Évaluation Comparative des Méthodes d'Imputation pour Données Neuropsychologiques

Cas d'Application : Base NS-Park

Moad Hani

Doctorant en Sciences Informatiques

Service Informatique, Logiciel et Intelligence Artificielle (ILIA)

Université de Mons, Belgique

moad.hani@umons.ac.be

Octobre 2025

Résumé

Dans le contexte clinique de l'évaluation neuropsychologique des patients parkinsoniens, la présence de données manquantes ($< 10\%$ des cas) constitue un défi majeur pour l'analyse et l'interprétation des résultats. Cette étude comparative évalue six méthodes d'imputation (Mean, Median, KNN, MICE, EM, MissForest) sur la base NS-Park contenant 1408 patients et 42 variables cognitives. Les performances sont mesurées selon trois mécanismes de missingness (MCAR, MAR, MNAR) et quatre taux de données manquantes (10%, 20%, 30%, 40%) avec des métriques rigoureuses (MAE, RMSE, R^2, similarité de distribution). Les résultats démontrent la supériorité de **MissForest** (MAE = 2.20, $R^2 = 0.91$) avec une robustesse exceptionnelle face aux différents mécanismes, suivi de **MICE** (MAE = 2.53) offrant un compromis optimal précision/temps d'exécution. Cette étude fournit des recommandations pratiques pour l'implémentation clinique et la validation méthodologique.

Mots-clés : Imputation de données, Maladie de Parkinson, Neuropsychologie, Machine Learning, MissForest, MICE, Évaluation cognitive

Table des matières

1 Introduction **3**
1.1 Contexte Clinique . . . 3
1.2 Base de Données NS-Park . . . 3
1.3 Problématique . . . 3
1.4 Objectifs de l'Étude . . . 3
1.4.1 Objectif Principal . . . 3
1.4.2 Objectifs Secondaires . . . 4

2 Méthodologie **4**
2.1 Simulation des Données Manquantes . . . 4
2.1.1 Mécanismes de Missingness . . . 4
2.1.2 Taux de Données Manquantes . . . 5
2.1.3 Configuration Expérimentale . . . 5
2.2 Méthodes d'Imputation Évaluées . . . 5
2.2.1 Méthodes Simples (Baseline) . . . 5
2.2.2 Méthodes Avancées . . . 6
2.2.3 Méthode Machine Learning . . . 7
2.3 Métriques d'Évaluation . . . 7
2.3.1 Métriques de Précision . . . 7
2.3.2 Métrique de Distribution . . . 7

3 Résultats **8**
3.1 Performance Globale . . . 8
3.2 Analyse par Mécanisme de Missingness . . . 8
3.3 Analyse de Sensibilité au Taux de Missingness . . . 9
3.4 Temps d'Exécution . . . 9

4 Discussion **9**
4.1 Supériorité de MissForest . . . 9
4.2 Compromis Précision/Temps . . . 10
4.3 Limites et Perspectives . . . 10
4.3.1 Limites Méthodologiques . . . 10
4.3.2 Perspectives Futures . . . 10

5 Conclusion **10**
5.1 Recommandations Pratiques . . . 11
5.2 Contribution Scientifique . . . 11

A Annexe A : Tableau Récapitulatif Complet **12**

B Annexe B : Configuration Technique **12**
B.1 Environnement de Calcul . . . 12
B.2 Hyperparamètres Optimaux . . . 12

1 Introduction

1.1 Contexte Clinique

Dans la pratique neuropsychologique, l'évaluation des fonctions cognitives chez les patients atteints de la maladie de Parkinson se heurte fréquemment au problème des données manquantes. Les études cliniques rapportent que **15 à 30% des évaluations** présentent des valeurs incomplètes [1, 2], dues à plusieurs facteurs :

- **Fatigue cognitive** : Épuisement du patient pendant l'évaluation prolongée
- **Contraintes temporelles** : Limitation du temps de consultation
- **Incapacité partielle** : Impossibilité de compléter certains tests (déficits cognitifs sévères)
- **Refus du patient** : Non-acceptation de tests jugés difficiles ou anxiogènes
- **Erreurs administratives** : Perte ou oubli de formulaires

Ces lacunes compromettent l'analyse statistique, réduisent la puissance des tests et peuvent introduire des biais dans les décisions cliniques. L'imputation de données manquantes devient donc une étape méthodologique cruciale.

