité de distribution entre les données réelles $\mathcal{D}_{\text{orig}}$ et générées $\mathcal{D}_{\text{syn}}$ en comparant les intervalles de confiance des coefficients directeurs de régression linéaire. Dans notre cas, les intervalles de confiance des coefficients β seront créés en appliquant la même régression linéaire sur la k-ième dimension de jeu de donnée en fonction du timecode des ensembles $\mathcal{D}_{\text{syn}}$ et $\mathcal{D}_{\text{orig}}$. Grâce aux distributions empiriques des données réelles et générées de la k-ième dimension d'un verbe en fonction du timecode ($f_{\text{rel, k}}$ et $f_{\text{orig, k}}$), le degré de recouvrement est calculé par la fonction I_k.

  $$I_k = \frac{1}{2}\left[\int_{L_{\text{syn, k}}}^{U_{\text{syn, k}}} f_{\text{orig, k}}(t)\,dt + \int_{L_{\text{orig, k}}}^{U_{\text{orig, k}}} f_{\text{syn, k}}(t)\,dt\right]$$

  avec $(U_{\text{rel, k}}, L_{\text{rel, k}})$ les bornes inférieures et supérieures de l'intervalle de confiance de $\mathcal{D}_{\text{syn}}$ pour la k-ième dimension d'un verbe. Afin d'obtenir une estimation globale du recouvrement de tous les verbes en fonction du timecode, Karr pose la fonction I telle que :

  $$I = \frac{1}{p}\sum_{i=1}^{p} I_k$$

Cette fonction a une valeur entre 1 et 0, 1 étant un recouvrement parfait et 0 un recouvrement nul. Elle moyenne les recouvrements unidimensionnels sans prendre en compte les interactions inter-dimensions. Dans le cadre de notre étude, il pourrait être intéressant d'observer la valeur de I en fonction, non seulement des k dimensions par verbe mais aussi par objet.

- *Chevauchement d'ellipsoïdes*

  Le chevauchement d'ellipsoïdes est une approche multidimensionnelle de la comparaison des distributions. En effet, l'une des principales critiques d'un chevauchement d'intervalles (IO) est le manque de représentation des relations plus complexes entre les données. L'objectif est de comparer les coefficients directeurs lors d'une régression linéaire multiple entre les dimensions k et le timecode. Cette méthode consiste en la création d'un ellipsoïde (espace multidimensionnel) dans lequel existe le coefficient d'une régression multiple d'un jeu de données. La probabilité que le coefficient d'un second jeu de données appartienne à cet espace est calculée avec une méthode de Monte-Carlo. De cette manière, Karr évalue le degré de chevauchement des surfaces de probabilités des valeurs des coefficients des distributions $\mathcal{D}_{\text{syn}}$ et $\mathcal{D}_{\text{orig}}$ (Karr et al. 2006).

  Afin de construire l'ellipsoïde qui recouvre β d'un jeu de données, Karr délimite une surface avec une loi de Fisher.

$$\frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{p\hat{\sigma}^2} \leq F(\alpha; p, n-p)$$

  Ici, $\hat{\beta}$ représente l'estimation du maximum des vrais coefficients β, $\hat{\sigma}^2$ est la variance résiduelle estimée, et $F(\alpha; p; n-p)$ représente la valeur critique d'une distribution de Fisher avec p et $n-p$ degrés de liberté.

  Dans le cas étudié, les bornes de l'ellipsoïde de $\mathcal{D}_{\text{syn}}$ (E_{syn}) sont obtenues en fixant $\hat{\beta} = \hat{\beta}_{\text{syn}}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{syn}}$, et $X = X_{\text{syn}}$. E_{orig} est obtenu de manière similaire. La mesure d'utilité EO est ensuite calculée, très similaire à la mesure I_k, comme la moyenne de deux probabilités :

$$EO = \frac{1}{2}\left(P(\beta \in E_{\text{orig}} | \mathcal{D}_{\text{syn}}) + P(\beta \in E_{\text{syn}} | \mathcal{D}_{\text{orig}})\right)$$

  Autrement dit, nous examinons la probabilité que β de $\mathcal{D}_{\text{syn}}$ appartienne à l'ellipsoïde qui englobe β de $\mathcal{D}_{\text{orig}}$. Cette probabilité, moyennée avec son complémentaire, donne l'indice de superposition des ellipsoïdes.

