FACULTÉ DES SCIENCES ET TECHNOLOGIES | Université de Lille

Le problème des données manquantes

Reda Mdair - Victor Naninck - Théo Dugauguez

Sous la direction de Charlotte Baey

4 Mars 2025

Sommaire

Mécanismes des données manquantes

Méthodes classiques de gestion des incomplétudes

Méthodes avancées d'imputation

Applications à des jeux de données réels

Conclusion et perspectives

Introduction

Introduction

Une idée naïve : croire que les jeux de données sont toujours complets et structurés.

- En pratique, **les données sont souvent incomplètes**.
- Un traitement inadapté des données manquantes peut conduire à **des biais et des conclusions erronées**.

Mécanismes des données manquantes

Pourquoi s'intéresser au type de données manquantes ?

Comprendre l'origine des valeurs manquantes est essentiel pour choisir une méthode de traitement adaptée.

- **Typologie** : Trois mécanismes principaux expliquent l'apparition des valeurs manquantes :
  - 1. MCAR : *Missing Completely At Random*
  - 2. MAR : *Missing At Random*
  - 3. MNAR : *Missing Not At Random*

Modélisation des valeurs manquantes

Définition formelle : Soit une matrice de données X de taille $n \times d$, avec :

- X_{obs} les valeurs observées.
- X_{mis} les valeurs manquantes.
- M une matrice indicatrice telle que :

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

Objectif : Comprendre comment M est générée afin d'adopter la meilleure stratégie de traitement.

1. MCAR : Missing Completely At Random

Définition : Dans le cas MCAR, les valeurs sont absentes **de manière totalement aléatoire**, indépendamment des autres variables.

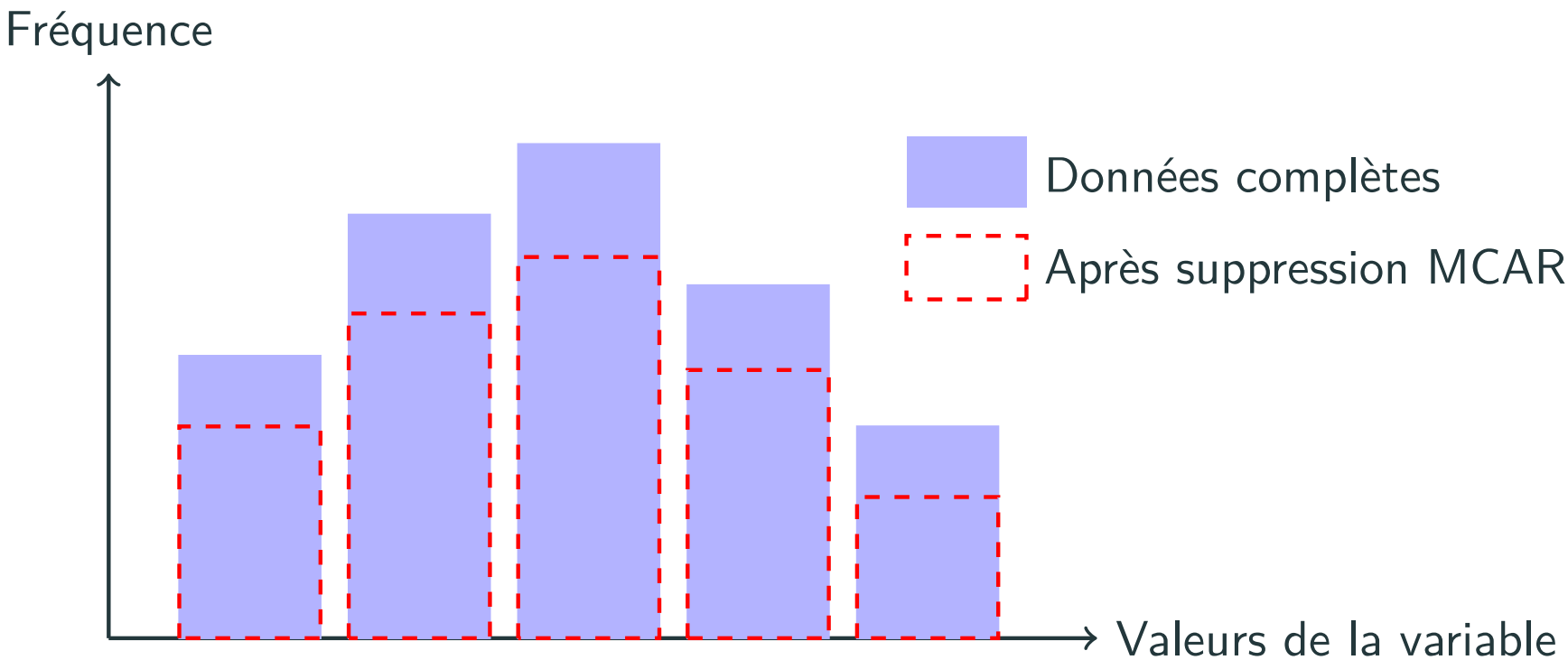
$$P(M \mid X) = P(M)$$

Conséquences :

- Les observations restantes sont un échantillon représentatif non biaisé.
- Suppression des valeurs manquantes possible sans introduire de distorsion.

1. MCAR : Missing Completely At Random

Visualisation typique de données manquantes MCAR



2. MAR : Missing At Random

Définition : Dans le cas MAR, la probabilité d'absence d'une donnée dépend uniquement des **valeurs observées**.

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

Enquête de satisfaction

Les clients ayant effectué peu d'achats sont moins susceptibles d'y répondre. L'absence de réponse dépend d'une variable observée (le nombre d'achats).

Conséquences :

- L'absence d'une valeur peut être prédite à partir des variables observées.
- Utilisation possible de méthodes d'imputation avancées.

3. MNAR : Missing Not At Random

Définition : Dans le cas MNAR, la probabilité d'une valeur manquante dépend de la valeur elle-même ou d'une variable **non observée**.

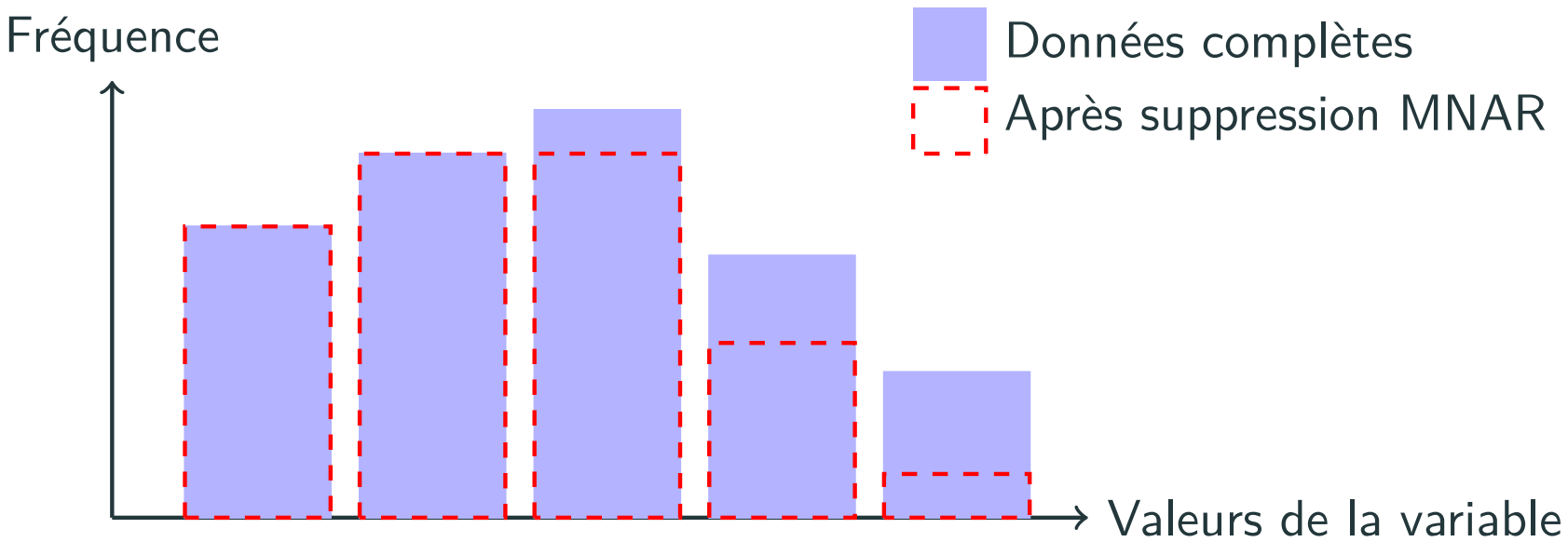
$$P(M \mid X) = P(M \mid X_{obs}, X_{mis}) \neq P(M \mid X_{obs})$$

Conséquences :

- **Problématique** : Le processus de génération des valeurs manquantes est biaisé.
- **Difficile à traiter** sans hypothèses supplémentaires ou modèles spécifiques.