1.2 Base de Données NS-Park

Cette étude s'appuie sur la base **NS-Park**, une cohorte neuropsychologique longitudinale constituée de :

- **N = 1408 patients** parkinsoniens évalués en consultation
- **42 variables cognitives** issues de batteries standardisées
- **Dataset quasi-complet** : 99,998% de complétude initiale (seulement 2 valeurs manquantes sur 59 136 cellules)

Les domaines cognitifs couverts incluent :

1. **Évaluation cognitive globale** : MoCA (Montreal Cognitive Assessment)
2. **Mémoire** : Empans, rappel immédiat et différé
3. **Fonctions exécutives** : TMT (Trail Making Test), Stroop, fluences verbales
4. **Langage** : Dénomination (BNT), compréhension
5. **Fonctions visuospatiales** : Copie de figures complexes
6. **Attention soutenue** : Tests chronométrés
7. **Variables démographiques** : Âge, sexe, durée de scolarité

1.3 Problématique

La question de recherche centrale est la suivante :

> *"Quelle méthode d'imputation offre les meilleures performances pour gérer les données manquantes en neuropsychologie, tout en préservant l'intégrité statistique des analyses et la validité des conclusions cliniques ?"*

1.4 Objectifs de l'Étude

1.4.1 Objectif Principal

Comparer six méthodes d'imputation (simples, avancées et machine learning) pour identifier la plus adaptée à une application clinique en neuropsychologie.

1.4.2 Objectifs Secondaires

O1. Évaluer la robustesse des méthodes selon le **mécanisme de missingness** (MCAR, MAR, MNAR)
O2. Analyser la **sensibilité** au taux de données manquantes (10% à 40%)
O3. Mesurer le **compromis précision/temps d'exécution** pour l'implémentation clinique
O4. Évaluer la **préservation des distributions** statistiques originales
O5. Formuler des **recommandations pratiques** pour la validation clinique

2 Méthodologie

2.1 Simulation des Données Manquantes

Pour évaluer les méthodes d'imputation de manière contrôlée et reproductible, nous avons simulé trois mécanismes de missingness distincts sur le dataset NS-Park complet.

2.1.1 Mécanismes de Missingness

MCAR (Missing Completely At Random) La probabilité qu'une donnée soit manquante est **indépendante** de toutes les variables (observées ou non).

$$P(\text{missing}_{ij}) = p \quad \forall i, j \tag{1}$$

où p est une constante (taux de missingness).
Exemple clinique : Erreur administrative aléatoire, perte de formulaire.
Implémentation :

```
mask = np.random.random((n_rows, n_cols)) < rate
X_missing[mask] = np.nan
```

MAR (Missing At Random) La probabilité de données manquantes **dépend de variables observées** (ex : âge).

$$P(\text{missing}_{ij} \mid \hat{\text{Age}}_i) = 0.5p + p \cdot \frac{\hat{\text{Age}}_i - \hat{\text{Age}}_{\min}}{\hat{\text{Age}}_{\max} - \hat{\text{Age}}_{\min}} \tag{2}$$

Exemple clinique : Les patients plus âgés abandonnent plus souvent les tests longs en raison de fatigue accrue.
Implémentation :

```
age_normalized = (Age - Age.min()) / (Age.max() - Age.min())
prob_missing = np.clip(rate * 0.5 + rate * age_normalized, 0, 0.5)
mask = np.random.random(n_rows) < prob_missing
```

MNAR (Missing Not At Random) La probabilité de données manquantes **dépend de la valeur manquante elle-même**.

$$P(\text{missing}_{ij} \mid X_{ij}) = p \cdot \left(1 - \frac{X_{ij} - X_{\min}}{X_{\max} - X_{\min}}\right) \tag{3}$$

Exemple clinique : Patients avec scores cognitifs très bas (MoCA < 10) incapables de terminer les tests.

Implémentation :

```
vals_normalized = (X - X.min()) / (X.max() - X.min())
prob_missing = rate * (1 - vals_normalized)
prob_missing = np.clip(prob_missing, rate * 0.3, rate * 1.5)
```

Réalisme clinique : MNAR est le mécanisme le plus fréquent en neuropsychologie.

2.1.2 Taux de Données Manquantes

Quatre taux de missingness ont été évalués : **10%, 20%, 30%, 40%**.

TABLE 1 – Justification des taux de missingness

Taux	Sévérité	Justification
10%	Faible	Taux minimal observé en pratique courante
20%	Modéré	Taux standard en études neuropsychologiques longitudinales
30%	Élevé	Taux critique nécessitant imputation robuste
40%	Extrême	Limite supérieure pour tester la robustesse des méthodes

2.1.3 Configuration Expérimentale

— **3 mécanismes** $\times$ **4 taux** = **12 scénarios**
— **3 itérations** par scénario (stabilité statistique)
— **Total : 36 évaluations** par méthode d'imputation
— **Split train/test** : 80%/20% avec stratification
— **Pas de data leakage** : Imputation réalisée uniquement sur le train set

2.2 Méthodes d'Imputation Évaluées

Six méthodes d'imputation couvrant différents paradigmes ont été comparées.

