Traitement des données manquantes pour des capteurs de bâtiments connectés

Ahmed Es Sabar*[1], Marie-Lise Pannier[1], Alain Godon[2], David Bigaud[1]

[1] Univ Angers, LARIS, SFR MATHSTIC, F-49000 Angers, France
62 avenue Notre Dame du Lac, 49000 Angers
[2] Univ Angers, Département Systèmes Automatisés et Génie Informatique, Polytech Angers, F-49000 Angers, France
62 avenue Notre Dame du Lac, 49000 Angers
***ahmed.es-sabar@etud.univ-angers.fr**

RÉSUMÉ. Le traitement des données issues de bâtiments connectés doit permettre d'optimiser leur gestion énergétique, tout en garantissant un niveau de confort aux occupants. Les données collectées peuvent toutefois être incomplètes du fait de défaillances lors de l'acquisition des mesures. Cela compromet le traitement de l'information. Des méthodes, dites d'imputation, sont alors à appliquer pour consolider les données. Cet article propose un état de l'art sur les méthodes de gestion des données manquantes et d'évaluation de la qualité de l'imputation. Neuf méthodes d'imputation sont ensuite appliquées au cas de données d'ambiance d'un appartement T2, pour lequel le statut de présence de l'occupant est connu. Les méthodes sont comparées, d'une part en étudiant la qualité de l'imputation sur ces séries temporelles multivariées et d'autre part en évaluant la performance des méthodes sur la tâche finale, i.e. la classification du statut de présence. Il ressort de ces comparaisons que la performance de la tâche finale est peu affectée par la performance des méthodes d'imputation dans notre cas.

MOTS-CLÉS : Bâtiments connectés, Imputation de données, Occupation.

ABSTRACT. The processing the data from smart building is a way to both optimise their energy management and provide a high level of comfort to the occupants. Because of various possible failures in the data collection process, the information gathered can be incomplete or incorrect. In such case, so-called data imputation methods should be used in order to make possible the data processing. This article reviews methods to deal with missing data and to assess the performance of the imputation. Nine of these methods are applied to the data collected from an indoor environment-monitoring sensor located in an apartment for which the presence status of the occupant is known. The methods are compared based on the imputation quality of the multivariate time series as well as based on the performance of the final classification task, i.e. classifying the occupancy status. For this case study, it turns out that the performance imputation task has little impact on the performance of the final task.

KEYWORDS : Smart Building, Data Imputation, Occupancy.

1. INTRODUCTION

L'utilisation de dispositifs connectés est une piste de plus en plus souvent explorée pour optimiser la gestion énergétique des bâtiments tout en garantissant un niveau de confort aux occupants. Dans les bâtiments intelligents, un grand nombre de données est collecté à l'aide de capteurs connectés. Ces données mesurées sont transmises et analysées, par exemple à l'aide de méthodes d'apprentissage automatique, afin de mieux comprendre les habitudes des usagers (confort, gestion énergétique) et d'anticiper leurs actions adaptatives. Connaissant l'effet de l'occupation sur les consommations

énergétiques du bâtiment, des stratégies de régulations sont enfin définies et appliquées à l'aide d'actionneurs connectés agissant sur les systèmes. Des problèmes peuvent toutefois survenir lors de la collecte de données. En cas de panne de capteurs ou de défaillances lors de la transmission des mesures, les données peuvent être incomplètes ou aberrantes. Cela rend difficile, voire impossible, le traitement de l'information, en particulier lorsqu'il repose sur l'utilisation de méthodes d'apprentissage automatique ou profond (Hasan et al. 2021). Il convient alors de mettre en œuvre des méthodes d'imputation pour consolider les données préalablement à leur traitement final. Un ensemble de méthodes d'imputation est décrit dans la littérature et cet article présente tout d'abord une synthèse de ces travaux.

Dans le cadre du projet BIoT (*Building Internet of Things* – https://biot.u-angers.fr/), des capteurs connectés sont mis en place dans des logements et des salles de classe afin de détecter l'occupation, d'identifier des ouvertures de fenêtres et de prédire les consommations d'électricité. Dans ce travail, différentes méthodes d'imputation sont sélectionnées et appliquées. Leurs performances sur la tâche d'imputation et sur la tâche finale (classification de l'état de présence) sont ensuite comparées.