Visualisations typiques de données manquantes MNAR

- **Données manquantes aux extrêmes droits**

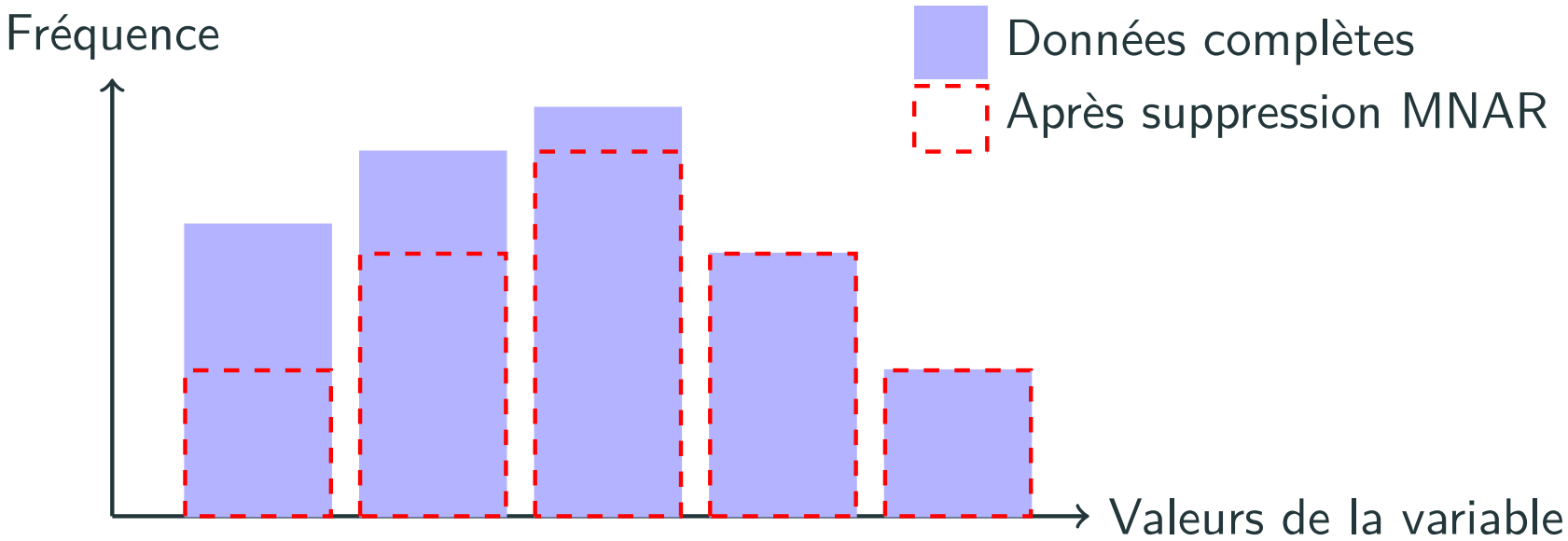


Exemple

Dans un sondage sur les revenus, les personnes ayant un gros salaire sont susceptibles de ne pas répondre.

3. MNAR : Missing Not At Random

- **Données manquantes aux extrêmes gauches**



Exemple

Dans une enquête sur la satisfaction des clients sur un service, ceux qui sont insatisfaits peuvent ne pas vouloir investir de leur temps pour y répondre.

Effet sur les statistiques descriptives

Conséquences sur les mesures de tendance centrale et de dispersion :

- Les moyennes, médianes et écarts-types peuvent être biaisés si les absences ne sont pas aléatoires.
- Si certaines catégories sont plus touchées que d'autres, les statistiques globales sont faussées.

Effet sur l'inférence statistique

Conséquences sur les estimations et tests statistiques :

- **Sous MCAR** : L'échantillon reste représentatif, mais la précision diminue.
- **Sous MAR/MNAR** : Les paramètres peuvent être biaisés, faussant les conclusions.

Exemple

Dans une étude clinique, si les patients ressentant des effets secondaires abandonnent l'étude, l'efficacité du traitement risque alors d'être surestimée.

Difficulté d'identification du mécanisme

Problème : il est souvent impossible de savoir si une donnée est MAR ou MNAR.

Outils de détection :

- **Le test de Little (1988)** : basé sur une statistique χ^2, permet de vérifier si les données sont MCAR.
- **Approche exploratoire** : Collecter des variables supplémentaires pour comprendre les absences.

Même avec des tests, il reste difficile de distinguer MAR et MNAR sans hypothèses fortes.

Méthodes classiques de gestion des incomplétudes

Méthodes classiques de gestion des données manquantes

Caractéristiques des méthodes classiques :

- Faciles à mettre en œuvre, largement utilisées en pratique.
- Peuvent cependant introduire des **biais** et réduire la fiabilité des analyses.
- Leur efficacité dépend du **mécanisme des valeurs manquantes** (MCAR, MAR, MNAR).

Stratégies abordées :

1. Suppression des observations.
2. Imputation par des valeurs fixes.

Suppression des observations

Principe : Éliminer les données incomplètes pour ne conserver que les observations complètes.

Méthodes :

- **Suppression complète (listwise deletion)** :
  $\rightarrow$ Seules les observations entièrement renseignées sont conservées.
- **Suppression partielle (pairwise deletion)** :
  $\rightarrow$ Exploite un maximum d'informations en utilisant les observations incomplètes, sauf celles ayant une case manquante pour une variable importante.

Problème : Peut biaiser l'analyse si les valeurs manquantes ne sont pas MCAR.

Imputation par des valeurs fixes

Principe : Remplacer chaque valeur manquante x_{ij} par une valeur déterminée à partir des données existantes.

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

Choix fréquents de $f(X_j)$:

- **Moyenne** : $\bar{X}_j = \frac{1}{n_j} \sum_i x_{ij}$
- **Médiane** : Plus robuste aux valeurs extrêmes.
- **Mode** : Pour les variables catégorielles.

Avantages : facile à mettre en œuvre, évite la suppression des observations.

Limites de l'imputation par des valeurs fixes

Pourquoi cette méthode peut poser problème ?

- **Réduction artificielle de la variance** :
  - Attribution d'une même valeur à plusieurs observations.
- **Altération des relations entre variables** :
  - L'imputation ne prend pas en compte les interactions entre les variables.

Performance au tennis

Si les performances de certains joueurs sont manquantes et qu'on les remplace par la moyenne de tous les joueurs, les écarts de niveau sont effacés, empêchant d'identifier les talents ou les joueurs en difficulté.

Méthodes avancées d'imputation

Méthodes avancées d'imputation

Pourquoi utiliser des méthodes avancées ?

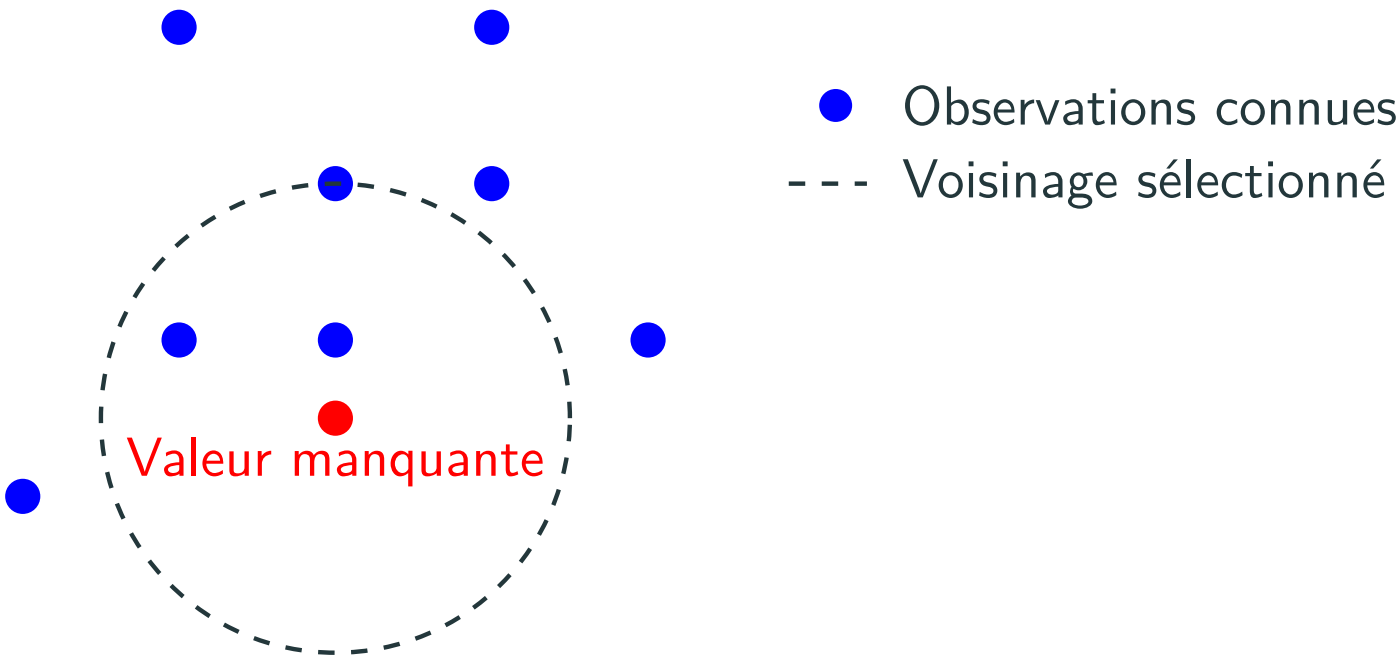
- Les méthodes classiques (suppression, imputation par moyenne) sont souvent trop simplistes.
- Les méthodes avancées exploitent **les relations entre variables** pour une meilleure estimation.

Trois méthodes étudiées :

1. **Imputation par les k-plus proches voisins (k-NN)**.
2. **Imputation multiple par équations chaînées (MICE)**.
3. **Imputation par MissForest**.

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

Principe : Les individus similaires partagent des caractéristiques communes.



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

Étapes de l'algorithme :

1. Identifier les individus ayant une valeur non observée dans la variable cible X_j.
2. Calculer la distance entre l'individu à imputer et les autres.
3. Sélectionner les k plus proches voisins (choix optimal : $k = \sqrt{n_{full}}$ où n_{full} est le nombre d'observations complètes).
4. Imputer la valeur manquante :
   - **Valeur définie en fonction des valeurs observées** chez les k plus proches voisins (si X_j est numérique).
   - **Modalité la plus fréquente** (si X_j est catégorielle).

Avantages et limites du k-NN

Avantages :

- Approche **non paramétrique** (aucune hypothèse sur la distribution des données).
- Exploite les similarités entre individus pour une estimation plus précise.
- Applicable aux **données mixtes** (numériques et catégorielles).

Limites :

- **Coût computationnel élevé** pour de grands jeux de données.
- **Sensibilité aux valeurs aberrantes**.
- **Dépendance au choix de la distance** et du paramètre k.

Imputation multiple par équations chaînées (MICE)

Pourquoi MICE ?

- Contrairement aux imputations déterministes, MICE capture **l'incertitude** liée aux valeurs manquantes.
- Génère **plusieurs versions** des données imputées pour obtenir des analyses plus robustes.
- Particulièrement adapté aux données **MAR** (Missing At Random).

Idée principale :

- Chaque variable avec des valeurs manquantes est modélisée en fonction des autres variables.
- L'imputation est réalisée en plusieurs étapes itératives.

Imputation multiple par équations chaînées (MICE)

Étapes de l'algorithme :

1. **Imputation initiale** : Remplacement des valeurs manquantes par une estimation simple (*moyenne, médiane, ou valeur aléatoire*).
2. **Réinitialisation** : Une seule variable est sélectionnée à chaque étape.
3. **Imputation conditionnelle** : La variable X_j est imputée selon un modèle statistique choisi au préalable.
4. **Répétition des étapes 2 et 3** : Chaque variable est imputée une par une, en utilisant les valeurs mises à jour.
5. **Convergence** : Le processus continue jusqu'à stabilisation des imputations.
6. **Combinaison des résultats** : Moyenne des imputations obtenues à chaque étape.

Imputation multiple par équations chaînées (MICE)

Avantages :

- Capture **l'incertitude** liée aux données manquantes.
- **Conserve les relations entre variables**, contrairement aux imputations simples.
- Flexible : différents modèles peuvent être utilisés pour différentes variables.

Limites :

- **Coût computationnel élevé** : multiple imputations nécessitent du temps de calcul.
- **Dépendance aux modèles d'imputation** : un modèle mal spécifié peut biaiser les résultats.
- **Choix du nombre d'imputations** m : souvent fixé entre 5 et 20 imputations pour un compromis entre précision et efficacité.

Imputation par MissForest

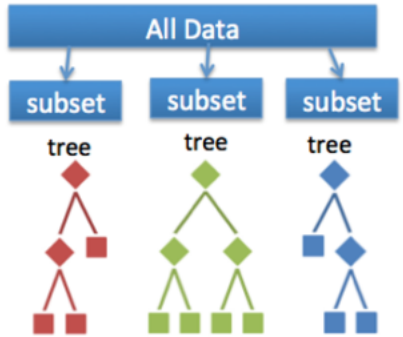
Pourquoi MissForest ?

- Méthode **non paramétrique** basée sur les **forêts aléatoires**.
- Applicable aux **données mixtes** (numériques et catégorielles).

Approche :

- Apprend à prédire les valeurs manquantes en utilisant les variables complètes.
- Répète ce processus itérativement jusqu'à convergence.

MICE +



$\sim$

MissForest

L'algorithme MissForest diffère de MICE lors de l'imputation conditionnelle : au lieu d'utiliser des modèles statistiques classiques (régression linéaire ou logistique), il s'appuie sur des forêts aléatoires, capturant mieux les relations complexes et les interactions non linéaires entre variables.

Avantages et limites de MissForest

Avantages :

- S'adapte aux **structures complexes** des données (relations non linéaires).
- Ne nécessite **aucune hypothèse** sur la distribution des données.

> Stekhoven et Bühlmann (2012) ont montré que MissForest réduit l'erreur d'imputation jusqu'à 50% par rapport aux méthodes classiques.

Limites :

- Performances réduites si la proportion de valeurs manquantes est très élevée.

Applications à des jeux de données réels

1/ Simulation et imputation de données sportives

Simulation et imputation de données sportives

- Jeu de données complet.
- Répertorie le nombre de médailles obtenues de chaque métal, le total, le PIB et la population ($\sim$ nombre d'habitants).

Aperçu des premières observations

	country	country_code	gold	silver	bronze	total	gdp	gdp_year	population
1	United States	USA	40	44	42	126	81695.19	2023	334.9
2	China	CHN	40	27	24	91	12614.06	2023	1410.7
3	Japan	JPN	20	12	13	45	33834.39	2023	124.5
4	Australia	AUS	18	19	16	53	64711.77	2023	26.6
5	France	FRA	16	26	22	64	44460.82	2023	68.2

- **Objectif** : introduire aléatoirement 10% de données manquantes sur le nombre de médailles entre les différents pays puis les imputer par diverses méthodes.

Quelques statistiques

Catégorie	Moyenne	Médiane
Gold	3.644	1
Silver	3.633	1
Bronze	4.256	2
Total	11.53	5

On remarque les limites des méthodes classiques d'imputation pour ce contexte. Par exemple, imputer les médailles d'or des USA et de la Chine par la moyenne (=3.644) ou la médiane (=1) serait très loin du compte.

Simulation et imputation de données sportives

	ctry	gold	silv	bron	tot
1	USA	40	44	42	126
2	CHN	40	27	24	91
3	JPN	20	12	13	45
4	AUS	18	19	16	53
5	FRA	16	26	22	64
6	NLD	15	7	12	34
7	GBR	14	22	29	65
8	KOR	13	9	10	32
9	ITA	12	13	15	40
10	DEU	12	13	8	33

Valeurs réelles à imputer

	ctry	gold	silv	bron	tot
1	USA	40	44	42	126
2	CHN	40	27	24	91
3	JPN	20	4	13	37
4	AUS	18	19	16	53
5	FRA	16	10	22	48
6	NLD	15	7	12	34
7	GBR	14	22	29	65
8	KOR	13	9	10	32
9	ITA	2	13	3	18
10	DEU	12	13	8	33

2-NN

	ctry	gold	silv	bron	tot
1	USA	40	44	42	126
2	CHN	40	27	24	91
3	JPN	20	4	13	37
4	AUS	18	19	16	53
5	FRA	16	20	22	58
6	NLD	15	7	12	34
7	GBR	14	22	29	65
8	KOR	13	9	10	32
9	ITA	1	13	3	17
10	DEU	12	13	8	33

2-NN Pop.

	ctry	gold	silv	bron	tot
1	USA	40	44	42	126
2	CHN	40	27	24	91
3	JPN	20	1	13	34
4	AUS	18	19	16	53
5	FRA	16	13	22	51
6	NLD	15	7	12	34
7	GBR	14	22	29	65
8	KOR	13	9	10	32
9	ITA	8	13	2	23
10	DEU	12	13	8	33

MICE

	ctry	gold	silv	bron	tot
1	USA	40	44	42	126
2	CHN	40	27	24	91
3	JPN	20	11	13	44
4	AUS	18	19	16	53
5	FRA	16	20	22	58
6	NLD	15	7	12	34
7	GBR	14	22	29	65
8	KOR	13	9	10	32
9	ITA	13	13	10	36
10	DEU	12	13	8	33

MissForest

Deux méthodes d'évaluation de l'imputation utilisées :

- exactitude (proportion de données correctement imputées)
- précision (proximité moyenne avec les données réelles)

Résultats obtenus selon la répartition des données manquantes

	country	country_code	gold	silver	bronze	total
1	United States	USA	*NA*	44	*NA*	*NA*
2	China	CHN	40	27	*NA*	*NA*
3	Japan	JPN	*NA*	12	13	*NA*
4	Australia	AUS	18	19	16	53
5	France	FRA	16	*NA*	22	*NA*

