Imputing Missing Values in 6G Datasets with Tabular Methods

Kenneth Chan
Supervisors: Yuandou Wang and Rihan Hai

TU Delft

01 · INTRODUCTION

6G is only as smart as its data.

- Deployments introduce **missing data** — blockage, fading, sensor dropout.
- This missingness is **structured**, not random.
- 6G is temporal, but flattened to **tabular form** for downstream models.
- **Gap:** no systematic comparison of tabular imputers on 6G data.

02 · RESEARCH QUESTIONS

How do tabular imputation techniques compare on missing values in 6G datasets?

- **RQ1** Reconstruction error vs. rate (10/30/50%) and mechanism.
- **RQ2** Preserving distribution properties (mean, variance, joint structure)?
- **RQ3** Impact of imputation technique on the downstream task?

03 · DATASETS

DeepSense 6G — real-world vehicle measurements pairing GPS location with 60 GHz wireless beam power readings.

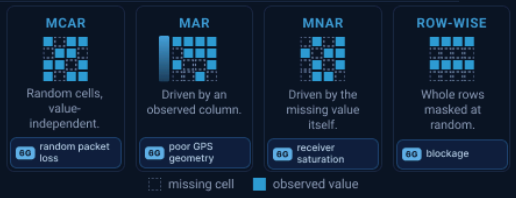
- **Scenario 5** — urban setting · n = 2,300 · 29 sequences.
- **Scenario 33** — city street setting · n = 3,644 · 21 sequences.

04 · EXPERIMENTAL PIPELINE

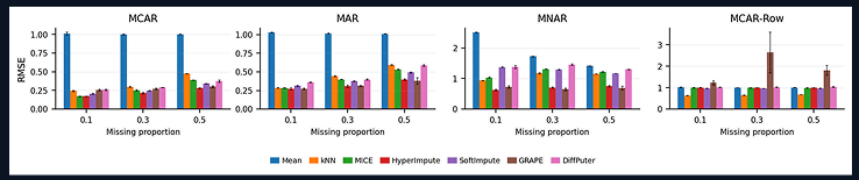
Clean data → Inject missingness → 7 imputers → Compute metrics → Compare & analyse

4 mechanisms × 3 missing rates × 7 methods × 3 seeds × 2 datasets

05 · MISSINGNESS MECHANISMS



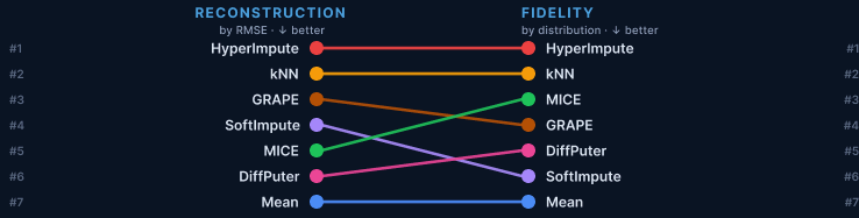
06 · RQ1 - RECONSTRUCTION ERROR



1. **No universal winner.** Cell-wise (MCAR/MAR), all learned methods beat the baseline and HyperImpute leads.
2. **Row-wise breaks nearly everyone.** Only kNN survives; GRAPE destabilises (RMSE 2.65). MNAR is the hardest cell-wise case.

07 · RQ2 - DISTRIBUTIONAL FIDELITY

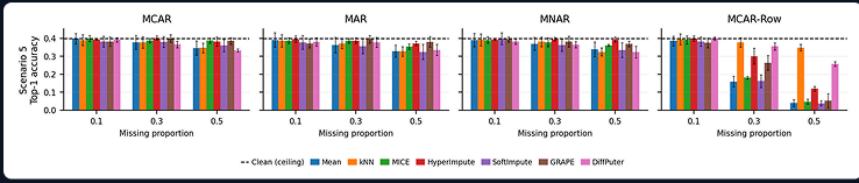
Each method's MEAN rank averaged over all 24 conditions (2 datasets × 4 mechanisms × 3 rates)



1. **Fidelity ≠ reconstruction.** Rankings correlate but don't match — SoftImpute drops from 4th to 6th, MICE climbs from 5th to 3rd. The chart tells you what RMSE alone cannot.

08 · RQ3 - DOWNSTREAM BEAM PREDICTION

The task: predict the optimal 60 GHz beam from GPS position alone (64-beam codebook). Imputers touch only training labels; scored Top-1 vs. a clean-data ceiling.



1. **Cell-wise: no effect.** Even mean fill stays near the clean ceiling — the arg-max beam label tolerates per-cell damage.
2. **Row-wise: decisive.** Only kNN (and partly DiffPuter) preserve the label signal; the rest fall to chance.

IMPUTATION METHODS

Mean	Column average.
kNN	k most similar rows.
MICE	Iterative regression per column.
SoftImpute	Low-rank SVD matrix completion.
HyperImpute	AutoML over per-column models.
GRAPE	GNN over a row–feature bipartite graph.
DiffPuter	Diffusion model refined by EM.

EVALUATION METRICS

LENS	METRIC	PURPOSE
Point-wise	**RMSE, MAE**	*Reconstruction error*
Distributional	**Wasserstein-1**	*Marginal similarity*
Moments	**Mean shift, var. ratio**	*Statistical preservation*
Joint structure	**Correlation Frobenius**	*Dependency preservation*
Downstream	**Top-K beam accuracy**	*Task impact*

09 · CONCLUSION

- **No method dominates** — the missingness **mechanism** governs performance.
- **kNN is the most robust default**: the only method that never collapses.
- **HyperImpute** leads on both reconstruction and distributional fidelity — but only under cell-wise missingness.
- The three evaluation axes **disagree**: low error ≠ faithful distribution ≠ good downstream.

10 · LIMITATIONS & FUTURE WORK

Limitations: 2 datasets · 3 seeds · synthetic missingness · one downstream model.

Future work:

- Head-to-head against **time-series imputers** under row-wise loss.
- Extract **real blockage/dropout patterns** from traces, re-inject on complete data.