FACULTÉ DES SCIENCES ET TECHNOLOGIES | Université de Lille

Master ISN

Le problème des données manquantes

Reda Mdair - Victor Naninck - Théo Dugauguez

Sous la direction de Charlotte Baey

Février 2025

Table des matières

1 Introduction

Dans toute analyse de données, une hypothèse implicite semble couler de source : disposer d'un jeu de données complet, structuré et fiable. Pourtant, la réalité est bien différente. Un statisticien, qu'il travaille sur des sondages ou des bases de données de tout type, est régulièrement confronté à des données incomplètes ou incohérentes. C'est par exemple le cas de réponses non renseignées dans une enquête, de mesures perdues lors de la transmission d'un signal ou encore de valeurs corrompues par des erreurs de saisie. Cette absence d'informations constitue un défi préoccupant en science des données.

Longtemps considérées comme une simple contrainte technique, les données manquantes ont progressivement suscité un intérêt croissant en statistique. Little et Rubin, dans leur ouvrage *Statistical Analysis with Missing Data*, ont mis en évidence que des données absentes, ignorées ou mal traitées, peuvent fausser les inférences statistiques et nuire à la fiabilité des modèles prédictifs. En effet, éliminer systématiquement les observations incomplètes, bien qu'intuitif, peut biaiser une analyse et affaiblir la robustesse d'un modèle. À l'inverse, imputer ces valeurs par des méthodes adaptées permet d'exploiter pleinement l'information disponible et d'améliorer la qualité des inférences.

Ce rapport propose une exploration synthétique des notions essentielles liées aux données manquantes. Après avoir exposé les principaux mécanismes expliquant leur apparition, nous présenterons les approches classiques de gestion de ces valeurs, avant d'introduire des méthodes plus avancées permettant une reconstruction plus fine des données absentes. Enfin, nous illustrerons ces concepts à travers deux applications concrètes sur des jeux de données réels, qui nous donneront l'occasion d'offrir quelques réflexions prospectives sur ce sujet.
L'objectif est de fournir une approche claire et pragmatique des stratégies d'imputation, sans entrer dans des développements mathématiques approfondis.

2 Mécanismes des données manquantes

Comprendre la manière dont les valeurs manquantes apparaissent est essentiel pour choisir des méthodes adaptées à leur traitement.

2.1 Typologie des données manquantes

On peut formaliser la classification des données manquantes en trois catégories distinctes : *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) et *Missing Not At Random* (MNAR). Chacune de ces catégories repose sur la relation entre la présence des valeurs manquantes et les autres variables du jeu de données. Cette typologie est cruciale car elle conditionne les choix méthodologiques en matière d'imputation et d'inférence.

Pour formaliser ces notions, considérons pour la suite du rapport une matrice de données X de taille $n \times d$, où n représente le nombre d'observations et d le nombre de variables. On définit une matrice indicatrice des valeurs manquantes $M = (m_{ij})_{1 \leq i \leq n, 1 \leq j \leq d}$, avec :

$$m_{ij} = \begin{cases} 1 & \text{si } x_{ij} \text{ est manquante,} \\ 0 & \text{sinon.} \end{cases}$$

On note également X_{obs} les données observées et X_{mis} les données manquantes. Nous décrivons dans ce qui suit les trois principaux mécanismes de valeurs manquantes en fonction de leur distribution.

2.1.1 Données manquantes complètement aléatoirement (MCAR)

Le mécanisme MCAR (*Missing Completely At Random*) correspond au cas où la probabilité d'absence d'une donnée est totalement indépendante des autres valeurs, qu'elles soient observées ou non. Cela signifie que le processus de génération des valeurs manquantes est entièrement aléatoire et ne suit aucun schéma particulier.

Formellement, on a :

$$P(M \mid X) = P(M)$$

où la probabilité que x_{ij} soit manquante ne dépend ni des valeurs observées X_{obs}, ni des valeurs manquantes X_{mis}.

Dans ce contexte, les données restantes peuvent être considérées comme un échantillon aléatoire non biaisé de la population d'origine. Une analyse effectuée uniquement sur les observations complètes reste valide, bien qu'une perte d'information puisse entraîner une diminution de la précision statistique.

Exemple : Dans un sondage, chaque personne décide de répondre ou non à une question en lançant un dé, et refuse de répondre si un certain numéro apparaît. Ce processus de réponse est totalement indépendant des caractéristiques des individus, illustrant ainsi un cas d'incomplétude MCAR.

2.1.2 Données manquantes aléatoirement (MAR)

Le mécanisme MAR (*Missing At Random*) est défini par le fait que la probabilité de manquer une valeur dépend uniquement des variables observées et non des valeurs elles-mêmes.