2.2.1 Méthodes Simples (Baseline)

Mean Imputation Remplacement par la moyenne de la variable.

$$\hat{x}_{ij} = \frac{1}{n_{\text{obs}}} \sum_{i:x_{ij} \neq \text{NA}} x_{ij} \tag{4}$$

Avantages : Rapidité (< 0.01s), simplicité d'implémentation.
Limites : Réduction de variance, biais sur corrélations, ne préserve pas la distribution.

Median Imputation Remplacement par la médiane de la variable.

$$\hat{x}_{ij} = \text{médiane}(x_{\cdot j}) \tag{5}$$

Avantages : Robuste aux outliers.
Limites : Idem Mean, perte d'information structurale.

2.2.2 Méthodes Avancées

KNN (K-Nearest Neighbors) Imputation par moyenne pondérée des $k = 5$ plus proches voisins.

$$\hat{x}_{ij} = \frac{\sum_{k=1}^{K} w_k \cdot x_{kj}}{\sum_{k=1}^{K} w_k} \tag{6}$$

où $w_k = \frac{1}{d(x_i, x_k)}$ avec $d(\cdot)$ distance euclidienne.
Avantages : Préserve structure locale, rapide ($\sim$0.5s).
Limites : Sensible aux échelles, choix de k critique.

MICE (Multiple Imputation by Chained Equations) Imputation itérative par modèles de régression chainés.

Algorithme :

1. Initialisation : Imputation naïve (moyenne)
2. Pour $t = 1, \ldots, T$ (10 itérations) :
   - Pour chaque variable j avec missing :
     - Régresser X_j sur autres variables X_{-j}
     - Prédire valeurs manquantes $\hat{X}_j$
3. Convergence si $\|\hat{X}^{(t)} - \hat{X}^{(t-1)}\| < \epsilon$

Avantages : Modèle relationnel entre variables, théorie solide.
Limites : Suppose MAR, temps modéré ($\sim$2s).

EM (Expectation-Maximization) Optimisation itérative de la vraisemblance sous hypothèse de normalité.

Algorithme :

1. **E-step :** Calculer espérance des valeurs manquantes conditionnellement aux paramètres actuels
2. **M-step :** Estimer paramètres (μ, Σ) par maximum de vraisemblance
3. Itérer jusqu'à convergence

$$\hat{x}_{ij}^{(t+1)} = \mu_j^{(t)} + \Sigma_{j,-j}^{(t)} (\Sigma_{-j,-j}^{(t)})^{-1} (x_{i,-j} - \mu_{-j}^{(t)}) \tag{7}$$

Avantages : Théoriquement optimal sous normalité multivariée.
Limites : Convergence lente (50 itérations), sensible aux hypothèses.

2.2.3 Méthode Machine Learning

MissForest Imputation itérative par Random Forest.

Algorithme :

1. Initialisation : Imputation naïve (médiane)
2. Pour $t = 1, \ldots, T$ (3 itérations) :
   - Trier variables par nombre de valeurs manquantes (croissant)
   - Pour chaque variable j :
     - Entraîner Random Forest : $f_j : X_{-j} \to X_j$
     - Prédire valeurs manquantes : $\hat{X}_j = f_j(X_{-j})$
3. Convergence si erreur OOB se stabilise

Configuration :

- **n_estimators** = 50 arbres
- **max_iter** = 3 itérations
- **Parallélisation** : n_jobs = -1 (tous CPU disponibles)

Avantages : Capture relations non-linéaires, robuste aux outliers, sans hypothèse distributionnelle.

Limites : Coût computationnel élevé (∼265s par run).

2.3 Métriques d'Évaluation

2.3.1 Métriques de Précision

MAE (Mean Absolute Error)

$$\text{MAE} = \frac{1}{|\Omega_{\text{miss}}|} \sum_{(i,j)\in\Omega_{\text{miss}}} |x_{ij}^{\text{true}} - \hat{x}_{ij}| \tag{8}$$

où Ω_{miss} = ensemble des valeurs manquantes simulées.

