Validation d'une méthode d'imputation de données manquantes pour la reconstitution des séries de température.

Islam BOUSRI [1]* , Sahabi Abed Salah [1], Benamara Mohamed Arab[1]

Abstract
Les données climatiques ont une importance capitale pour l'étude du changement climatique notamment les canicules et les vagues de chaleur. Parfois les séries de ces données présentent des lacunes sur des périodes temporaires aléatoirement réparties dans le temps, ceci est dû à plusieurs facteurs : panne de l'instrument de mesure, indisponibilité de site, problème de transmission de la donnée, problème d'archivage ...etc. Pour reconstituer ces données manquantes, les chercheurs font recours à des méthodes d'imputation. Dans ce travail, nous avons comparé les performances de 4 méthodes d'imputation multiple des données manquantes sur la température à savoir : missForest, MICE, KNN et ACP. Les résultats des calculs de la moyenne absolue des biais de l'estimation (MAB) à trois niveaux : 5%, 10% et 20% ont révélé que missForest a été la méthode la plus performante pour le traitement de données manquantes de température.

Keywords
données manquantes, MAB, missForest, MICE, plus proches voisins (KNN), ACP.

[1] *Office national de la météorologie , Dar El Beida, Alger*
***Correspondant**: bousri.islam@gmail.com

Contents

1. Introduction

L'imputation multiple est une approche générale du problème des données manquantes. Elle vise à tenir compte de l'incertitude concernant les données manquantes en créant plusieurs ensembles différents de données imputées et en combinant de manière appropriée les résultats obtenus. Dans le domaine de la climatologie,
les méthodes d'imputation de données manquantes sont suffisamment développées et ceci vu le besoin permanent en ces données notamment pour des études de changement climatique et de variabilité climatique [1].

En Octobre 2014, Turrado a évalué l'imputation multivariée de données manquantes ou erronées d'un capteur de rayonnement solaire à l'échelle de dix minutes en utilisant la méthode Multiple Imputation by Chained Equations (MICE)[2]. A. Ilin et A. Kaplan (2009) ont reconstitué les températures globales historiques de surface de la mer (SST) pour la période 1982-1991 en appliquant la méthode d'analyse en composantes principales (ACP) bayésienne [3].

L'utilisation de ces méthodes a été élargie à d'autres secteurs d'activité. Dans l'industrie, M. Wang a proposé, en 2019, un algorithme amélioré de remplissage des données basé sur la Méthode des plus proches voisins (KNN). L'application de cet algorithme sur un échantillon de données de production d'une zone de puits dans le champ pétrolifère de Daqing est doublement bénéfique. Il a permis, non seulement le remplissage des données manquantes, mais aussi l'amélioration de la précision [4]. Dans le domaine énergétique, A. Sundararajan et Arif I. SarwatEmail ont analysé, en 2019, le mécanisme d'imputation de données manquantes d'un système photovoltaïque distribué (PV) raccordé au réseau de Miami qui utilise trois paramètres essentiels : l'irradiation, la température ambiante et la température des modules. Ils ont comparé les performances d'imputation de différentes méthodes : imputation aléatoire, imputation multiple par maximisation des attentes, KNN et forêts aléatoires, en utilisant des mesures d'erreur et d'effet de taille. Les valeurs imputées sont utilisées, ensuite, dans un perceptron multicouche pour prédire et comparer la production de PV avec les valeurs observées. Les résultats ont montré que les valeurs imputées à l'aide de KNN et de forêts aléatoires présentent les plus faibles différences de proportions et aident les services publics à faire des prévisions plus précises de la production pour la planification de la distribution [5]. P. Dixneuf, dans ses travaux en 2019, a étudié la performance de la méthode missForest et son application au problème des données manquantes en environnement. Il a comparé cette approche avec deux autres

méthodes (multivariate imputation by chained equations (MICE) et K-nearest neighbors (KNN)). L'étude a montré l'efficacité de la méthode d'imputation missForest par rapport aux autres méthodes pour le traitement de données manquantes en environnement [**?**]. Deux travaux sur le rattrapage des données ont été déjà publiés dans la revue JAMA. Le premier est celui de K. Soltani et M. Haouari (2017) qui ont utilisé la méthode d'ACP pour la reconstitution des séries mensuelles de températures maximales et minimales sur l'ouest Algérien [1]. Le second travail est celui de F. Kertali (2019) qui a utilisé la méthode de double masse pour rattraper des donnéen=nombre de valeurs imputées.s manquantes d'une série de données pluviométriques [6]. Table 2. Algorithme de K-nearest neighbors (KNN).

Dans cet article, nous souhaitons valider une méthode d'imputation multiple des données manquantes pour le paramètre de température en faisant la comparaison entre 4 approches à savoir : l'analyse en composantes principales (ACP), K plus proches voisins (KNN) , imputation multiple par équations chaînées (MICE) et Forêts aléatoires (missForest).

2. Description des méthodes utilisées

Sont décrites, ci-dessous, les différentes méthodes utilisées dans cet article.

MissForest

D.J.Stekhoven et P.Bühlmann (2011) ont proposé une méthode de complétion basée sur les forêts aléatoires appelée MissForest [7]. Dans cette approche, les variables contenant des valeurs manquantes sont initialisées pour l'imputation en remplaçant les cellules manquantes par des valeurs

moyennes correspondantes (pour les variables continues), ou par la catégorie la plus fréquente (pour les variables catégorielles). Une variable en cours d'imputation est ensuite divisée en deux parties distinctes: la partie observée qui ne contient aucune valeur manquante et la partie manquante qui sert d'ensemble de prédiction. Une forêt aléatoire est ajustée en utilisant la partie observée comme réponse et les valeurs correspondantes des autres variables prédicteurs, et la partie manquante est remplacée par les valeurs prédites de la forêt aléatoire. L'algorithme passe ensuite à la variable suivante à imputer. L'itération s'arrête lorsque la différence entre les valeurs actuelles et les valeurs précédemment imputées augmente ou si le nombre maximal d'itérations est atteint.

- Algorithme :

Notée X, notre matrice n×p pour l'imputation γ . Le critère d'arrêt (il fonctionne en itérations, s'arrêtant lorsque la différence entre l'itération i et i + 1 des trames de données imputées commence à augmenter pour les variables catégorielles et numériques).
X_s Avec s = 1,..... , S dont les valeurs manquantes sont indexées par $I^i_{mis} \subseteq \{1,.....n\}$ on définit :

y^s_{obs} sont les valeurs observées dans X_s
y^s_{mis} sont les valeurs manquantes dans X_s
X^s_{mis} sont les régresseurs observés pour $i^s_{obs} = \{1,.....n\} \setminus i^s_{mis}$
X^s_{mis} sont les régresseurs manquants pour I^s_{mis}

Table 1. Algorithme de MissForest .

Step 1. Faire une première estimation pour toutes les valeurs catégorielles/numériques manquantes (par exemple, la moyenne, le mode)
Step 2. k← vecteur des indices de colonnes en X, triés par ordre croissant du pourcentage de données manquantes.
Step 3. Tant que γ n'est pas atteint **faire :**

$X^{imp}_{old} \leftarrow$ stocker la matrice précédemment imputée **pour** S dan k **faire**:
Ajuster une forêt aléatoire en prédisant les valeurs non manquantes de X_s : $Y^{(s)}_{obs} \sim X^{(s)}_{obs}$
Utilisez ceci pour prédire les valeurs manquantes de X_s :prévoir $Y^{(s)}_{obs}$ en utilisant $X^{(s)}_{obs}$
$X^{imp}_{new} \leftarrow$ mettre à jour la matrice imputée, en utilisant les $Y^{(s)}_{mis}$
fin pour
mise à jour γ
fin Tant que
renvoyer la matrice finale imputée X^{imp}

Méthode des plus proches voisins (KNN)

KNN est une méthode de compilation utile pour faire correspondre un point avec ses k voisins les plus proches dans un espace multidimensionnel. Initialement introduite en 2001 par **O. Troyanskaya** pour l'étude de l'expression des gènes [8]. Elle peut être, aussi, utilisée pour des données continues, discrètes, ordinales et catégorielles. Ce qui la rend particulièrement utile pour traiter toutes sortes de données manquantes. L'hypothèse sous-jacente à l'utilisation de KNN pour les valeurs manquantes est qu'une valeur ponctuelle peut être approchée par les valeurs des points qui lui sont les plus proches, en fonction d'autres variables. L'imputation par KNN consiste à suivre l'algorithme suivant:

Algorithme des k plus proches voisins (k-nn)

- Step 1. Choisir k : 1 ≥ k ≥ n.
- Step 2. Calculer les distances entre $Y_i * i.i$
- Step 3. Garder les k observations $Y_{(i1)}..,Y_{(it)}$ pour lesquelles ces distances sont les plus petites.

- Step 4. Affecter aux valeurs manquantes la moyenne des valeurs des k voisins :
  $(Y_i\, j)miss = Y_i * i * i.i = \frac{1}{k}(Y_{(i1)} + ::: + Y_{(i1)})$
  Avec :
  k : le nombre de voisins .
  Y : matrice des données observées .
  Y * : matrice des données manquantes.

Multiple Imputation by Chained Equations (MICE) :

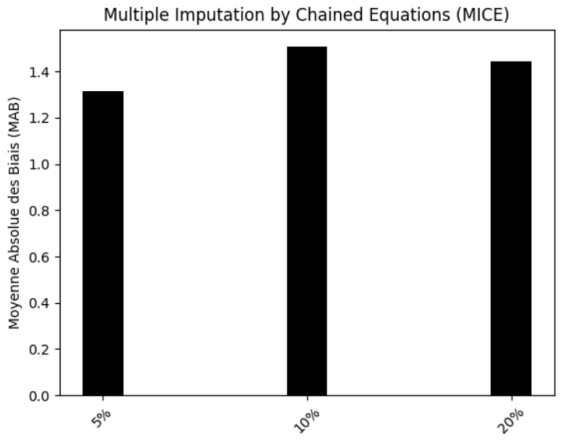
Elle est basée sur un algorithme Monte-Carlo Markov Chain (MCMC). Dans cette technique d'imputation, de nombreux modèles de régression sont exécutés de telle sorte que la variable à laquelle il manque des données est modélisée en fonction d'autres variables de l'ensemble de données [9]. Chaque variable est modélisée en tenant compte du type de variable. Par exemple, la régression logistique est utilisée pour modéliser des variables binaires, alors que la correspondance prédictive de la moyenne est utilisée pour les variables continues [9]. Selon Melissa et al (2011), le processus d'équation enchaînée est décomposé en quatre étapes principales qui sont répétées jusqu'à ce que les résultats optimaux soient atteints [9]. La première étape consiste à remplacer toutes les données manquantes par la moyenne des valeurs observées pour la variable, qui fait office d'indicateur. La deuxième étape consiste à remettre ces imputations moyennes à la valeur "manquante". Dans la troisième étape, les valeurs observées d'une variable (par exemple, "x") sont régressées sur les autres variables de sorte que "x" est la variable dépendante et le reste sont des variables indépendantes.

Analyse en composantes principales (ACP régulé) :

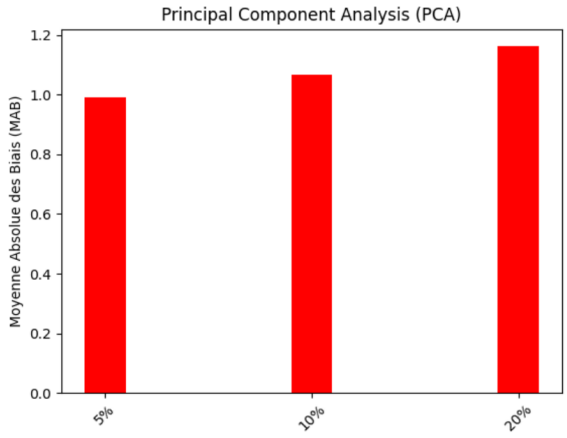
paramètres et d'imputation des valeurs manquantes à l'aide de la matrice adaptée (régularisée) sont itérées jusqu'à la convergence. Dans le logiciel R [10] le nombre de composants utilisés dans l'algorithme peut être, facilement, trouvé en utilisant des critères de validation croisée dans la fonction estim_ncpPCA .

3. Application et validation:

Pour réaliser cette étude, nous avons utilisé la réanalyse ERA5 des températures de surface depuis 1979 jusqu'à 2015 et un échantillon de données enregistrées par la station d'observation de Dar el Beida pour la même période . cet échantillon est composé de paramètres suivants : températures moyennes quotidiennes, températures minimales journalières et températures maximales journalière mesurées à deux mètres au-dessus du sol,
Nous avons, ensuite, procédé aléatoirement à l'enlèvement progressive d'un quota de données de 5%, 10% puis 20%. Puis, nous avons imputé les données manquantes en utilisant les quatre méthodes décrites précédemment.
Ensuite, nous avons calculé la moyenne absolue des biais



(a)

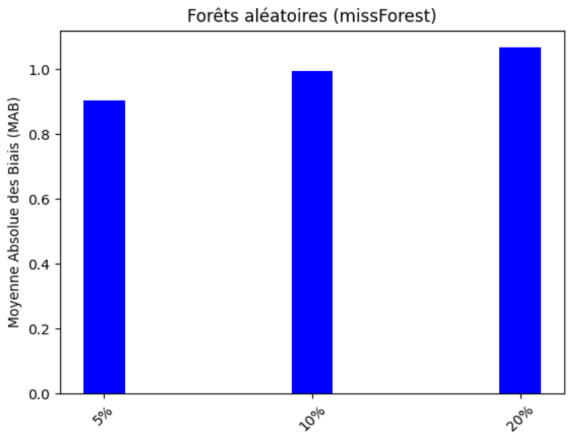


(b)

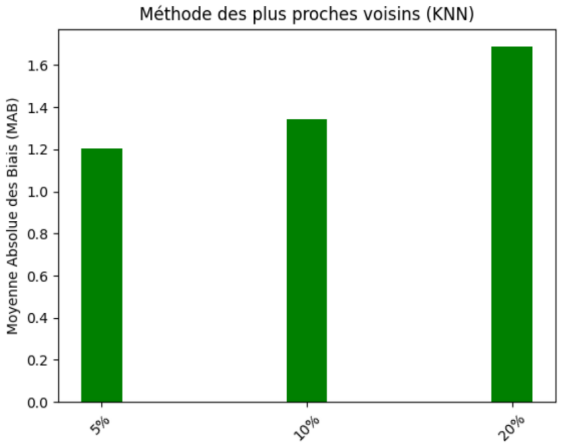
de l'estimation (MAB) pour chaque niveau d'enlèvement 5%, 10% puis 20%. Et enfin, nous nous validerons, à chaque fois la méthode qui minimise la MAB.
Tous les algorithmes de calcul relatifs aux quatre méthodes utilisées dans cette étude sont référés aux packages du logiciel R [10] .

$$MAB = \frac{\sum_{i=1}^{n} |valeurréelle - valeurimputée|}{n} \quad (1)$$

- n=nombre de valeurs imputées.



(c)



(d)

Fig.1 : Moyennes absolues des biais (MAB) obtenues par les quatre méthodes : (a) ; méthode MICE, (b) ; méthode (ACP), (c)méthode missForest, et (d) ; méthode KNN

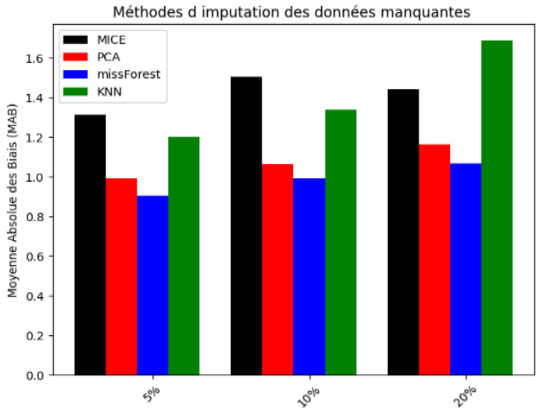


Fig.2 : MAB pour les 4 méthodes .