Exemple : Lors d'une enquête de satisfaction sur un service de transport, les passagers disposant d'un abonnement mensuel répondent majoritairement aux questions sur la fréquence d'utilisation, tandis que ceux qui achètent des tickets à l'unité laissent souvent cette question sans réponse. Ici, l'absence de réponse dépend d'une variable observée (le type de billet utilisé) et non directement de la fréquence d'utilisation elle-même.

Mathématiquement, ce cas de données incomplètes s'écrit :

$$P(M \mid X) = P(M \mid X_{obs})$$

Autrement dit, conditionnellement aux valeurs observées, l'apparition des valeurs manquantes est indépendante des valeurs non observées.

Ce cas est plus fréquent que MCAR et permet d'utiliser des méthodes d'imputation avancées pour estimer les valeurs manquantes sans induire de biais majeur. En particulier, si les relations entre les valeurs manquantes et les valeurs observées sont bien modélisées, les résultats restent valides après correction.

2.1.3 Données manquantes non aléatoirement (MNAR)

Le mécanisme MNAR (*Missing Not At Random*) est le plus complexe à traiter, car la probabilité d'une donnée manquante dépend directement de la valeur de cette donnée ou d'une variable non observée.

Exemple : Dans une enquête sur les revenus, les personnes ayant un salaire très élevé peuvent refuser de le déclarer, créant une absence de données qui est directement liée à la valeur non observée.

On exprime cette typologie comme :

$$P(M \mid X) = P(M \mid X_{obs}, X_{mis}) \neq P(M \mid X_{obs})$$

ce qui signifie que même en connaissant toutes les valeurs observées, les valeurs manquantes restent liées à des informations inconnues ou à la valeur manquante elle-même.

Ce cas est problématique car aucune imputation classique ne peut totalement corriger le biais introduit. Il nécessite généralement des analyses de sensibilité, des hypothèses fortes ou des méthodes spécifiques pour reconstruire l'information absente.

2.2 Impact sur l'analyse statistique

Les données manquantes influencent profondément les résultats d'une analyse statistique. Leur présence peut modifier les estimations des paramètres, biaiser les conclusions et réduire la fiabilité des modèles prédictifs. L'ampleur de ces effets dépend du mécanisme sous-jacent (MCAR, MAR ou MNAR) et de la proportion de valeurs absentes. Comprendre ces impacts permet d'anticiper les risques et d'adopter des stratégies adaptées.

2.2.1 Effet sur les statistiques descriptives

La première conséquence des données manquantes concerne les mesures de tendance centrale et de dispersion. Si les valeurs manquantes ne sont pas réparties de manière aléatoire, elles peuvent décaler les moyennes, les médianes et d'autres indicateurs.

Considérons un exemple dans le tennis professionnel : supposons que l'on analyse les performances des joueurs sur une saison, mais que certains d'entre eux ne jouent pas certains tournois en raison de blessures ou de choix stratégiques. Si ces absences concernent majoritairement les meilleurs joueurs (préférant se retirer en fin de saison pour se préserver), alors les statistiques des matchs disputés donneront une fausse impression d'équilibre entre les joueurs.
Par exemple, si un compétiteur n'affronte jamais de joueurs du Top 10, son taux de victoire sera surestimé, le faisant paraître plus performant qu'il ne l'est réellement.

2.2.2 Impact sur l'inférence statistique

L'absence de certaines valeurs peut aussi fausser les estimations des paramètres et diminuer la puissance des tests statistiques.

- **Sous MCAR**, l'échantillon restant est toujours représentatif de la population globale, bien que l'estimation perde en précision.
- **Sous MAR ou MNAR**, le problème est plus compliqué : certains sous-groupes sont sous-représentés en information, ce qui fausse les estimations et peut conduire à des résultats trompeurs.

Prenons l'exemple d'une étude clinique cherchant à comparer l'effet de deux traitements. Si les patients ayant ressenti des effets secondaires abandonnent l'étude et que ces absences ne sont pas corrigées, l'efficacité du traitement peut être surévaluée.

Par ailleurs, les tests d'hypothèse nécessitent souvent que les données soient complètes. Une diminution du nombre d'observations réduit la puissance statistique, rendant plus difficile la détection d'effets réels.

2.2.3 Difficulté d'identification du mécanisme sous-jacent

L'un des plus grands défis liés aux données manquantes est qu'il est rarement possible d'identifier directement leur catégorisation.
On peut vérifier si les données sont MCAR à l'aide du **test de Little (1988)**, basé sur une statistique du χ^2. Un résultat non significatif suggère que les données sont manquantes complètement au hasard. En revanche, la distinction entre MAR et MNAR est beaucoup plus complexe, car elle nécessite une connaissance approfondie du contexte des données. Une stratégie consiste à collecter des variables supplémentaires susceptibles d'expliquer l'absence de certaines valeurs.

3 Méthodes classiques de gestion des incomplétudes

Les approches classiques pour traiter les valeurs manquantes reposent sur des solutions simples à appliquer, mais qui peuvent induire des biais et impacter la fiabilité des analyses. Ces méthodes sont répandues par leur facilité d'application. Néanmoins, leur efficacité dépend du mécanisme des données manquantes.

Nous présentons ici deux stratégies courantes : la suppression des observations ou variables incomplètes et l'imputation par des valeurs fixes.

3.1 Suppression des observations ou des variables

Une approche intuitive consiste à supprimer les observations ou les variables contenant des données incomplètes. Cette méthode, appelée **analyse des cas complets** (*Complete Case Analysis*), est souvent utilisée par défaut dans de nombreux logiciels statistiques. On distingue deux variantes principales :
- **Suppression complète** (*listwise deletion*) : cette approche élimine toute observation x_i contenant au moins une valeur manquante, induisant :

$$X_{\text{complet}} = \{x_i \in X \mid \sum_{j=1}^{d} m_{ij} = 0\}$$

où $m_{ij} = 1$ si x_{ij} est manquante et 0 sinon. Cela signifie que seules les lignes entièrement renseignées sont conservées.

- **Suppression partielle** (*pairwise deletion*) : plutôt que d'éliminer une observation entière, cette méthode conserve le contenu des informations en utilisant les valeurs disponibles pour chaque estimation. Par exemple, dans le calcul de la covariance entre deux variables X_j et X_k, seules les observations où ces deux variables sont présentes sont utilisées :

$$\hat{\text{Cov}}(X_j, X_k) = \frac{1}{n_{jk}} \sum_{i \in \mathcal{O}_{jk}} (x_{ij} - \bar{X}_j)(x_{ik} - \bar{X}_k)$$

où $\mathcal{O}_{jk}$ est l'ensemble des observations complètes pour les variables X_j et X_k, et n_{jk} leur nombre total.

Cette approche permet d'exploiter un maximum d'informations, mais peut biaiser les résultats si les valeurs manquantes sont liées aux données observées (cas MAR ou MNAR).

Prenons un cas concret dans l'analyse d'archives historiques. Supposons que l'on étudie l'évolution des prix immobiliers depuis le XIXe siècle. Les transactions anciennes sont souvent incomplètes, avec des informations manquantes sur la surface, l'état ou la localisation des biens. En supprimant ces observations (X_t), on réduit l'échantillon à X'_t, dont l'espérance conditionnelle devient :

$$E(X'_t \mid M_t) \neq E(X_t).$$

L'analyse repose donc principalement sur les ventes récentes, faussant l'interprétation des tendances historiques. On pourrait alors conclure, à tort, que les prix ont toujours été élevés, simplement parce que les biens anciens ont disparu des données.

3.2 Imputation par des valeurs fixes

Une autre approche classique consiste à remplacer chaque valeur manquante x_{ij} par une valeur fixe issue des données existantes. L'imputation peut être définie de la manière suivante :

$$x_{ij}^{\text{imputé}} = \begin{cases} x_{ij} & \text{si } m_{ij} = 0, \\ f(X_j) & \text{si } m_{ij} = 1. \end{cases}$$

où f représente une fonction d'imputation, souvent la **moyenne** ($\bar{X}_j = \frac{1}{n_j} \sum_i x_{ij}$) ou la **médiane** pour les variables quantitatives. À noter que la médiane est un choix plus pertinent que la moyenne en présence de valeurs extrêmes. Dans le cas des variables catégorielles, on peut attribuer la modalité la plus fréquente (**mode**).

Cette méthode présente deux avantages majeurs : elle est facile à mettre en œuvre et évite la perte de données liée à la suppression des observations. Toutefois, elle a des limites importantes. Tout d'abord, elle réduit artificiellement la variance des données : en attribuant une seule valeur à tous les points manquants, on lisse la distribution, ce qui peut fausser les résultats. De plus, cette approche détruit les relations entre variables, ce qui peut compromettre la validité des analyses.

Un exemple concret peut être trouvé dans le tennis : si certaines performances de joueurs manquent et qu'on les remplace par la moyenne des autres joueurs, cela néglige les écarts de niveau et empêche d'identifier les véritables talents ou les joueurs en difficulté.

