

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Exploration de la génération de données synthétiques d'apprentissage profond pour partager des données sensibles liées à des services publics

par Benjamin Santos, Rafik Chemli et Julian Templeton

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Exploration de la génération de données synthétiques d'apprentissage profond pour partager des données sensibles liées à des services publics

Benjamin Santos, Rafik Chemli et Julian Templeton¹

Résumé

Les services publics ont des renseignements essentiels sur la consommation d'énergie et les caractéristiques des bâtiments, des renseignements que les organismes gouvernementaux peuvent utiliser pour améliorer leurs analyses correspondantes. Ces données, de nature confidentielle, sont toutefois associées aux enregistrements de clients. Les données sur les bâtiments et la consommation d'énergie peuvent être trop délicates pour être partagées. Souvent, les versions agrégées de ces données sont échangées en vertu de contrats très restrictifs, ce qui limite le nombre de statistiques qui peuvent en être extraites. Avec l'avancement des techniques génératives d'apprentissage automatique, Statistique Canada et Ressources naturelles Canada ont évalué s'il était possible d'utiliser ces modèles pour produire des versions synthétiques des données sur les services publics qui peuvent être échangées intégralement avec les organismes qui en font la demande. Un service public peut créer ces ensembles de données synthétiques dans le cadre d'un programme exécuté localement. Les extraits peuvent être approuvés avant d'être envoyés. Ces travaux ont permis d'établir que les services publics peuvent utiliser certains modèles génératifs pour produire de nouvelles versions d'un ensemble de données, ainsi que de recenser les problèmes qui doivent être résolus avant de mettre en œuvre le tout en pratique. Des modèles tabulaires et de séries temporelles ont été mis à l'essai pour différents scénarios d'échange de données. Le modèle TimeGAN a permis de saisir les pics et les creux de consommation d'énergie au cours d'une journée donnée en respectant des exigences de calcul raisonnables. Bien que ce processus prenne plusieurs jours avant que les quantités d'énergie annuelles totalisent des milliers d'enregistrements de clients, de nouvelles initiatives d'échange de données peuvent être mises en œuvre entre les services publics et les organismes nationaux de statistique, tout en gérant les risques liés à la protection des renseignements personnels. Au fur et à mesure que les travaux avanceront dans le cadre des futures phases qui se dérouleront auprès de vrais partenaires des services publics, la confiance envers ces approches peut se renforcer. On pourra commencer à mettre à l'essai celles-ci en utilisant des données réelles fournies par de réels détenteurs de données.

Mots clés : Données synthétiques; série temporelle; modèles génératifs; échange de données; confidentialité; données des services publics.

1. Introduction

Les données des services publics, essentielles pour réaliser une analyse des bâtiments, sont confidentielles et sont souvent restreintes en raison de questions liées à la protection des renseignements personnels. Ressources naturelles Canada (RNC) cherche à recueillir ces données auprès de différentes sources et à les utiliser pour améliorer leurs études et analyses. Cependant, la protection de la vie privée des clients, dont les enregistrements se trouvent dans les ensembles de données des services publics, fait partie des nombreuses préoccupations des détenteurs de données. Il existe donc des obstacles importants quand vient le temps d'accéder à ces données, ce qui peut entraîner des retards considérables lors de la réalisation de toute analyse nécessitant l'utilisation de ces données. Pour relever ces défis, on suggère d'utiliser des modèles modernes de génération de données synthétiques (DS) d'apprentissage profond. Cette approche offre la possibilité d'établir un équilibre entre la création de valeur émanant de l'inférence statistique et la protection de la vie privée, ce qui permet d'améliorer l'analyse sans compromettre la confidentialité des clients. Dans cet article, nous allons donner un aperçu des travaux réalisés dans le cadre de ce projet, en plus de fournir des points de discussion précis découlant des essais effectués. La première section de l'article décrira clairement la portée du projet et ses limites. Ensuite, nous décrirons tous les renseignements de base entourant les données et les modèles utilisés. Nous allons ensuite faire état des essais effectués. L'article se termine en résumant les travaux et en décrivant les étapes futures.