2. ÉTAT DE L'ART SUR LE TRAITEMENT DES DONNÉES MANQUANTES

Les méthodes d'imputation ont été employées pour consolider des données mesurées ou simulées dans de nombreux domaines : médical (Zhang 2016; Esteban et al. 2017; Che et al. 2018), génétique (de Souto et al. 2015), transport (Rafsunjani et al. 2019; Huang et al. 2021), réseaux (Osman et al. 2018; Ruggles et al. 2020). Le domaine du bâtiment ne fait pas exception. Des méthodes d'imputation y ont par exemple été appliquées pour traiter des données sur les ambiances thermique et lumineuse, et sur la consommation énergétique de systèmes (Chong et al. 2016; Pazhoohesh et al. 2019; Cho et al. 2020).

Généralement, les auteurs travaillant sur le traitement des données manquantes suivent un processus en trois étapes. Premièrement, la catégorie de données manquantes est déterminée. Deuxièmement, plusieurs méthodes d'imputation sont appliquées sur la base de données. Troisièmement, les méthodes sont comparées en calculant des indicateurs de performance. Ces trois étapes sont détaillées ci-après.

2.1. CLASSIFICATION DES DONNÉES MANQUANTES

Little et Rubin (1987) ont identifié trois mécanismes d'absence de données. Le mécanisme MCAR (pour *missing completely at random*) correspond au cas où la probabilité d'absence reste la même pour chaque observation. Dans le mécanisme MAR (pour *missing at random*), la probabilité d'absence pour une observation sur une variable dépend des valeurs d'autres variables. Enfin, pour le dernier mécanisme MNAR (pour *missing not at random*), la probabilité d'absence d'une observation dépend de la valeur de cette observation (e.g. en dehors de sa plage de mesure, un capteur ne renvoie pas de données).

Paradoxalement, les auteurs partent souvent d'une base de données complète, i.e. sans données manquantes, et en suppriment un pourcentage plus ou moins important (5 à 40 %), en appliquant l'un des trois mécanismes, pour obtenir une base incomplète (Chong et al. 2016; Pazhoohesh et al. 2019; Cho et al. 2020; Okafor et Delaney 2021). Cette approche permet, lors du calcul des indicateurs de performance, de comparer les écarts entre les versions complète et incomplète de la base.

2.2. MÉTHODES D'IMPUTATION DES DONNÉES

Une liste non exhaustive des méthodes d'imputation utilisées dans la littérature est disponible dans le Tableau 1. Le principe de fonctionnement de ces méthodes est décrit, entre autres, par Hasan et al. (2021) et Weerakody et al. (2021).

Classe de méthode	Méthode
Méthodes traditionnelles	Suppression des observations contenant des données manquantes Remplacement par des valeurs fixes : zéros ; moyenne, médiane ou mode ; dernière donnée observée (LOCF) ou prochaine donnée observée (NOCB) Remplacement par des valeurs aléatoirement échantillonnées dans [min ;max]
Méthodes statistiques et méthodes d'apprentissage automatique ML (*machine learning*)	Régression linéaire, polynomiale ou logistique Analyse en composantes principales Méthode de Monte-Carlo par Chaîne de Markov MCMC Modèle autorégressif et moyenne mobile ARMA ou ses extensions Imputation multiple par équations chaînées MICE ou par régression additive Méthode basée sur l'espérance-maximisation EM Décomposition en valeurs singulières SVD et factorisation matricielle MF Complétion de matrices MC Optimisation par algorithme génétique Arbre de décision DT ou forêt aléatoire RF Partitionnement de données (*clustering*) Séparateur à vaste marge SVM Méthode des K plus proches voisins KNN
Méthodes d'apprentissage profond DL (*deep learning*)	Réseau de neurones récurrents RNN avec des cellules LSTM ou GRU Réseau antagoniste génératif GAN avec des cellules LSTM ou GRU Auto-encodeur variationnel (VAE)

Tableau 1 : Type de méthodes d'imputation des données utilisées dans la littérature.

Dans les méthodes traditionnelles, les observations pour lesquelles la valeur d'une variable est absente sont, soit supprimées, soit remplacées par des valeurs fixes, c.a.d. identiques à chaque observation. Ces méthodes sont toutefois critiquées dans la littérature, car elles peuvent introduire un biais dans le traitement des données (Luo et al. 2018; Osman et al. 2018). Des méthodes d'imputation plus complexes, basées sur des approches statistiques, de ML et de DL ont alors été proposées.

