Forecasting with Deep Learning

Beyond Average of Average of Average Performance

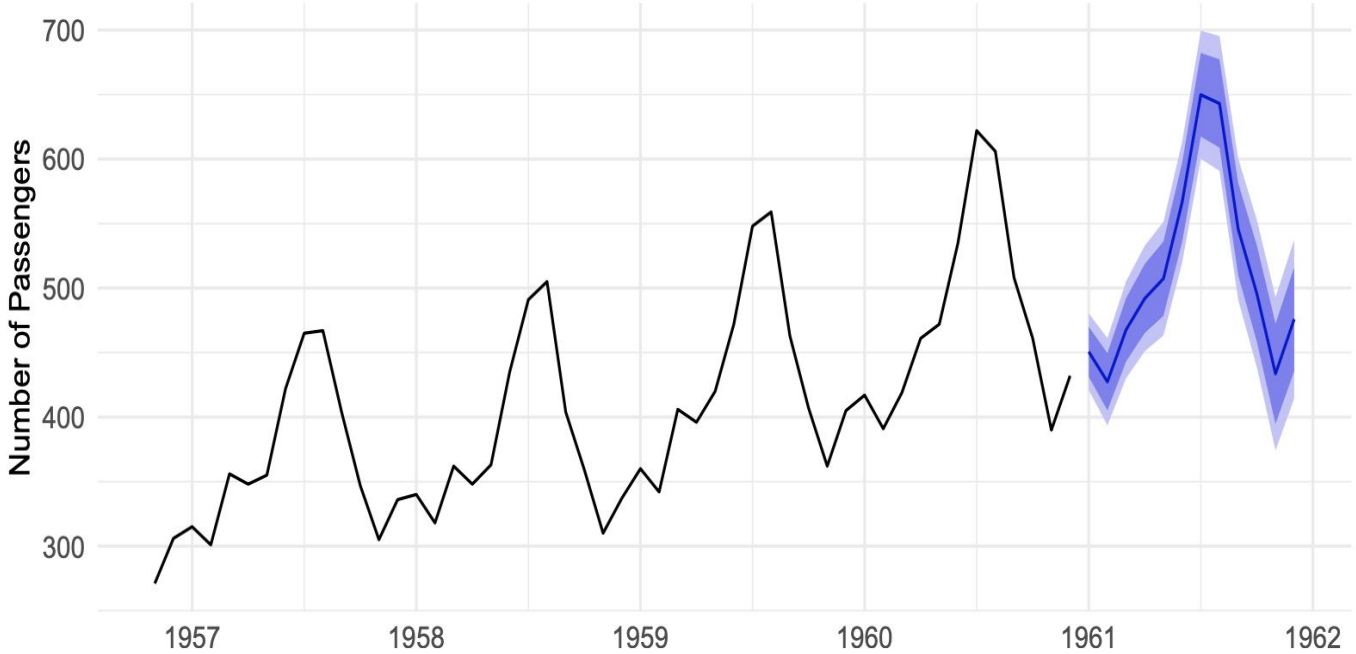
Vitor Cerqueira[1,2], Luis Roque[1,2], Carlos Soares[1,2,3]

cerqueira.vitormanuel@gmail.com

1 Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
2 Laboratory for Artificial Intelligence and Computer Science (LIACC), Portugal
3 Fraunhofer Portugal AICOS, Portugal

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU
Center for Responsible AI

Problem definition: Univariate time series forecasting



Relevant task in various application domains, e.g. retail, industry.

Motivation

Adequate evaluation is important

- Getting reliable performance estimates
- Accurate model selection
- Ensuring reliable predictions

Current evaluation practices

- Focus on single summary statistics
  - e.g., average SMAPE
- Dilute information under specific conditions
- Fail to highlight models' limitations

Table 1: Performance on the M4, M3, TOURISM test sets, aggregated over each dataset. Evaluation metrics are specified for each dataset; lower values are better. The number of time series in each dataset is provided in brackets.

M4 Average (100,000)			M3 Average (3,003)		TOURISM Average (1,311)	
	sMAPE	OWA		sMAPE		MAPE
Pure ML	12.894	0.915	Comb S-H-D	13.52	ETS	20.88
Statistical	11.986	0.861	ForecastPro	13.19	Theta	20.88
ProLogistica	11.845	0.841	Theta	13.01	ForePro	19.84
ML/TS combination	11.720	0.838	DOTM	12.90	Stratometrics	19.52
DL/TS hybrid	11.374	0.821	EXP	12.71	LeeCBaker	19.35
N-BEATS-G	11.168	0.797		12.47		**18.47**
N-BEATS-I	11.174	0.798		12.43		18.97
N-BEATS-I+G	**11.135**	**0.795**		**12.37**		18.52

The above are averages over several:

- Horizons
- Forecast origins
- Time series

Table source: Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations, 2020*.

Towards an Aspect-based Evaluation

Current evaluation practices

- Focus on single summary statistics
  - e.g., average SMAPE
- Dilute information under specific conditions
- Fail to highlight models' limitations

Controlling forecasting accuracy for several dimensions

- Helps identify conditions where models excel or struggle
- Enable more informed model selection
  - Better understanding of trade-offs between different models
- Increased overall responsible use of forecasting in diverse problems

Proposed Aspect-based Evaluation Framework

Analyses performance across multiple dimensions

- Sampling frequency
- Forecasting horizon
- Problem *difficulty*
- In anomalies – observations outside 99% prediction interval

Uses several summarisation techniques

- Overall performance
- Expected shortfall – performance in worst 5% of cases
- Win/loss ratios – for a non-parametric comparison
- All of the above, while incorporating region of practical equivalence

Experiments

Use case: comparing NHITS with classical forecasting approaches

Classical approaches:

- ARIMA
- ETS
- Seasonal Naive
- Random walk with drift
- Simple exponential smoothing
- Theta

NHITS:

- State-of-the-art neural network based on stacks of MLP’s
- Highly efficient (50 times faster than transformers)
- Several works have shown it to be more accurate than several transformers and RNNs

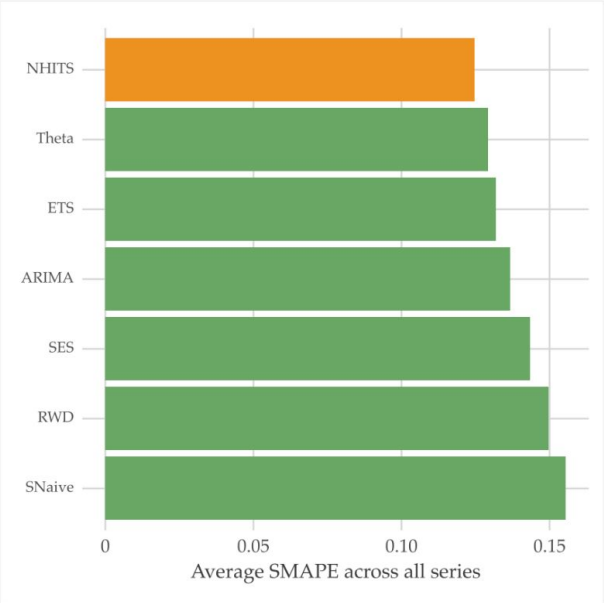
Experiments

On three benchmark datasets

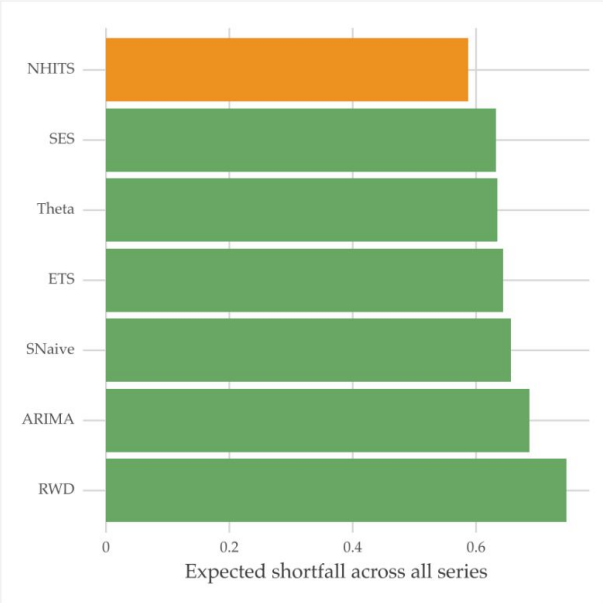
Table 1: Summary of the datasets: number of time series, number of observations, forecast horizon, number of lags, and frequency.

