Master Systèmes Dynamiques et Signaux

Mémoire de master

Prétraitement des données mesurées et traitement des données manquantes pour des capteurs de bâtiments connectés

---

Auteur :
M. Ahmed ES-SABAR

Jury :
Pr. Anne HEURTIER
Dr. Marie-Lise PANNIER
Dr. Mirvana HILAL
Dr. Remy GUYONNEAU
Dr. Sébastien LAGRANGE

Président :
Pr. Laurent. HARDOUIN

Version du
30 juin 2022

Table des matières

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.

Je tiens ici à remercier l'ensemble des personnes qui m'ont permis de mener à bien mon travail au sein du laboratoire LARIS de l'université d'ANGERS.

En premier lieu, je remercie monsieur David BIGAUD, directeur du laboratoire de m'avoir accueilli dans ce dernier.

Je remercie infiniment madame Marie-Lise PANNIER, mon encadrante de stage, pour sa disponibilité, son encadrement et ses conseils tout au long de cette période.

Je souhaite également adresser mes remerciements à monsieur Laurent HARDOUIN, président du jury et à tous les membres du jury pour la correction et l'évaluation de ce travail.

Enfin, je remercie l'ensemble des personnes avec lesquelles j'ai échangé au sein du laboratoire et tout particulièrement monsieur Alain GODON, enseignant-chercheur à Polytech.

Liste des illustrations et des tableaux

Chapitre 1. Présentation du sujet d'étude

1.1 Vers une gestion intelligente des bâtiments connectés

Le bâtiment doit aujourd'hui maîtriser son impact environnemental : réduire la consommation énergétique, limiter la pollution, réduire l'empreinte carbone. Il est aussi important de veiller à la qualité de l'air intérieur et les bonnes conditions de confort de ces occupants. +Or, en France, le secteur du bâtiment est le plus gros consommateur d'énergie. Il pèse 70 millions de tonnes d'équivalent pétrole, soit 43% de l'énergie finale totale [1]. La politique de réduction des consommations énergétiques et des émissions de gaz à effet de serre constitue une priorité du Grenelle de l'Environnement.

La France s'est dotée entre autres de réglementations pour la construction de bâtiments neufs, ainsi que pour les travaux de rénovation. Ces dispositifs visent à améliorer les performances des bâtiments en agissant sur l'enveloppe de celui-ci (isolation) et sur ses systèmes (chauffage, éclairage, eau chaude, système de refroidissement, ventilation).

Cependant, la manière d'occuper le bâtiment a un impact considérable sur la consommation énergétique, sur la qualité de l'air intérieur et sur le confort de ces occupants. En effet, l'occupation (nombre d'occupants présents, température de consigne demandée, actions de l'occupant sur le bâtiment, usage d'appareils...) est l'un des facteurs les plus importants ayant un impact considérable sur la consommation énergétique et sur le confort dans le bâtiment. Une bonne connaissance de l'occupation est un prérequis pour améliorer la gestion énergétique des bâtiments et éviter des consommations inutiles. En effet, il ne suffit pas de concevoir un bâtiment performant pour assurer une bonne performance énergétique. Des recherches scientifiques ont montré que le comportement des usagers (ouverture de fenêtres et des stores, actions sur les systèmes de chauffage ou de climatisation) du bâtiment est responsable d'un écart substantiel de consommation énergétique entre des bâtiments identiques [2] [3].

Il est donc primordial d'évaluer et de prédire l'occupation pour améliorer la performance énergétique des bâtiments, tout en satisfaisant au confort des usagers. Pour ce faire, nous envisageons d'explorer des méthodes d'apprentissage automatique (*Machine Learning* ML) en exploitant des données mesurées dans un appartement de type T2 et dans deux salles de classe instrumentées à Polytech Angers.

Par ailleurs, l'apprentissage automatique exige des quantités de données importantes et de bonnes qualité. Or, les données collectées peuvent être incomplètes, mal formatées, sujettes à des erreurs, présentes en faible quantité ou nécessitent un prétraitement particulier.

Mon projet de stage vise à réaliser un prétraitement des données mesurées et traiter les données manquantes sur deux cas d'étude : un appartement T2 connecté et deux salles de classe connectées de l'école Polytech d'Angers.

1.2 Cas d'étude

1.2.1 Un appartement T2 connecté

Dans ce cas d'étude, les données traitées 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[1], placé dans la pièce de vie, 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. Il s'agit donc d'une tâche de classification.

Figure 1: Capteur de qualité de l'air intérieur Netatmo placé dans le T2

1.2.2 Des salles de classe connectées à Polytech

Dans le cadre du projet RFI Wise BIoT, https://biot.u-angers.fr/, deux salles de l'école Polytech Angers ont été fortement instrumentées pour étudier à la fois l'optimisation de l'emplacement des capteurs connectés ainsi que l'évaluation de l'occupation dans le bâtiment à travers des approches d'intelligence artificielle.

Ces salles sont équipées d'un ensemble de cartes multicapteurs (Figure 2), répartis dans la pièce, ainsi que de capteurs d'ouverture de fenêtres et de puissance électrique appelée. En plus, une station météo a été installée sur le toit de l'école.

Figure 2 Carte électronique multicapteurs

Dans la salle 219, 14 cartes électroniques multicapteurs ont été placées selon un maillage comme indiqué sur la Figure 3, à une hauteur de 2,5 m, dans le but de déterminer les meilleurs emplacements vis-à-vis d'une tâche finale. Chaque carte permet de mesurer la température, l'humidité, le CO_2, le COV (les composés organiques volatiles), la luminosité et le bruit.

L'ensemble des 14 cartes partage la même source d'alimentation électrique. De plus, ce réseau de multicapteurs est constitué de petits groupes (pour éviter la perte de données transmises par les cartes) de 3 à 4 cartes électroniques reliées entre elles d'une manière filaire pour transmettre les données mesurées par les capteurs. Une des cartes de chaque groupe joue le rôle de maître (interroge les cartes de son groupe pour récupérer les mesures). Chaque maître renvoie les données récupérées au serveur via une connexion sans fil (Wi-Fi). Les données sont sauvegardées dans une base de données avec celles des capteurs de fenêtres qui communiquent

[1] Capteur de Qualité de l'Air Intérieur Intelligent | Netatmo

également en réseau sans fil. Deux capteurs permettent de mesurer la puissance générale appelée par la salle et la puissance de l'éclairage. Leurs données sont également sauvegardées dans une autre base de données (communication Wi-Fi). La station météo permet de mesurer plusieurs grandeurs physiques, dont la température extérieure, l'humidité, la pression et le rayonnement. On se retrouve donc avec un total de 93 capteurs et plus de 22000 observations sur la période d'étude allant du 15 mars au 31 mai 2022.

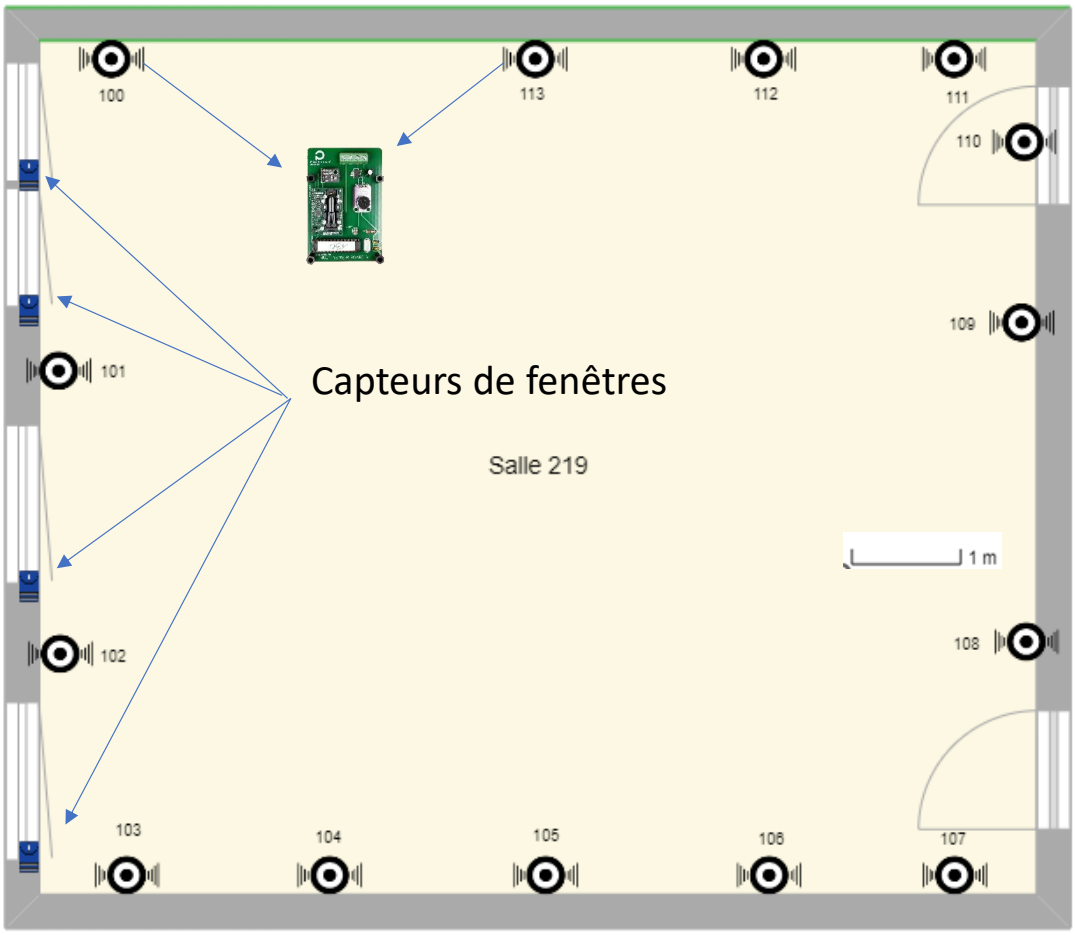


Figure 3 Emplacement des capteurs de la salle 219

Dans la salle 114, nous avons la même configuration avec 12 cartes : 3 capteurs de fenêtres et 4 capteurs mesurant la puissance électrique (général, vidéoprojecteur, prises et l'éclairage).

J'ai participé avec l'équipe projet à des opérations de confection de capteurs (soudage des capteurs sur ces cartes électroniques, test de fonctionnement, calibrage...) et à l'instrumentation des salles.

1.3 Vers des modèles basés sur les données

Traditionnellement, les bâtiments ont été modélisés avec des modèles basés sur la physique, on les appelle des modèles à boîte blanche, de haute dimension tels que les modèles résistance-capacité (RC) [1], TRNSYS [2] et EnergyPlus [3]. Ces modèles reposent sur la thermique du bâtiment et modélisent le transfert de chaleur entre les composants des bâtiments. Ils exigent la connaissance des paramètres détaillés du bâtiment et des équations de bilan thermique. Il s'agit du concept du problème direct qui décrit comment les paramètres du modèle se traduisent en effets observables. L'avantage de ces modèles est leur capacité à modéliser la température

avec une granularité fine, mais ils ont un inconvénient qui est leur dimensionnalité élevée qui les rend coûteux en calcul. Ainsi, il est difficile d'agir en temps réel sur les systèmes du bâtiment.

Ces modèles sont essentiellement utilisés en phase de conception pour modéliser le bâtiment et ainsi dimensionner ses systèmes comme le chauffage, la ventilation et la climatisation (CVC). Cependant, si on s'intéresse à agir sur ces systèmes en phase d'exploitation pour des fins d'optimisation, ces modèles sont inadaptés pour fournir des informations utiles à l'optimisation énergétique centrée sur l'usage. En effet, ces informations sont les entrées de ces modèles physiques. En d'autres termes, nous sommes en présence d'une approche de problème inverse, qui consiste à tenter de déterminer les causes d'un phénomène à partir des observations de ses effets.

D'autre part, les chercheurs tentent d'identifier des modèles de dimension inférieure, basés sur des données, on les appelle des modèles à boîte noire. Contrairement aux modèles physiques, ces modèles ne nécessitent pas de construire des équations d'équilibre thermique ; par conséquent, nul besoin de disposer d'informations physiques sur le bâtiment. Ils sont basés sur des relevés historiques pour déduire la relation cachée entre des variables explicatives et une variable cible (sans trop se soucier des processus sous-jacents). Les variables explicatives correspondent par exemple à la météo, aux données d'ambiance ou encore aux actions des occupants. La variable cible peut être la consommation d'énergie du bâtiment, le niveau d'occupation, ou l'état des fenêtres.

C'est, pour cette raison, que nous considérons le *Machine Learning* comme alternative à ces modèles physiques. En effet, le ML, repose sur des travaux qui ont été réalisés en mathématiques et en physique tels que le problème inverse, les statistiques et la théorie de l'information.

Les modèles pilotés par les données ont l'avantage d'être construits rapidement et évalués en termes d'erreur : une grande quantité de données est utilisée pour construire le modèle (jeu d'entrainement) et une autre partie des données est utilisée pour l'évaluer (jeu de test).

En somme, les modèles basés sur les données sont bien appropriés pour les bâtiments dont les paramètres physiques (composition des parois) sont inaccessibles ou les conditions d'occupations sont incertaines. En plus, ces modèles axés sur les données permettent aisément d'évaluer ou de prédire des caractéristiques importantes pour optimiser les performances énergétiques et environnementales en étant centré sur l'usager. Parmi ces caractéristiques, on peut retrouver : la consommation énergétique, les conditions de confort (thermique, qualité d'air intérieur) et des variables stochastiques comme l'occupation et les actions de l'occupant.

Exemple simple de modèle piloté par les données :

Un modèle de régression linéaire est un modèle piloté par les données qui établit une relation entre une variable cible et un ensemble de variables explicatives. Ce modèle de régression peut ensuite être utilisé pour comprendre la relation entre les variables et leurs impacts sur la variable cible. Dans d'autres cas, il peut également être utilisé pour réaliser des prédictions.

Chapitre 2. Étude bibliographique et objectifs du stage

2.1 Rappel de l'étude bibliographique

Le premier semestre a été consacré à la réalisation d'une étude bibliographique sur le traitement des données manquantes. Les paragraphes suivants résument cette étude.

Les méthodes d'imputation ont été employées pour consolider des données mesurées ou simulées dans de nombreux domaines : médical [4]–[6], génétique [7], transport [8], [9], réseaux [10]-[11]. 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 [12]–[14].

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.

2.1.1 Classification des données manquantes

Little et Rubin [15] 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 [12]–[14], [16]. Cette approche permet, lors du calcul des indicateurs de performance, de comparer les écarts entre les versions complète et incomplète de la base.

2.1.2 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. [17] et Weerakody et al. [18].

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 ML	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.-à-d. identiques à chaque observation. Ces méthodes sont toutefois critiquées dans la littérature, car elles peuvent introduire un biais dans le traitement des données [10], [19]. Des méthodes d'imputation plus complexes, basées sur des approches statistiques, de ML et de DL ont alors été proposées.

Osman et al. [10] 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 [5], [6], [9], [16], [18], [20]

Ces mêmes critères de choix des méthodes ont été identifiés dans le domaine du bâtiment. Pour les données MAR [14], 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. [12] 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. [13], 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.

2.1.3 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. 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.

2.1.4 Bilan

Concernant le traitement des données manquantes dans les problématiques du bâtiment, nous avons identifié deux pistes de recherche :

- d'une part, nous nous interrogeons sur le choix des méthodes d'imputation ;
- et d'autre part, nous questionnons les synergies entre les méthodes d'évaluation de la performance basées sur la tâche d'imputation et sur la tâche finale.

2.2 Les objectifs du stage

L'objectif de mon stage était de sélectionner des méthodes d'imputation, et de comparer et d'analyser leurs performances sur les données collectées dans les deux salles de Polytech Angers et dans un appartement de type T2. Pour ce faire, des opérations supplémentaires ont été identifiées comme des objectifs intermédiaires à réaliser. Il s'agit de :

- Développer des outils permettant d'automatiser des opérations de prétraitement de données.
- Réaliser des opérations d'exploration de données (représentation graphique, corrélations entre les variables disponibles)
- Identifier des algorithmes d'apprentissage automatique pour former des modèles de tâches finales (classification et régression). Ceci pour pouvoir apprécier les performances des méthodes d'imputation au regard de la tâche finale.
- Réaliser de la sélection des variables vis-à-vis d'une tâche finale : réduction de la dimensionnalité.
- Enrichir les données des séries temporelles pour la tâche finale : extraction des caractéristiques.
- Mettre en place la méthodologie de traitement de données manquantes identifiée lors de la recherche bibliographique (générer ou simuler des données manquantes selon le mécanisme MCAR et MAR).
- Rédiger un article de conférence.

Chapitre 3. Travaux réalisés et résultats

3.1 Méthodologie

Tous les projets de ML suivent un processus similaire, décrit au § 3.1.1. Pour répondre aux objectifs spécifiques de notre projet, la méthodologie présentée au § 3.1.2 est suivie. Cette méthodologie est présentée en détail dans les paragraphes suivants. Les étapes préliminaires de préparation de données (fusion et visualisation ...) ne sont développées dans ce manuscrit.

3.1.1 Comment marche un projet ML ?

Cette partie décrit les différentes étapes d'un projet d'apprentissage automatique. Il existe des étapes standard à suivre pour accomplir un projet de science des données. Tout d'abord, il faut collecter les données en fonction du problème à traiter. L'étape suivante consiste à réaliser des opérations de prétraitement de données. La dernière étape consiste à construire, évaluer et déployer un modèle d'apprentissage automatique.

Dans ce projet, nous avons travaillé sur tous ces étapes, mais, nous avons plus focalisé nos efforts sur le prétraitement des données et notamment sur le traitement des données manquantes.

D'ailleurs, le prétraitement des données dans l'apprentissage automatique est une étape incontournable qui contribue à améliorer la qualité des données afin de pouvoir comprendre et capturer l'information utile à partir des données. Il fait référence à des opérations de nettoyage et d'organisation des données brutes pour les rendre adaptées à la construction d'un modèle d'apprentissage automatique. En d'autres termes, ces opérations consistent à transformer les données brutes en un format compréhensible et lisible.

Pour ce projet de stage, il a été nécessaire de réaliser plusieurs fonctions et algorithmes sous Python pour permettre de répondre aux objectifs intermédiaires cités précédemment. Ce sont des opérations chronophages et sont souvent propres à chaque projet. Elles nécessitent un investissement important en temps (environ 8 semaines) et en développement d'outils adaptés à chaque projet. Les fonctionnalités développées sont disponibles sur le dépôt suivant : https://github.com/Ah-essabar/LARIS01.git. Les scripts utilisant ces fonctionnalités pour traiter les données manquantes et prédire les variables cibles sont regroupés au dépôt suivant (en cours d'organisation) : https://github.com/Ah-essabar/Master_SDS_Angers.git.

La Figure 4 synthétise ce processus d'élaboration d'un modèle basé sur les données.

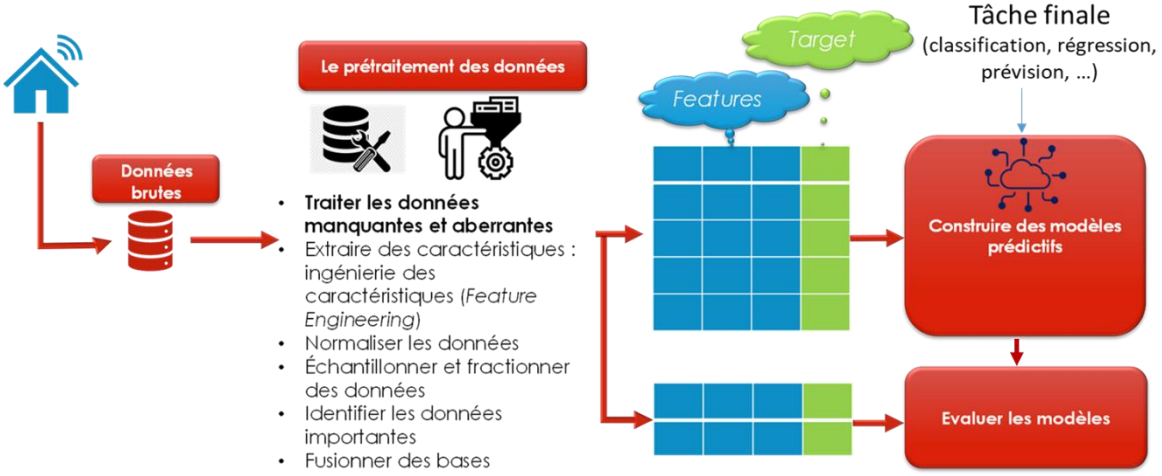


Figure 4 Processus d'élaboration d'un modèle basé sur les données

3.1.2 Méthodologie globale

Dans cette partie, nous allons présenter la méthodologie appliquée pour comparer des méthodes d'imputation pour les cas d'étude 1 (appartement T2) et 2 (classe Polytech). Celle-ci est similaire pour ces deux cas d'étude. Les variantes indiquées en gras et * concernent uniquement le cas d'étude 2 (voir partie sélection des variables, et étapes 1 et 2 de la Figure *5*).

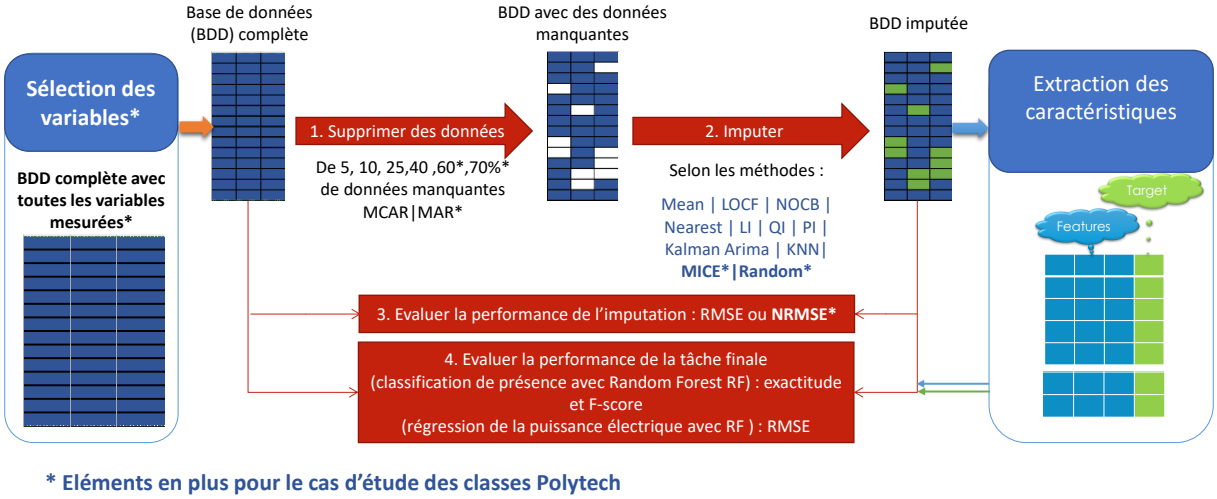


Figure 5 : Méthodologie suivie pour évaluer les performances des méthodes d'imputation.

3.1.3 Prétraitement - La sélection des variables

L'opération de sélection des variables est utilisée uniquement sur le cas d'étude 2 (salle de classe de Polytech). Nous disposons pour chaque classe de Polytech, d'une centaine de variables de grandeurs physiques différentes. Parmi les objectifs du projet figure l'optimisation de l'emplacement des capteurs. Nous avons proposé la méthode de la sélection des variables comme technique pour répondre à cet objectif. Cette étape n'est pas nécessaire pour le cas d'étude 1 (appartement T2) du fait du faible nombre de variables et de la présence d'un seul appareil de mesure.

Cette opération est devenue l'objet de nombreuses recherches et particulièrement dans les domaines d'applications où de nombreuses variables sont disponibles (des dizaines à des centaines de milliers). Lorsque le nombre de variables augmente, le traitement de données fait

face à un phénomène connu sous le terme « le fléau de la dimension ou la malédiction de la dimension ». Ce terme a été introduit par le mathématicien Richard Bellman en 1961. Il désigne divers problèmes qui ont lieu seulement dans les espaces de grande dimension.

Par ailleurs, Verleysen et François [1] précisent que les espaces de haute dimension présentent des propriétés géométriques surprenantes et contre-intuitives qui exercent une grande influence sur les performances des outils d'analyse de données. Ils considèrent que les normes euclidiennes et les noyaux gaussiens, qui sont couramment utilisés dans les modèles d'apprentissage automatique, deviennent inappropriés dans les espaces de haute dimension à cause « de la concentration du phénomène normatif ».

Parmi les solutions à ce phénomène, on trouve les méthodes de réduction de la dimension. Dans notre projet, la réduction de la dimension, par l'élimination de certaines variables, présente un intérêt supplémentaire dans l'optimisation de l'emplacement des capteurs : Réduire le nombre de capteurs revient à identifier indirectement l'emplacement des capteurs qui maximisent les performances pour les tâches finales (prédiction de la puissance électrique appelée, évaluation du nombre d'occupants dans le bâtiment ou la détection des ouvertures des fenêtres).

Parmi l'ensemble des variables disponibles, il est clair qu'il peut y avoir des données utiles, inutiles ou redondantes. Donc, il est nécessaire de distinguer les caractéristiques pertinentes pour construire des modèles prédictifs robustes. Pour ce faire, des techniques de sélection de variables sont employées. Elles visent à maximiser la pertinence et à minimiser la redondance pour obtenir un petit sous-ensemble en éliminant les variables non essentielles [21]. Par conséquent, les modèles d'apprentissage deviennent plus robustes et permettent de gagner en temps de calcul.

Dans la littérature, on distingue deux principales approches : enveloppe (*Wrapper*) et filtre :

- Une approche de filtrage consiste à sélectionner les variables en fonction de mesures statistiques. Elle est indépendante de l'algorithme d'apprentissage et nécessite moins de temps de calcul. Le gain d'information, le test du khi-deux, le score de Fisher, le coefficient de corrélation et le seuil de variance sont quelques-unes des mesures statistiques utilisées pour classer l'importance des variables.
- Une approche *Wrapper* consiste à rechercher, évaluer et comparer des combinaisons de variables pour former un modèle d'apprentissage donné. En effet, contrairement aux approches de filtrage, un modèle prédictif est nécessaire pour évaluer une combinaison de fonctionnalités et attribuer des scores de performance du modèle.

L'approche retenue est la technique *Wrapper* à l'aide de la classe *SelectfromModel* (métamodèle) de *sklearn*. Le modèle utilisé est *RandomForestRegressor* (RFR). Il s'agit d'entraîner le métamodèle avec RFR en lui fournissant l'ensemble des variables et la variable cible. Dans notre cas, il s'agit de la puissance électrique appelée. Après l'entraînement, le modèle nous fournit les caractéristiques sélectionnées qui permettent de mieux prédire la puissance électrique appelée (tâche finale). 16 variables sélectionnées parmi 93 disponibles vont être considérées par la suite pour comparer les méthodes d'imputation pour la salle 219.

3.1.4 Prétraitement – Gestion de l'imputation et comparaison des performances d'imputation

3.1.4.1 Cas d'étude 1 : Appartement T2

Partant d'une base de données complète, les données ont préalablement été normalisées pour les rendre adimensionnelles (variations entre -1 et 1 pour chaque variable). Ensuite, un pourcentage de données a été supprimé selon le mécanisme MCAR (étape 1 de la Figure 5). 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 (étape 2 de la Figure 5), 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). Des informations complémentaires sur certaines méthodes sont disponibles en Annexe 1. La plupart de ces méthodes sont univariées, c.-à-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 (étape 3 de la Figure 5) 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).

3.1.4.2 Cas d'étude 2 : Salle de classe 219 de Polytech Angers

La méthodologie est la même que celle décrite pour le cas d'étude 1. Néanmoins, voici les changements ou les suppléments réalisés :

- Etape 1 : les données sont supprimées (5 à 70%) selon les mécanismes MAR et MCAR ;
- Etape 2 : MICE (méthode présentée en Annexe 1) et une méthode aléatoire (tirage des valeurs aléatoirement entre les valeurs minimales et maximales pour chaque variable) sont étudiées en plus des méthodes citées pour le cas d'étude 1 ;
- Etape 3 : la performance de l'imputation est évaluée avec NRSME et RMSE.

3.1.5 Traitement des données – Tâche finale et comparaison des performances de l'imputation sur la tâche finale

Dans ce projet, les tâches finales correspondent à la détection de l'occupation dans l'appartement de type T2 (tâche de classification de l'état de présence) et à la prédiction de la puissance électrique appelée dans la salle 219 (tâche de régression). Nous avons choisi les modèles de forêts aléatoires pour réaliser ces tâches. En effet, les RF ont montré des bonnes performances lors tests préalablement menés.

3.1.5.1 Cas d'étude 1: Appartement T2

L'algorithme de forêt aléatoire (RF) avec 100 arbres a été appliqué pour la classification de l'état de présence 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). Cela constitue l'étape 4 de la Figure 5. 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. Il s'agit de l'opération d'extraction des caractéristiques (*feature extraction* FE) de la Figure 5. Les caractéristiques suivantes ont été ajoutées : heure du jour, jour de la semaine et variables décalées sur trois pas de temps (pour les autres variables mesurées).

3.1.5.2 Cas d'étude 2 : Salle de classe 219 de Polytech Angers

Le même processus que pour le cas d'étude 1 est appliqué. Un algorithme RF a été appliqué pour la régression afin de prédire la puissance électrique dans la salle 219. La performance de cette tâche a été mesurée avec le RMSE. 80% des données sont utilisées pour l'entrainement et 20% pour le test. La manière d'échantillonner ces deux ensembles a fait l'objet des multiples essais. Le choix final est présenté dans les résultats.

Dans un premier temps, les méthodes d'imputation ont été comparées vis-à-vis de la tâche finale en utilisant les variables sélectionnées précédemment. Dans un second temps, une opération d'extraction des caractéristiques a été menée pour améliorer les performances de la tâche finale. Pour ce faire, des caractéristiques statistiques (moyenne, maximum, minium, écart-type ...) ont été calculées toutes les 30 min et données en entrée du modèle de forêt aléatoire.

Notons que les variables n'ont pas été normalisées en amont pour ce cas d'étude pour faciliter l'interprétation des erreurs (qui sont ainsi dans l'unité de la grandeur à prédire. De plus, il a été observé pour le cas d'étude 1 que la normalisation n'a pas d'effet sur les performances des méthodes d'imputation.

3.2 Résultats

La méthodologie a été appliquée sur les deux cas d'étude présentés au §1.2.

3.2.1 Appartement T2

Les résultats présentés dans cette partie ont fait l'objet d'une publication pour la conférence IBPSA France 2022 à Châlons-en-Champagne. Les travaux ont été acceptés pour une présentation orale. L'article de conférence correspondant est disponible au lien suivant : Traitement des données manquantes pour des capteurs de bâtiments connectés (archives-ouvertes.fr).

3.2.1.1 Présentation des données brutes

Un aperçu de l'évolution des cinq grandeurs mesurées dans le T2 est disponible à la Figure 6.

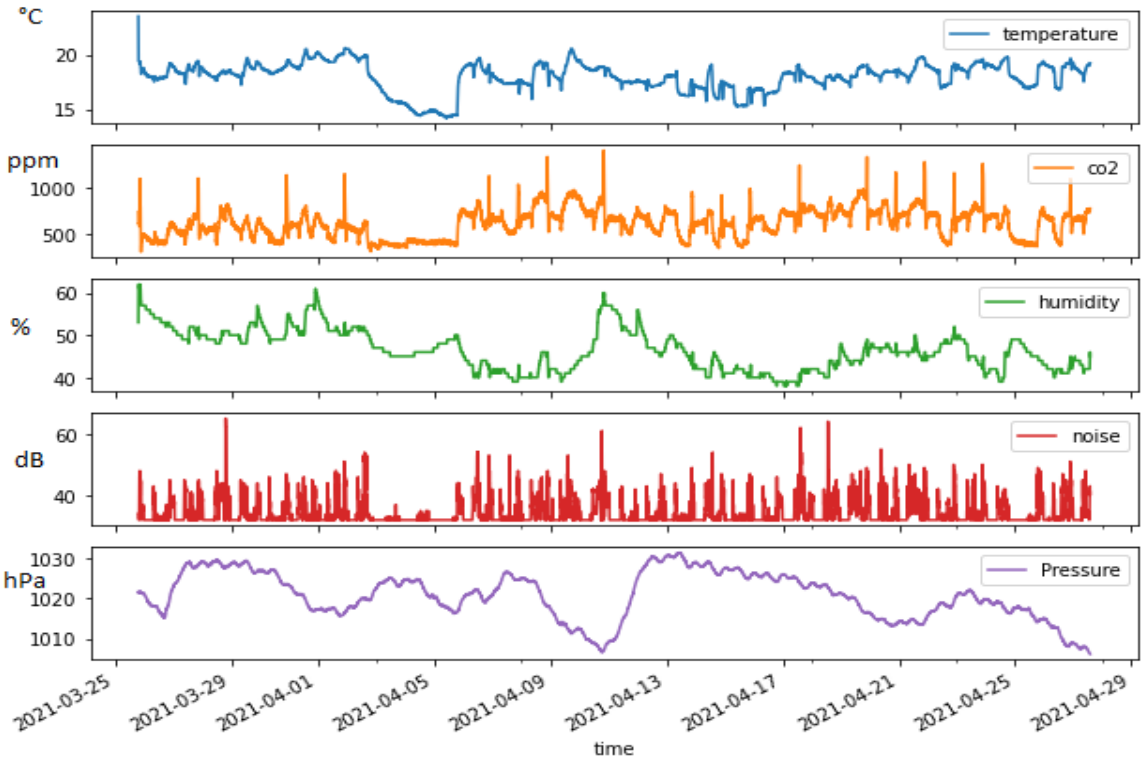


Figure 6 : Évolution des grandeurs mesurées dans l'appartement T2

3.2.1.2 Résultats

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 7. 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.

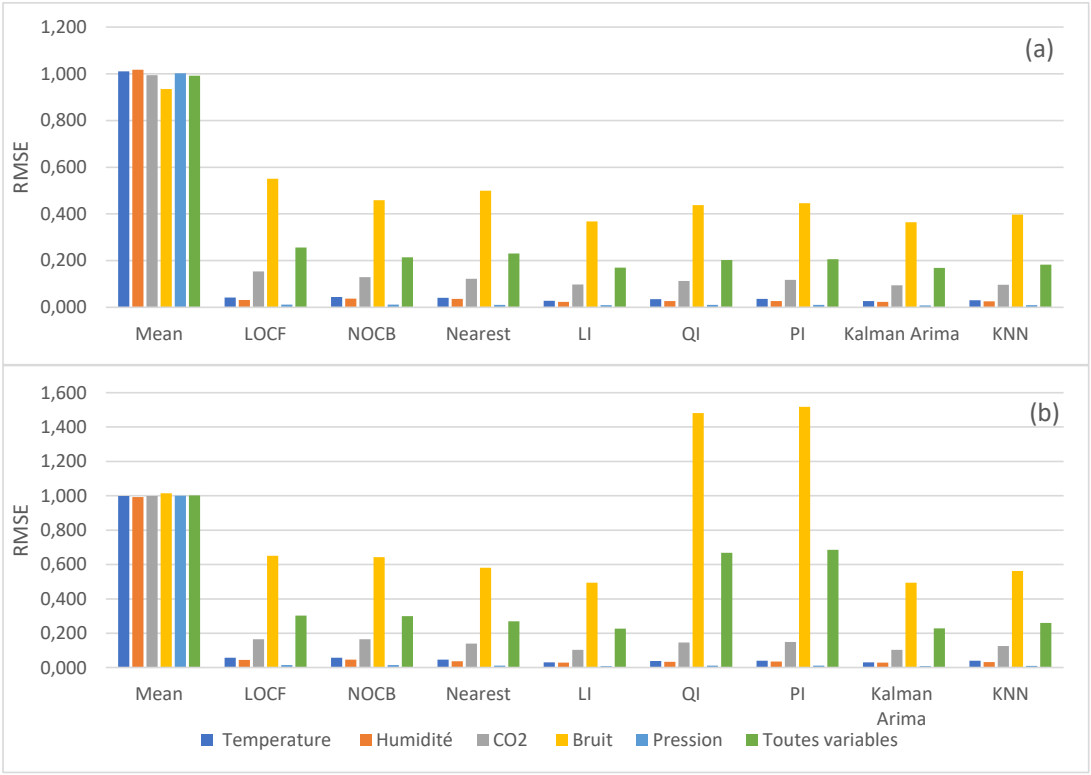


Figure 7 : RMSE pour la tâche d'imputation avec (a) 5 % et (b) 40 % de données manquantes.

Les performances de la tâche finale, c.-à-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). C'est ce qui est présenté dans la partie 3.2.2.

3.2.2 Polytech

L'instrumentation étant complètement opérationnelle dans les salles depuis la mi-mars, la période utilisée pour le traitement des données dans cette étude va du 15 mars au 31 mai 2022. Cette période est caractérisée par une faible occupation des salles, notamment à partir de la mi-avril. En effet, les étudiants de 5ème année étaient déjà partis en entreprise ; les étudiants de 3ème année sont partis en stage dès la fin mars ; et les étudiants de 4ème année fin avril. L'occupation sur le mois de mai a été ponctuelle. Cette faible occupation a une répercussion sur les conclusions, comme nous le verrons dans la suite.

Les salles 114 et 219 ont été fortement instrumentées avec des capteurs multiphysiques. L'objectif était de trouver les meilleurs emplacements pour obtenir des informations sur l'usage : détection des ouvertures de fenêtre ; estimation du niveau d'occupation ; estimation de la consommation d'électricité ; évaluation du confort (hygrothermique et qualité de l'air) des occupants.

Pour identifier les capteurs fournissant le plus d'information, une étape de sélection de variables a été réalisée pour le cas de l'estimation de la puissance électrique appelée en salle 219. Les résultats de la sélection de variables sont présentés au § 3.2.2.1. À l'issue de cette tâche, les quelques capteurs sélectionnés sont utilisés pour tester et comparer les méthodes d'imputation (§3.2.2.2), et pour prédire la puissance électrique appelée dans la salle (§3.2.2.3).

3.2.2.1 Sélection des variables

Nous avons utilisé la technique *SelectfromModel* expliquée au § 3.1.3 pour réduire le nombre de variables de notre jeu de données de la salle 219.

Parmi les 93 variables disponibles, 16 ont été sélectionnées par cette technique.

Nous avons six variables de CO_2, trois de température, une de bruit, une d'humidité relative, une de luminosité et une variable d'état des fenêtres (notons que les états des 4 sont agrégés en une seule variable nommée *windows*). En plus de ces variables, nous avons aussi la température, la pression et l'humidité extérieures.

Les emplacements des variables intérieurs sont indiqués dans la Figure 8.

On constate que les capteurs de CO_2 sont plus proches des ouvertures : fenêtres ou portes.

Le constat est le même pour les capteurs de température sélectionnés. Les capteurs 105 et 113 se trouvent sur l'axe central respectivement à l'arrière et au-devant de la salle. Le 105 est sélectionné, mais pas le 113. Deux hypothèses sont plausibles pour expliquer cette sélection : 113 est redondant avec 105, ou bien 105 est plus proche des étudiants qui ont tendance à occuper les places de derrières (plus de variabilité dans les données mesurées).

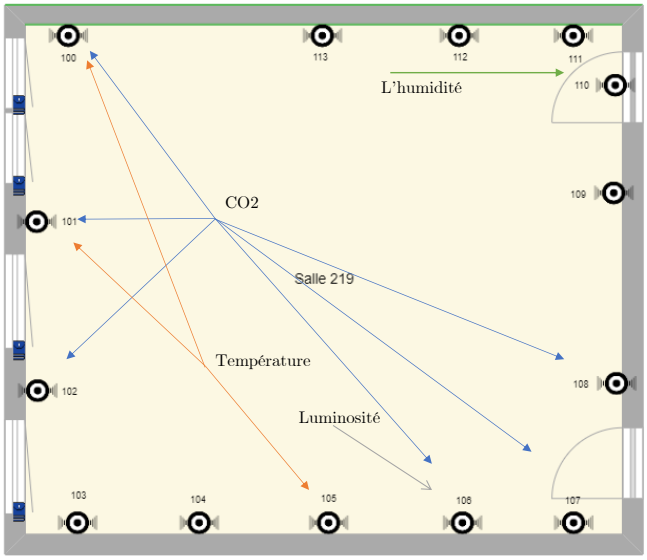


Figure 8 : Emplacements des variables intérieures sélectionnées

Enfin, on observe que lorsque deux capteurs sont placés dans des configurations similaires (proximité des ouvertures et emplacement dans la salle), seul l'un des deux est sélectionné.

L'évolution dans le temps des 16 variables sélectionnées est représentée dans la Figure 9

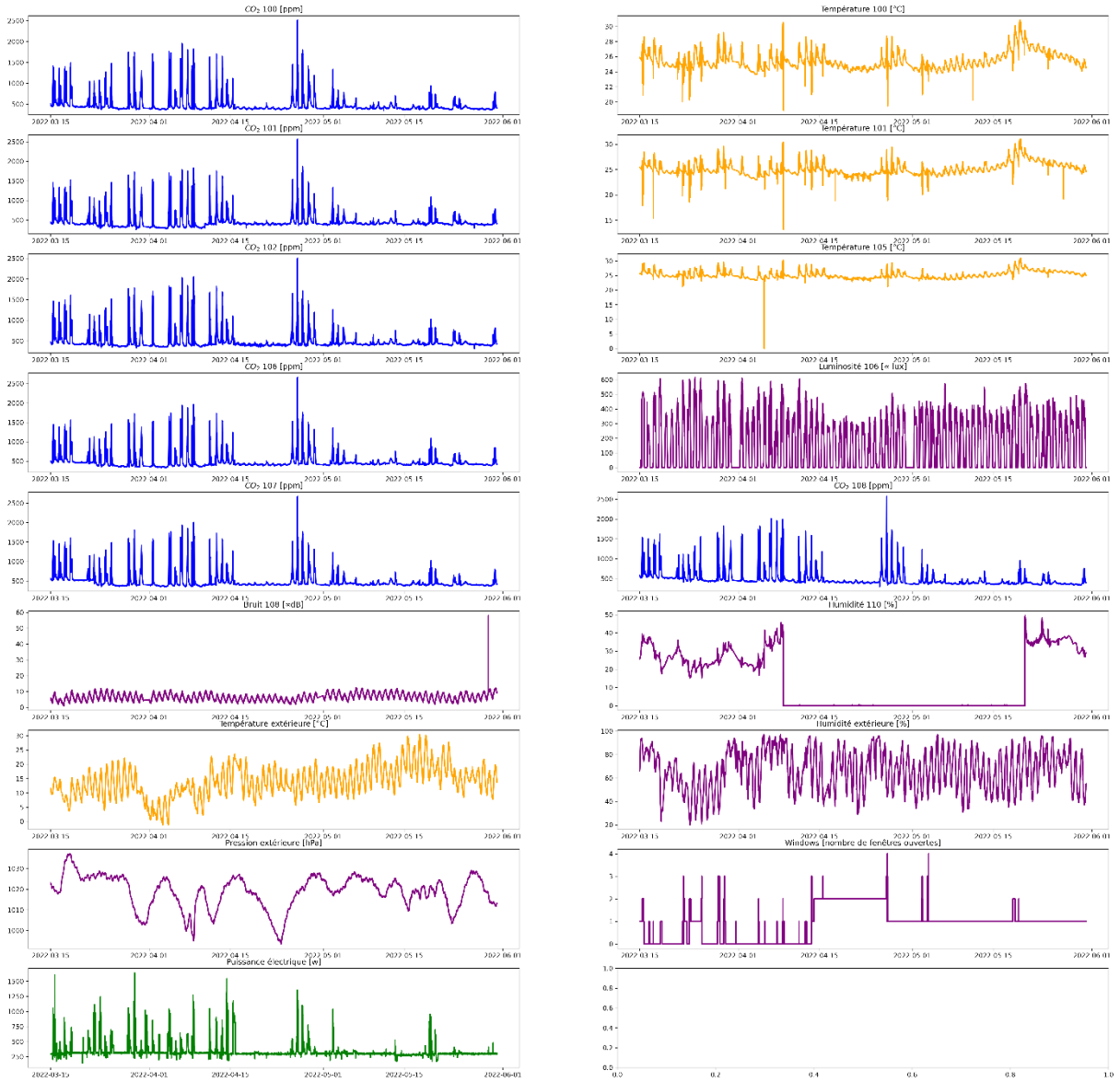


Figure 9 : Évolution des variables sélectionnées et de la variable cible (puissance électrique) sur la période de l'étude

3.2.2.2 Tâche d'imputation

Les RMSE ou NRMSE de la tâche d'imputation sont présentés dans les paragraphes suivants. Ils sont regroupés par type de variable. Par souci de concision, seuls les cas 10 et 60% de données manquantes sont analysés pour les mécanismes MCAR et MAR.

3.2.2.2.1 Comparaison des RMSE pour le CO_2

Avec 10% de données manquantes, les performances d'imputation sont similaires pour MAR et MCAR (voir Figure 10 a et b). Toutes les méthodes offrent de bonnes performances d'imputation, à l'exception de la moyenne (RMSE de l'ordre de 200 ppm) et de la méthode d'imputation par des valeurs aléatoires (non représentée ici et qui présente des RMSE de l'ordre de 1200 ppm).

Pour le mécanisme MCAR, les performances d'imputation décroient lorsqu'on passe de 10 à 60% de données manquantes (voir Figure 10 a et c). Toutefois, le RSME dépasse rarement 50 ppm pour toutes les méthodes (à l'exception de la moyenne et de la méthode aléatoire). Les méthodes QI, LI, PI et Kalman Arima sont les plus performances, aussi bien avec 10 qu'avec

60 % de données manquantes. Cela rejoint les conclusions faites sur le cas d'étude précédent. Au contraire, la méthode MICE, qui présentait de bonnes performances à 10% de données manquantes, voit ses performances fortement dégradées à 60%. Cette dégradation significative du RMSE pour MICE peut s'expliquer par le fonctionnement de cette méthode multivariée : en l'imputation pour une grandeur s'effectue en se basant sur les valeurs des autres grandeurs.

Avec 60% de données manquantes, pour le mécanisme MAR, les performances varient d'un capteur à l'autre (voir Figure 10 d). Pour certains capteurs de CO_2, le RMSE est très important (capteurs 100 et 107), tandis que pour d'autres, le RMSE reste du même ordre de grandeur que pour le MAR10. Cela s'explique par la répartition des données manquantes en fonction des capteurs comme on peut le voir au Tableau 4. Pour les capteurs 100 et 107, la dispersion des données manquantes consécutives est différente (écart-type et durée maximale d'absence de données plus importants que pour les autres).

Tableau 4 : Information sur les valeurs consécutives manquantes pour MAR60 pour le CO_2

Numéro de capteur de CO_2	Moyenne de la durée d'absence de données consécutive [min]	Durée maximale consécutive d'absence de données [h]	Écart-type d'absence de données consécutive [min]
100	15,2	5,6	20,3
101	16,3	2,3	14,7
102	0,0	0,0	0,0
106	16,2	2,1	14,2
107	14,8	5,3	19,0
108	15,9	1,9	13,9

La méthode MICE présente de bonnes performances pour le mécanisme MAR, quel que soit le pourcentage de données manquantes. En effet, cette méthode multivariée tire parti des données présentes en plus grand nombre sur les autres variables. Puisque les données sont complètes pour le capteur de CO_2 102 avec MAR60, la méthode MICE peut facilement imputer les autres valeurs de CO_2 aux autres points de la pièce. C'est pour cela que cette méthode MICE est meilleure que toutes les autres méthodes pour MAR60.

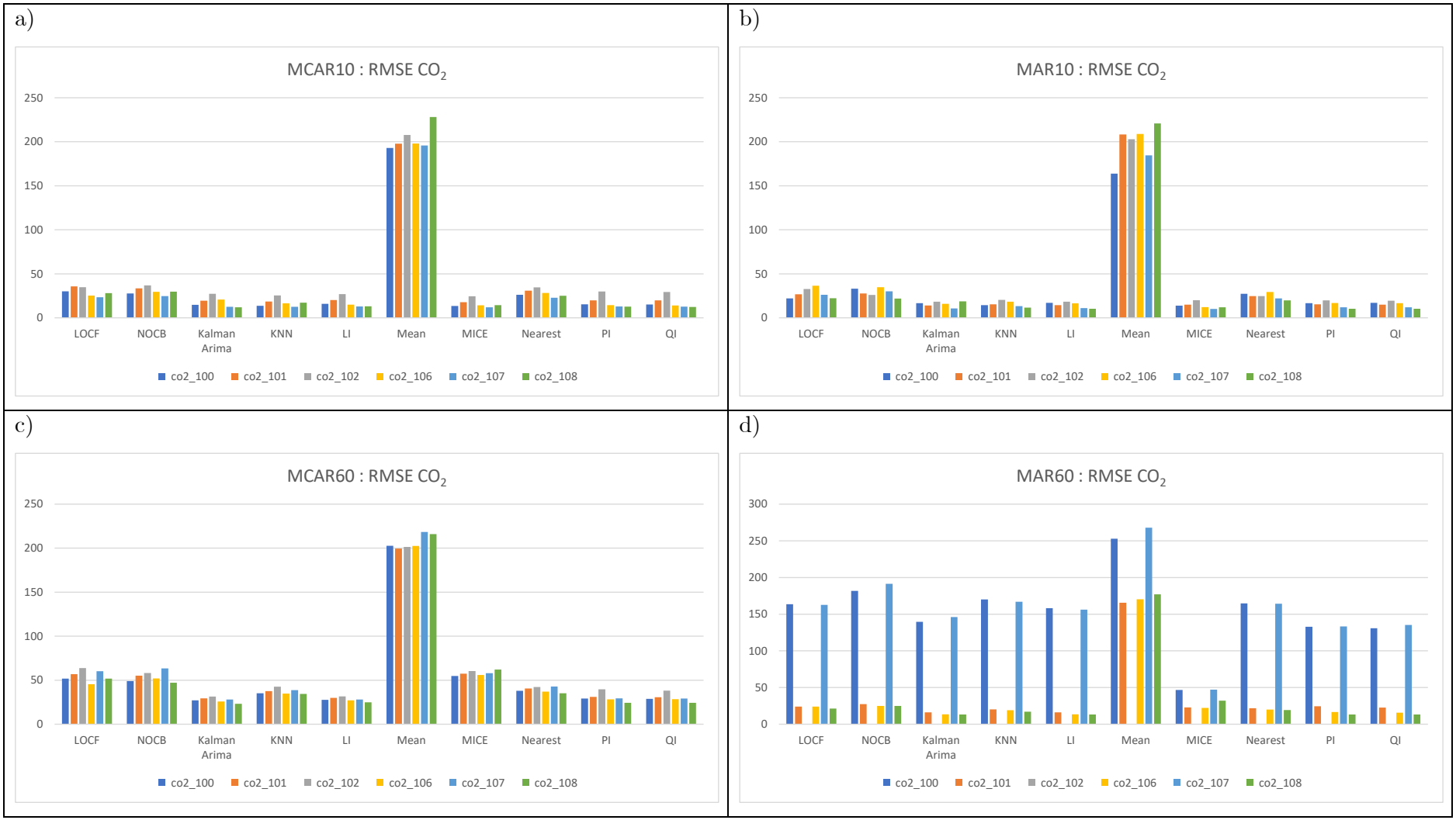


Figure 10 : Comparaison des RMSE pour les capteurs de CO_2 avec les mécanismes MCAR et MAR, pour 10 et 60% de données manquantes.

3.2.2.2.2 Comparaison des RMSE et NRMSE pour la température

Les capteurs de température sont à l'intérieur de la salle à l'exception du capteur 'weather_out' qui représente la température extérieure. L'écart-type sur le capteur de température extérieur est cinq fois plus grand pour ceux des capteurs intérieurs. Cela s'explique par une plus grande amplitude de température à l'extérieur. En étudiant le RMSE (comme fait précédemment pour le CO_2), les conclusions peuvent être biaisées à cause de cette différence d'écart-type. Ainsi, dans la Figure 19 (en Annexe 2), on observe que les performances de prédiction de température extérieure sont moins bonnes. Cela est flagrant pour la méthode MICE, où le RMSE est au moins trois fois plus grand pour la température extérieure. Notons que pour MAR10 (Figure 19 b), le capteur de température extérieur n'est pas visible (le RMSE vaut 0 puisqu'il n'y a pas de données manquantes).

Pour ne pas biaiser les conclusions, il vaut mieux considérer le NRMSE qui correspond au RMSE de chaque variable, divisé par son écart-type. C'est ce qu'on observe à la Figure 11.

Afin de mieux visualiser les résultats, certaines méthodes ne sont pas affichées dans les graphiques. Ainsi, la méthode aléatoire n'est pas affichée dans la Figure 11. Pour le NRMSE ; et les méthodes moyenne et aléatoire ne sont pas affichées dans la Figure 19 pour le RMSE. Cela permet d'éviter un problème d'effet d'échelle sur les autres méthodes.

Pour le mécanisme MAR, les mêmes conclusions que pour le CO_2 sont observées : les performances de l'imputation varient fortement d'un capteur à l'autre. Cela s'explique par la répartition des données manquantes en fonction des capteurs comme on peut le voir au Tableau 5. Pour les capteurs 101 et 105, l'écart-type et durée maximale d'absence de données sont plus importants que pour les autres.

Tableau 5 : Information sur les valeurs consécutives manquantes pour MAR60 pour la température

Numéro du capteur de température	Moyenne de la durée d'absence de données consécutive [min]	Durée maximale consécutive d'absence de données [h]	Écart-type d'absence de données consécutive [min]
100	16,29	2,00	14,38
101	15,18	5,58	20,20
105	14,85	5,25	19,46
Température extérieure	16,32	2,08	13,80

En observant les graphiques du mécanisme MCAR, on peut supposer un léger effet de l'emplacement des capteurs sur les performances d'imputation. Les performances pour le capteur 105 (au milieu de la salle), sont souvent meilleures que celles des capteurs 100 et 101 (proches des fenêtres). Les amplitudes de température des capteurs 100 et 101 est sans doute plus importantes au moment des ouvertures de fenêtres, que pour le capteur 105.

Globalement (quels que soient le mécanisme et le pourcentage de données manquantes), la méthode d'interpolation linéaire donne de bonnes performances pour imputer les capteurs de température.

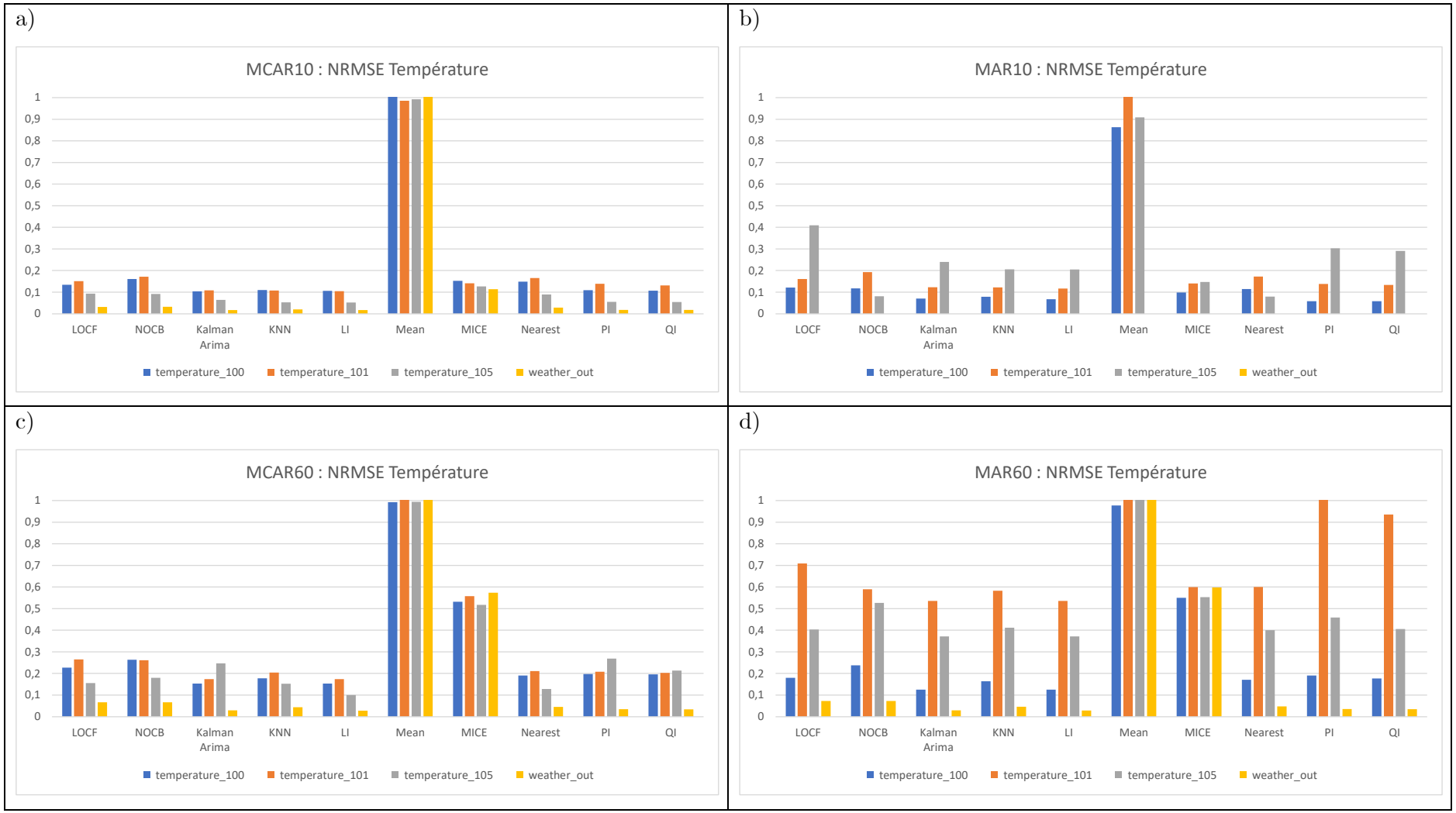


Figure 11 : Comparaison des NRMSE pour les capteurs de température avec les mécanismes MCAR et MAR, pour 10 et 60% de données manquantes.

3.2.2.2.3 Comparaison NRMSE pour les autres capteurs

Les NRMSE sont analysés conjointement pour tous les autres capteurs dans la Figure 12. La méthode aléatoire, qui présente des NRMSE pouvant aller jusqu'à 12 sur certains capteurs, n'est représentée dans la Figure 12, dans un souci de clarté sur le graphique. Notons que les capteurs « *weather_ bar* » et « »*weather_ hum* » désignent respectivement la pression et l'humidité extérieures.

Notons que certaines variables se caractérisent par des discontinuités. Pour la luminosité, l'allumage des lampes créé des discontinuités. De même, l'arrivée des étudiants dans la salle génère un bruit et donc une discontinuité pour le capteur de son. L'état d'ouverture des fenêtres s'exprime par des valeurs discrètes (de 0 à 4). Enfin, l'humidité relative à l'extérieure peut varier brusquement en cas de pluie. Les performances d'imputation sur ces grandeurs physiques sont généralement moins bonnes que pour les autres grandeurs qui ont une évolution continue.

Visuellement, on remarque que les méthodes aléatoire, moyenne, et la méthode MICE sont moins bonnes que les autres. Pour les variables discontinues (en particulier la luminosité et l'humidité extérieure), Kalman présente aussi de moins bonnes performances. Les autres méthodes fournissent de bonnes performances d'imputation et qui sont d'un même ordre de grandeur.

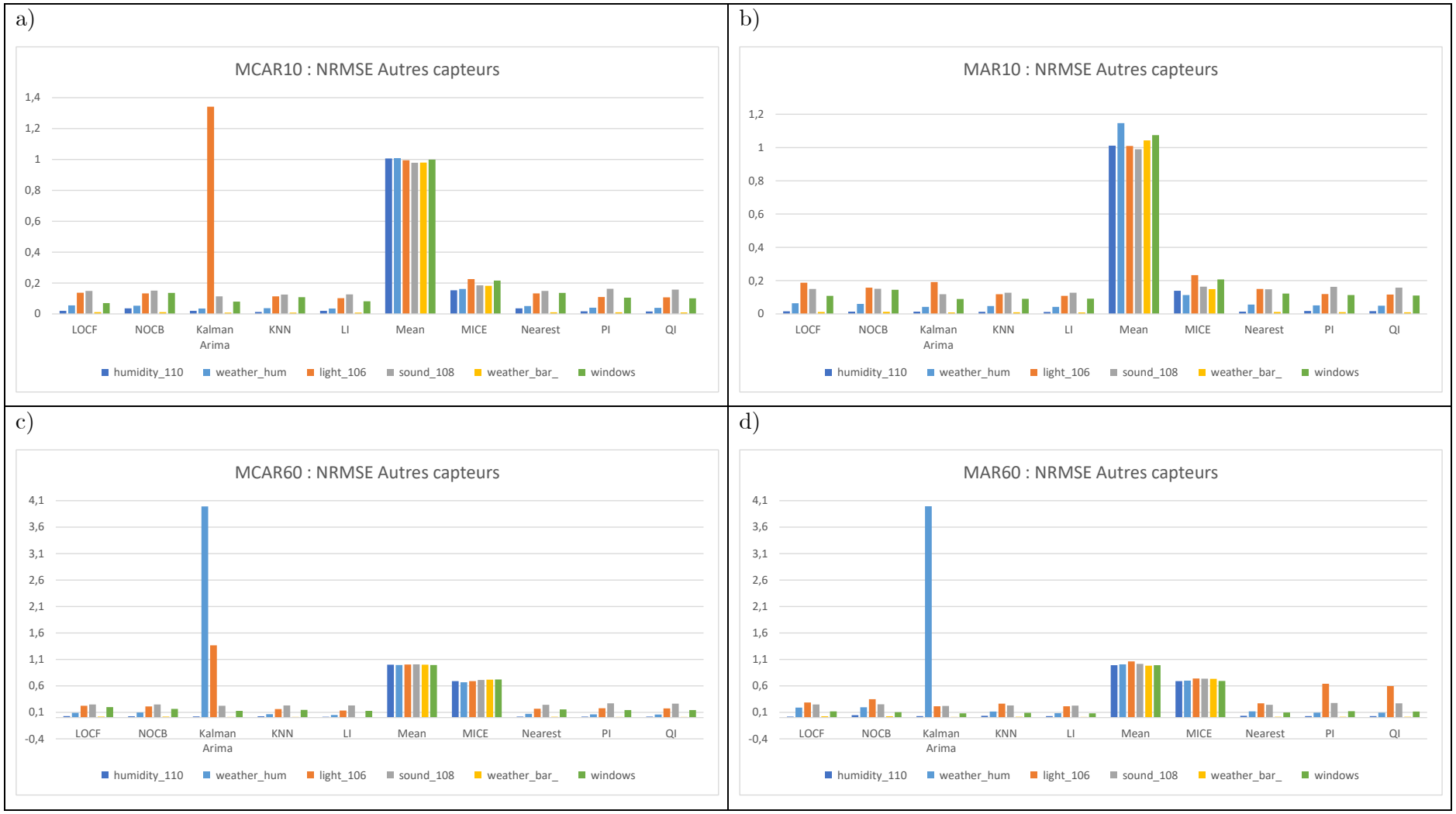


Figure 12 : Comparaison des NRMSE pour les capteurs d'humidité, bruit, luminosité, pression et état des fenêtres avec les mécanismes MCAR et MAR, pour 10 et 60% de données manquantes.

3.2.2.2.4 Conclusion sur la tâche d'imputation

En considérant l'imputation sur l'ensemble des 16 capteurs, nous avons observé que les méthodes LI, PI et QI sont performantes, quel que soit le pourcentage de données manquantes et le mécanisme de leur génération. La méthode aléatoire au contraire, et dans une moindre mesure la méthode moyenne, fournissent des RMSE éloignés des autres méthodes.

La méthode MICE, qui n'avait pas été étudiée dans le cas d'étude 1, a été sélectionnée pour ce cas d'étude pour sa capacité à imputer des données multivariées. Il s'avère que cette méthode n'est pas toujours performante sur nos données, et notamment lorsque les variables sont moins corrélées, ou contiennent beaucoup de données manquantes.

Enfin, la méthode de Kalman Arima montre ses limites pour les variables présentant de fortes discontinuités et dont le nombre de données consécutives manquantes est élevé.

3.2.2.3 Tâche finale

Dans la tâche finale, il s'agit, comme précisé plus haut, de prédire la puissance électrique des salles. Cette information permet d'obtenir la puissance en évitant l'installation d'un capteur couteux. De plus, l'apprentissage peut être transposé (avec des méthodes de *transfert learning*) à d'autres salles d'informatiques. Enfin, dans les salles informatiques, la consommation peut informer sur le niveau d'occupation.

La méthode de forêt aléatoire est utilisée pour prédire la puissance électrique. La Figure 13 montre un aperçu d'une partie d'arbre constituant la forêt aléatoire. Le taux de CO_2 du capteur 106 (proche de la porte) est le facteur discriminant pour cet arbre. Les nœuds suivants sont créés à partir des valeurs de taux de CO_2, de température et d'ouverture des fenêtres.

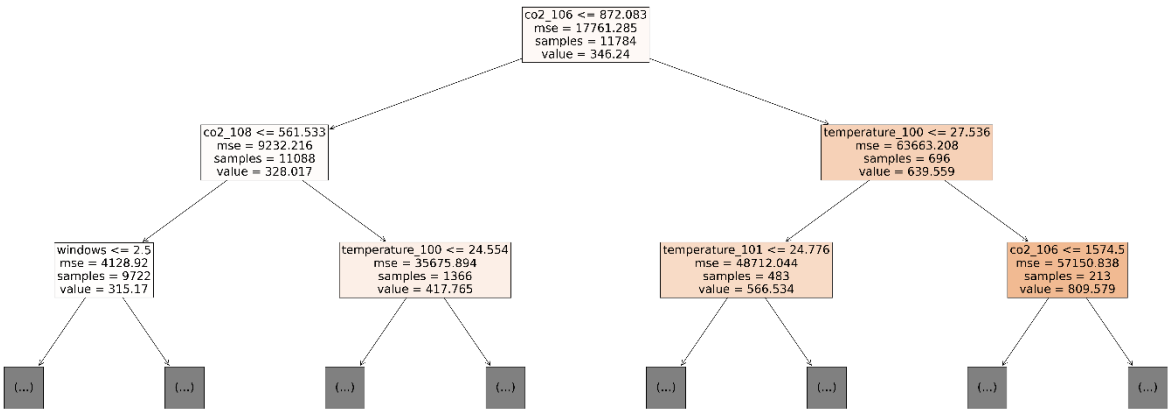


Figure 13 : Aperçu du début d'un arbre de décision de la forêt aléatoire

Les résultats présentés dans la suite sont le fruit d'un travail itératif sur la formation du modèle de la tâche finale.

3.2.2.3.1 Séparation des données d'entrainement et de test

Les données mesurées varient fortement entre les mois de mars et de mai de fait de l'occupation d'une part, et de la saisonnalité d'autre part :

- Comme discuté plus haut, le taux d'occupation des salles est faible et disparate sur la période d'étude (presque plus d'étudiants au mois de mai). Sur les 2 mois et demi, la salle n'est occupée que 10% du temps environ (taux de CO_2 supérieur à 600 ppm, 11,8% du temps).

- De plus, l'augmentation de la température extérieure au cours de la période d'étude conduit à un changement de comportement des occupants (ouverture plus fréquente des fenêtres) induisant une aération plus importante. La concentration en CO_2 peut être faible suite à l'aération, alors même qu'il y a beaucoup d'étudiants dans la salle.

L'utilisation des méthodes classiques de séparation entre les ensembles d'entrainement et de test (répartition chronologique, ou mélange aléatoire) peut conduire à ne pas avoir un échantillonnage représentatif sur les ensembles d'entrainement et de test. Pour limiter ce problème, une méthode de séparation des ensembles d'entrainement et de test a été développée. Elle consiste à grouper les données par jours et à choisir de manière aléatoire parmi les jours disponibles, au lieu de choisir aléatoirement parmi toutes les observations. Ainsi 80% des jours entiers sont utilisés pour l'entrainement et 20% pour le test. Intuitivement, le fait de grouper les données par jour permet d'augmenter les chances d'obtenir des jours d'occupation et des jours de différentes saisons parmi les 20% de test.

Deux exemples de séparation entre les données d'entraînement (*train* en bleu) et de test (en orange) sont visibles à la Figure 14. Les graphiques montrent l'évolution de la puissance appelée en W en fonction du temps.

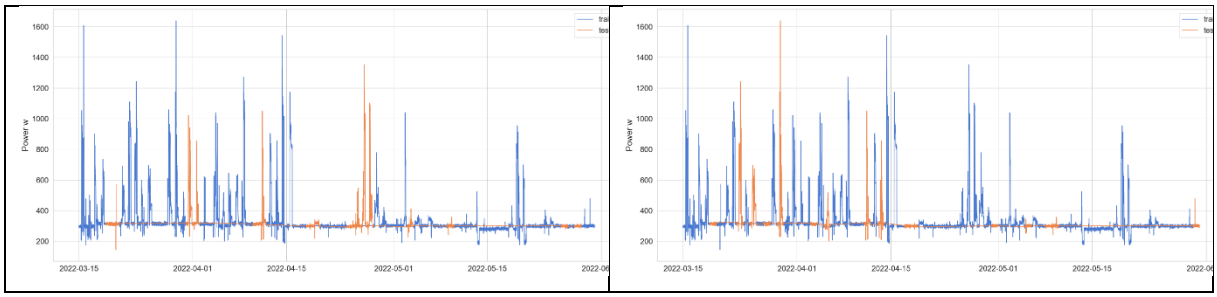


Figure 14 : Séparation des données entre l'entraînement (en bleu) et le test (en orange).

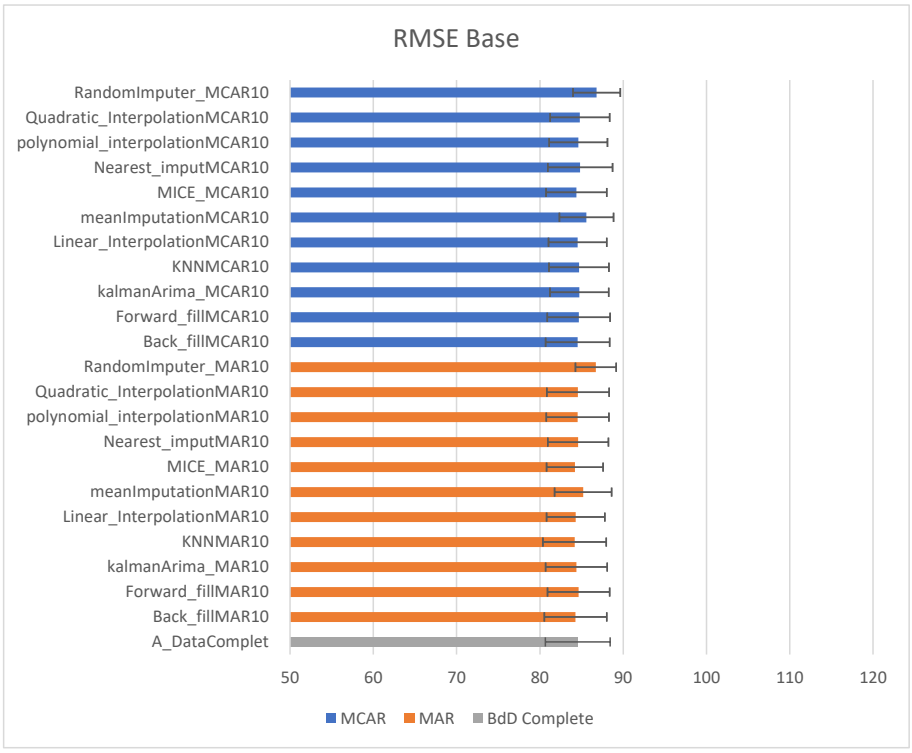
3.2.2.3.2 Limitation du surapprentissage et de la variabilité des RMSE

Les premiers résultats obtenus montraient : i) un surapprentissage important (RMSE de l'entrainement beaucoup plus faible que les RMSE obtenus pour le test) ; ii) et n'étaient pas reproductibles (variance importante des résultats en fonction des données sélectionnées pour le test et pour l'entrainement). Pour solutionner ces deux problèmes, plusieurs options ont été explorées. Les hyperparamètres du modèle de forêt aléatoire ont été réglés pour réduire le surapprentissage. Toutefois, la variance des RMSE restait importante. En fonction de la répartition obtenue entre les données d'entrainement et de test, les RMSE variaient de 80 à 100 W instantanés avec un écart-type pour chaque méthode de 20 à 30 W. La solution proposée pour réduire cette variabilité a été de sélectionner des graines permettant de réduire le surapprentissage sur la base complète. Pour ce faire, une série de 100 tirages aléatoires d'ensemble d'entraînement et de test a été réalisée. Pour chaque tirage, le modèle de forêt aléatoire a été construit et les RMSE ont été observés sur les jeux d'entraînement et de test. Les 15 graines permettant d'avoir les plus faibles écarts entre les RMSE d'entraînement et de test ont été choisies, car dans ce cas le modèle est plus généralisable (mêmes ordres de grandeur des erreurs à l'entraînement et au test). Ainsi, les méthodes d'imputation peuvent être comparées dans les mêmes conditions.

Dans la suite, les résultats montrent le RSME moyen obtenu en utilisant les 15 graines. C'est ce que l'on peut voir à la Figure 15. Le barre d'incertitude représente l'écart-type sur les 15 valeurs de RMSE. Avec 10% de données manquantes, il n'est pas possible de conclure sur le choix d'une meilleure méthode : les performances de toutes les méthodes sont comparables et

sont dans la marge d'incertitude. Avec 60% de données manquantes, les méthodes aléatoires et moyennes montrent de moins bonnes performances, aussi bien le mécanisme MAR que pour le MCAR.

a)



b)

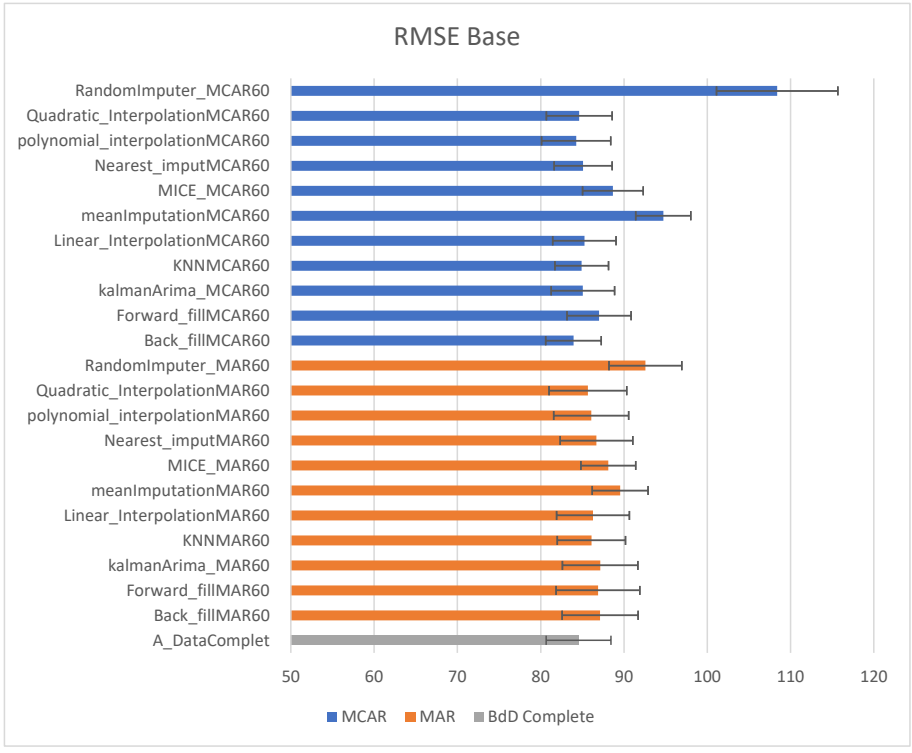


Figure 15: RMSE de la tâche finale avec 10 et 60% de données manquantes

Pour mieux comprendre la répartition des erreurs dans le temps, le graphique de la Figure 16 a été tracé. On observe que la prédiction est correcte en cas d'absence (faibles valeurs de CO_2). En revanche, en cas de présence, la prédiction est faussée. On distingue trois cas de figure.

Dans un premier cas (zone 1), la puissance prédite est inférieure à la puissance réelle et le taux de CO_2 est important : dans ce cas, il y a beaucoup d'occupants dans la salle utilisant des ordinateurs. La difficulté de prédiction peut être liée au manque de données d'occupation sur la période d'étude. Dans un second (zone 2), la puissance prédite est toujours inférieure à la puissance réelle, mais le taux de CO_2 est faible : on peut supposer qu'un grand nombre d'occupants est présent et donc que beaucoup d'ordinateurs sont allumés, mais qu'une ou plusieurs fenêtres sont ouvertes, ce qui complexifie la prédiction. Dans le dernier cas (zone 3), le taux de CO_2 montre une occupation faible, mais aucun poste informatique n'est allumé (puissance électrique réelle au minimum). Le modèle prédit alors un usage d'électricité erroné.

Même si ce modèle présente des erreurs sur la prédiction de la puissance électrique appelée, il pourrait efficacement informer sur le niveau d'occupation de la salle.

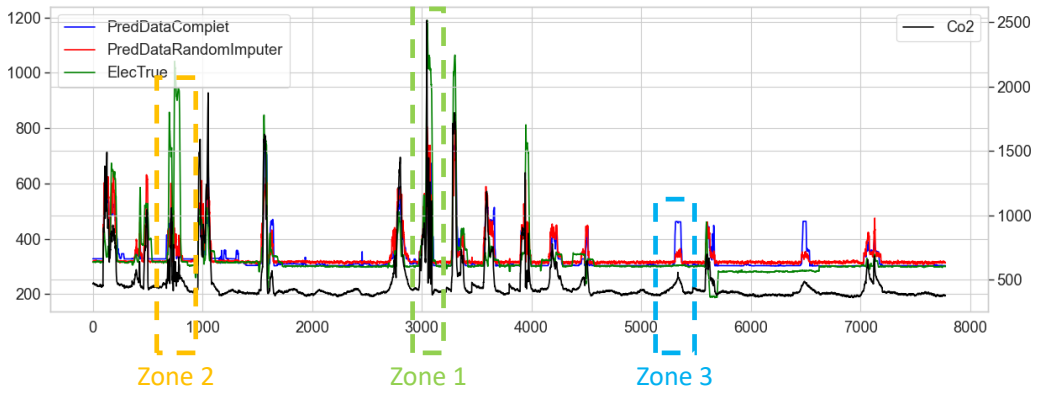


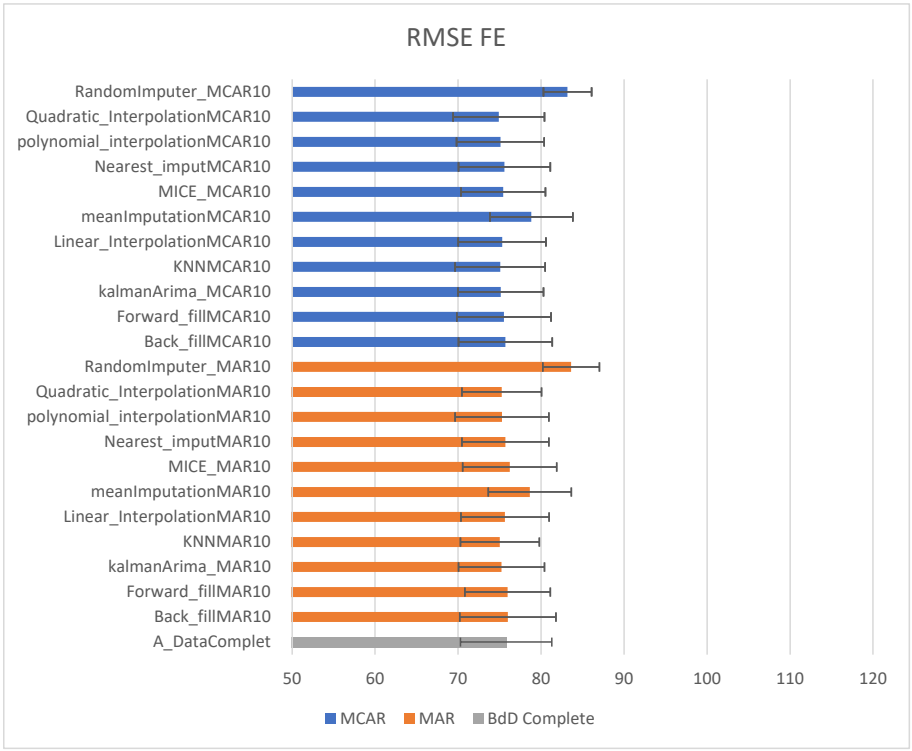
Figure 16: Évolution de la puissance appelée réelle (en vert) et prédite sur la base imputée aléatoirement (en rouge), et prédite sur la base complète (en bleu).

3.2.2.3.3 Réduction de l'erreur du modèle par extraction des caractéristiques

Les séries temporelles des variables ont été regroupées à un pas de temps de 30 min. À chacun de ces pas de temps et pour toutes les variables, la valeur moyenne, minimale, maximale, l'écart-type et la somme ont été calculés. Ces nouvelles séries sont utilisées pour prédire la puissance moyenne appelée sur le pas de temps de 30 min.

Les résultats pour la base complète et pour les bases imputées sont donnés dans la Figure 17 pour 10 et 60% de données manquantes. Ce traitement a permis d'améliorer d'environ 10 W le RMSE sur toutes les méthodes. Comme précédemment, il reste difficile d'identifier une meilleure méthode compte tenu des incertitudes. Toutefois, la méthode aléatoire montre ses limites.

a)



b)

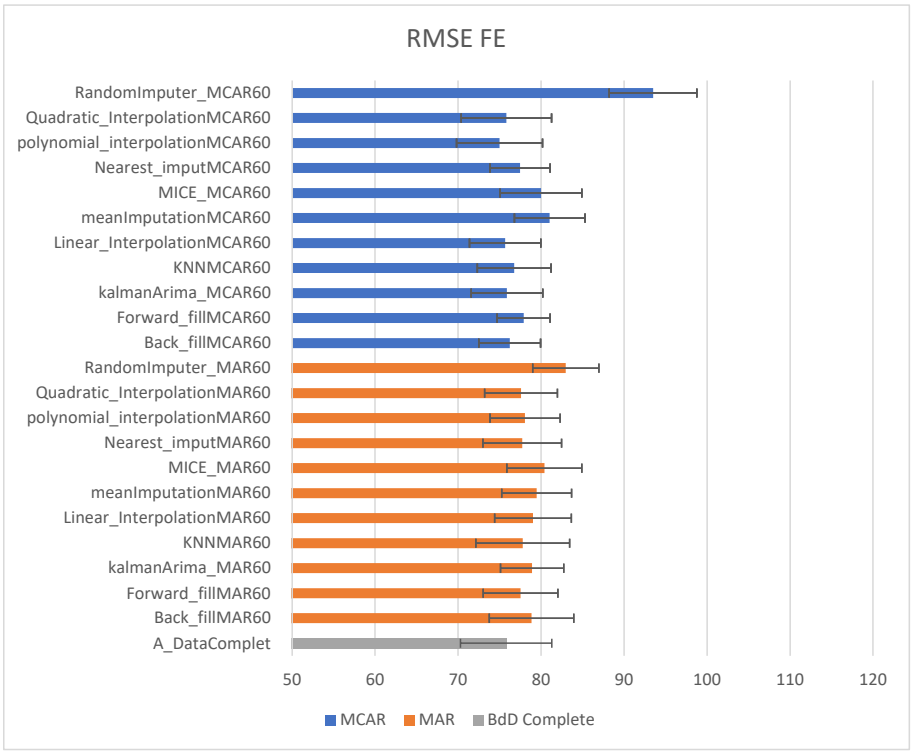


Figure 17: RMSE de la tâche finale avec 10 et 60% de données manquantes après extraction des caractéristiques.

3.2.2.3.4 Conclusion sur la tâche d'imputation

Comme pour le cas d'étude précédent, la performance de la tâche finale s'est avérée peu sensible à la méthode d'imputation. Quelle que soit la méthode d'imputation utilisée, les performances de la tâche finale sont similaires. Ainsi, il ressort qu'il ne faut pas se limiter à l'étude des performances de la tâche d'imputation, qui ne se répercutent pas forcément sur la tâche finale.

Il semble que pour améliorer les performances de la tâche finale, d'une manière plus significative que par le choix d'une méthode d'imputation, l'opération d'extraction des caractéristiques est une alternative à considérer.

Toutefois, ces constats restent à confirmer en traitant les données sur une période de temps plus conséquente (augmentation de la quantité de données) et montrant davantage de diversité (occupation plus fréquente, avec des groupes de taille différente et pour plusieurs saisons).

Chapitre 4. Conclusions et perspectives

4.1 Conclusions

Dans ce mémoire de Master, je me suis intéressé à la fois au prétraitement des données mesurées et au traitement des données manquantes dans les bâtiments connectés.

L'objectif de ces travaux était de comparer et analyser les performances d'imputation de plusieurs méthodes sur deux cas d'étude : (i) un appartement où le statut de présence de l'occupant est connu, (ii) et une salle de classe fortement instrumentée de Polytech Angers.

Les principales conclusions de l'état de l'art mené au premier semestre, sur les méthodes de traitement des données manquantes et d'évaluation de la qualité de l'imputation, ont été rappelées. Après le développement d'outils pour automatiser le prétraitement des données, onze méthodes d'imputation ont été comparées : 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), la méthode des plus proches voisins (KNN), la méthode d'imputation multiple par équations chaînées (MICE) et enfin une méthode aléatoire d'imputation.

Les méthodes sont comparées, d'une part en étudiant la qualité de l'imputation sur ces séries chronologiques multivariées. D'autre part, la comparaison est effectuée en évaluant les performances des méthodes sur la tâche finale, c'est-à-dire la classification du statut de présence pour le premier cas d'étude, et la prédiction de la puissance électrique appelée pour la deuxième étude de cas. Sur la tâche d'imputation, les performances des méthodes se distinguent selon le pourcentage de données manquantes, le mécanisme générateur d'absence, et le caractère continu ou linéaire des variables. De manière générale, les méthodes LI, QI et PI sont les plus performantes sur l'imputation. En revanche, les méthodes aléatoire et moyenne fournissent des performances médiocres. Les différences observées entre les méthodes sur la tâche d'imputation ne se traduisent pas sur la tâche finale. Les écarts entre les méthodes sur la tâche finale restent dans la marge d'erreur. Il ressort donc de ces comparaisons que la performance de la tâche finale est peu affectée par la performance de l'imputation. Ces résultats devraient être confirmés par d'autres études avec des données plus nombreuses et diversifiées. Pour améliorer les performances de la tâche finale, d'autres stratégies peuvent être envisagées, telles que procéder à une extraction des caractéristiques des séries temporelles.

Sur le second cas d'étude, un travail préalable au traitement des données manquantes a été réalisé. Il s'agit d'une sélection de variables d'une part pour réduire la dimensionnalité (très nombreuses variables mesurées dans de cas d'étude), et d'autre part pour aider au choix des meilleurs emplacements pour les capteurs vis-à-vis d'une tâche finale.

Ce stage a été l'occasion pour moi, de réaliser un projet de manière autonome, tout en gardant un lien avec ma tutrice de stage qui a su me guider dans les différentes étapes de ce projet. J'ai pu développer des compétences sur le prétraitement de données et dans le Machine Learning. En somme, je me suis familiarisé avec le domaine des data science ainsi qu'avec le domaine de la recherche. Cela me sera utile dans la suite de ma carrière au laboratoire EASE de l'Université Gustave Eiffel.

4.2 Perspectives

Les perspectives à ces travaux sont nombreuses. Des points à explorer ont été identifiés, à la fois sur le traitement des données manquantes et sur la façon de construire et d'exploiter les modèles prédictifs.

Traitement des données manquantes ;

Nous envisageons d'améliorer la méthodologie relative à l'imputation des données. Il serait intéressant de procéder à l'imputation en deux temps : imputer les données qui serviront à l'entrainement de la tâche finale dans un premier temps, et imputer les données de test indépendamment de la première phase d'imputation. Cela permettrait d'éviter les effets potentiels des fuites des données pouvant être générées par les méthodes d'imputation et en particulier pour les méthodes qui ont besoin d'information statistique sur le jeu de données. Par exemple, pour la méthode moyenne appliquée sur le jeu complet, la valeur utilisée pour l'imputation contient une partie de l'information présente dans les données de test. Or l'information des données de test est censée être inconnue au moment de l'entrainement dans les applications réelles.

Nous envisageons également d'évaluer les méthodes ensemblistes. En d'autres termes, il s'agit d'agréger (en réalisant une moyenne pondérée) les imputations données par plusieurs méthodes et de voir si la performance de l'imputation s'en trouve améliorée.

Construction et exploitation des modèles prédictifs

Pour cette étude sur le traitement des données manquantes, nous nous sommes limités à l'étude de méthodes de ML classiques, ne prenant pas en compte la temporalité des séries. Il serait opportun d'explorer d'autres approches, telles que l'apprentissage profond (*deep learning* avec des réseaux de neurones récurrents LSTM et GRU) pour mieux prédire ou classifier les variables cibles. En complément, d'autres métriques de performances spécifiques aux séries temporelles, telles que la déformation temporelle dynamique (DTW), pourraient être considérées.

Pour le second cas d'étude, les données actuellement traitées sont peu diversifiées et déséquilibrées. Pour rééquilibrer les données entre occupation et absence, des méthodes de rééchantillonnage ou des données d'augmentation de données peuvent être appliquées. Pour amener plus de diversité, des expérimentations contrôlées (nombre de personnes et ouvertures prédéfinies) pourraient être réalisées.

D'autres tâches finales pourraient être étudiées sur les données du second cas d'étude : prédiction des ouvertures de fenêtres, du niveau d'occupation, qualité de l'air, conditions de confort hygrothermique. La prédiction du niveau d'occupation est complexe du fait de l'absence de données labellisées fiables. Il serait envisageable d'exploiter le modèle de prédiction de la puissance électrique pour prédire le niveau d'occupation sur la base des quelques données labellisées fiables. Le modèle formé sur les données de la salle 219 pourrait ensuite être transposé pour prédire le niveau d'occupation en salle 114. Les données de la salle 114 pour les mêmes emplacements de capteurs seraient alors utilisées en entrée du modèle développé en 219, après un recalage avec des données d'occupation étiquetées en 114.

À plus long terme, les connaissances sur l'occupation, acquises par le traitement de données de capteurs, seront utilisées dans une perspective d'optimisation de la consommation des bâtiments connectés centrée sur l'usager.

Bibliographie

[1] « Expertises, Bâtiment », *ADEME*. https://www.ademe.fr/expertises/batiment (consulté le 27 juin 2021).

[2] Z. Yang et B. Becerik-Gerber, « How Does Building Occupancy Influence Energy Efficiency of HVAC Systems? », *Energy Procedia*, vol. 88, p. 775-780, juin 2016, doi: 10.1016/j.egypro.2016.06.111.

[3] É. Vorger, « Étude de l'influence du comportement des habitants sur la performance énergétique du bâtiment », p. 475.

[4] Z. Zhang, « Missing data imputation: focusing on single imputation », *Ann. Transl. Med.*, vol. 4, n° 1, p. 9, janv. 2016, doi: 10.3978/j.issn.2305-5839.2015.12.38.

[5] C. Esteban, S. L. Hyland, et G. Rätsch, « Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs », *ArXiv170602633 Cs Stat*, déc. 2017, Consulté le: 31 janvier 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.02633

[6] Z. Che, S. Purushotham, K. Cho, D. Sontag, et Y. Liu, « Recurrent Neural Networks for Multivariate Time Series with Missing Values », *Sci. Rep.*, vol. 8, n° 1, Art. n° 1, avr. 2018, doi: 10.1038/s41598-018-24271-9.

[7] M. C. de Souto, P. A. Jaskowiak, et I. G. Costa, « Impact of missing data imputation methods on gene expression clustering and classification », *BMC Bioinformatics*, vol. 16, n° 1, p. 64, févr. 2015, doi: 10.1186/s12859-015-0494-3.

[8] S. Rafsunjani, R. S. Safa, A. A. Imran, S. Rahim, et D. Nandi, « An Empirical Comparison of Missing Value Imputation Techniques on APS Failure Prediction », *Int. J. Inf. Technol. Comput. Sci.*, 2019, doi: 10.5815/IJITCS.2019.02.03.