  Ces probabilités sont obtenues par des simulations de Monte Carlo. Dans le cas de $P(\beta \in E_{\text{orig}} | \mathcal{D}_{\text{syn}})$, Karr tire des valeurs de β de l'espace $\mathcal{D}_{\text{syn}}$, la proportion de ceux-ci qui appartiennent à E_{orig} est la probabilité étudiée.

  L'EO excelle dans la prise en compte des corrélations entre les paramètres, cruciales dans les scénarios où les interdépendances entre les paramètres influencent l'utilité des données. Bien qu'exigeante en termes de calculs par l'estimation des probabilités, l'EO représente une mesure robuste, garantissant une évaluation complète de la fidélité des données synthétiques à travers différentes dimensions.

L'analyse des chevauchements d'IO et d'EO bien qu'utilisée pour évaluer l'utilité des données synthétiques, ressemble beaucoup à une mesure de la fidélité. Cette similarité est due au fait que de manière générale, l'évaluation de la fidélité (la précision avec laquelle les données synthétiques imitent les données réelles) est intrinsèquement liée à leur utilité (la capacité de ces données à être utilisées de manière interchangeable avec les données réelles). Ainsi, bien que classée sous l'utilité, cette analyse fournit également des informations essentielles sur la fidélité des données synthétiques.

3.3.3 Performances des modèles

L'objectif ici est de comparer la performance d'un même modèle lorsqu'il est entraîné une fois sur des données réelles et une autre fois sur des données synthétiques. Si la performance du modèle sur les données synthétiques est comparable à celle sur les données réelles, cela indique que les données synthétiques sont utiles pour l'entraînement et peuvent servir efficacement à la modélisation dans des situations similaires.

Validation croisée

La validation croisée est massivement utilisée afin de paramétrer au mieux les modèles et obtenir les meilleurs résultats, mais nécessite d'être modifié pour pouvoir être applicable aux séries temporelles. Les défis rencontrés peuvent inclure des questions telles que la prise en compte de la dépendance temporelle dans les données, la gestion de l'ordre temporel lors de la séparation des données en ensembles d'entraînement et de test, et la manière d'éviter le biais dans l'évaluation du modèle dû à ces dépendances temporelles (Bergmeir and Benítez 2012). Plusieurs méthodes ont été élaborés ("time series split" ou "forward chaining" par exemple), impliquant la division des données en plusieurs sous-ensembles, chacun contenant une séquence temporelle de points de données. Contrairement à la validation croisée classique, où l'ordre des données n'est pas crucial, la validation croisée pour séries temporelles maintient la chronologie des observations, préservant ainsi les dépendances temporelles naturelles des séries. Le processus d'entraînement est ensuite classique. L'avantage de cette approche réside dans sa capacité à tenir compte des dépendances temporelles au sein des données, améliorant ainsi la pertinence des évaluations de performance du modèle dans des contextes temporels réels. Cependant, la mise en œuvre de la validation croisée pour séries temporelles peut être intensive en ressources, notamment avec des ensembles de données volumineux, car elle nécessite la manipulation de multiples séquences temporelles. Une fois la validation croisé effectué, le score du modèle entraîné sur les données d'origines $\mathcal{D}_{org}$ est comparé à celui obtenu sur les données synthétiques $\mathcal{D}_{syn}$ si les scores sont équivalents on en conclut à une utilité importante de $\mathcal{D}_{syn}$ démontrant sa qualité. Il est aussi possible de moyenner les estimations de coefficients et d'évaluer leurs écarts-types à l'issue de la procédure de validation croisée pour comparer leurs valeurs en fonction de la méthode de synthèse utilisée.

Analyse des Prédictions

1. MSE

   De plus, il existe différentes mesures afin d'évaluer la pertinence de $\mathcal{D}_{syn}$. La plus connu étant la MSE, déjà énoncé précédemment peut être utilisé dans le but de calculer la distance euclidienne carrée entre les prédictions des modèles issus de $\mathcal{D}_{orig}$ et celle issus de $\mathcal{D}_{syn}$ (Y^{orig}, Y^{syn}).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}(y_i^{\text{syn}} - y_i^{\text{orig}})^2$$

Avec $(y_i^{\text{syn}} - y_i^{\text{orig}})^2$ la distance euclidienne carrée de la i-ème prédiction de Y^{syn} et Y^{orig}.

