Center for Responsible AI

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Online Data Augmentation for Forecasting with Deep Learning

Vitor Cerqueira[1], Moisés Santos[1], Luis Roque[1], Yassine Baghoussi[3], Carlos Soares[1,2]

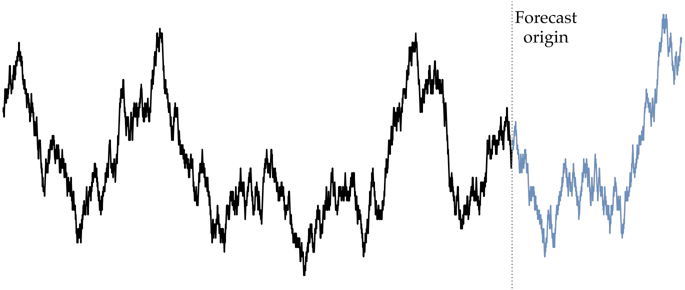
1. Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

2. Fraunhofer Portugal AICOS, Portugal

3. INESC TEC, Portugal

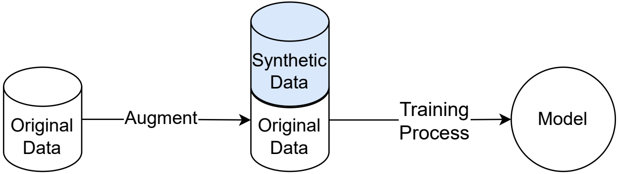
October 1, 2025

Problem setting



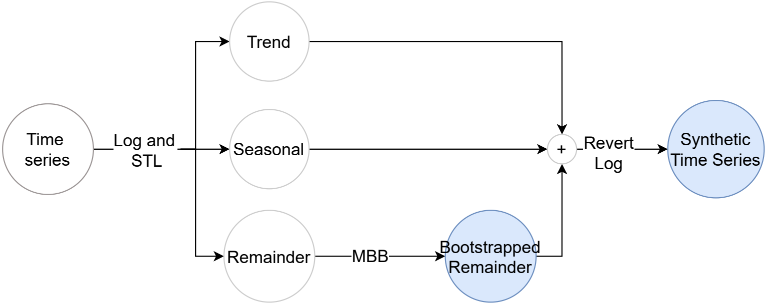
- Univariate Time Series Forecasting
  - Datasets with multiple time series
  - Limited data for training

Data Augmentation



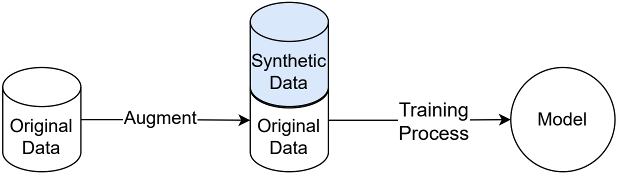
- Augmenting training datasets with synthetic data
  - Typically done apriori (before model training)
  - Need to store synthetic observations

Method Example - STL+MBB



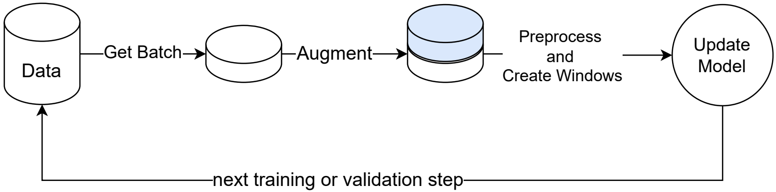
- Decomposition-based moving block bootstrapping
- Sampling with replacement ensures variability

Research Questions



- RQ: Can we do the augmentation process online?
  - Is it effective relative to an offline approach?
  - Does it work for different architectures and augmentation methods?

Online Augmentation



- Augmentation batch by batch
  - During both training and validation
  - Agnostic to architecture or augmentation strategy
  - Using standardized time series
  - Each batch has a 50/50 mix of real and synthetic instances

Dataset

Center for Responsible AI

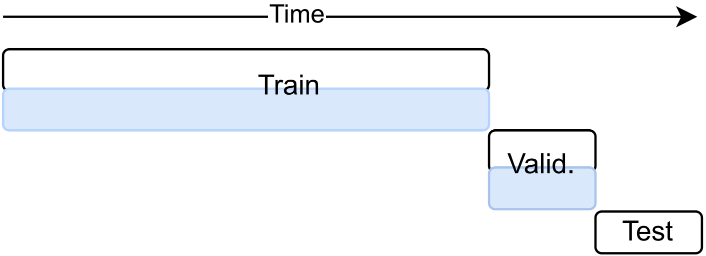
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Table 1 Summary of the datasets: average value, number of time series, number of observations, seasonal period, and forecasting horizon.