¹ Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 (benjamin.santos@statcan.gc.ca)

1.1 Portée

Cette étude cherche à répondre à trois questions sur l'utilisation de techniques de génération de DS d'apprentissage profond dans le cadre des données confidentielles des services publics. 1) Quelles sont les méthodes accessibles pour produire des DS de qualité? 2) En se fondant sur le test d'un sous-ensemble de ces modèles, est-il possible de les calculer et sont-ils capables de saisir les répartitions générales et les tendances (c'est-à-dire les pics et les creux énergétiques) des données originales? 3) D'après les résultats des tests et les discussions tenues, quels sont les principaux facteurs à prendre en compte pour mettre en pratique cette recherche, si elle produit de bons résultats?

Parmi les limites associées à ces travaux, mentionnons les suivantes : 1) Nous nous sommes concentré uniquement sur la faisabilité de l'exécution des méthodes, sur la question à savoir si les DS saisissent la répartition générale des données originales et si les pics et les creux énergétiques peuvent être saisis pour chaque jour; 2) toutes les techniques recherchées n'ont pas pu être mises à l'essai; et 3) en fonction de la puissance de calcul accessible, seule une optimisation de base est effectuée.

2. Contexte

2.1 Données synthétiques et protection de la vie privée

Le domaine de recherche des DS est vaste comprenant de nombreuses techniques et de nombreux critères d'évaluation définis. Dans cet article, nous définissons un ensemble de données synthétiques qui maintiendra les mêmes propriétés statistiques et comportements que l'ensemble de données initial, mais qui représente un ensemble de données totalement unique (voir Données entièrement synthétiques dans CEE-ONU, 2022). Chaque enregistrement synthétique représente un enregistrement de client fictif. Idéalement, l'ensemble de données synthétiques complet et les tendances observées ressembleront à l'ensemble de données initial, même s'il s'agit de nouveaux points de données. Les modèles d'apprentissage profond génératifs cherchent à produire un nouvel ensemble de données qui se fonde sur les propriétés statistiques apprises des données initiales. Ils peuvent s'appliquer à des données de séries temporelles et peuvent être configurés de manière appropriée pour réaliser des tâches particulières, comme le maintien des pics et des creux en ce qui concerne les données sur l'énergie. Ce nouvel ensemble de données sera naturellement différent des données initiales, ajoutant un degré de protection de la vie privée au processus. Malgré leur potentiel, ces méthodes ont un coût. En général, elles nécessitent des unités centrales graphiques (UCG) ou des accélérateurs et suffisamment de mémoire vive (RAM). En outre, le temps nécessaire pour entraîner un modèle en fonction de mesures horaires d'une journée donnée peut varier, allant de 20 minutes à plusieurs heures, en fonction de la configuration et du type de modèle utilisé. En outre, l'évaluation des DS est difficile à mesurer et exige une excellente compréhension de leur utilisation (pour en savoir davantage, voir CEE-ONU, 2022). Il existe donc de nombreux compromis à analyser lorsque l'on décide du type de modèle génératif à utiliser. Malgré la complexité du modèle et le fait que les résultats constituent un nouvel ensemble de données, l'utilisation de ces modèles peut offrir des possibilités dans un scénario « **tout ou rien** ». C'est le cas lorsque RNCAN ne peut pas récupérer les données originales telles quelles et ne veut pas seulement des analyses de base à partir des données d'un service public. Bien que les DS soient différentes des données originales, elles pourraient être utiles pour réaliser une analyse approfondie, en plus d'être utiles pour effectuer une analyse exploratoire des données et pour réaliser des essais et assurer l'apprentissage. Il peut s'agir d'une initiative de grande envergure, mais qui peut également ouvrir la voie à de nouvelles initiatives et perspectives qu'il aurait été impossible d'obtenir autrement.

Puisque les préoccupations en matière de protection de la vie privée sont au premier plan de cette étude, il est essentiel de comprendre les protections au chapitre de la vie privée fournies par les solutions envisagées. Comme mentionné précédemment, les DS produites par les modèles dans le cadre de ces travaux sont différentes des points de données originaux, et la qualité des données produites varie en fonction du modèle sélectionné et des hyperparamètres qu'il utilise. En outre, certains modèles génératifs mettent en application la confidentialité différentielle (CD) dans le cadre du processus d'entraînement, ce qui offre de solides garanties mathématiques au chapitre du degré de confidentialité, représenté sous forme de bruit, ajouté aux données. Les modèles utilisant la CD, comme DPGAN (Xie, 2018), fourniront davantage de garanties en matière de protection de la vie privée, afin de mieux protéger les enregistrements des clients à un coût correspondant à leur valeur d'inférence statistique. À part le service public, RNCAN ne serait pas

en mesure de dire avec certitude si les données sont différentes ou non des données originales, et dans quelle mesure. Il est difficile de trouver un équilibre entre la protection de la vie privée et la valeur d'inférence statistique des DS produites, mais cela permet aux services publics d'être plus certain que les données de leurs clients sont protégées.

2.2 Données tabulaires et de série temporelle

Différents modèles génératifs sont conçus pour utiliser divers types de données; il est donc essentiel de comprendre les données qui serviront de données d'entrée et ce que les données de sortie doivent saisir. Dans le cadre de ce projet, nous avons utilisé des données de série temporelle simulées associées aux enregistrements de la consommation d'énergie des clients pour différents bâtiments. Par conséquent, si l'objectif est de saisir l'ensemble de la série temporelle, le modèle utilisé doit être conçu pour accepter des séries temporelles de consommation d'énergie comme données d'entrée et produire des données sous forme de série temporelle. Dans ce cas-ci, les données sont pour quelques heures d'une journée pour chaque jour d'une année arbitraire. Par conséquent, le modèle apprend à partir des séries temporelles d'entrée et produit une nouvelle série temporelle qui crée séquentiellement des échantillons correspondant à la quantité totale utilisée comme données d'entrée. Par exemple, si un modèle apprend un seul jour d'enregistrements de clients, il sera en mesure de produire exactement un jour d'enregistrements pour le même nombre d'enregistrements utilisés comme données d'entrée. Au cours de la période saisie, les renseignements des périodes précédentes sont utilisés comme contexte pour prédire les résultats des périodes futures. La tâche se révèle plus ardue que de travailler avec des données tabulaires qui n'exigent pas que le temps soit saisi dans les données d'entrée elles-mêmes. Le tableau 2.2.1-1 propose un exemple illustrant la forme que les données peuvent prendre dans une série temporelle.

Tableau 2.2.1-1

Exemple de points de données d'une série chronologique (extraction)

datapoint_id	:dcv_type	:ext_wall_cond	hour	energy_elec	energy_gas
a	Occupancy_based	0,278	0	0,252	0,001
a	Occupancy_based	0,278	1	0,211	0,532
a	Occupancy_based	0,278	2	0,157	0,741
b	No_DCV	0,183	0	0,211	0,532
b	No_DCV	0,183	1	0,157	0,741

Dans le tableau 2.2.1-1, on peut constater que ces données correspondent à un ensemble de caractéristiques statiques (caractéristiques du bâtiment) et de données de série chronologique (énergie). Même si les données représentent une série chronologique, elles peuvent aussi être utilisées comme données tabulaires en ignorant l'aspect temporelle. Les modèles génératifs utilisant un modèle tabulaire apprennent comment produire des enregistrements de clients individuels sans fournir de contexte sur le temps. En utilisant ces modèles, on ne saisit pas de renseignements des séries temporelles, ce qui peut convenir si les données d'entrée ne contiennent pas de tendances dépendant du temps. Il peut tout de même être utile de travailler dans un cadre tabulaire pour saisir des renseignements qui peuvent être difficiles à obtenir dans un modèle de séries temporelles. Dans le cadre de cette étude, on peut utiliser un modèle tabulaire pour prédire un seul résultat par enregistrement de client lorsque cet enregistrement contient des renseignements auxiliaires de la période résumée. Par exemple, si on sait que la consommation d'énergie pour l'électricité et le gaz connaîtra un seul pic et un seul creux en journée, on peut faire en sorte que le modèle apprenne à produire des versions de DS des enregistrements pour chaque jour, y compris les heures et les valeurs des pics et des creux.

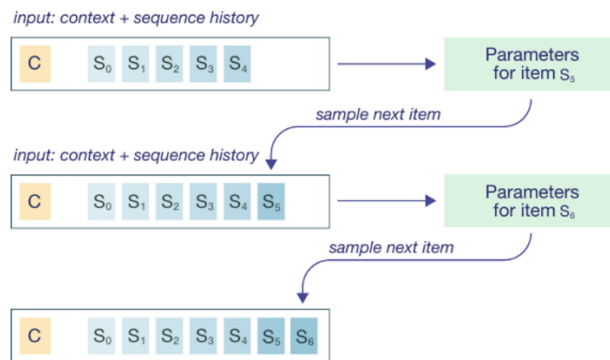
2.3 Modèle génératif autorégressif probabiliste conditionnel

Ce modèle, portant le nom de modèle génératif autorégressif probabiliste conditionnel ou CPAR (Zhang, 2022), cherche à produire des DS à partir des données réelles composées d'un contexte ou de caractéristiques statiques et d'une série chronologique. Ces données renferment un ensemble de types de données : numériques, catégoriques et chronologiques; dans le tableau 2.2.1-1 : « :dcv_type » est une caractéristique catégorique (renfermant un petit ensemble de valeurs possibles), « energy_elec » peut être considéré comme une valeur numérique (continue) et « hour » correspond à l'heure. Il faut également souligner que chaque série temporelle ou suite (bleu ou rouge) dispose d'un ensemble différent de caractéristiques statiques (contexte).

Tout d'abord, le modèle divise le contexte à partir de suites. Il met ensuite en application un modèle de copule gaussienne au contexte, afin de comprendre les corrélations entre les caractéristiques statiques (les paramètres du bâtiment). Plus précisément, les distributions marginales de chaque colonne sont extraites et transformées en distributions uniformes, puis la distribution conjointe est calculée sous forme de distribution normale multivariée qui saisira la relation entre les caractéristiques statiques. Depuis le modèle de copule gaussienne, un nouveau contexte est produit. Le modèle cherche ensuite à saisir les dépendances de la série temporelle en utilisant l'historique de la suite (S) et du contexte (C). Ainsi, il peut produire des données synthétiques de série temporelle. Cette idée est illustrée dans la figure 2.3.1 (Zhang, 2022). Un réseau neuronal est donc entraîné en utilisant les séries comme données d'entrée. Plus précisément, des unités récurrentes fermées sont utilisées dans ce cas-ci. En plus des unités de mémoire à court terme longues, il s'agit de composantes particulières dans la catégorie élargie de réseaux de neurones récurrents (RNR). Les RNR permettent aux données de sortie précédentes d'être traitées comme des données d'entrée.

Figure 2.3-1

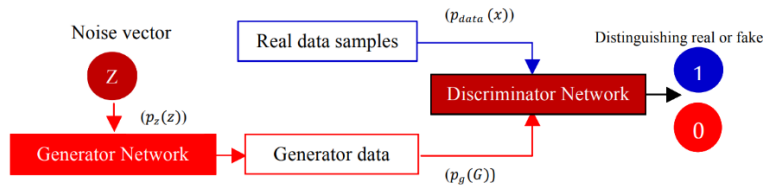
Exemple illustrant comment la série et le contexte sont utilisés dans le modèle CPAR



2.4 Modèle génératif tabulaire

Les réseaux antagonistes génératifs (RAG) sont des modèles génératifs constitués de deux parties. Ces deux parties prennent la forme de réseaux neuronaux. Le premier est un discriminateur (D) qui estime la probabilité qu'un échantillon donné vienne du véritable ensemble de données. Le deuxième est un générateur (G) qui produit des échantillons synthétiques à partir de la distribution uniforme $p_z(z)$. Voir la figure 2.4-1 (Jabbar, 2021). Autrement dit, D agit comme vérificateur et cherche à ne pas être trompé par G, alors que G est un faussaire qui tente de créer des échantillons réalistes pour tromper D. Pour améliorer le rendement de D et de G, les réseaux sont en concurrence les uns avec les autres au cours du processus d'entraînement. Ils participent donc à un jeu à somme nulle, dans le cadre duquel si l'un gagne, l'autre perd. Puisque l'objectif de G est d'apprendre la distribution de données réelles $p_{data(x)}$, et que l'objectif de D est de déceler les véritables échantillons (tirés d'échantillons réels et synthétiques $p_{g(G)}$), lors de l'entraînement, la fonction de perte doit tenir compte des pertes de G et de D pour améliorer les tâches.

Figure 2.4-1
Architecture générale d'un RAG



2.4.1 Réseau antagoniste génératif (RAG) tabulaire conditionnel (CTGAN)

La production de nouvelles images fait partie des premiers résultats obtenus en utilisant des RAG. Plusieurs variantes de modèles de RAG ont fait surface pour la production de données tabulaires, l'un d'entre eux étant le CTGAN (Xu, 2019) qui a apporté des améliorations pour gérer le mélange de types trouvé dans les données tabulaires, le déséquilibre au chapitre des caractéristiques catégoriques, l'apprentissage à partir de vecteurs codés one-hot et la production conditionnelle d'échantillons. Comme les RAG les plus simples ne peuvent pas traiter ces problèmes, on a ajouté le CTGAN aux expériences de cette étude.

2.4.2 Anonymisation au moyen de la synthèse de données à l'aide de RAG (ADSGAN) et de DPGAN

Le modèle ADSGAN s'appuie sur les idées des RAG conditionnels et prétend surpasser le modèle DPGAN, qui est un modèle RAG entraîné avec des garanties de CD. Par CD, on entend ce qui suit : pour tous les ensembles de données D_1 et D_2 qui diffèrent au chapitre d'un seul point de données, et pour tous les sous-ensembles S pour un algorithme randomisé A (ou un modèle d'AA ou un générateur de DS) (Yoon, 2020) :

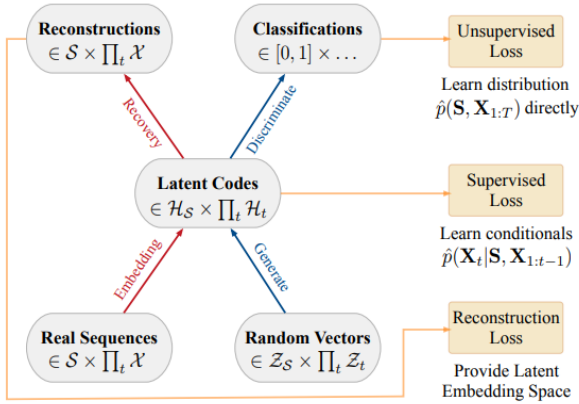
$$P(A(D_1) \in S) \leq e^\epsilon P(A(D_2) \in S)$$

L'algorithme ou le générateur A fournit un ϵ -CD, si la différence entre la distribution de probabilité du générateur entraîné avec l'ensemble de données D_1 et la distribution de probabilité du générateur entraîné avec le même ensemble de données à l'exception d'un point de données est inférieure à e^ϵ . C'est pourquoi on affirme que le modèle DPGAN produit des DS différenciellement confidentielles (DSDC). Selon les auteurs du modèle ADSGAN (Yoon, 2020), les DSDC pourraient être de mauvaise qualité lorsque vient le temps d'avoir une distribution conjointe semblable à l'ensemble de données original (voir le modèle de copule gaussienne ci-dessus), et la confidentialité différentielle ne peut pas être calculée directement avec des ensembles de données finis. Pour outrepasser ces limites, les auteurs ont créé un concept différent appelé ϵ -identifiabilité qui offre une protection moindre que la confidentialité différentielle, mais va de pair avec leur cas d'utilisation sur la synthèse des dossiers médicaux.