Table 1. Résultat de l'imputation et la validation.

Méthode Taux de données manquantes	MICE	ACP	missForest	KNN	Validation
5%	1,315625	0,9909400	0,9048437	1,340625	missForest
10%	1,5067708	1,0658378	0,9927916	1,6869791	missForest
20%	1,4447916	1,1619766	1,0663802	1,2026874	missForest

Position	5%	10%	20%
1	missForest	missForest	missForest
2	ACP	ACP	ACP
3	MICE	MICE	KNN
4	KNN	KNN	MICE

Tableau 4 :Tableau de classement des méthodes par performance en fonction du quota de données enlevées.

La Figure 01 représente des histogrammes des moyennes absolues des biais obtenues par les quatre méthodes. Pour les trois niveaux de quotas d'enlèvement, la méthode missForest occupe la première position en termes de performance avec des moyennes absolues des biais qui se situent au voisinage de 1°C (Tableau 3). Pour cette méthode, la valeur de MAB est proportionnelle au nombre de données manquantes.

En deuxième position, nous retrouvons la méthode ACP avec des valeurs de MAB très proches de celles de missForest. À l'instar de la méthode missForest, les performances de la méthode ACP diminuent avec l'augmentation du nombre de données enlevées.

Les moyennes absolues des biais obtenues pour les deux méthodes MICE et KNN sont relativement loin de celles de MissForest. Pour les deux niveaux d'enlèvement 5% et 10%, les performances de la méthode MICE sont meilleures par rapport à celles de KNN. Par contre, KNN est mieux placée par rapport à MICE pour le niveau 20%. Il est à noter que la méthode KNN présente une particularité par rapport aux autres méthodes. En effet, la diminution des performances de cette méthode n'est pas proportionnelle au nombre de données manquantes. Au contraire, par exemple l'estimation des moyennes absolues des biais obtenues pour le niveau de données manquantes 20% est meilleure pour le cas de 5%.

4. Conclusion

Nous avons réalisé une étude comparative entre quatre méthodes d'imputation multiple des données manquantes de température à savoir : missForest, MICE, KNN et ACP. Pour cette étude, nous avons utilisé un échantillon de données d'observation enregistrées par la station de Dar el Beida depuis 1936 jusqu'à 2015. Les résultats des calculs de la moyenne absolue des biais de l'estimation (MAB) à trois niveaux 5%, 10% et 20% ont révélé que missForest a été la méthode la plus performante pour le traitement de données manquantes de température. Cette méthode est alors recommandée pour combler les lacunes des séries de températures.

References

[1] Julie Josse, François Husson, et al. Reconstitution des séries mensuelles de températures maximales et minimales sur l'ouest algérie. *journal Algérien de Météorologie Appliquée*, 0:85–90, 2017.

[2] Concepción Crespo Turrado, María del Carmen Meizoso López, Fernando Sánchez Lasheras, Benigno Antonio Rodríguez Gómez, José Luis Calvo Rollé, and Francisco Javier de Cos Juez. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors*, 14(11):20382–20399, 2014.

[3] Alexander Ilin and Alexey Kaplan. Bayesian pca for reconstruction of historical sea surface temperatures. In *2009 international joint conference on neural networks*, pages 1322–1327. IEEE, 2009.

[4] Mei Wang, Dong Li, Kaiyuan Qi, Chenglong Xue, and Erlong Yang. Sknn algorithm for filling missing oil data based on knn. In *IOP Conference Series: Materials Science and Engineering*, volume 612, page 032099. IOP Publishing, 2019.

[5] Aditya Sundararajan and Arif I Sarwat. Evaluation of missing data imputation methods for an enhanced distributed pv generation prediction. In *Proceedings of the Future Technologies Conference*, pages 590–609. Springer, 2019.

[6] Kertali. Étude de comblement de lacunes : Cas des séries pluviométriques observées du réseau de l'onm. In *Journal Algérien de Météorologie Appliquée*, pages 49–58, 2019.

[7] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[8] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[9] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[10] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.