Dataset	Average value	# time series	# observations	Period	h
M1 Monthly	72.7	617	44892	12	12
M1 Quarterly	40.9	203	8320	4	8
M3 Monthly	117.3	1428	167562	12	12
M3 Quarterly	48.9	756	37004	4	8
Tourism Monthly	298.5	366	109280	12	12
Tourism Quarterly	99.6	427	42544	4	8
Total	-	3797	409602	-	-

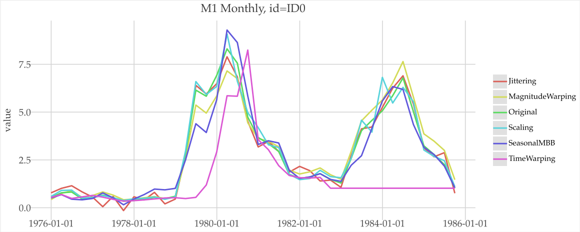
- 6 benchmark datasets
- Low sampling frequency
  - Often limited in sample size

Validation split approach



- Last h observations for testing. h observations before those for validation
  - per time series

Architectures and evaluation

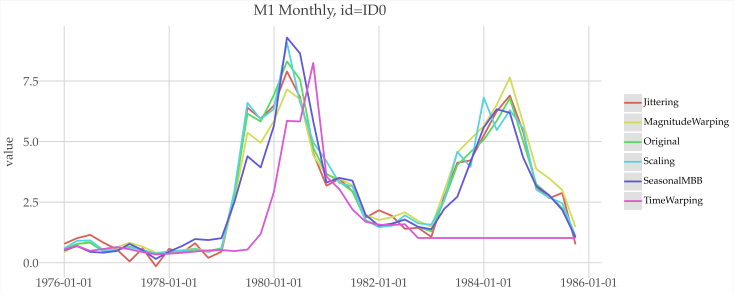


- MLP
- KAN
- NHITS

- Evaluated with:

$$\text{MASE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{n-m}\sum_{i=m+1}^{n}|y_i - y_{t-m}|}$$

Synthetic data generation methods



- 7 augmentation techniques: Jittering, Magnitude warping, Time warping, MBB, Scaling, DBA, TSMixup

Synthetic data generation methods

Table 2 Parameters of the time series synthetic generation methods.

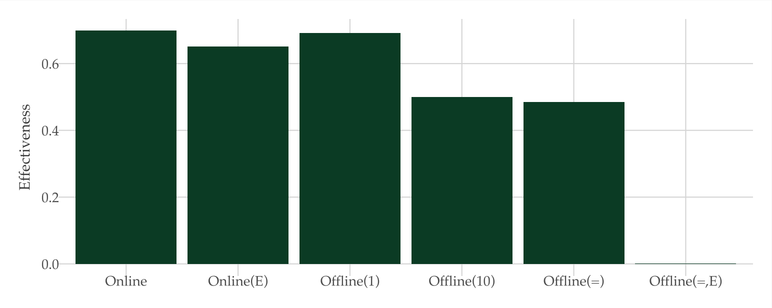
Method	Parameter	Values
`MBB`	log	{**True**, False}
`Jittering`	s	{**0.03**, 0.05, 0.1, 0.15, 0.2, 0.3}
`Scaling`	σ scaling factor in $\mathcal{N}(1, \sigma^2)$	{0.03, 0.05, **0.1**, 0.15, 0.2, 0.3}
`M-Warp`	σ scaling factor in $\mathcal{N}(1, \sigma^2)$ # knots	{0.05, **0.1**, 0.15} {3, **4**, 5}
`T-Warp`	σ scaling factor in $\mathcal{N}(1, \sigma^2)$ # knots	{0.05, **0.1**, 0.15} {3, **4**, 5}
`DBA`	Max # time series Dirichlet concentration	{5, 7, **10**, 15} {**1.0**, 1.5, 2.0}
`TSMixup`	Max # time series Dirichlet concentration	{5, 7, **10**, 15} {**1.0**, 1.5, 2.0}

- Parameter configuration pool for synthetic data generators

Augmentation strategies

- `Online`: The proposed online data augmentation scheme with a batch size of 32 time series.
- `Online(E)`: Parameters of the data generation methods are randomly sampled;
- `Offline(1)`: An approach that does data augmentation before the fitting process;
- `Offline(10)`: `Offline(1)`, but creating 10 synthetic time series instead of 1;
- `Offline(=)`: Creating a number of synthetic time series to match the synthetic sample size created by the `Online` approach.
- `Offline(=, E)`: `Offline(=)` + Parameters of the data generation methods are randomly sampled

Overall effectiveness



- Probability of outperforming the baseline (no augmentation)
  - Online variants are competitive with Offline(1)

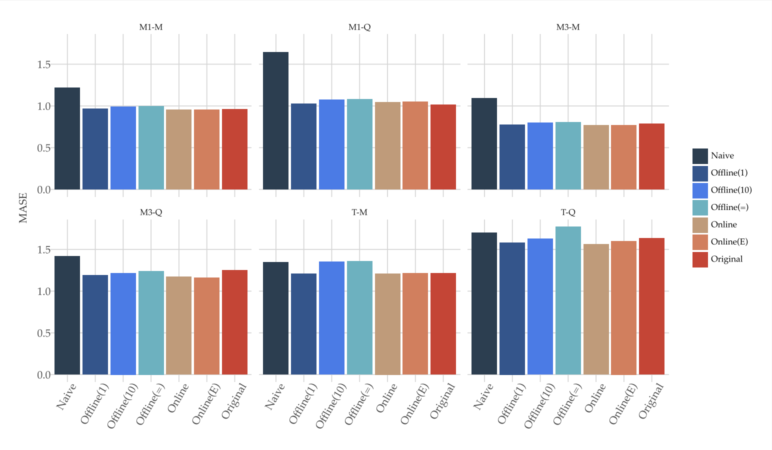
Results per dataset

Center for Responsible AI

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Online Data Augmentation

Cerqueira et al.



- Online variants are competitive in all datasets

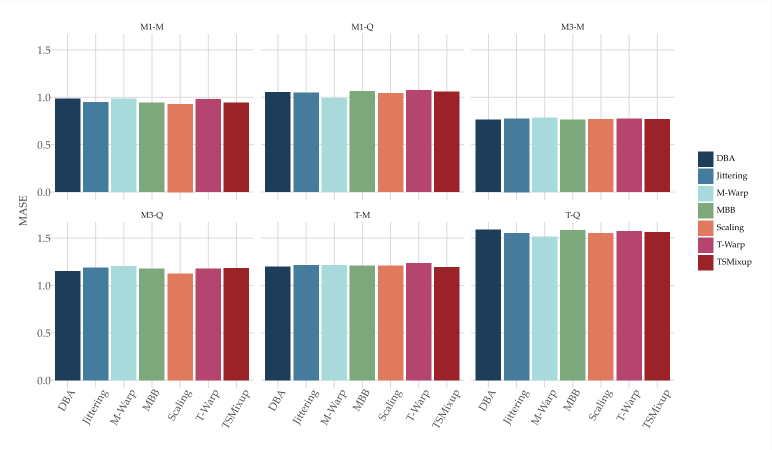
Results per synthetic data generator

Center for Responsible AI

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

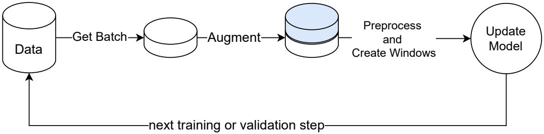
Online Data Augmentation

Cerqueira et al.



- Varying relative performance on synthetic data generators

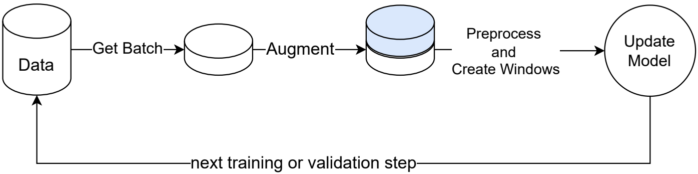
Conclusions



- Online data augmentation is an effective approach for addressing limited data problems
  - In neural-based time series forecasting
  - Consistent results across datasets

- Synthetic data generators show varying relative performance

Next steps



- Can we learn what's the best augmentation during training?
  - Curriculum learning

metaforecast


```python
from metaforecast.synth.callbacks import OnlineDataAugmentationCallback
from metaforecast.synth import SeasonalMBB

tsgen = SeasonalMBB(seas_period=12)

augmentation_cb = OnlineDataAugmentationCallback(generator=tsgen)
```

```python
from neuralforecast import NeuralForecast
from neuralforecast.models import NHITS

models = [NHITS(input_size=horizon,
                h=horizon,
                start_padding_enabled=True,
                accelerator='mps'),
          NHITS(input_size=horizon,
                h=horizon,
                start_padding_enabled=True,
                accelerator='mps',
                callbacks=[augmentation_cb])]

nf = NeuralForecast(models=models, freq='ME')
```

github.com/vcerqueira/metaforecast

Q&A

Thank you!

Responsible AI Research | Privacy, Auditing, and Compliance Tools

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU