

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^a

^a 1 Office national de la météorologie , Dar El Beida, Alger

Correspondant: bousri.islam@gmail.com

Résumé

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. 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 multi-variée 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'appli-

cation de cet algorithme sur un échantillon de données de production d'une zone de puits dans le champ pétrolier 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