→

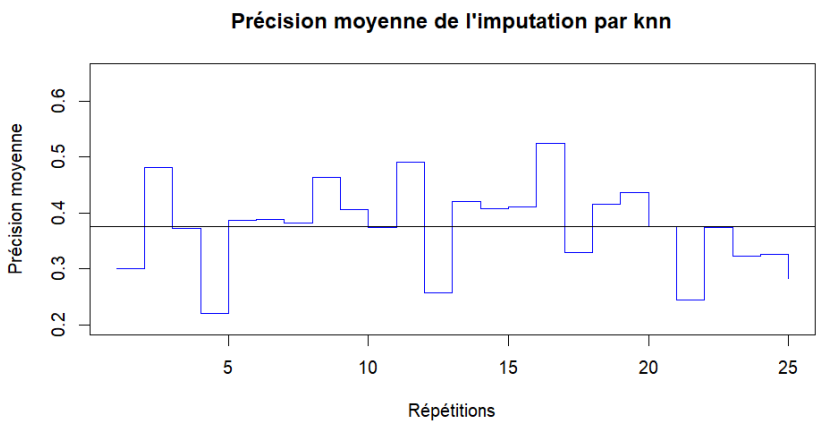
	2-nn	2-nn avec population	MICE	MissForest
Précision méthode 1	2,1 %	6,3 %	12,5 %	10,4 %
Précision méthode 2	31,4 %	29,4 %	47,7 %	53,9 %

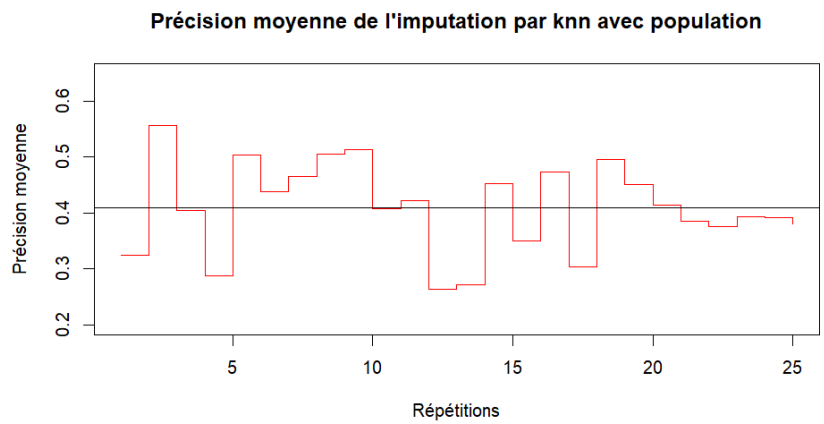
	country	country_code	gold	silver	bronze	total
1	United States	USA	40	44	42	126
2	China	CHN	40	27	24	91
3	Japan	JPN	20	12	13	45
4	Australia	AUS	18	19	16	53
5	France	FRA	16	26	22	64
6	Netherlands	NLD	15	7	*NA*	*NA*

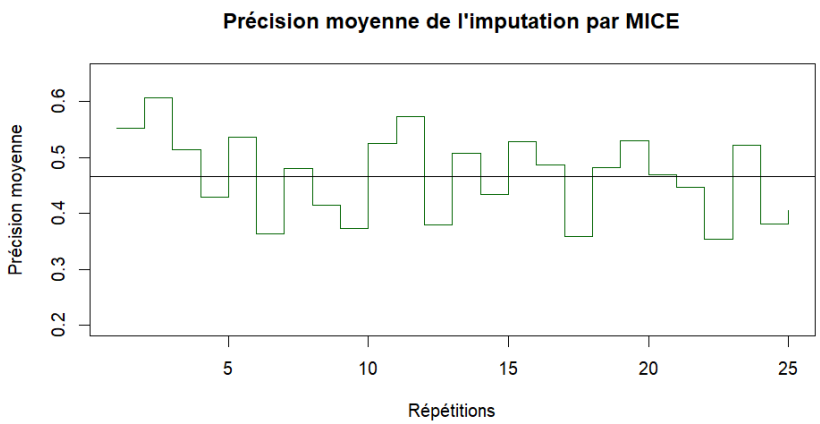
→

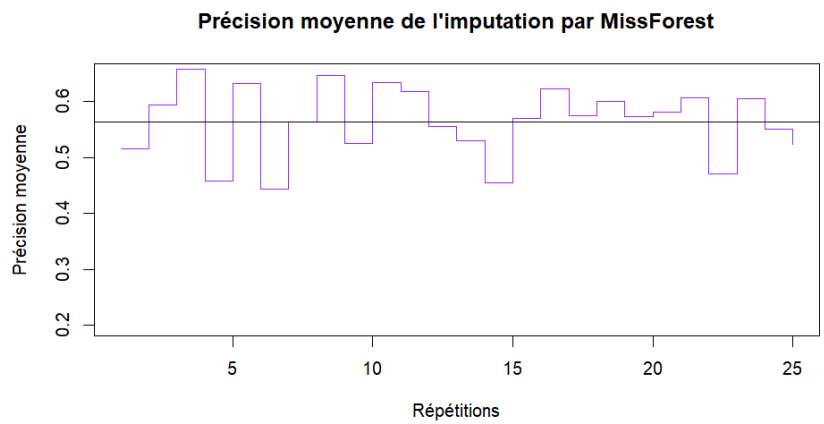
	2-nn	2-nn avec population	MICE	MissForest
Précision méthode 1	10 %	18,3 %	20 %	20 %
Précision méthode 2	36,6 %	42,6 %	47,2 %	49,6 %

Étude de la stabilité des méthodes sur la précision en réitérant les imputations









2/ Imputation de données météorologiques manquantes

Présentation de la base de données météorologique

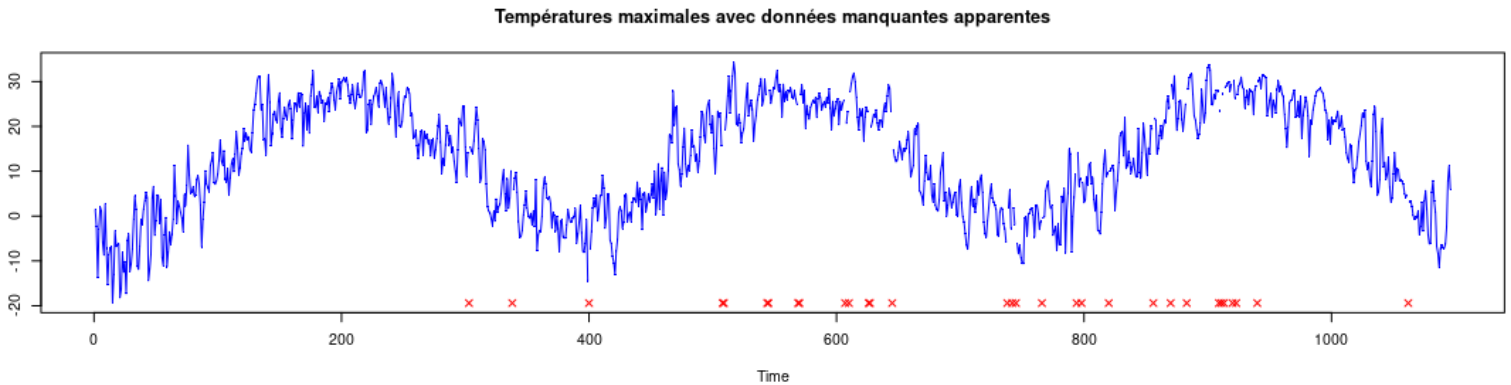
Caractéristiques générales

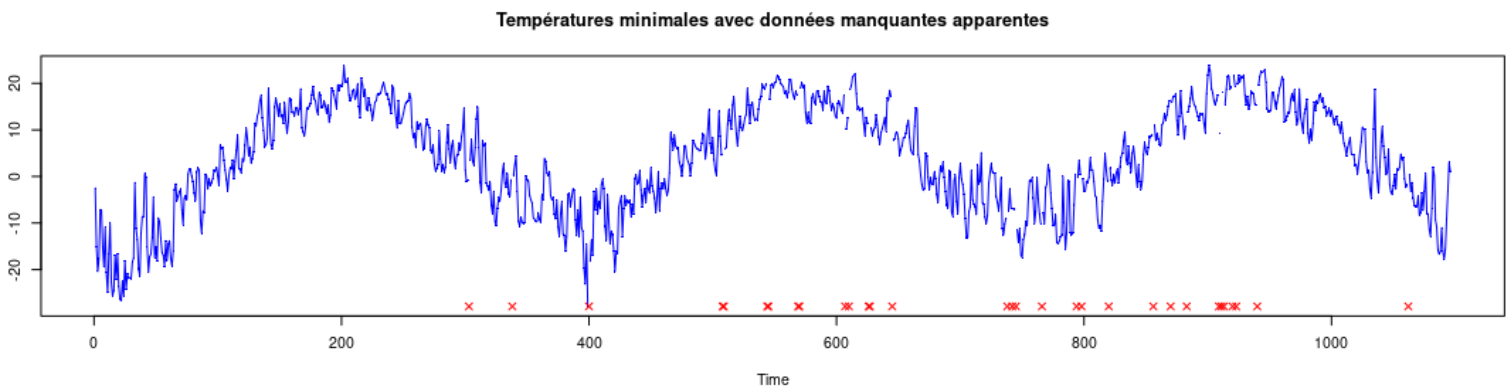
- **Nombre d'observations** : 976
- **Fréquence des données** : journalière (sur 3 années)
- **Type des variables** : continu