4 Méthodes avancées d'imputation

Afin de mieux exploiter l'information disponible, des approches plus avancées ont été développées, permettant une estimation plus fine des valeurs manquantes en tenant compte des relations entre variables.

Dans cette partie, nous étudions trois méthodes avancées : l'imputation par les k-plus proches voisins (k-NN), l'imputation multiple par équations chaînées (MICE) et l'imputation par MissForest.

4.1 Imputation par k-plus proches voisins (K-NN)

L'imputation par les k-plus proches voisins repose sur l'hypothèse que les observations similaires (en caractéristiques observées) partagent des points communs. Lorsqu'une valeur est manquante dans une variable X_j pour un individu, elle peut être estimée à partir des observations les plus proches en termes de distance via les étapes suivantes :

1. Identifier l'ensemble des individus ayant une valeur observée dans la variable X_j.
2. Calculer les distances entre ces individus et celui ayant la donnée non observée.
3. Sélectionner les k observations les plus proches.
4. Estimer la valeur manquante en prenant la moyenne (si X_j est numérique) ou le mode (si X_j est catégorielle) des k plus proches voisins.

Cette méthode est non paramétrique, ce qui signifie qu'elle ne suppose aucun modèle probabiliste sous-jacent aux données.

4.1.1 Algorithme

L'imputation d'une valeur manquante x_{ij} se fait en identifiant les k observations les plus proches, définies par un critère de distance $d(i, i')$. La distance la plus couramment utilisée est la distance euclidienne :

$$d(i, i') = \sqrt{\sum_{j=1}^{q} (x_{ij} - x_{i'j})^2}$$

où q représente le nombre de variables complètement observées utilisées pour le calcul de la distance. Le calcul utilise les variables pour lesquelles les deux individus ont des valeurs non manquantes. D'autres distances sont parfois utilisées pour mieux gérer les échelles différentes ou les données mixtes (numériques et catégorielles).

Une fois les k plus proches voisins sélectionnés, l'imputation se fait comme suit :
- Pour une **variable numérique** X_j, la valeur imputée est donnée par la moyenne :

$$x_{ij}^{\text{imputé}} = \frac{1}{k} \sum_{i' \in \mathcal{V}_{ij}} x_{i'j} \qquad \text{(cas le plus courant)}$$

où $\mathcal{V}_{ij}$ est l'ensemble des k plus proches voisins de l'individu i pour la variable j.

- Pour une **variable catégorielle**, on choisit la modalité la plus fréquente :

$$x_{ij}^{\text{imputé}} = \arg\max_{c \in \mathcal{C}_j} \sum_{i' \in \mathcal{V}_{ij}} \mathbb{1}_{x_{i'j}=c}$$

où $\mathcal{C}_j$ est l'ensemble des catégories possibles de X_j.

Le choix du paramètre k est crucial pour garantir la robustesse de l'imputation. Une valeur trop faible peut rendre l'imputation instable, tandis qu'une valeur trop grande risque de lisser excessivement les données. Plusieurs approches de sélection sont proposées par les statisticiens, dont une courante consiste à choisir :

$$k = \sqrt{n_{\text{complet}}}$$

où n_{complet} est le nombre moyen d'observations complètes des variables utilisées.

4.1.2 Avantages et limites

Bien que la méthode soit flexible et exploratrice du dataset, elle présente quelques inconvénients :
- **Coût computationnel élevé** : Pour de grands jeux de données, le calcul des distances peut être long.
- **Sensibilité aux données aberrantes** : Un voisinage mal défini peut fausser l'imputation.
- **Choix de la distance** : Une distance inappropriée peut altérer la sélection des voisins.

Malgré ces limites, l'imputation par k-NN reste une méthode puissante et largement utilisée, en particulier lorsque les relations entre variables sont complexes et difficiles à modéliser avec des approches plus rigides.

4.2 Imputation multiple par équations chaînées (MICE)

L'imputation multiple par équations chaînées (MICE) est une méthode avancée qui permet de traiter les données manquantes en générant plusieurs jeux de données imputés. Contrairement aux méthodes déterministes, elle capture l'incertitude liée à l'imputation en proposant différentes versions des valeurs incomplètes.

L'idée centrale est de modéliser chaque variable avec des valeurs manquantes en fonction des autres variables. Plutôt que de remplacer définitivement une valeur absente par une seule estimation, MICE effectue plusieurs imputations pour chaque donnée manquante, permettant ainsi d'obtenir des résultats statistiquement plus robustes. Cette approche est particulièrement adaptée aux mécanismes MAR (Missing At Random), où les absences sont liées aux variables observées.

4.2.1 Algorithme

L'algorithme MICE repose sur un processus itératif en plusieurs étapes :

1. **Imputation initiale** : chaque valeur manquante est d'abord remplacée par une estimation simple (moyenne, médiane ou valeur aléatoire).

2. **Réinitialisation** : on bascule les valeurs précédemment imputées vers des données manquantes pour une seule variable.

3. **Imputation conditionnelle** : pour cette dernière variable X_j en question, on utilise un modèle statistique basé sur les autres variables observées X_{-j} pour estimer les valeurs absentes :

$$X_j^{(t)} \sim P(X_j \mid X_{-j})$$

où $X_j^{(t)}$ est la version imputée à l'itération t.

4. **Répétition pour toutes les variables** : chaque variable est imputée une à une, en tenant compte des imputations précédentes. Une itération complète correspond à un cycle d'imputation pour toutes les variables avec valeurs manquantes.

5. **Convergence** : ce processus est répété plusieurs fois jusqu'à stabilisation des imputations.

6. **Combinaison des résultats** : au terme de m imputations, les analyses sont effectuées sur chaque jeu de données imputé, puis les résultats sont réunis pour obtenir une estimation finale :

$$\bar{\theta} = \frac{1}{m} \sum_{t=1}^{m} \theta^{(t)}$$

où le paramètre $\theta^{(t)}$ désigne l'estimation de la valeur d'intérêt obtenue lors de l'itération t. La variance totale prend en compte deux variabilités :

$$T = V_{intra} + \left(1 + \frac{1}{m}\right) V_{inter}$$

où : - $V_{intra} = \frac{1}{m} \sum_{t=1}^{m} V^{(t)}$ est la variance intra-imputation.
- $V_{inter} = \frac{1}{m-1} \sum_{t=1}^{m} (\theta^{(t)} - \bar{\theta})^2$ est la variance inter-imputation.
Le terme $(1 + \frac{1}{m})$ s'ajoute afin d'inclure l'hypothèse que les estimations de θ ne proviennent que d'un nombre fini d'itérations m.

4.2.2 Avantages et limites

La méthode MICE permet une meilleure prise en compte de l'incertitude et évite les biais introduits par des imputations trop simplistes (comme le remplacement par la moyenne). De plus, elle offre des imputations cohérentes avec les relations entre variables.

En revanche, cette méthode est plus coûteuse en calcul que les approches classiques et nécessite de bien spécifier les modèles d'imputation sous-jacents. En particulier, si les relations entre variables sont complexes, des modèles simplistes peuvent introduire des erreurs. Enfin, le choix du nombre d'imputations m a une influence sur la qualité des résultats : en pratique, m est souvent choisi entre 5 et 20 imputations. Au-delà, les gains en précision sont négligeables.

4.3 Imputation par MissForest

MissForest est une méthode d'imputation non paramétrique basée sur les forêts aléatoires. Contrairement aux approches comme MICE ou k-NN, elle n'exige aucune relation entre variables.

L'algorithme suit une approche itérative en plusieurs étapes :

1. **Initialisation** : Remplacement des valeurs manquantes par une première estimation naïve (moyenne/médiane pour les variables numériques, mode pour les catégoriques) :

$$X_{ij}^{(0)} = \begin{cases} x_{ij}, & \text{si } (i,j) \notin M, \\ \frac{1}{|S_j|} \sum_{i \in S_j} x_{ij}, & \text{si } (i,j) \in M, \end{cases} \tag{1}$$

   où M est l'ensemble des indices des valeurs manquantes et S_j l'ensemble des observations disponibles pour la variable j.
2. **Ordonnancement des variables** : Classement des variables selon leur taux de valeurs manquantes en ordre croissant.
3. **Modélisation séquentielle** : Pour chaque variable X_s contenant des valeurs manquantes, réinitialisation des valeurs imputées à "manquantes" et ajustement d'un modèle de forêt aléatoire f_s sur les données complètes (X_{-s}) pour prédire X_s :

$$f_s : X_{-s}^{\text{obs}} \to X_s^{\text{obs}} \tag{2}$$

4. **Prédiction et mise à jour** : Remplacement des valeurs manquantes via une fonction de prédiction $\hat{f}_s$ issue du modèle :

$$X_s^{\text{mis}} = \hat{f}_s(X_{-s}^{\text{mis}}) \tag{3}$$

5. **Itération** : Répétition du processus pour toutes les variables contenant des valeurs manquantes jusqu'à atteindre une itération complète.
6. **Convergence** : Répétition du processus itératif jusqu'à ce que la différence entre les deux dernières matrices imputées cesse de diminuer :

$$\Delta(X^{(t)}, X^{(t-1)}) < \epsilon \tag{4}$$

   où $\Delta(\cdot, \cdot)$ est une mesure de dissimilarité entre les deux matrices imputées successives et ϵ un seuil fixé.

Avantages et limites MissForest est efficace pour les données MAR et s'adapte bien aux structures complexes, tout en conservant les distributions statistiques d'origine. Il est particulièrement utile lorsque les relations entre variables sont non linéaires et qu'une approche non paramétrique est préférable. Son principal inconvénient est son coût computationnel plus élevé que les méthodes plus simples (k-NN, MICE), en particulier pour les grandes bases de données. Stekhoven et Bühlmann (2012) ont montré que MissForest surpasse les approches classiques en termes de précision d'imputation, réduisant l'erreur jusqu'à 50% sur divers jeux de données publics.

5 Applications pratiques sur des jeux de données réels

5.1 Simulation et imputation de données sportives incomplètes

Nous avons également manipulé un jeu de données sur le thème du sport, répertoriant le nombre de médailles de chaque métal ainsi que le nombre total de médailles obtenues lors des Jeux Olympiques de Paris en 2024, par chaque pays ayant remporté au moins une médaille. On y trouve aussi, pour chaque pays, la population, ainsi que le PIB (GDP en anglais) et l'année au cours de laquelle le PIB a été mesuré.

Une première brève étude a montré que le jeu de données était complet, et que 90 pays ont obtenu au moins une médaille lors des Jeux Olympiques 2024, pour un total de 328 médailles d'or, 327 médailles d'argent et 383 médailles de bronze remportées (dans certaines disciplines, plusieurs athlètes peuvent obtenir une médaille du même métal).

Le jeu étant complet, nous avons ensuite introduit les données manquantes nous-mêmes. Cela étant dit, la complétude du jeu de données permet l'évaluation de chaque méthode d'imputation implémentée, en comparant les valeurs imputées aux valeurs initialement présentes dans le jeu de données. L'introduction des données manquantes a été réalisée de la manière suivante : définition de 10 % de valeurs manquantes pour chacune des 3 variables représentant le nombre de médailles d'or, d'argent et de bronze obtenues, réparties aléatoirement parmi les pays présents dans la table. Pour chaque pays pour lequel il manque la donnée du nombre de médailles obtenues d'au moins un métal, on introduit une valeur manquante au nombre total de médailles obtenues. Ainsi, la répartition des valeurs manquantes suit un cadre MAR, ce qui permet d'utiliser la plupart des méthodes d'imputation.

Après l'introduction des valeurs manquantes, nous avons procédé à leur imputation, d'abord par la méthode des k-plus proches voisins, puis par la méthode MICE (méthode d'imputation par équations chaînées) et via MissForest.

La précision des méthodes a été calculée selon deux approches différentes :

- Proportion de valeurs imputées avec la bonne valeur, pour chaque méthode d'imputation ;
- Précision calculée par pourcentage pour chaque valeur à imputer, puis calcul de la moyenne de ces précisions, à nouveau pour chaque méthode d'imputation.

Après quelques simulations, il est apparu que le faible nombre de variables empêche une réelle stabilité des méthodes d'imputation. Afin que nos résultats ne soient pas excessivement biaisés par le schéma de répartition des données manquantes, nous avons procédé à une série de 25 simulations de chaque méthode, en générant à chaque reprise une nouvelle distribution des valeurs manquantes. Les calculs des précisions sont alors basés sur la moyenne des précisions observées lors de chacune des 25 répétitions.

La première méthode de calcul de précision a fournit les résultats suivants :

- 20,4 % de valeurs correctement imputées par méthode des 2-plus-proches voisins, avec une distance basée sur le nombre de médailles obtenues de chaque métal ;
- 20,4 % de valeurs correctement imputées par méthode des 2-plus-proches voisins, avec une distance basée sur le nombre de médailles obtenues de chaque métal ainsi que la population ;
- 24,5 % de valeurs correctement imputées avec la méthode d'imputation par équations chaînées ;
- 26,5 % de valeurs correctement imputées suivant la méthode d'imputation MissForest.

Il est à noter que la méthode d'imputation MissForest se distingue nettement des autres grâce à sa capacité supérieure à imputer correctement les valeurs manquantes.

La seconde méthode de calcul de la précision, basée davantage sur la cohérence de la valeur imputée avec la valeur réelle (une valeur incorrectement imputée mais proche de celle réelle est moins pénalisante qu'une valeur très éloignée), donne les résultats suivants :

Méthode	2-nn	2-nn avec population	MICE	MissForest
Précision	37,6 %	40,9 %	46,6 %	56,5 %

Là encore, la méthode d'imputation MissForest se distingue des autres méthodes : les valeurs prédites sont en moyenne nettement plus proches des valeurs réelles que suite à une imputation par une autre méthode.

5.2 Imputation de données météorologiques manquantes

Dans cette section, nous avons cherché à imputer des données manquantes sur un ensemble de trois jeux de données provenant du site du gouvernement du Canada, qui représentent les données météorologiques enregistrées par la station météo Montréal INTL A entre le premier janvier 2022 et le 31 décembre 2024.
La particularité de ce jeu de données est qu'il représente des données continues sur des séries temporelles. En particulier, on a pris 3 années consécutives pour pouvoir observer la saisonnalité.

Le jeu de données est composé de 31 variables et de 1096 individus. Sur ces 31 variables, 11 sont des variables "indicatrices" qui donnent parfois des précisions sur la nature de la donnée : M si elle est manquante ; T si on a observé une trace, c'est-à-dire qu'on a observé de la pluie mais en trop petite quantité pour être mesurée.
Au final, il n'y a que 11 colonnes avec des variables quantitatives ainsi que 3 colonnes *date*, on ne conservera que la variable "Date.Heure" qui donne la date complète et dont on se servira comme indice de temps.

Les variables sont donc :

1. - Temp max.(°C) : la température maximale enregistrée au jour j.

2. - Temp min.(°C) : la température minimale enregistrée au jour j.

3. - Temp moy.(°C) : la température moyenne enregistrée au jour j.

4. - DJC (°C) : la différence entre 18 et Temp moy.(°C) (0 si < 0).

5. - DJR (°C) : la différence entre Temp moy.(°C) et 18 (0 si < 0).

6. - Pluie tot. (mm) : la totalité des précipitations liquides recensées le jour j.

7. - Neige tot. (cm) : la totalité des précipitations solides recensées le jour j.

8. - Précip. tot. (mm) : la somme des deux colonnes précédentes.

9. - Neige au sol (cm) : la moyenne de mesures de l'épaisseur de la neige au sol à un moment t.

10. - Dir. raf. max. (10s deg) : la direction d'où provenait la rafale de vent la plus forte enregistrée le jour j.

11. - Vit. raf. max. (km/h) : la vitesse de la plus forte rafale de vent enregistrée le jour j.

DJC, *DJR*, et *Précipitation totale* ne sont que la somme ou la différence d'autres variables. On ne va ainsi pas les conserver dans l'étude afin de ne pas fausser les résultats.
On ne garde pas non plus la variable *Neige au sol* car celle-ci est majoritairement vide entre avril et octobre, ce qui fausse le nombre de données manquantes.
Enfin, les deux colonnes relatives au vent ont beaucoup de données manquantes. Cependant, la majorité de celles-ci est due au fait que la donnée n'est pas recensée pour les rafales inférieures à 30 km/h. Lors d'un essai d'imputation, ce paramètre n'a pas réussi à être pris en compte, c'est pourquoi nous n'avons pas non plus conservé ces deux variables.

On a donc 5 variables continues qui représentent les températures et les précipitations journalières pendant 3 années consécutives. Le jeu contient déjà des données manquantes naturellement pour une raison inconnue. On va alors regarder comment se déroule l'imputation dans le cas d'une régression sur une série temporelle.

A l'aide de la fonction *vis_miss* du package *naniar* du logiciel R on peut obtenir une visualisation de la répartition des données manquantes :

On constate en particulier que très peu de données sont manquantes, et qu'elles sont réparties de manière aléatoire. La littérature conseille, dans le cas où il y a peu de données manquantes, de les supprimer. Comme nous étudions des données météorologiques journalières, nous ne voulons pas appliquer cette solution mais bien chercher à imputer ces données. Nous allons alors appliquer des méthodes d'imputation classiques et comparer les résultats avec ceux trouvés par des méthodes d'imputation avancées.

Comme nous ne connaissons pas la valeur des données, on ne peut utiliser de critère pour valider notre imputation. Nous avons d'abord comparé les statistiques descriptives avant et après imputation pour vérifier que l'imputation n'est pas biaisée et qu'elle est cohérente avec les données déjà présentes. Cependant, comme le nombre de valeurs manquantes est trop petit, aucune des méthodes n'a significativement fait varier les statistiques descriptives. Nous avons alors comparé graphiquement l'évolution de la série pour voir si les données étaient cohérentes avec celles qui en étaient proches dans le temps. Nous avons également comparé les corrélations entre les variables, les distributions et la moyenne mobile pour voir si la série après imputation de nos données est cohérente avec la série d'origine.

Au final, nous avons trouvé que les méthodes MICE et Miss Forest renvoient des données peu cohérentes (on peut voir à l'œil nu de forts écarts de température entre la donnée imputée et les données proches), de même pour l'imputation par la moyenne.
En revanche, on trouve des données plutôt cohérentes avec la méthode Arima qui cherche de façon automatique le meilleur modèle ARIMA, avec une moyenne mobile sur une période d'une semaine, et avec la méthode k-NN sur les 5 jours les plus proches.
Cela se justifie par le fait que les méthodes ARIMA et k-NN exploitent la structure temporelle des données, contrairement à MICE et MissForest qui traitent chaque observation indépendamment.

Ainsi, lorsqu'il y a peu de données manquantes, les méthodes les plus efficaces pour étudier un cas de série temporelle sont les méthodes prenant en compte les individus proches dans le temps pour imputer les données.

6 Conclusion et perspectives

Ce travail a permis d'explorer différentes approches pour identifier, traiter et imputer des valeurs absentes, en mettant en lumière les avantages et les limites de chaque méthode.

L'étude a d'abord rappelé les trois principaux mécanismes sous-jacents aux données manquantes (*MCAR, MAR, MNAR*), qui guident sur le choix des stratégies les plus adaptées à leur gestion. Ensuite, nous avons comparé des approches classiques, comme la suppression des observations incomplètes ou l'imputation par la moyenne, avec des techniques plus avancées telles que MICE, MissForest et k-NN. Les résultats ont clairement montré que le choix de la méthode dépend du contexte des données et de la structure des valeurs manquantes.

Dans le cas des séries temporelles météorologiques, les méthodes exploitant la proximité temporelle, comme ARIMA et k-NN, ont donné des résultats plus cohérents que les approches plus généralistes comme MissForest. À l'inverse, pour les données sportives, où les relations entre variables sont plus complexes, l'imputation par forêts aléatoires s'est révélée plus efficace, surpassant nettement les autres méthodes en termes de précision.

Ces observations soulignent l'importance d'une approche réfléchie et contextualisée lorsqu'il s'agit de traiter un problème ayant une base de données incomplète. Aucune méthode n'est universelle : l'efficacité d'une approche dépend non seulement de la nature des données, mais aussi de leur structure et du mécanisme sous-jacent de l'absence d'information. Une mauvaise gestion des valeurs manquantes peut induire des biais significatifs et nuire à la qualité des analyses.

En perspective, plusieurs pistes peuvent être envisagées pour affiner ces résultats. Tout d'abord, l'exploration de techniques hybrides, combinant plusieurs méthodes d'imputation, pourrait permettre d'améliorer la fiabilité des estimations.
Une autre piste, plus technique, serait d'intégrer un processus d'évaluation automatique. Celui-ci analyserait d'abord les caractéristiques du jeu de données (série temporelle, variables corrélées, présence de valeurs extrêmes, etc.) puis appliquerait la méthode d'imputation la plus pertinente en fonction de ces éléments. Une telle approche permettrait de maximiser la cohérence des valeurs imputées sans nécessiter d'expertise avancée pour choisir manuellement la méthode. Ce type d'algorithme adaptatif semble particulièrement pertinent pour des analyses à grande échelle, où la diversité des données rend difficile l'application d'une seule et même stratégie.

Références

[1] Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

[2] Wikipédia. *Données manquantes, Imputation (statistique).*

[3] Crettaz de Roten, F. & Helbling, J.-M. (1996). *Données manquantes et aberrantes : le quotidien du statisticien analyste de données.* Revue de Statistique Appliquée, tome 44 (n°2), p. 105-115.

[4] Imbert, A. & Vialaneix, N. (2018). *Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes.* Journal de la Société Française de Statistique, 159 (2), pp.1-55.

[5] Cnam – UE RCP208. *Cours - Données manquantes.*

[6] Wikistat. *Imputation de données manquantes.*

[7] Stekhoven, D. J. & Bühlmann, P. (2012). *MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics, Volume 28, Issue 1, p. 112-118.

[8] Yosef, M. (2024). *2024 Olympics Medals and Economic Status.* Kaggle. Disponible via le lien : Dataset - Olympics 2024

[9] NAVCAN. (2022, 2023, 2024). *Rapport de données quotidiennes - Montréal, Québec.* Environnement Canada.
Disponible via le lien : Dataset Météo - Montréal