[9] T. Huang, P. Chakraborty, et A. Sharma, « Deep convolutional generative adversarial networks for traffic data imputation encoding time series as images », *Int. J. Transp. Sci. Technol.*, nov. 2021, doi: 10.1016/j.ijtst.2021.10.007.

[10] M. Osman, A. Abu-Mahfouz, et P. Page, « A Survey on Data Imputation Techniques: Water Distribution System as a Use Case », *IEEE Access*, vol. 6, p. 63279-63291, oct. 2018, doi: 10.1109/ACCESS.2018.2877269.

[11] T. H. Ruggles, D. J. Farnham, D. Tong, et K. Caldeira, « Developing reliable hourly electricity demand data through screening and imputation », *Sci. Data*, vol. 7, n° 1, p. 155, déc. 2020, doi: 10.1038/s41597-020-0483-x.

[12] A. Chong, K. P. Lam, W. Xu, O. T. Karaguzel, et Y. Mo, « IMPUTATION OF MISSING VALUES IN BUILDING SENSOR DATA », p. 8.

[13] M. Pazhoohesh, Z. Pourmirza, et S. Walker, « A Comparison of Methods for Missing Data Treatment in Building Sensor Data », in *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, août 2019, p. 255-259. doi: 10.1109/SEGE.2019.8859963.

[14] B. Cho *et al.*, « Effective Missing Value Imputation Methods for Building Monitoring Data », in *2020 IEEE International Conference on Big Data (Big Data)*, déc. 2020, p. 2866-2875. doi: 10.1109/BigData50022.2020.9378230.

[15] R. J. A. Little et D. B. Rubin, *Statistical Analysis With Missing Data*. Wiley, 1987.

[16] N. U. Okafor et D. T. Delaney, « Missing Data Imputation on IoT Sensor Networks: Implications for on-site Sensor Calibration », *IEEE Sens. J.*, p. 1-1, 2021, doi: 10.1109/JSEN.2021.3105442.

[17] Md. K. Hasan, Md. A. Alam, S. Roy, A. Dutta, Md. T. Jawad, et S. Das, « Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021) », *Inform. Med. Unlocked*, vol. 27, p. 100799, janv. 2021, doi: 10.1016/j.imu.2021.100799.

[18] P. B. Weerakody, K. W. Wong, G. Wang, et W. Ela, « A review of irregular time series data handling with gated recurrent neural networks », *Neurocomputing*, vol. 441, p. 161-178, juin 2021, doi: 10.1016/j.neucom.2021.02.046.

[19] Y. Luo, X. Cai, Y. ZHANG, J. Xu, et Y. xiaojie, « Multivariate Time Series Imputation with Generative Adversarial Networks », in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Consulté le: 22 décembre 2021. [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper/2018/hash/96b9bff013acedfb1d140579e2fbeb63-Abstract.html

[20] N. Fouladgar et K. Främling, « A Novel LSTM for Multivariate Time Series with Massive Missingness », *Sensors*, vol. 20, nº 10, p. 2832, mai 2020, doi: 10.3390/s20102832.

[21] A. Hashemi, M. B. Dowlatshahi, et H. Nezamabadi-pour, « VMFS: A VIKOR-based multi-target feature selection », *Expert Syst. Appl.*, vol. 182, p. 115224, nov. 2021, doi: 10.1016/j.eswa.2021.115224.

[22] P. Subramaniam et M. Kaur, « Review of Security in Mobile Edge Computing with Deep Learning », mars 2019, p. 1-5. doi: 10.1109/ICASET.2019.8714349.

[23] « SVM | Support Vector Machine Algorithm in Machine Learning », *Analytics Vidhya*, 12 septembre 2017. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (consulté le 3 août 2021).

Annexes

Annexe 1. Détail sur les méthodes d'apprentissage automatique

Méthodes d'imputation des données

Les méthodes d'imputation utilisée, pour lesquels le non n'est pas explicite, sont présentées dans la suite.

Kalman est une méthode d'imputeTS qui effectue un lissage de Kalman en utilisant la représentation d'espace d'état du modèle ARIMA (*Autoregressive integrated moving average*). La méthode utilisée est na_kalman avec le modèle auto.arima.

K- plus proches voisins (*K-Nearest Neighbors*) (KNN) est un algorithme d'apprentissage automatique supervisé. L'algorithme se base sur l'intégralité du jeu d'entraînement pour fournir une prédiction. Son concept se fonde sur une mesure de similarité entre une nouvelle observation et l'ensemble des données fournies dans la phase d'apprentissage. Selon cette mesure de similarité (distance entre deux points de données), l'algorithme fait ses prédictions.

L'imputation multiple par équations chaînées MICE est une méthode multivariée qui consiste à imputer les données manquantes dans un ensemble de données au moyen d'une série itérative de modèles prédictifs. Dans chaque itération, chaque variable dans le jeu de données est imputée à l'aide des autres variables du jeu de données. Les itérations sont exécutées jusqu'à atteindre la convergence.

Modèle pour la tâche finale

Une forêt aléatoire (*Random Forest*) est une technique de Bagging qui consiste à combiner la prédiction de plusieurs arbres. La raison du recours à l'utilisation de plusieurs arbres alors que la prédiction pourrait être réalisée à l'aide d'un seul est que l'un des principaux inconvénients de l'arbre de décision est le surajustement qui affecte les performances du modèle. L'algorithme de Random Forest introduit un caractère aléatoire dans le choix des variables explicatives et les échantillons pour former les arbres individuels, ce qui permet d'obtenir un modèle prédictif performant en limitant le phénomène de surajustement et de réduire la variance. Pour classer une nouvelle observation, chaque arbre fournit une prédiction. On considère ainsi que l'arbre "vote". La forêt agrège en un seul résultat l'ensemble des votes de tous ses arbres. En cas de régression, l'algorithme prend la moyenne des résultats des différents arbres.

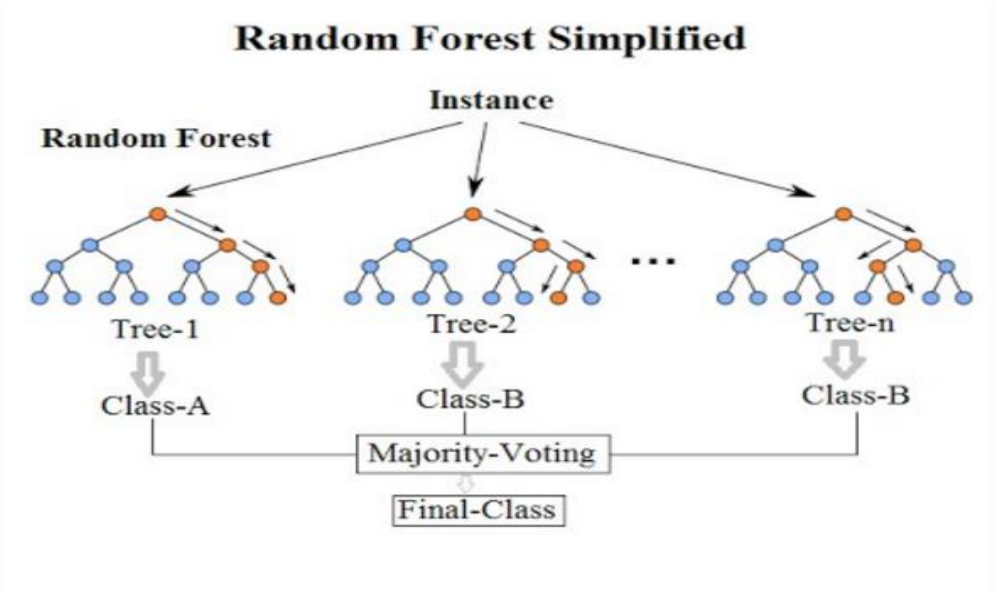


Figure 18 Représentation simplifiée de la forêt aléatoire/ source [23]

Annexe 2. Résultats du cas d'étude Polytech

Tâche d'imputation

RMSE des capteurs de température

En plus de la valeur de NRMSE présentée pour les capteurs de température au § 3.2.2.2.2, nous montrons à la Figure 19 les RMSE pour ces mêmes capteurs, à titre informatif.

Notons que l'échelle des ordonnées des graphiques avec 10 % de données manquantes (Figure 19 a et b), est différente de celle des graphiques pour 60% (Figure 19 c et d).

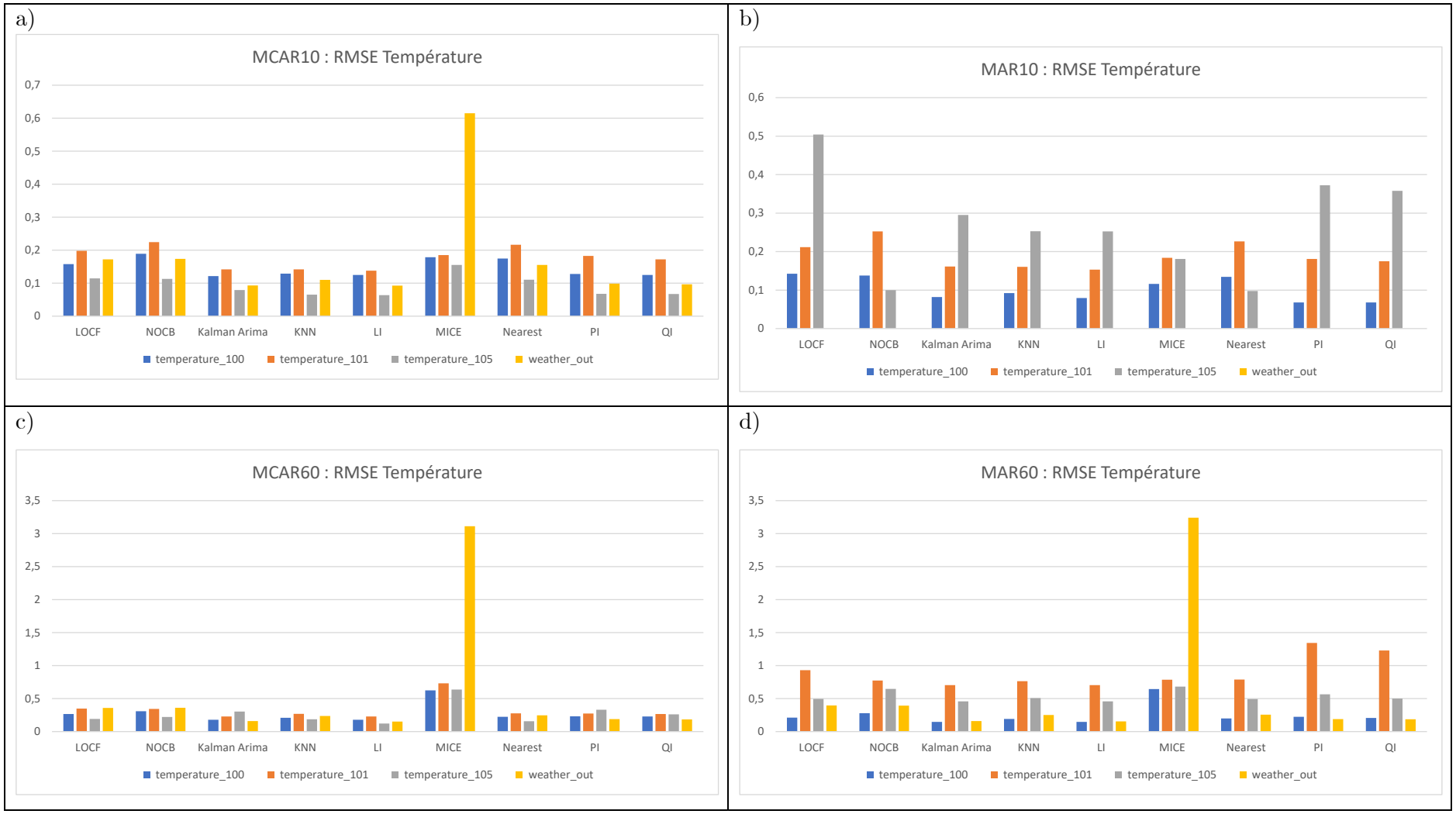


Figure 19 : Comparaison des RMSE pour les capteurs de température avec les mécanismes MCAR et MAR, pour 10 et 60% de données manquantes.

Résumé — The processing of data from smart building allows both to optimise their energy management and to provide a high level of comfort to the occupants. Because of various possible failures in the data collection process, the collected information can be incomplete, incorrect, or poorly structured. In this case, so-called data imputation methods must be used in order to make data processing possible. A state of the art on the methods of processing missing data and evaluating of the quality of the imputation was carried out. After the development of tools to automate data preprocessing, nine (resp. eleven) imputation methods are applied for two study cases : (i) an apartment where the occupnat's presence status is known, (ii) and data collected in a well instrumented classroom of Polytech Angers. The methods are compared, on the one hand by studying the quality of the imputation on these multivariate time series. On the other hand, a comparison is performed by evaluating the performances of the methods on the final task, i.e. the classification of the presence status for the first case of study, and the prediction of the electrical power demand for the second case study. It appears from these comparisons that the performance of the final task is little affected by the performance of the imputation methods in our cases. These resuls should be confirmed by further studies.

Mots clés : Smart Building, Data Imputation, Occupancy classification, Electricity power prediction.

Polytech Angers
62, avenue Notre Dame du Lac
49000 Angers