Osman et al. (2018) proposent un logigramme pour aider à choisir le type de méthode à appliquer à un problème. Selon eux, le pourcentage de données manquantes et le mécanisme d'absence de données sont les principaux critères discriminants. Par ailleurs, un autre critère de choix concerne la capacité d'une méthode à traiter des séries temporelles multivariées, telles que rencontrées dans le bâtiment. Les modèles de type ARMA sont adaptés à ces données particulières. Une autre option pour traiter les séries temporelles consiste à ajouter en entrées des méthodes ML des variables décalées, c.à.d. des observations des pas de temps précédents. Les méthodes DL sont de plus en plus souvent utilisées et se révèlent particulièrement efficaces pour traiter les séries temporelles (Esteban et al. 2017; Che et al. 2018; Fouladgar et Främling 2020; Huang et al. 2021; Okafor et Delaney 2021; Weerakody et al. 2021).

Ces mêmes critères de choix des méthodes ont été identifiés dans le domaine du bâtiment. Pour les données MAR de Cho et al. (2020), l'interpolation linéaire était la méthode la plus efficace lorsque les séries temporelles contenaient moins de 8 valeurs consécutives manquantes, mais KNN fournissait de

meilleurs résultats jusqu'à 48 données consécutives manquantes. Chong et al. (2016) recommandent l'utilisation de régression linéaire ou de SVM (en cas de relations non linéaires) pour leurs données MAR. En complément, ils conseillent d'utiliser des variables décalées pour améliorer les performances des méthodes. Le paramétrage des méthodes dépend aussi du pourcentage de données manquantes selon Pazhoohesh et al. (2019), qui ont observé que le nombre de voisins de la méthode KNN doit augmenter avec la part de données manquantes pour conserver de bonnes performances pour leurs données MCAR.

2.3. ÉVALUATION DE LA PERFORMANCE DES MÉTHODES

Lorsque la base de données complète (ou vérité terrain) est disponible, des indicateurs de performance mesurant les écarts entre la vérité terrain et les valeurs imputées sont calculés. Les auteurs cherchent la méthode minimisant l'erreur absolue moyenne MAE, le carré moyen des erreurs MSE, l'erreur quadratique moyenne RMSE ou encore l'erreur quadratique moyenne normalisée NRMSE. La MSE et la RMSE sont préférées pour pénaliser les écarts importants. La normalisation est utile pour adimensionner les résultats. Notons que les auteurs ne mentionnent que rarement à quelles fins, c'est-à-dire pour quelle tâche finale, la comparaison de méthodes d'imputation est réalisée.

Dans le cas où la vérité terrain est inaccessible, la performance de l'imputation est évaluée sur la tâche finale (Weerakody et al. 2021). Les indicateurs cités plus hauts sont alors calculés pour des tâches de régression, tandis que l'exactitude, la précision, le rappel ou le F-score sont évalués pour des tâches de classification.

Nous nous interrogeons dans cet article d'une part sur le choix des méthodes d'imputation, et d'autre part sur les synergies entre les méthodes d'évaluation de la performance basées sur la tâche d'imputation et sur la tâche finale.

3. CAS D'ÉTUDE

Les données étudiées sont des séries temporelles multivariées enregistrées dans un appartement de type T2 situé à Angers, sur une période allant du 1er avril au 21 mai 2021. Un capteur de qualité de l'air intérieur Netatmo a mesuré la température intérieure, le taux d'humidité, la concentration en CO_2, le niveau de bruit et la pression. Les données ont été agrégées à un pas de temps de 5 min. Le jeu de données étant complet, plus de 14 600 observations sont disponibles. Parallèlement à ces mesures, l'occupant de l'appartement a complété toutes les 15 min un carnet d'activité lorsqu'il était présent. L'objet de l'étude est de prédire la présence de l'occupant dans l'appartement à partir des mesures.

4. MÉTHODOLOGIE

L'objectif est de déterminer la méthode d'imputation utilisée pour les données et de vérifier si la performance des méthodes reste la même lorsqu'on s'intéresse à la tâche d'imputation ou à la tâche finale. Pour cela, la méthodologie suivante a été appliquée (Figure 1).