Le MSE est utilisé pour mesurer la différence directe entre les prédictions des modèles basés sur les données originales et synthétiques.

Une seconde mesure d'évaluation plus générale est la divergence de Kullback-Liebler déjà énoncé plus haut. Appliquée aux ensembles de prédictions, KL serais une mesure pertinente de même.

2. pMSE

   La pMSE est une méthode statistique permet de quantifier la capacité d'un modèle à distinguer une observation synthétique d'une observation réelle (Henri Chhoa 2023).

   Il compare la probabilité d'appartenir aux données synthétiques $\hat{p}_i$ estimée par un classeur et le pourcentage des données synthétique globale (0.5%).

$$\text{pMSE} = \frac{1}{2N} \sum_{i=1}^{2N} (\hat{p}_i - c)^2$$

   où c représente la proportion de données synthétiques ($c = 0.5$).

   Le cas idéal correspond au scénario où toutes les prédictions $\hat{p}_i$ sont égales à c. La pMSE du classeur serais donc proche de 0.

3. F1 score

   Le score F1 est une mesure utilisée afin d'évaluer les erreurs d'un algorithme de classification.

$$\text{F1 score} = 2 \times \left(\frac{\text{Precision} \times \text{Rappel}}{\text{Precision} + \text{Rappel}} \right)$$

$$\text{Precision} = \left(\frac{\text{Vrais Positifs (TP)}}{\text{Vrais Positifs (TP)} + \text{Faux Positifs (FP)}} \right)$$

$$\text{Rappel} = \left(\frac{\text{Vrais Positifs (TP)}}{\text{Vrais Positifs (TP)} + \text{Faux Négatifs (FN)}} \right)$$

4. MAPE

   La MAPE est calculée en prenant la moyenne des erreurs absolues en pourcentage entre les valeurs réelles réelle y_i et les valeurs prédites $\hat{y}_i$. Plus la MAPE est petite, meilleur est la prédiction du modèle.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Visualisation

Les techniques de visualisation dans l'étude de la qualité des données synthétiques ne se limitent pas uniquement à évaluer la fidélité, mais peuvent également être utilisées pour mesurer l'utilité des modèles. Par exemple, en visualisant comment les modèles formés sur des données synthétiques se comportent par rapport à ceux formés sur des données réelles, on peut obtenir des insights visuels

sur leur efficacité. Ces visualisations peuvent révéler des différences subtiles dans les performances des modèles, offrant ainsi une compréhension plus nuancée de l'utilité des données synthétiques dans diverses applications pratiques.

4 Autre méthode utilisé pour l'évaluation des données générés

L'évaluation humaine

L'évaluation humaine fournit des informations qualitatives sur la fidélité, la variabilité et la pertinence des séries temporelles synthétiques, ainsi que sur leur capacité à préserver les caractéristiques des séries temporelles du monde réel à partir desquelles elles ont été générées. Elle nécessite l'implication d'experts humains dans l'évaluation des données synthétisées sur la base de critères spécifiques. Les méthodes d'évaluation humaine comprennent des évaluations qualitatives, dans lesquelles les experts fournissent des évaluations détaillées sur la cohérence, la pertinence et d'autres aspects subjectifs des données (Stenger et al. 2024).
De plus, des évaluations quantitatives subjectives peuvent être intégrées à travers des échelles de notation ou des questionnaires pour quantifier des aspects tels que la clarté, la variété, et l'originalité des données générées. Les retours des évaluateurs humains jouent un rôle essentiel dans le processus d'amélioration. Le feedback constructif fourni par les experts guide les ajustements nécessaires dans les modèles de génération, permettant une amélioration itérative de la qualité des données produites. Cette rétroaction humaine contribue également à affiner les critères d'évaluation, rendant le processus d'évaluation plus pertinent.

5 Perspective et application dans le projet

En conclusion de cette étude, il serait opportun de développer une méthodologie d'évaluation des données synthétiques en utilisant les trois critères mentionnés précédemment, à savoir la confidentialité, la fidélité et l'utilité. Cette approche consisterait à définir un score pour chaque critère, représentant la moyenne des mesures spécifiques à chaque critère.