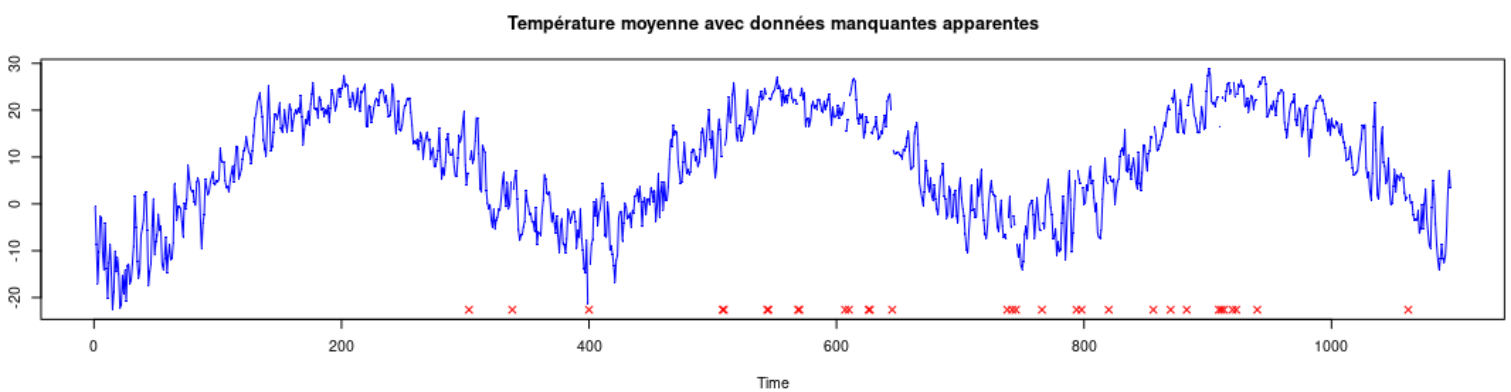
Variables enregistrées

1. Température maximale (°C)
2. Température minimale (°C)
3. Température moyenne (°C)
4. Précipitations liquides totales (mm)
5. Précipitations solides totales (cm)

Imputation de données météorologiques manquantes



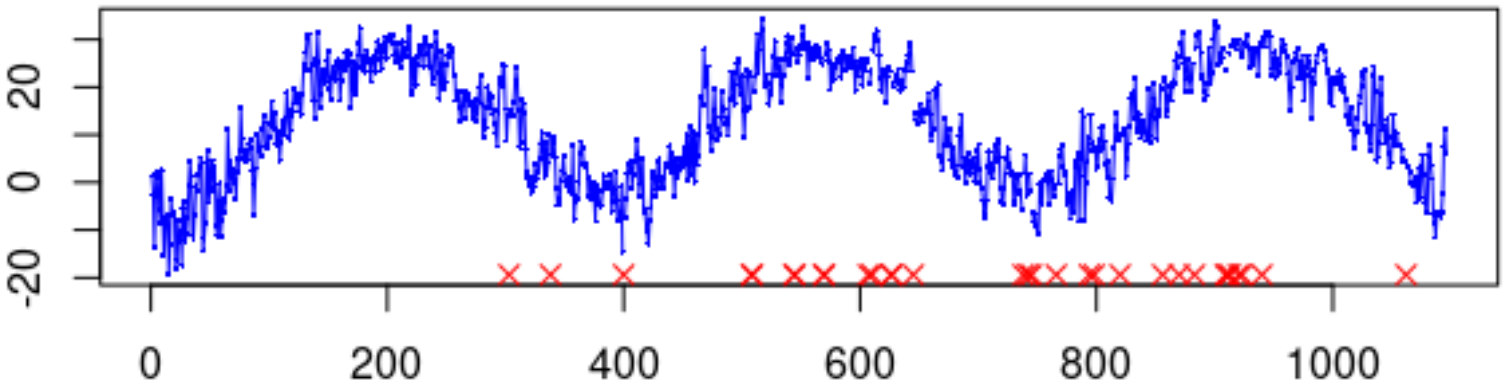




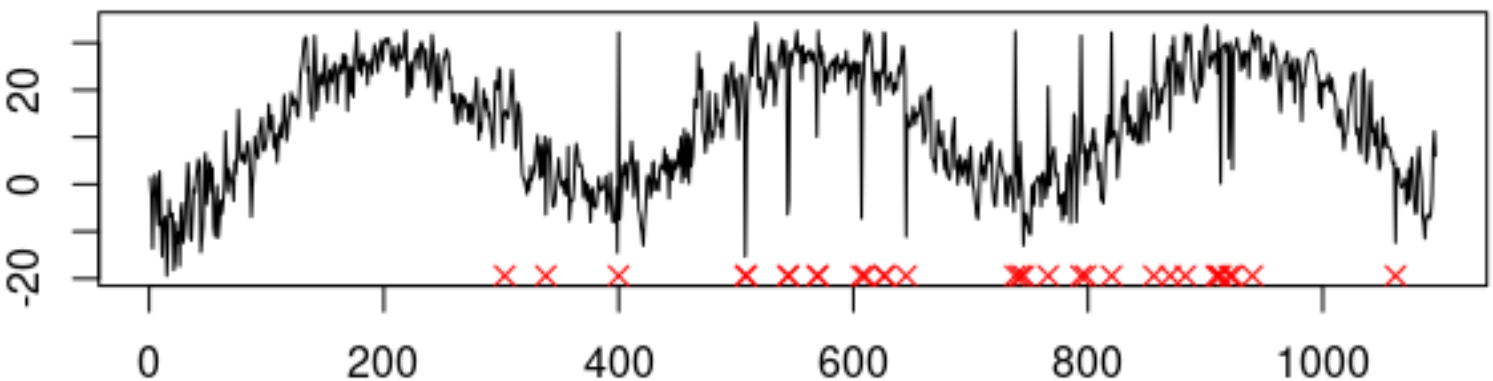
Répartition des données manquantes entre les variables

Un exemple de mauvaise imputation (cas MICE)

Températures maximales précédant l'imputation



Températures maximales suite à l'imputation via MICE



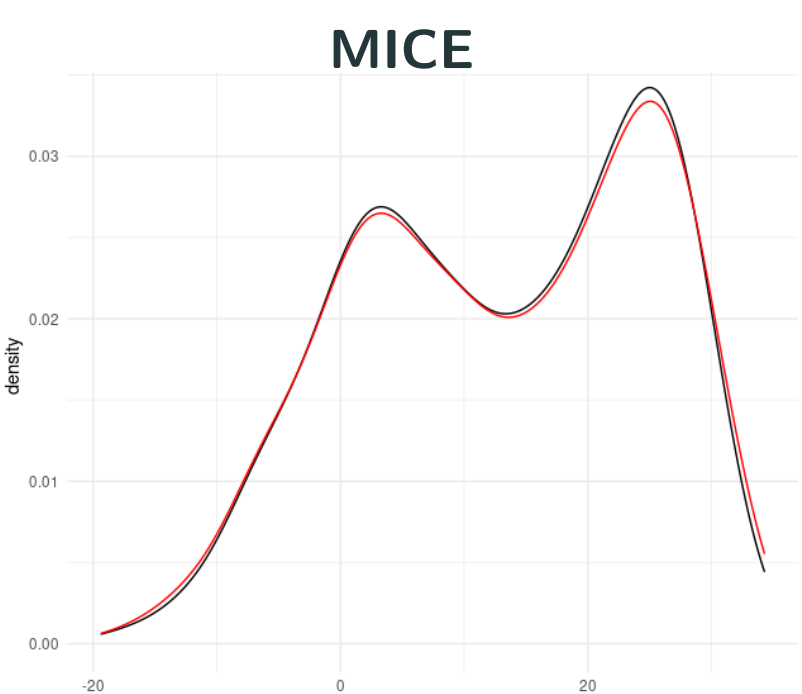
Un exemple de mauvaise imputation (cas MICE)