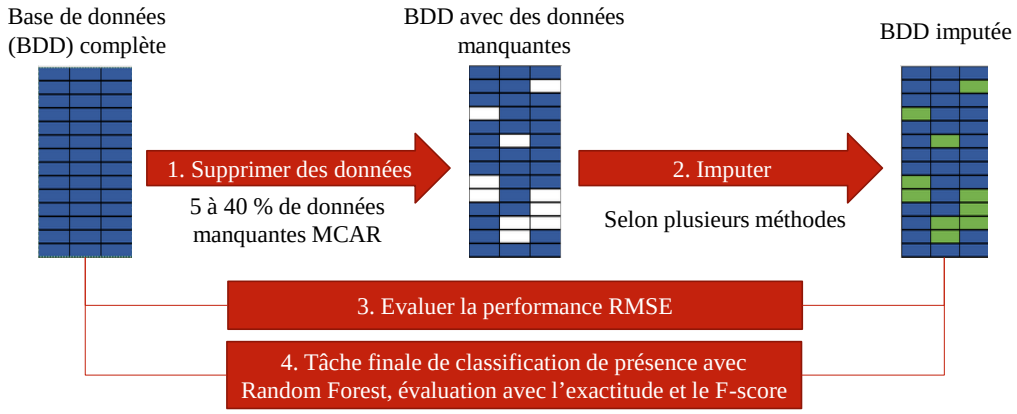

Figure 1 : Méthodologie suivie pour évaluer les performances des méthodes d'imputation.

Partant d'une base de données complète, la première étape a consisté à normaliser les données pour les rendre adimensionnelles (variations entre -1 et 1 pour chaque variable). Ensuite, un pourcentage de données a été supprimé selon le mécanisme MCAR. Ce mécanisme a été choisi, car les valeurs manquantes sont indépendantes des autres valeurs observées dans notre cas. Quatre bases de données avec respectivement 5, 10, 25 et 40 % de données manquantes sur les mesures sont construites. Pour chaque variable, il peut y avoir jusqu'à 11 valeurs consécutives manquantes, soit près d'une heure sans mesures. Nous ne considérons pas d'absence de données sur la variable cible (présence de l'occupant).

Dans une seconde étape, neuf méthodes d'imputation ont été testées : le remplacement par la moyenne (*Mean*), le remplacement par la dernière valeur connue (LOCF), par la prochaine valeur connue (NOCB), ou par une combinaison des deux méthodes précédentes (Nearest), l'interpolation linéaire (LI), quadratique (QI) et polynomiale d'ordre 3 (PI), le filtre de Kalman utilisant le modèle ARIMA (K. ARIMA), et enfin la méthode des plus proches voisins (KNN). La plupart de ces méthodes sont univariées, c.a.d. que l'imputation d'une variable se fait sans utiliser les valeurs des autres variables. Ces méthodes univariées sont bien adaptées à notre cas où les grandeurs mesurées sont peu corrélées entre elles (Es Sabar 2021). Seule KNN est capable de traiter le cas multivarié. Dans ces travaux préparatoires, aucune méthode de DL n'est employée. Le DL pourra être appliqué par la suite, en fonction des premiers résultats obtenus. Les méthodes utilisées sont disponibles dans les librairies Pandas et Scikit-Learn de Python, ou dans la librairie imputeTS de R.

La troisième étape consiste à évaluer la performance de la tâche d'imputation pour les différents pourcentages de valeurs manquantes. Pour cela, l'indicateur RMSE a été choisi. Il est calculé sur chaque variable ainsi que sur l'ensemble des données imputées (toutes variables confondues).

Enfin, lors de la quatrième étape, l'algorithme de forêt aléatoire (RF) avec 100 arbres a été appliqué pour la classification de l'état de présence (tâche finale) pour la base de données complète d'origine et pour chaque base de données imputée (avec les méthodes d'imputation suscitées et pour différents pourcentages de données manquantes). La création du modèle a été répétée 10 fois, en prenant les données du 1er avril au 10 mai (80 %) pour l'entrainement et les 20 % restants pour le test. Les performances de RF ont été évaluées en calculant la moyenne μ et l'écart-type σ de l'exactitude et du F-score sur les 10 répétitions. La RF avait montré de bonnes performances de classification dans une précédente étude sur les mêmes données (Es Sabar 2021). Pour améliorer la performance, des variables complémentaires ont été ajoutées : heure du jour, jour de la semaine et variables décalées sur trois pas de temps (pour les autres variables mesurées).

5. RÉSULTATS ET DISCUSSIONS

La méthodologie a été appliquée au cas d'étude. Les résultats sont présentés dans la suite uniquement pour 5 % et 40 % de données manquantes dans un souci de concision.

Les résultats sur la performance de la tâche d'imputation sont donnés dans la Figure 2. Les différences entre les méthodes d'imputation sont notables. L'imputation par la moyenne donne des valeurs de RSME bien plus importantes que les autres méthodes dans la plupart des cas. En revanche, la méthode du filtre de Kalman avec le modèle ARIMA (K. ARIMA) et la méthode d'interpolation linéaire (LI) donnent les meilleures performances : ce constat se vérifie pour chaque variable prise séparément ainsi que pour l'ensemble des variables. Selon la variable étudiée, les performances d'une même méthode varient. Si pour la température, l'humidité et la pression, les écarts entre les données réelles et les données imputées sont très faibles, ce n'est pas le cas pour le niveau de bruit et dans une moindre mesure pour la concentration en CO_2. Cela peut s'expliquer par une disparité dans la variabilité de la grandeur mesurée : la concentration en CO_2 et le niveau de bruit peuvent varier avec une amplitude forte d'une observation à la suivante (e.g. ouverture d'une fenêtre). De plus, le niveau de bruit se caractérise par une forte discontinuité, contrairement à toutes les autres grandeurs mesurées (e.g. allumage soudain d'un aspirateur). Enfin, la RMSE augmente mécaniquement avec le taux de données manquantes, ce qui est confirmé avec les taux de 10 % et 25 % non présentés ici.

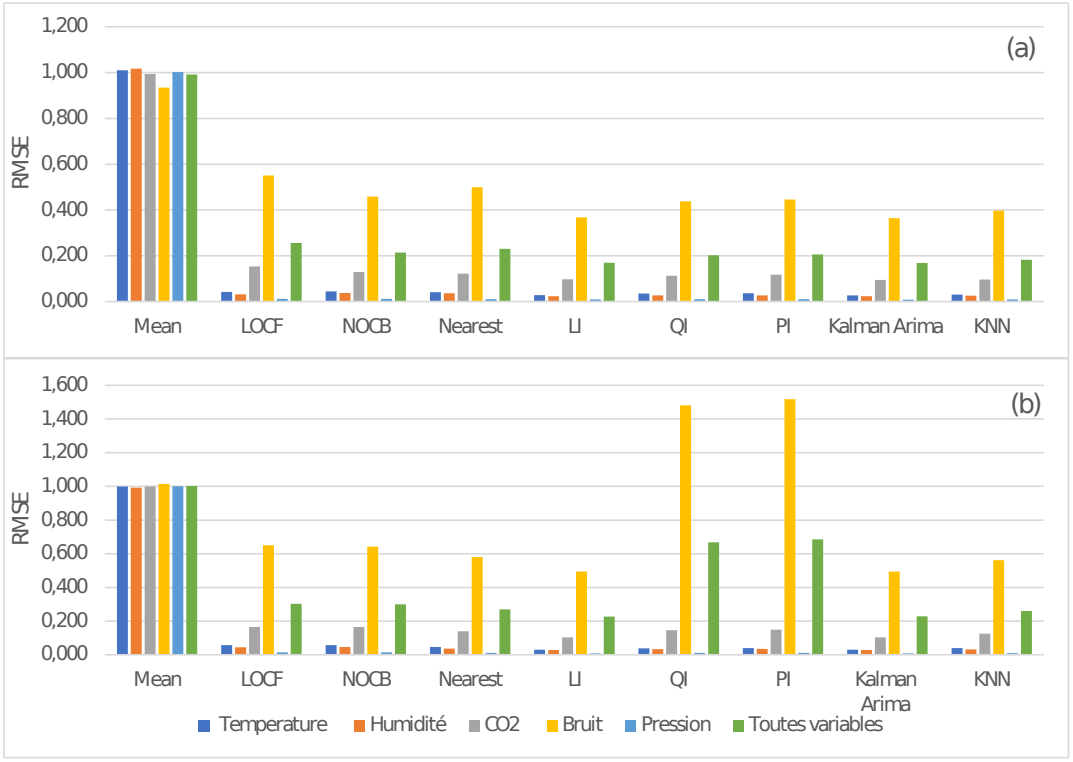

Figure 2 : RMSE pour la tâche d'imputation avec (a) 5 % et (b) 40 % de données manquantes.

Les performances de la tâche finale, c.a.d. de la classification entre présence et absence, sont données dans le Tableau 2 et le Tableau 3, pour 5 et 40 % de données manquantes respectivement. Il ressort de ces résultats que la performance de la tâche finale est très peu sensible à la méthode d'imputation utilisée et au pourcentage de données manquantes. En effet, il n'y a pas de différences significatives sur l'exactitude et le F-score calculés à partir des données d'origine et à partir des bases de données imputées. Il s'avère que l'erreur commise par les méthodes d'imputation n'impacte pas les

règles de décision du modèle dans notre cas. Notons que les variables les plus importantes pour la construction des arbres sont la concentration en CO_2 et la température sur les trois derniers pas de temps et l'heure du jour. Les résultats montrent toutefois une performance légèrement meilleure pour la base de données imputée par la moyenne. Cela peut s'expliquer par une fuite de données : la valeur moyenne imputée (moyenne de toutes les valeurs connues pour une variable) contient aussi une partie de l'information sur les données de test du problème de classification.

		Base complète	Bases de données imputées								
			Mean	LOCF	NOCB	Nearest	LI	QI	PI	K. ARIMA	KNN
Exactitude	**μ**	0,909	0,919	0,913	0,911	0,913	0,913	0,911	0,913	0,909	0,909
	σ	0,004	0,003	0,003	0,005	0,005	0,005	0,003	0,005	0,007	0,006
F-score	**μ**	0,896	0,905	0,898	0,897	0,898	0,896	0,896	0,896	0,894	0,894
	σ	0,005	0,004	0,003	0,007	0,006	0,005	0,003	0,005	0,009	0,008

Tableau 2 : Performance de la tâche de classification, pour 5 % de données manquantes.

		Base complète	Bases de données imputées								
			Mean	LOCF	NOCB	Nearest	LI	QI	PI	K. ARIMA	KNN
Exactitude	**μ**	0,909	0,927	0,902	0,909	0,912	0,910	0,910	0,910	0,915	0,906
	σ	0,004	0,001	0,003	0,003	0,003	0,006	0,005	0,006	0,003	0,007
F-score	**μ**	0,896	0,914	0,886	0,889	0,896	0,897	0,895	0,897	0,903	0,890
	σ	0,005	0,002	0,005	0,004	0,004	0,008	0,006	0,008	0,005	0,008

Tableau 3 : Performance de la tâche de classification, pour 40 % de données manquantes.

Au regard des résultats obtenus sur ce cas d'étude, il est recommandé de ne pas se focaliser uniquement sur les performances de la tâche d'imputation. Il n'y a pas forcément de corrélation entre les performances d'imputation et les performances de la tâche finale. Cependant, ces conclusions sont à confirmer en travaillant sur d'autres jeux de données et d'autres objectifs d'étude (classification multi-classe, régression, partitionnement de données ou encore prévision).

Dans l'objectif d'évaluer les effets potentiels des fuites des données générées par les méthodes d'imputation, il serait intéressant de procéder à l'imputation en deux temps : sur les données d'entrainement et validation de la tâche finale d'une part, et sur les données de test d'autre part. Par ailleurs, d'autres métriques de performances spécifiques aux séries temporelles, telles que la déformation temporelle dynamique (DTW), pourraient être utilisées.

6. PERSPECTIVES

Ce travail constitue un préalable à une étude plus complète menée dans le cadre de l'instrumentation des salles de classe de Polytech Angers. Des cartes multicapteurs (Figure 3) mesurant la température, le taux d'humidité, la concentration en CO_2 et en COV, la luminosité et le niveau de bruit sont installées dans deux salles. Les ouvertures de fenêtre et les consommations électriques des salles sont enregistrées en complément. Les objectifs sont d'évaluer le placement optimal des capteurs pour détecter l'occupation et identifier des ouvertures de fenêtres ainsi que de prédire les consommations d'électricité. Dans ce contexte, les données manquantes peuvent être liées à des pannes électriques ou

des défaillances de capteurs.

Figure 3 : Carte multicapteurs développée à Polytech Angers.

7. CONCLUSION

Cet article s'est intéressé à la problématique des données manquantes issues de capteurs de bâtiments connectés en présentant un état de l'art sur les méthodes d'imputation, et un cas d'étude dans lequel neuf méthodes sont comparées. Dans la littérature, les auteurs s'intéressent aux performances de l'imputation, sans mentionner pour quelle tâche finale les données sont consolidées. Cependant, pour nos données d'ambiance mesurées dans un T2, la performance de la tâche finale de classification de la présence s'est avérée peu sensible à la méthode d'imputation. Bien que ces premiers résultats restent à confirmer sur d'autres cas d'étude, il semble recommandé de ne pas se focaliser sur les performances de l'imputation. La connaissance de l'occupation acquise par le traitement de données de capteurs sera utilisée dans une perspective d'optimisation de la consommation des bâtiments connectés centrée sur l'usager.