RMSE (Root Mean Square Error)

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega_{\text{miss}}|} \sum_{(i,j)\in\Omega_{\text{miss}}} (x_{ij}^{\text{true}} - \hat{x}_{ij})^2} \tag{9}$$

R^2 (Coefficient de Détermination)

$$R^2 = 1 - \frac{\sum_{(i,j)}(x_{ij}^{\text{true}} - \hat{x}_{ij})^2}{\sum_{(i,j)}(x_{ij}^{\text{true}} - \bar{x}_j)^2} \tag{10}$$

2.3.2 Métrique de Distribution

Test de Kolmogorov-Smirnov Pour chaque variable j :

$$D_j = \sup_x |F_j^{\text{true}}(x) - F_j^{\text{imp}}(x)| \tag{11}$$

Score de similarité : $S_j = 1 - D_j$ (plus proche de 1 = meilleur).

Score global :

$$S_{\text{global}} = \frac{1}{p} \sum_{j=1}^{p} S_j \tag{12}$$

3 Résultats

3.1 Performance Globale

Le tableau 2 présente les performances moyennes sur l'ensemble des 36 évaluations (3 mécanismes × 4 taux × 3 itérations).

TABLE 2 – Performances globales des méthodes d'imputation

Rang	Méthode	MAE	RMSE	R^2	Similarité	Temps (s)
1	**MissForest**	**2.20**	**6.18**	**0.91**	**0.77**	265
2	MICE	2.53	6.89	0.88	0.69	2.0
3	KNN	2.70	7.24	0.85	0.74	0.5
4	EM	3.76	9.12	0.68	0.40	3.0
5	Mean	3.76	9.15	0.67	0.40	0.01
6	Median	3.55	8.96	0.70	0.58	0.01

Observations clés :

- **MissForest** surpasse systématiquement toutes les méthodes :
  - Réduction MAE de **15%** vs MICE
  - Réduction MAE de **33%** vs KNN
  - Réduction MAE de **42%** vs méthodes simples
- **MICE** offre un **excellent compromis précision/temps** (MAE=2.53, 2s)
- **KNN** : Solution rapide avec précision acceptable (MAE=2.70, 0.5s)
- **Méthodes simples** inadaptées (MAE > 3.5)

3.2 Analyse par Mécanisme de Missingness

Le tableau 3 détaille les performances selon le mécanisme.

TABLE 3 – Performances par mécanisme de missingness (MAE)

Méthode	MCAR	MAR	MNAR	Écart-type
MissForest	**2.18**	**2.26**	**2.15**	**0.06**
MICE	2.47	2.54	2.60	0.07
KNN	2.70	2.70	2.78	0.05
EM	3.71	3.71	3.84	0.08
Mean	3.71	3.71	3.88	0.10
Median	3.56	3.58	3.51	0.04

Observations :

— **MissForest** : Performances homogènes sur tous mécanismes (écart-type = 0.06)
— Meilleure performance sur **MNAR** (2.15) pourtant le plus difficile
— **MICE** : Légère dégradation sur MNAR (+5% vs MCAR)
— **Méthodes simples** : Échouent systématiquement (> 3.5)

3.3 Analyse de Sensibilité au Taux de Missingness

La figure **??** (non reproduite ici, voir résultats graphiques) illustre l'évolution des performances.

Tendances observées :

— **MissForest** : MAE croît linéairement avec le taux
  — 10% : MAE = 2.08
  — 20% : MAE = 2.20
  — 30% : MAE = 2.28
  — 40% : MAE = 2.35 (+13% seulement)
— **MICE** : Dégradation modérée (+12% de 10% à 40%)
— **Méthodes simples** : Dégradation forte (+25-30%)

Robustesse démontrée jusqu'à 40% de données manquantes.

3.4 Temps d'Exécution

Table 4 – Temps d'exécution total (36 runs)

Méthode	Temps/run (s)	Temps total (min)	% du total
MissForest	265.0	159.0	98.0%
EM	3.0	1.8	1.1%
MICE	2.0	1.2	0.7%
KNN	0.5	0.3	0.2%
Mean	0.01	0.0	0.0%
Median	0.01	0.0	0.0%
Total	-	**162.3**	100%

MissForest représente 98% du temps total mais offre une précision inégalée.