Par la suite, un score global serait calculé en agrégeant les trois scores individuels de chaque critère. Cette approche permettrait d'obtenir un critère de décision pour évaluer la qualité de ces données générées. Ainsi, il deviendrait plus aisé d'appréhender et de comparer les performances des données synthétiques selon ces critères clés.

References

Bergmeir, Christoph and José M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191. Data Mining for Software Trustworthiness, pp. 192–213. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2011.12.028.

Bergsma, Wicher (2013). "A bias-correction for cramer's v and tschuprow's t". In: *Journal of the Korean Statistical Society, 42(3):323–328.*

Bethencourt-Aguilar, Anabel et al. (2023). "Use of Generative Adversarial Networks (GANs) in Educational Technology Research". In.

Cano, Isaac and Vicenç Torra (2009). "Generation of synthetic data by means of fuzzy c-Regression". In: *2009 IEEE International Conference on Fuzzy Systems.* IEEE, pp. 1145–1150.

Commission Nationale de l'Informatique et des Libertés (CNIL) (Jan. 2022). *Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation.*

Dash, Saloni et al. (2020). "Medical time-series data generation using generative adversarial networks". In: *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*. Springer, pp. 382–391.

Defays, D and MN Anwar (1998). "Masking microdata using micro-aggregation". In: *Journal of Official Statistics* 14.4, p. 449.

Grigoryev, Timofey, Andrey Voynov, and Artem Babenko (2022). "When, why, and which pretrained GANs are useful?" In: *arXiv preprint arXiv:2202.08937*.

Hazra, Debapriya and Yung-Cheol Byun (2020). "SynSigGAN: Generative adversarial networks for synthetic biomedical signal generation". In: *Biology* 9.12, p. 441.

Henri Chhoa Salim Kabiri, Yann Huquet (2023). "Génération synthétique de données confidentielles - Comparaison de 3 algorithmes sur différents cas d'usage en banque et assurance". In: *Nexialog Consulting*.

Hirsch, H Guenter (2005). "Fant-filtering and noise adding tool". In: *Niederrhein University of Applied Sciences*.

Jäckle, Dominik et al. (2015). "Temporal MDS plots for analysis of multivariate data". In: *IEEE transactions on visualization and computer graphics* 22.1, pp. 141–150.

Jamin, Antoine and Anne Humeau-Heurtier (2020). "(Multiscale) Cross-Entropy Methods: A Review". In: *Entropy* 22.1. ISSN: 1099-4300. DOI: 10.3390/e22010045. URL: https://www.mdpi.com/1099-4300/22/1/45.

Karr, Alan F et al. (2006). "A framework for evaluating the utility of data altered to protect confidentiality". In: *The American Statistician* 60.3, pp. 224–232.

Kurtukova, Anna, Aleksandr Romanov, and Anasstasia Fedotova (2019). "De-Anonymization of the Author of the Source Code Using Machine Learning Algorithms". In: *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. IEEE, pp. 0612–0617.

Matthieu Herrmann Chang Wei Tan, Geoffrey I. Webb (2023). "Parameterizing the cost function of dynamic time warping with application to time series classification". In: *Data Mining and Knowledge Discovery* 37.

Morishita, Tetsuya (2021). "Time-dependent principal component analysis: A unified approach to high-dimensional data reduction using adiabatic dynamics". In: *J. Chem. Phys.* 155.134114.

Narayanan, Arvind and Vitaly Shmatikov (2008). "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, pp. 111–125.

Raghunathan, Trivellore E (2021). "Synthetic data". In: *Annual review of statistics and its application* 8, pp. 129–140.

Reiter, Jerome P and Robin Mitra (2009). "Estimating risks of identification disclosure in partially synthetic data". In: *Journal of Privacy and Confidentiality* 1.1.

Stenger, Michael et al. (2024). "Evaluation is key: a survey on evaluation measures for synthetic time series". In: *Journal of Big Data* 11.1, p. 66.

Vega-Márquez, Belén et al. (2020). "Creation of synthetic data with conditional generative adversarial networks". In: *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019) Seville, Spain, May 13–15, 2019, Proceedings 14*. Springer, pp. 231–240.

Yoon, Jinsung, Daniel Jarrett, and Mihaela van der Schaar (2019). "Time-series Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc.