Evaluating Tabular and Time-Series Data Augmentation for 6G-Relevant Network-Performance Regression

Quinton den Haan
Supervisors: Yuandou Wang, Rihan Hai

1 Background

6G requires large amounts of data

- Possibly expensive to obtain, limited in coverage, noisy or imbalanced
- Data augmentation as a solution
- Tabular and time-series data
- No systematic comparison between augmentation methods from different families on 6G data

2 Research Questions

How do different tabular and time-series augmentation techniques compare when addressing data scarcity in datasets relevant to future 6G systems?

- Downstream regression performance?
- Preservation statistical realism?

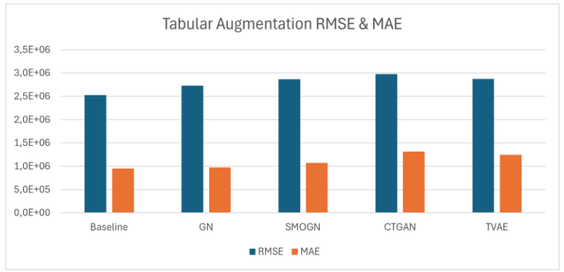
3 Experimental Pipeline

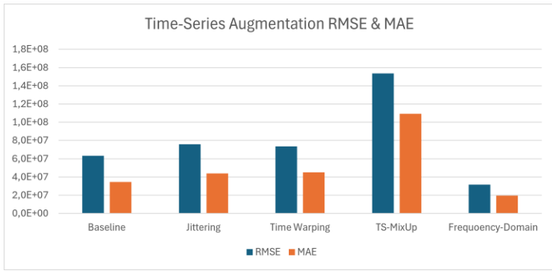
4 Datasets

Zenodo 5G network performance datasets

- AMF performance dataset
- Python web-server performance dataset
- Contains resource allocation, runtime behaviour and response-time measurements.

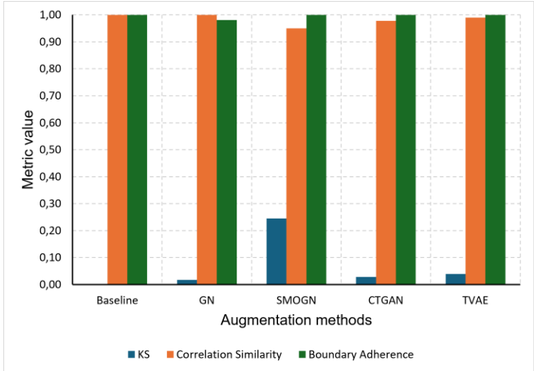
7 Downstream Regression

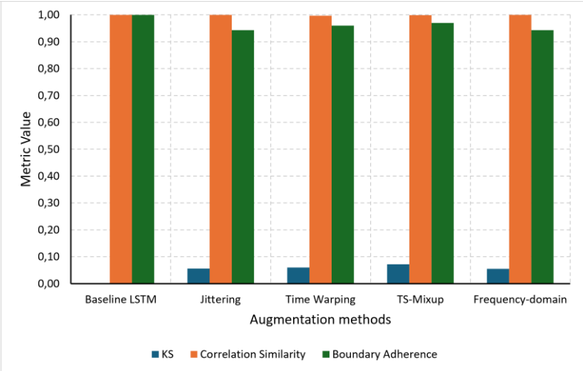




- No improvement over the baseline for tabular augmentation
- Only frequency-domain offers improvement, TS-MixUp significantly worse

8 Statistical Realism





- Realistic-looking augmented data did not necessarily improve prediction.

9 Augmentation Methods

Gaussian Noise	Adds small random feature noise	Tabular
SMOGN	Oversamples rare regression target regions	
CTGAN	Generates synthetic tabular rows adversarially	
TVAE	Samples tabular rows from latent space	
Jittering	Adds noise to each timestep	Time-Series
Time Warping	Stretches or compresses temporal patterns	
TS-MixUp	Linearly mixes windows and targets	
Frequency-domain	Perturbs Fourier coefficients of sequences	

10 Evaluation Metrics

RMSE, MAE, R^2	Downstream performance
Kolmogorov-Smirnov	Distribution shift
Correlation Similarity	Relationship preservation
Boundary Adherence	Range preservation

11 Conclusions

- Augmentation did not consistently improve regression performance.
- All tabular methods performed worse than the XGBoost baseline.
- Frequency-domain augmentation was the only time-series method that improved performance.
- High statistical realism ≠ better prediction.

12 Limitations

- Only two datasets evaluated. 5G/cloud datasets used as proxies.
- Only one downstream model per modality. LSTM baseline generalized poorly.
- Fixed hyperparameters and one seed.
- Metrics do not capture physical validity.

Delft University of Technology

CSE3000 Research Project

Q.T.denHaan@student.tudelft.nl