2.5 Modèle TimeGAN

Le modèle TimeGAN est une généralisation des générateurs pour les données tabulaires avec série temporelle (Yoon, 2019). La structure des données est la même que celle expliquée ci-dessus dans le cas du modèle CPAR. Pour un ensemble de données $D = \{(s_n, x_{n,1:T_n})\}_{n=1}^N$ de N points de données, on trouve un ensemble statique de caractéristiques S et de séries X . Le modèle TimeGAN cherche à atteindre deux objectifs. Tout d'abord, il faut apprendre une densité $p(S, X_{1:T})$ qui évalue mieux la distribution conjointe sous-jacente $p(S, X_{1:T})$ (cela tient compte de toutes les caractéristiques statiques et des séries). Ensuite, il faut apprendre la condition $p(X_t; X_{1:T-1})$. Le modèle est composé de quatre éléments : les fonctions de vectorisation et de récupération, le générateur de séries et le discriminateur de séries. Les deux premiers éléments sont entraînés ensemble, avec le générateur et le discriminateur (figure 2.5-1). Parmi les aspects importants de ce modèle, il faut considérer que les composantes « Vectorisation » et « Récupération » peuvent prendre n'importe quelle forme, à condition qu'elles soient autorégressives et qu'elles respectent un ordonnancement causal. C'est pourquoi il est possible d'utiliser les réseaux de neurones récurrents, les unités de mémoire à court terme longues, les unités récurrentes fermées et d'autres modèles autorégressifs pour entraîner les composantes conditionnelles.

Figure 2.5-1
Architecture du modèle TimeGAN



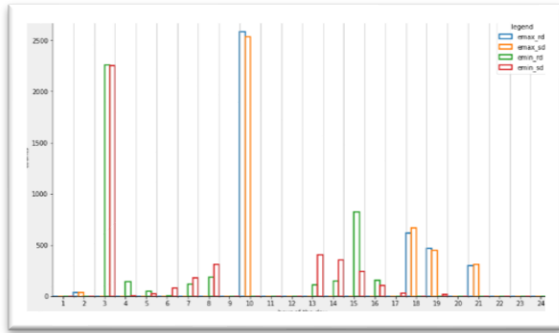
3. Résultats et conclusions

3.1 Expériences relatives au modèle génératif tabulaire

On recherche à vérifier si les générateurs peuvent saisir la relation, s'il y en a une, entre les caractéristiques du bâtiment (4 000 bâtiments différents) et les valeurs maximales et minimales de consommation avec le moment où elles se produisent (pics et creux). Nous avons mis l'accent sur une période de 24 heures, où (*temax*, *emax*) et (*temin*, *emin*) sont les heures associées au valeur maximum, et les heures associées au valeur minimum, respectivement, pour un service public précis. Nous avons également mis à l'essai plusieurs modèles génératifs à partir du paquet **synthcity** (Qian, 2023) : **CTGAN**, **ADSGAN**, **DPGAN** et **distributions marginales**. Le rapport détaillé renferme les mesures globales (Templeton, 2024). À partir de ces mesures, nous pouvons conclure que, de manière générale, le modèle **ADSGAN** offre le meilleur compromis en ce qui concerne les services publics et la protection de la vie privée, le modèle **DPGAN** altère les services publics et les **distributions marginales** altèrent les services publics et la protection de la vie privée. Cette situation est attendue, en raison de l'ajout de bruit de la CD à la phase d'entraînement dans le cas de **DPGAN**, alors que les **distributions marginales** établissent tout simplement un point de données en l'échantillonnant à partir des distributions marginales. Le modèle **CTGAN** affiche un bon rendement. Les mesures de la protection de la vie privée ne sont pas trop éloignées comparé à celle qui est offerte par les autres méthodes. En termes de rendement d'exécution, tous les générateurs, à l'exception des distributions marginales, prennent à peu près le même temps à entraîner : 10 minutes avec un processeur centrale **AMD** à 16 et au moins un processeur graphique avec une mémoire RAM de 2 Go. Le générateur de distributions marginales sert de modèle de référence. La rapidité de son entraînement est incomparable. Nous avons été surpris de constater que les modèles **ADSGAN** et **CTGAN** ont été en mesure de saisir les pics et creux (figure 3.1-1). Les mesures de classification (fig. 3.1-1) confirment également ces observations, alors que l'on peut observer que les pics d'électricité sont très bien représentés dans les DS, et que les creux pour les deux services publics n'ont pas été pris en compte. La moyenne de ces mesures est calculée pour toutes les catégories (moyenne macro).

Figure 3.1-1

(gauche) Comparaison des chiffres des heures de pointe réelles (DR) et synthétiques (DS) pour ADSCAN, (droite) mesures de classification des pics et des creux



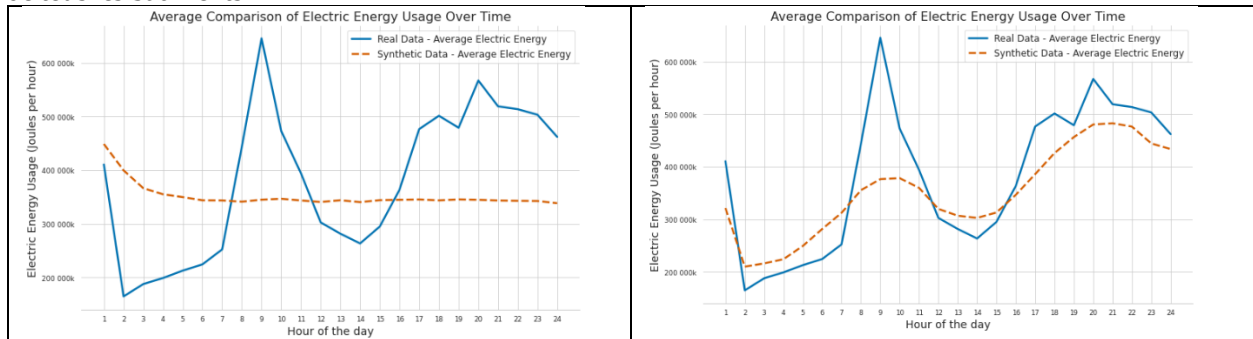
Extremum	Exactitude	Précision	Rappel	Score F1
Électricité max.	0,98	0,95	0,99	0,97
Électricité min.	0,72	0,35	0,41	0,32
Gaz max.	0,97	0,25	0,25	0,25
Gaz min.	0,72	0,25	0,27	0,20

3.2 Expériences liées à un modèle génératif de série chronologique

Dans la figure 3.2-2, on compare les données réelles et les DS de la moyenne horaire de l'électricité au cours d'une série temporelle de 24 heures au 1^{er} janvier pour obtenir des données sur l'électricité. Les DS du modèle CPAR (gauche) semblent s'écarter de la moyenne, ce qui signifie que la probabilité conditionnelle de la séquence n'a pas été prise en compte par le modèle. Il faut souligner qu'une évaluation hyperparamétrique est nécessaire pour examiner pleinement ce modèle, même si l'on ne s'attend pas à ce qu'il soit très efficace quand vient le temps de protéger la vie privée. À droite, on peut voir les résultats du modèle TimeGAN utilisant les réseaux de neurones récurrents (par défaut) du paquet *synthcity*, qui démontrent que les tendances sont mieux préservées. En ce qui concerne les extremums en matière d'énergie, on a constaté que l'erreur est supérieure au chapitre des chiffres relatifs aux heures dans les données synthétiques par rapport aux heures dans les données réelles. Le résultat est moins performant que celui observé lors des tests du modèle tabulaire.

Figure 3.2-2

Résultats du modèle CPAR par rapport aux résultats du modèle TimeGAN pour l'électricité horaire moyenne de tous les bâtiments



Avant de conclure, nous avons réalisé une analyse exploratoire, afin d'évaluer s'il est possible d'utiliser des modèles d'apprentissage profond génératifs pour produire des données synthétiques des données confidentielles des services publics. Nous avons commencé par évaluer les différentes méthodes qui peuvent être utilisées pour accomplir cette tâche et les comportements des données d'entrée utilisées pour les tests. En comprenant mieux les différentes approches et les données, on a effectué des tests avec les modèles CPAR, TimeGAN et tabulaires génératifs. Bien qu'on ait au départ observé des schémas clairs entre les distributions des ensembles de données et la distribution des données originales, des tests supplémentaires ont été effectués pour analyser si les modèles peuvent maintenir les pics et les creux de consommation d'énergie observés à partir des données originales. Il s'agit d'une tâche assidue, car, dans le cadre du projet, nous avons observé que le modèle TimeGAN et le modèle tabulaire génératif finissaient par maintenir les pics et les creux après un nombre considérable d'expériences. Ces modèles se sont donc révélés être des

options potentielles réalisables qui devront être évaluées de manière plus approfondie au cours des phases futures. Des considérations supplémentaires sont encore nécessaires, par exemple dans le cas de valeurs creuses de consommation de gaz (tous les bâtiments ne consomment pas de gaz), qui pourraient être traitées par un modèle génératif distinct. Dans l'ensemble, les tests ont prouvé la faisabilité de l'utilisation de modèles génératifs d'apprentissage profond pour produire des données synthétiques. Cependant, il ne s'agit que d'un aperçu préliminaire d'une initiative de grande envergure, mais importante.

Bibliographie

- Jabbar, A., Li, X., et Omar, B. (2021), "A survey on generative adversarial networks: Variants, applications, and training", *ACM Computing Surveys (CSUR)*, 54(8), p. 1-49.
- Qian, Z., Cebere, B. C., et Van der Schaar, M. (2023), "Synthcity: facilitating innovative use cases of SD in different data modalities", *arXiv preprint*, arXiv:2301.07573.
- Templeton, J., Santos, B., et Chemli, R. (2024), "Exploration of Deep Learning Synthetic Data Approaches to Help Retrieve Sensitive Utility Data", Statistique Canada, rapport interne.
- UNECE. (2022), "Synthetic Data for Official Statistics. A Starter Guide", Nations unies, [ECECESSTAT20226.pdf](#).
- Xie, L., Lin, K., Wang, S., Wang, F., et Zhou, J. (2018), "Differentially Private Generative Adversarial Network", *arXiv preprint*, arXiv:1802.06739.
- Xu, L., Skoularidou, M., Cuesta-Infante, A., et Veeramachaneni, K. (2019), "Modeling tabular data using conditional gan", *Advances in neural information processing systems*, 32, p. 7303-7313.
- Yoon, J., Jarrett, D., et Van der Schaar, M. (2019), "Time-series generative adversarial networks", *Advances in neural information processing systems*, 32, p. 5485-5495.
- Yoon, J., Drumright, L. N., et Van Der Schaar, M. (2020), "Anonymization through data synthesis using generative adversarial networks (ads-gan)", *IEEE journal of biomedical and health informatics*, 24(8), p. 2378-2388.
- Zhang, K., Patki, N., et Veeramachaneni, K. (2022), "Sequential models in the SD vault", *arXiv preprint*, arXiv:2207.14406.