8. REMERCIEMENTS

Ce travail a été réalisé dans le cadre du projet BIoT (*Building Internet of Things*) du programme RFI Wise (Recherche Formation Innovation en électronique et systèmes intelligents), des Pays de la Loire.

9. BIBLIOGRAPHIE

Che, Zhengping, Sanjay Purushotham, Kyunghyun Cho, David Sontag, et Yan Liu. 2018. « Recurrent Neural Networks for Multivariate Time Series with Missing Values ». *Scientific Reports* 8 (1): 6085. https://doi.org/10.1038/s41598-018-24271-9.

Cho, Brian, Teresa Dayrit, Yuan Gao, Zhe Wang, Tianzhen Hong, Alex Sim, et Kesheng Wu. 2020. « Effective Missing Value Imputation Methods for Building Monitoring Data ». In *2020 IEEE International Conference on Big Data (Big Data)*, 2866-75. https://doi.org/10.1109/BigData50022.2020.9378230.

Chong, Adrian, Khee Poh Lam, Weili Xu, Omer T. Karaguzel, et Yungjeong Mo. 2016. « Imputation of Missing Values in Building Sensor Data ». In *IBPSA-USA SimBuild 2016*, 8. Salt Lake City.

Es Sabar, Ahmed. 2021. « Evaluation de l'occupation dans le bâtiment à travers des algorithmes d'apprentissage automatique. Rapport de stage ». Travail de fin d'études, diplôme d'ingénieur de l'ENTPE. LARIS. https://acces.entpe.fr/public/tfe/2021/2021_09_10_ES-SABAR_Ahmed_RESUME_TFE_2021.pdf.

Esteban, Cristóbal, Stephanie L. Hyland, et Gunnar Rätsch. 2017. « Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs ». *arXiv:1706.02633 [cs, stat]*, décembre. http://arxiv.org/abs/1706.02633.

Fouladgar, Nazanin, et Kary Främling. 2020. « A Novel LSTM for Multivariate Time Series with Massive Missingness ». *Sensors* 20 (10): 2832. https://doi.org/10.3390/s20102832.

Hasan, Md. Kamrul, Md. Ashraful Alam, Shidhartho Roy, Aishwariya Dutta, Md. Tasnim Jawad, et Sunanda Das. 2021. « Missing Value Imputation Affects the Performance of Machine Learning: A Review and Analysis of the Literature (2010–2021) ». *Informatics in Medicine Unlocked* 27 (janvier): 100799. https://doi.org/10.1016/j.imu.2021.100799.

Huang, Tongge, Pranamesh Chakraborty, et Anuj Sharma. 2021. « Deep Convolutional Generative Adversarial Networks for Traffic Data Imputation Encoding Time Series as Images ». *International Journal of Transportation Science and Technology*, novembre. https://doi.org/10.1016/j.ijtst.2021.10.007.

Little, Roderick J. A., et Donald B. Rubin. 1987. *Statistical Analysis With Missing Data*. Wiley.

Luo, Yonghong, Xiangrui Cai, Ying ZHANG, Jun Xu, et Yuan Xiaojie. 2018. « Multivariate Time Series Imputation with Generative Adversarial Networks ». In *Advances in Neural Information Processing Systems*, 31:12. Curran Associates, Inc. https://papers.nips.cc/paper/2018/hash/96b9bff013acedfb1d140579e2fbeb63-Abstract.html.

Okafor, Nwamaka U., et Declan T. Delaney. 2021. « Missing Data Imputation on IoT Sensor Networks: Implications for on-Site Sensor Calibration ». *IEEE Sensors Journal* 21 (20): 22833-45. https://doi.org/10.1109/JSEN.2021.3105442.

Osman, Muhammad S., Adnan M. Abu-Mahfouz, et Philip R. Page. 2018. « A Survey on Data Imputation Techniques: Water Distribution System as a Use Case ». *IEEE Access* 6: 63279-91. https://doi.org/10.1109/ACCESS.2018.2877269.

Pazhoohesh, Mehdi, Zoya Pourmirza, et Sara Walker. 2019. « A Comparison of Methods for Missing Data Treatment in Building Sensor Data ». In *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, 255-59. https://doi.org/10.1109/SEGE.2019.8859963.