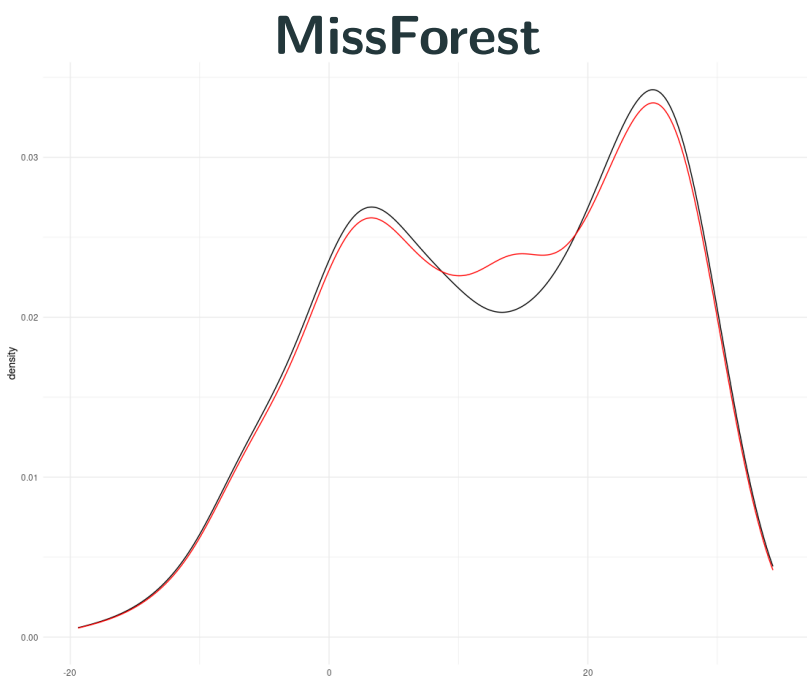
Comparaison des statistiques descriptives

```
> summary(df_clean)
   Date.Heure         Temp.max...C.     Temp.min...C.     Temp.moy...C.     Pluie.tot...mm.   Neige.tot...cm.
 Min.   :2022-01-01   Min.   :-19.40   Min.   :-27.900   Min.   :-22.600   Min.   :  0.00   Min.   : 0.0000
 1st Qu.:2022-10-01   1st Qu.:  3.20   1st Qu.: -3.700   1st Qu.: -0.300   1st Qu.:  0.00   1st Qu.: 0.0000
 Median :2023-07-02   Median : 14.10   Median :  4.600   Median :  9.500   Median :  0.00   Median : 0.0000
 Mean   :2023-07-02   Mean   : 13.01   Mean   :  4.023   Mean   :  8.529   Mean   :  2.56   Mean   : 0.5411
 3rd Qu.:2024-04-01   3rd Qu.: 23.80   3rd Qu.: 13.800   3rd Qu.: 18.800   3rd Qu.:  1.00   3rd Qu.: 0.0000
 Max.   :2024-12-31   Max.   : 34.30   Max.   : 23.800   Max.   : 28.800   Max.   :154.00   Max.   :21.8000
                      NA's   :31       NA's   :31        NA's   :31        NA's   :6        NA's   :6
> summary(df_imputed_mice)
   Date.Heure         Temp.max...C.     Temp.min...C.     Temp.moy...C.     Pluie.tot...mm.   Neige.tot...cm.
 Min.   :2022-01-01   Min.   :-19.40   Min.   :-27.900   Min.   :-22.60   Min.   :  0.000   Min.   : 0.0000
 1st Qu.:2022-10-01   1st Qu.:  3.10   1st Qu.: -3.525   1st Qu.: -0.20   1st Qu.:  0.000   1st Qu.: 0.0000
 Median :2023-07-02   Median : 14.05   Median :  4.600   Median :  9.50   Median :  0.000   Median : 0.0000
 Mean   :2023-07-02   Mean   : 13.03   Mean   :  4.108   Mean   :  8.58   Mean   :  2.556   Mean   : 0.5655
 3rd Qu.:2024-04-01   3rd Qu.: 24.02   3rd Qu.: 13.800   3rd Qu.: 18.82   3rd Qu.:  1.050   3rd Qu.: 0.0000
 Max.   :2024-12-31   Max.   : 34.30   Max.   : 23.800   Max.   : 28.80   Max.   :154.000   Max.   :21.8000
```

Un exemple de mauvaise imputation (cas MICE)

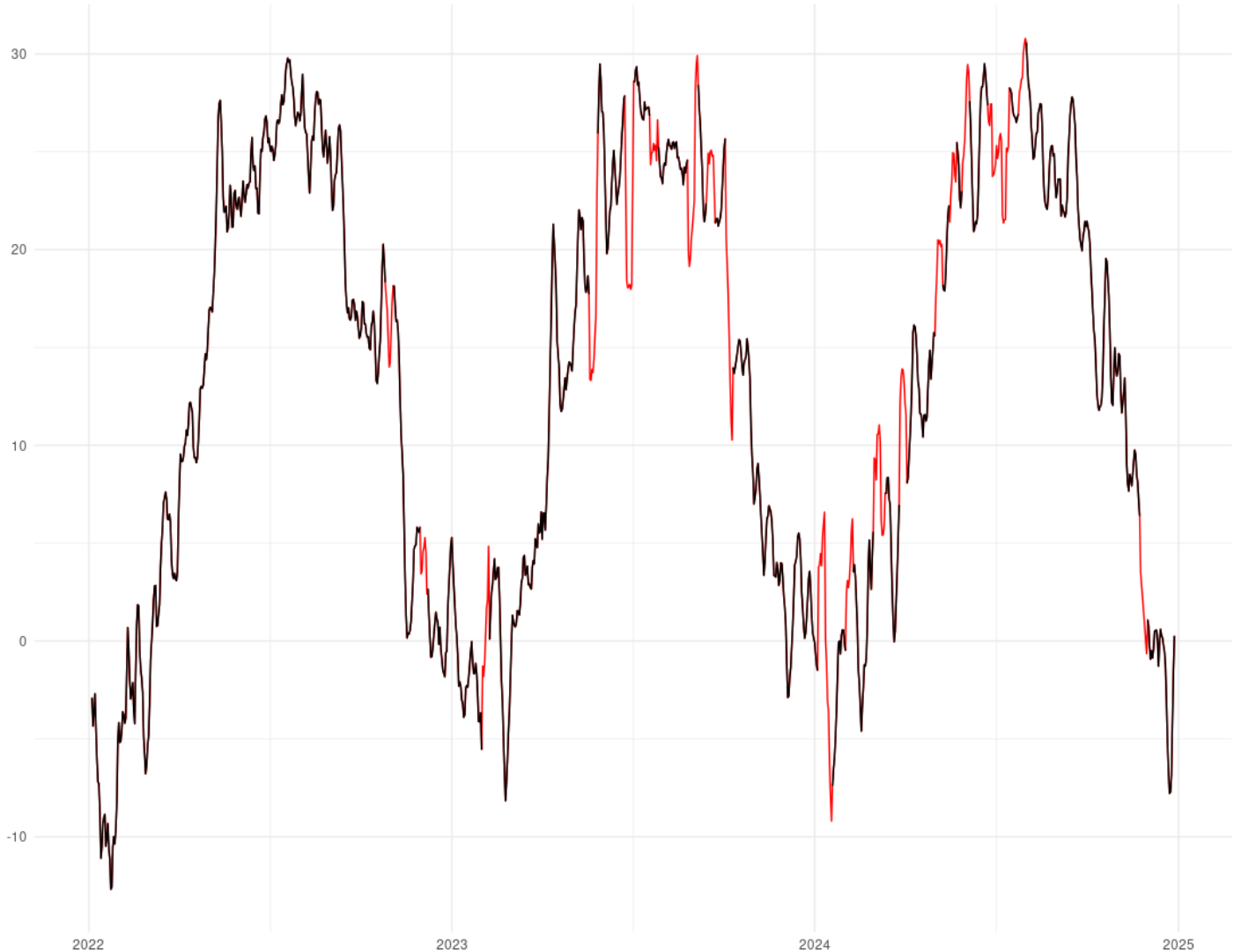
Comparaison des densités avant/après imputation (température maximale)





Un exemple de mauvaise imputation (cas MICE)

Moyenne mobile (7 jours) avant/après imputation (température maximale)



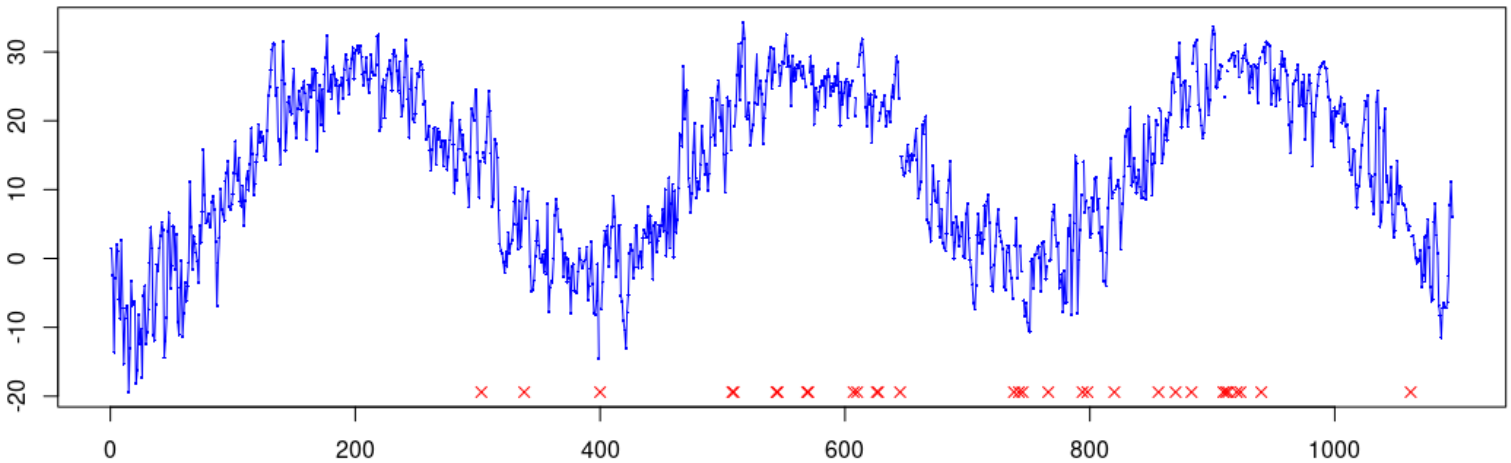
Un exemple de mauvaise imputation (cas MICE)