4 Discussion

4.1 Supériorité de MissForest

MissForest démontre une supériorité statistique significative sur toutes les métriques :

— **Précision maximale** : MAE = 2.20 (15% mieux que MICE)
— **Robustesse exceptionnelle** : Performance stable sur tous mécanismes (écart-type = 0.06)
— **Préservation des distributions** : Score = 0.77 (meilleur de tous)
— **Insensibilité au mécanisme** : Performant même sur MNAR (le plus difficile)

Cette supériorité s'explique par :

1. Capacité à capturer des **relations non-linéaires** complexes
2. **Robustesse aux outliers** (Random Forest = méthode ensembliste)
3. **Absence d'hypothèse distributionnelle** (vs EM qui suppose normalité)

4.2 Compromis Précision/Temps

Le choix de la méthode dépend du contexte d'utilisation :

Recherche approfondie → MissForest

— Précision maximale requise
— Temps d'exécution acceptable (quelques heures)
— Analyses critiques, publications scientifiques

Routine de recherche → MICE

— Compromis optimal (MAE = 2.53, 2s)
— Implémentation simple (scikit-learn)
— Études longitudinales avec multiples imputations

Application clinique temps-réel → KNN

— Rapidité critique (< 1s)
— Précision acceptable (MAE = 2.70)
— Outil décisionnel en consultation

4.3 Limites et Perspectives

4.3.1 Limites Méthodologiques

1. **Simulation contrôlée** : Mécanismes de missingness simplifiés vs réalité clinique complexe
2. **Dataset unique** : Validation nécessaire sur autres cohortes
3. **Mécanisme MAR/MNAR** : Modèles linéaires vs réalité non-linéaire

4.3.2 Perspectives Futures

1. **Méthodes avancées** : HyperImpute, autoencodeurs variationnels (VAEM)
2. **Validation clinique** : Test sur cohorte prospective avec vrais missing
3. **Implémentation** : Outil automatisé pour cliniciens
4. **Mécanismes réalistes** : Modélisation basée sur expertise neuropsychologique

5 Conclusion

Cette étude comparative rigoureuse sur 1408 patients et 42 variables cognitives démontre la supériorité de **MissForest** (MAE = 2.20, $R^2 = 0.91$) pour l'imputation de données neuropsychologiques, avec une robustesse exceptionnelle face à tous les mécanismes de missingness testés.

5.1 Recommandations Pratiques

R1. Adopter MissForest pour analyses critiques nécessitant précision maximale
R2. Privilégier MICE pour routine de recherche (compromis optimal)
R3. Utiliser KNN pour applications cliniques temps-réel
R4. Éviter méthodes simples (Mean/Median) en neuropsychologie
R5. Valider cliniquement sur cohortes prospectives avant implémentation

5.2 Contribution Scientifique

Cette étude apporte une **méthodologie rigoureuse d'évaluation** applicable à d'autres domaines de la neuropsychologie et fournit des **recommandations pratiques basées sur l'évidence** pour les chercheurs et cliniciens.

Références

[1] Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.

[2] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

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

[4] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice : Multivariate imputation by chained equations in R. *Journal of statistical software*, 45, 1-67.

[5] Troyanskaya, O., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.

A Annexe A : Tableau Récapitulatif Complet

TABLE 5 – Tableau récapitulatif complet des performances

Mécanisme	Méthode	MAE	RMSE	R^2	Similarité	Temps (s)
MCAR	MissForest	2.18	6.05	0.92	0.77	265
	MICE	2.47	6.72	0.89	0.69	2.0
	KNN	2.70	7.18	0.86	0.74	0.5
	EM	3.71	9.05	0.69	0.40	3.0
	Mean	3.71	9.08	0.68	0.40	0.01
	Median	3.56	8.89	0.71	0.58	0.01
MAR	MissForest	2.26	6.28	0.91	0.75	265
	MICE	2.54	6.95	0.88	0.70	2.0
	KNN	2.70	7.24	0.85	0.74	0.5
	EM	3.71	9.12	0.68	0.40	3.0
	Mean	3.71	9.15	0.67	0.40	0.01
	Median	3.58	8.96	0.70	0.58	0.01
MNAR	MissForest	2.15	6.22	0.90	0.73	265
	MICE	2.60	7.01	0.87	0.65	2.0
	KNN	2.78	7.31	0.84	0.69	0.5
	EM	3.84	9.20	0.67	0.36	3.0
	Mean	3.88	9.22	0.66	0.36	0.01
	Median	3.51	8.90	0.69	0.54	0.01

B Annexe B : Configuration Technique

B.1 Environnement de Calcul

- **Python** : 3.9
- **scikit-learn** : 1.0.2
- **NumPy** : 1.21.5
- **Pandas** : 1.4.2
- **CPU** : Intel Xeon 16 cores
- **RAM** : 64 GB

B.2 Hyperparamètres Optimaux

MissForest

- n_estimators = 50
- max_iter = 3
- n_jobs = -1

MICE

- max_iter = 10
- n_nearest_features = 10

KNN

— n_neighbors = 5
— weights = 'uniform'

EM

— max_iter = 50
— tolerance = 0.001