Rafsunjani, Siam, Rifat Sultana Safa, Abdullah Al Imran, Shamsur Rahim, et Dip Nandi. 2019. « An Empirical Comparison of Missing Value Imputation Techniques on APS Failure Prediction ». *International Journal of Information Technology and Computer Science(IJITCS)* 11 (2): p21-29. https://doi.org/10.5815/ijitcs.2019.02.03.

Ruggles, Tyler H., David J. Farnham, Dan Tong, et Ken Caldeira. 2020. « Developing Reliable Hourly Electricity Demand Data through Screening and Imputation ». *Scientific Data* 7 (1): 155. https://doi.org/10.1038/s41597-020-0483-x.

Souto, Marcilio CP de, Pablo A. Jaskowiak, et Ivan G. Costa. 2015. « Impact of missing data imputation methods on gene expression clustering and classification ». *BMC Bioinformatics* 16 (1): 64. https://doi.org/10.1186/s12859-015-0494-3.

Weerakody, Philip B., Kok Wai Wong, Guanjin Wang, et Wendell Ela. 2021. « A Review of Irregular Time Series Data Handling with Gated Recurrent Neural Networks ». *Neurocomputing* 441 (juin): 161-78. https://doi.org/10.1016/j.neucom.2021.02.046.

Zhang, Zhongheng. 2016. « Missing Data Imputation: Focusing on Single Imputation ». *Annals of Translational Medicine* 4 (1): 9-9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

Rappel du résumé soumis pour cet article

Les bâtiments sont de plus en plus souvent munis de capteurs connectés. Disposant d'un grand nombre et type de données mesurées dans un bâtiment, des méthodes d'apprentissage automatique peuvent être utilisées pour mieux comprendre l'occupation et anticiper des actions adaptatives des occupants. Ainsi, les bâtiments connectés doivent permettre une gestion énergétique optimisée, tout en garantissant un niveau de confort aux occupants. Toutefois, les capteurs connectés peuvent présenter des défaillances (données collectées aberrantes ou absences de données) qu'il convient de diagnostiquer et de corriger pour obtenir des modèles d'occupation plus fiables.

Dans le cadre du projet RFI Wise BIoT, un état de l'art sur les méthodes d'intelligence artificielle permettant de traiter les données manquantes est réalisé. Ces méthodes doivent s'adapter, dans la mesure du possible, au cas des séries temporelles multivariées. Il ressort de cette analyse de la littérature que, parmi les nombreuses méthodes employées, les interpolations (linéaires ou polynomiales) sont efficaces lorsqu'il s'agit de remplacer les quelques valeurs des pas de temps indisponibles par des valeurs réalistes. Dès lors qu'un grand nombre de données consécutives manquent, la méthode des k plus proches voisins (KNN) ou l'algorithme *missForest* basée sur des forêts aléatoires deviennent néanmoins plus performants. Le choix de la méthode et des valeurs de ses hyperparamètres dépend ainsi de la quantité de données manquantes. Par ailleurs, les réseaux de neurones récurrents à portes (GRU) sont pertinents lorsque les séries temporelles ne sont pas régulières, c'est-à-dire lorsque le pas de temps de mesures n'est pas contant d'une variable à une autre ou pour une même variable.

Particulièrement appropriées pour les séries temporelles multivariées, les méthodes *missForest* et GRU sont mises en œuvre sur des données collectées dans deux salles de classes fortement instrumentées de Polytech Angers. Ces salles sont équipées d'un ensemble de capteurs de température, d'humidité, de CO_2, de COV, de luminosité et de bruit, réparti dans la pièce, ainsi que de capteurs d'ouverture de fenêtres et de puissance électrique appelée. En parallèle, une station météorologique mesure les conditions extérieures. Une partie des informations collectées est retirée du jeu de données, selon des règles prédéfinies, afin de représenter les données manquantes. Différentes tailles d'échantillon manquant sont ainsi testées. Les performances des méthodes à s'approcher des données réelles (préalablement retirées) sont évaluées en utilisant l'erreur quadratique moyenne et la déformation temporelle dynamique.

Une fois les données manquantes efficacement remplacées par des valeurs réalistes, les séries temporelles seront analysées dans la suite du projet afin de mieux connaître et de prédire : les actions des occupants (présence et ouverture de fenêtres) ; les consommations énergétiques ; et les conditions les plus confortables pour les usagers. Ces informations sont utiles dans une perspective d'optimisation de la consommation énergétique des bâtiments connectés centrée sur l'usager.