Comparaison des matrices de corrélation

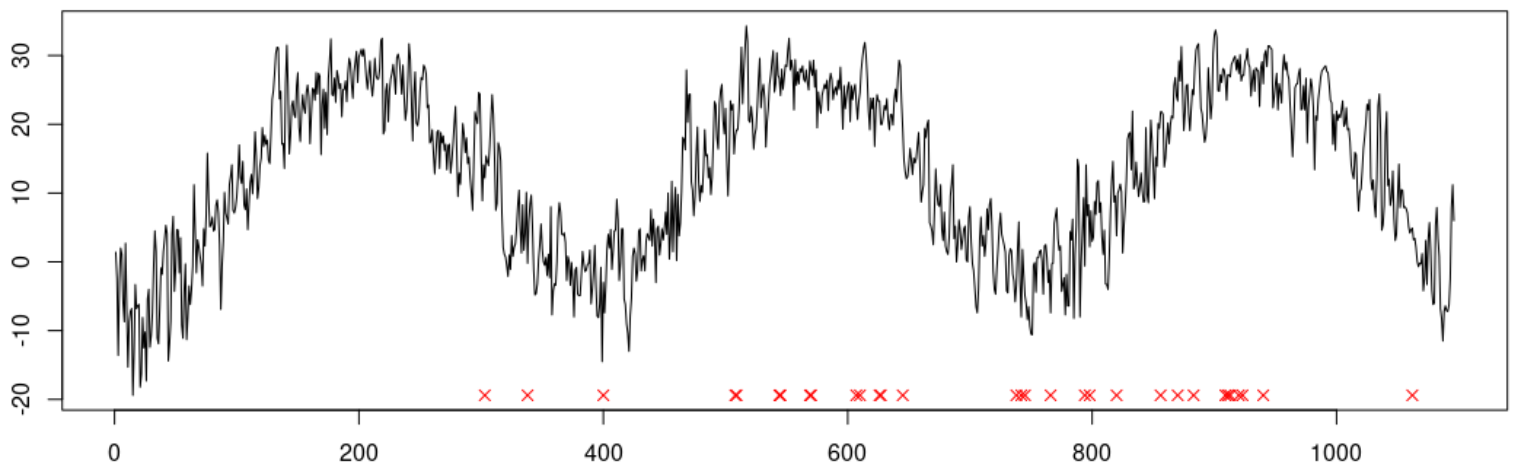
```
                Temp.max...C. Temp.min...C. Temp.moy...C. Pluie.tot...mm. Neige.tot...cm.
Temp.max...C.       1.0000000     0.9532599     0.9890203      0.12150448     -0.27614837
Temp.min...C.       0.9532599     1.0000000     0.9874389      0.17841043     -0.24799578
Temp.moy...C.       0.9890203     0.9874389     1.0000000      0.15083911     -0.26578818
Pluie.tot...mm.     0.1215045     0.1784104     0.1508391      1.00000000     -0.02803502
Neige.tot...cm.    -0.2761484    -0.2479958    -0.2657882     -0.02803502      1.00000000
> print(cor_mice)
                Temp.max...C. Temp.min...C. Temp.moy...C. Pluie.tot...mm. Neige.tot...cm.
Temp.max...C.       1.00000000     0.9204896     0.9816296      0.09528611     -0.26983810
Temp.min...C.       0.92048962     1.0000000     0.9781292      0.18101691     -0.24396883
Temp.moy...C.       0.98162957     0.9781292     1.0000000      0.13913508     -0.26288296
Pluie.tot...mm.     0.09528611     0.1810169     0.1391351      1.00000000     -0.03000434
Neige.tot...cm.    -0.26983810    -0.2439688    -0.2628830     -0.03000434      1.00000000
```

Un exemple de bonne imputation (méthode k-NN)

Températures maximales précédant l'imputation



Températures maximales suite à l'imputation via k-NN



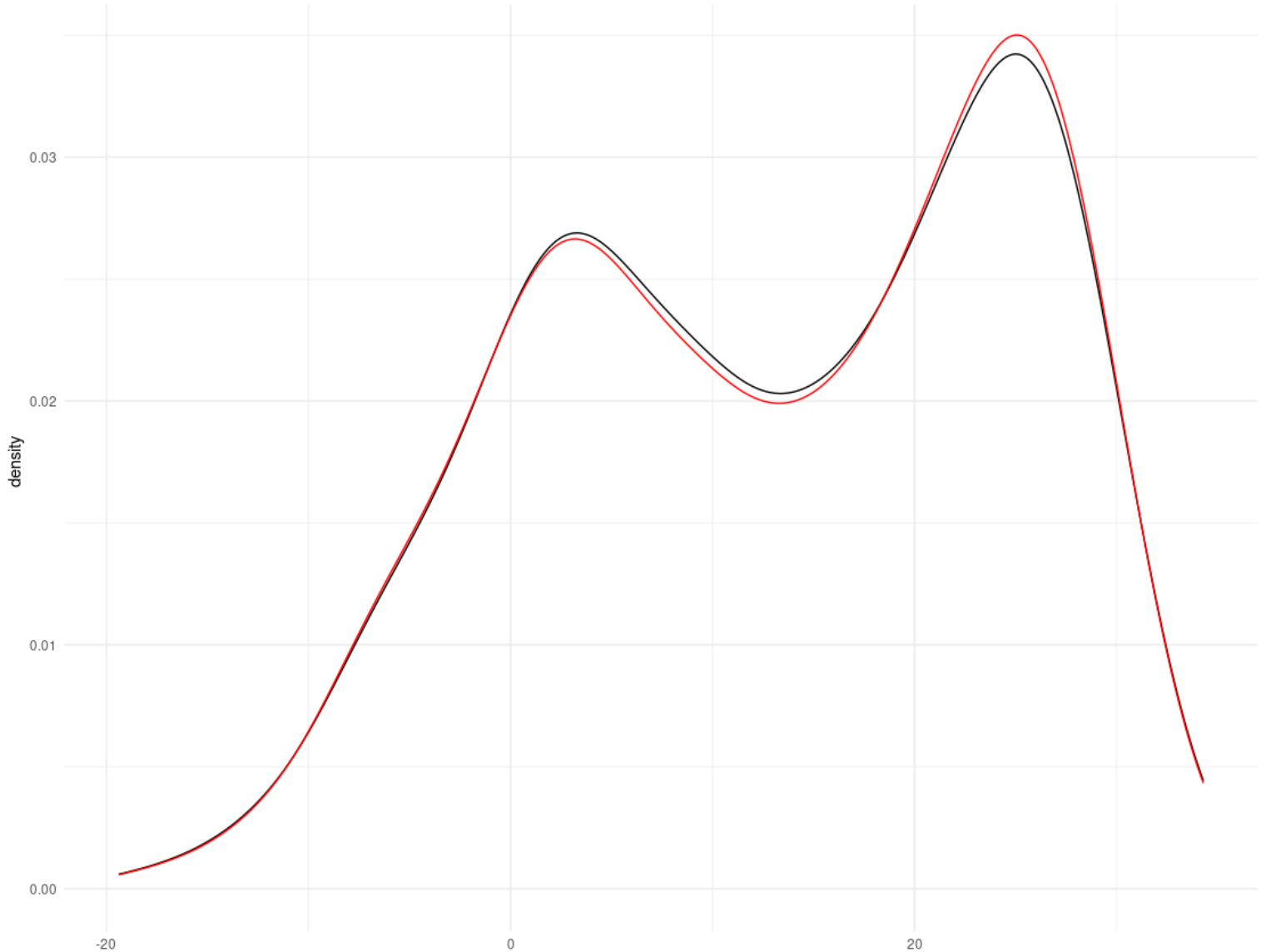
Un exemple de bonne imputation (méthode k-NN)

Comparaison des statistiques descriptives

```
> summary(df_clean)
   Date.Heure          Temp.max...C.      Temp.min...C.      Temp.moy...C.      Pluie.tot...mm.   Neige.tot...cm.       Jour_tot
 Min.   :2022-01-01   Min.   :-19.40   Min.   :-27.900   Min.   :-22.600   Min.   :  0.00   Min.   : 0.0000   Min.   :  1.0
 1st Qu.:2022-10-01   1st Qu.:  3.20   1st Qu.: -3.700   1st Qu.: -0.300   1st Qu.:  0.00   1st Qu.: 0.0000   1st Qu.: 92.0
 Median :2023-07-02   Median : 14.10   Median :  4.600   Median :  9.500   Median :  0.00   Median : 0.0000   Median :183.0
 Mean   :2023-07-02   Mean   : 13.01   Mean   :  4.023   Mean   :  8.529   Mean   :  2.56   Mean   : 0.5411   Mean   :183.2
 3rd Qu.:2024-04-01   3rd Qu.: 23.80   3rd Qu.: 13.800   3rd Qu.: 18.800   3rd Qu.:  1.00   3rd Qu.: 0.0000   3rd Qu.:274.2
 Max.   :2024-12-31   Max.   : 34.30   Max.   : 23.800   Max.   : 28.800   Max.   :154.00   Max.   :21.8000   Max.   :366.0
                      NA's   :31       NA's   :31        NA's   :31        NA's   :6        NA's   :6
> summary(df_knn)
   Date.Heure          Temp.max...C.      Temp.min...C.      Temp.moy...C.      Pluie.tot...mm.   Neige.tot...cm.       Jour_tot
 Min.   :2022-01-01   Min.   :-19.40   Min.   :-27.900   Min.   :-22.600   Min.   :  0.000   Min.   : 0.0000   Min.   :  1.0
 1st Qu.:2022-10-01   1st Qu.:  3.20   1st Qu.: -3.725   1st Qu.: -0.300   1st Qu.:  0.000   1st Qu.: 0.0000   1st Qu.: 92.0
 Median :2023-07-02   Median : 14.10   Median :  4.700   Median :  9.600   Median :  0.000   Median : 0.0000   Median :183.0
 Mean   :2023-07-02   Mean   : 13.08   Mean   :  4.081   Mean   :  8.597   Mean   :  2.546   Mean   : 0.5394   Mean   :183.2
 3rd Qu.:2024-04-01   3rd Qu.: 24.00   3rd Qu.: 13.800   3rd Qu.: 18.900   3rd Qu.:  1.000   3rd Qu.: 0.0000   3rd Qu.:274.2
 Max.   :2024-12-31   Max.   : 34.30   Max.   : 23.800   Max.   : 28.800   Max.   :154.000   Max.   :21.8000   Max.   :366.0
```

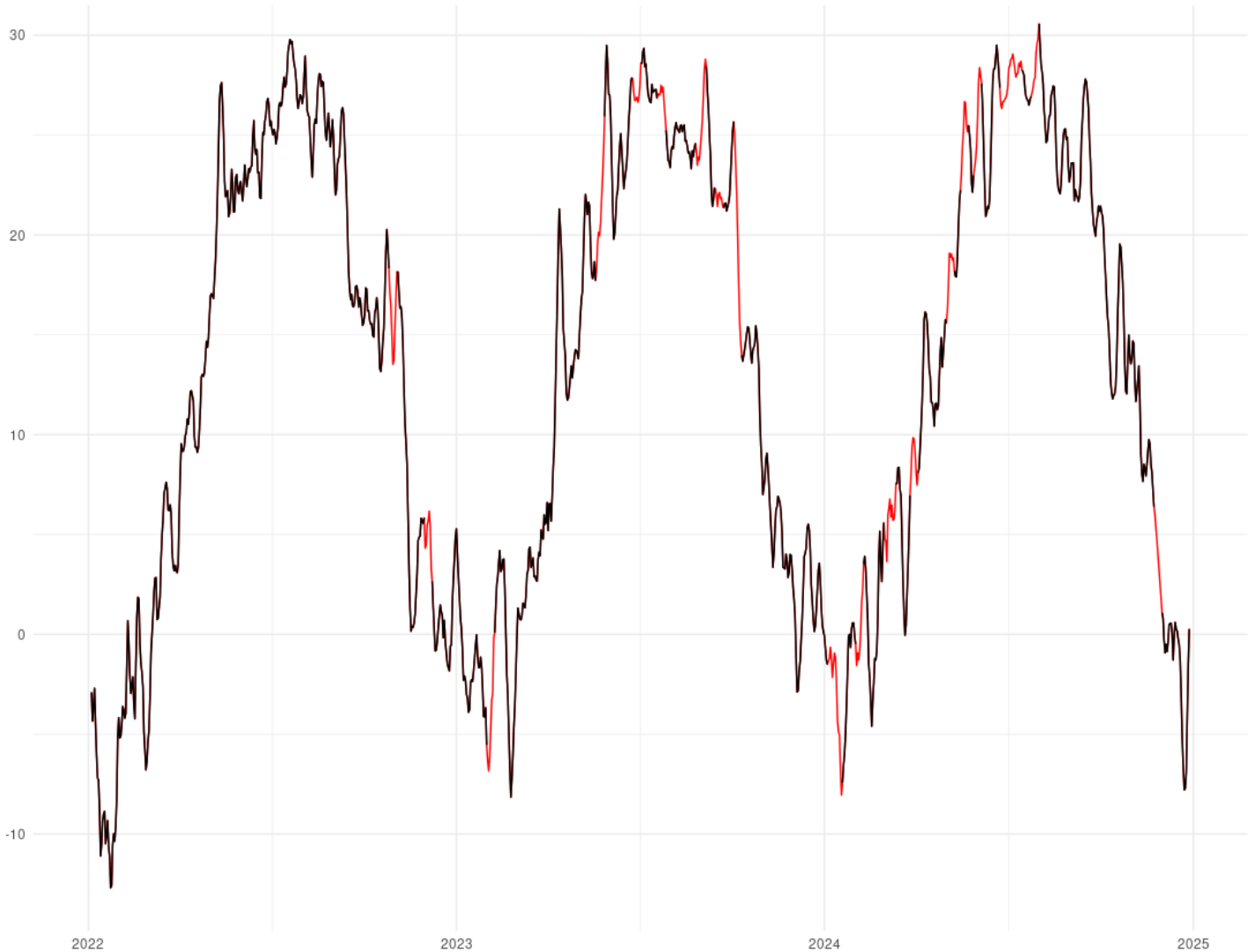
Un exemple de bonne imputation (méthode k-NN)

Densité avant/après imputation via k-NN



Un exemple de bonne imputation (méthode k-NN)

Moyenne mobile (7 jours) avant/après imputation (température maximale)



Un exemple de bonne imputation (méthode k-NN)

Comparaison des matrices de corrélation

```
> print(cor)
               Temp.max...C. Temp.min...C. Temp.moy...C. Pluie.tot...mm. Neige.tot...cm.
Temp.max...C.      1.0000000     0.9532599     0.9890203      0.12150448     -0.27614837
Temp.min...C.      0.9532599     1.0000000     0.9874389      0.17841043     -0.24799578
Temp.moy...C.      0.9890203     0.9874389     1.0000000      0.15083911     -0.26578818
Pluie.tot...mm.    0.1215045     0.1784104     0.1508391      1.00000000     -0.02803502
Neige.tot...cm.   -0.2761484    -0.2479958    -0.2657882     -0.02803502      1.00000000
> print(cor_knn)
               Temp.max...C. Temp.min...C. Temp.moy...C. Pluie.tot...mm. Neige.tot...cm.
Temp.max...C.      1.0000000     0.9538642     0.9891686      0.12537885     -0.27257925
Temp.min...C.      0.9538642     1.0000000     0.9874920      0.17732938     -0.24518379
Temp.moy...C.      0.9891686     0.9874920     1.0000000      0.15262882     -0.26262970
Pluie.tot...mm.    0.1253788     0.1773294     0.1526288      1.00000000     -0.02864907
Neige.tot...cm.   -0.2725792    -0.2451838    -0.2626297     -0.02864907      1.00000000
```

Bilan sur le jeu de données météorologique

Récapitulatif sur l'efficacité des méthodes

Non adapté	Adapté
MissForest	k-NN ($k = 5$)
Moyenne	Moyenne mobile (ordre 7)
MICE	ARIMA

Conclusion et perspectives

Conclusion et perspectives

Ce projet a permis :

- Compréhension des mécanismes des données manquantes (**MCAR, MAR, MNAR**).
- Comparaison des méthodes classiques (suppression, imputation moyenne) et avancées (**MICE, MissForest, k-NN**).
- Compréhension de l'importance d'adapter la méthode en fonction du contexte et de la structure des données.

Perspectives :

- **Approches hybrides** : combiner plusieurs méthodes pour améliorer la robustesse de l'imputation.
- **Automatisation** : développer un algorithme adaptatif sélectionnant la meilleure stratégie en fonction des caractéristiques des données.
- **À grande échelle** : intégrer ces techniques pour traiter efficacement des bases de données massives.

Bibliographie

[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, 44(2), 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), 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, 28(1), 112-118.
[8] Yosef, M. (2024). *2024 Olympics Medals and Economic Status*. Kaggle.
[9] NAVCAN. (2022, 2023, 2024). *Rapport de données quotidiennes - Montréal, Québec*. Environnement Canada.