		# time series	# observations	H	p	Frequency
M3	Monthly	1428	167562	18	23	12
	Quarterly	756	37004	8	10	4
	Yearly	645	18319	6	8	1
M4	Monthly	48000	11246411	18	23	12
	Quarterly	24000	2406108	8	10	4
	Yearly	23000	858458	6	8	1
Tourism	Monthly	366	109280	18	23	12
	Quarterly	427	42544	8	10	4
	Yearly	518	12678	6	8	1
Total		99140	14898364	-	-	-

Results - Overall performance



(a) Average SMAPE



(b) SMAPE expected shortfall

Fig. 1: Average SMAPE (a) and expected shortfall (b) for each model across all time series

NHITS shows the best overall performance, and also when considering worst-case scenarios

Results - Controlling for forecasting horizon

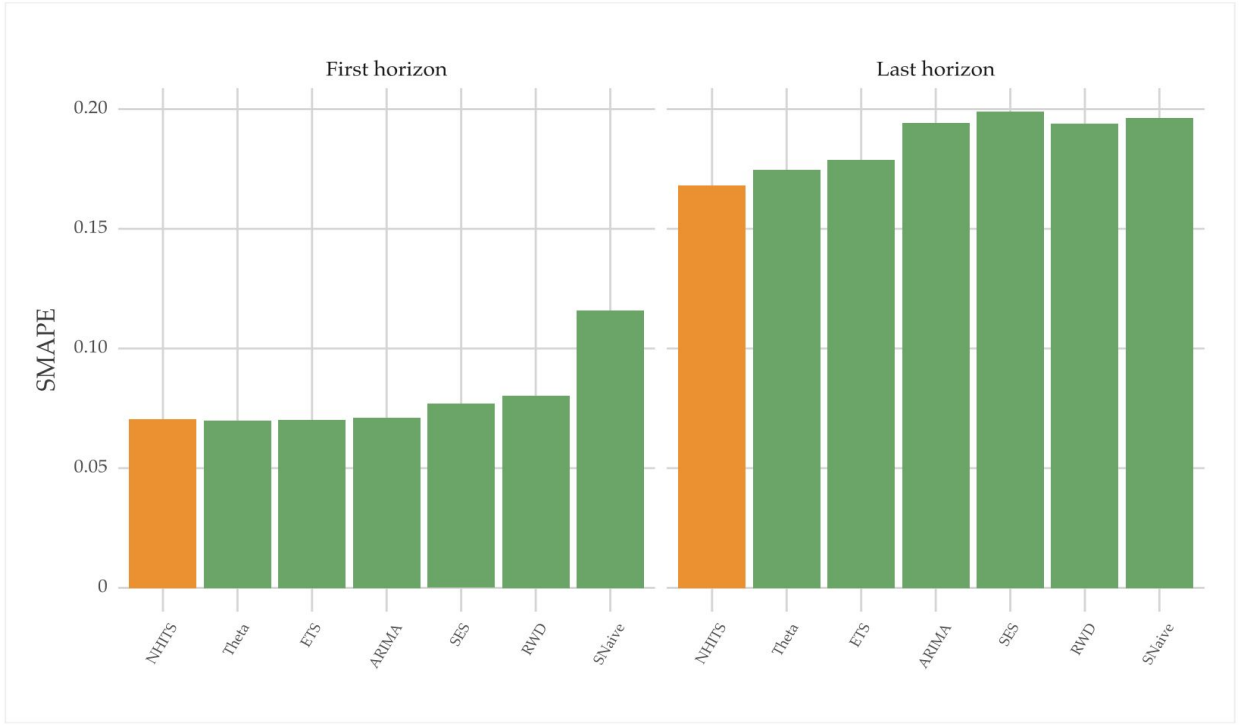
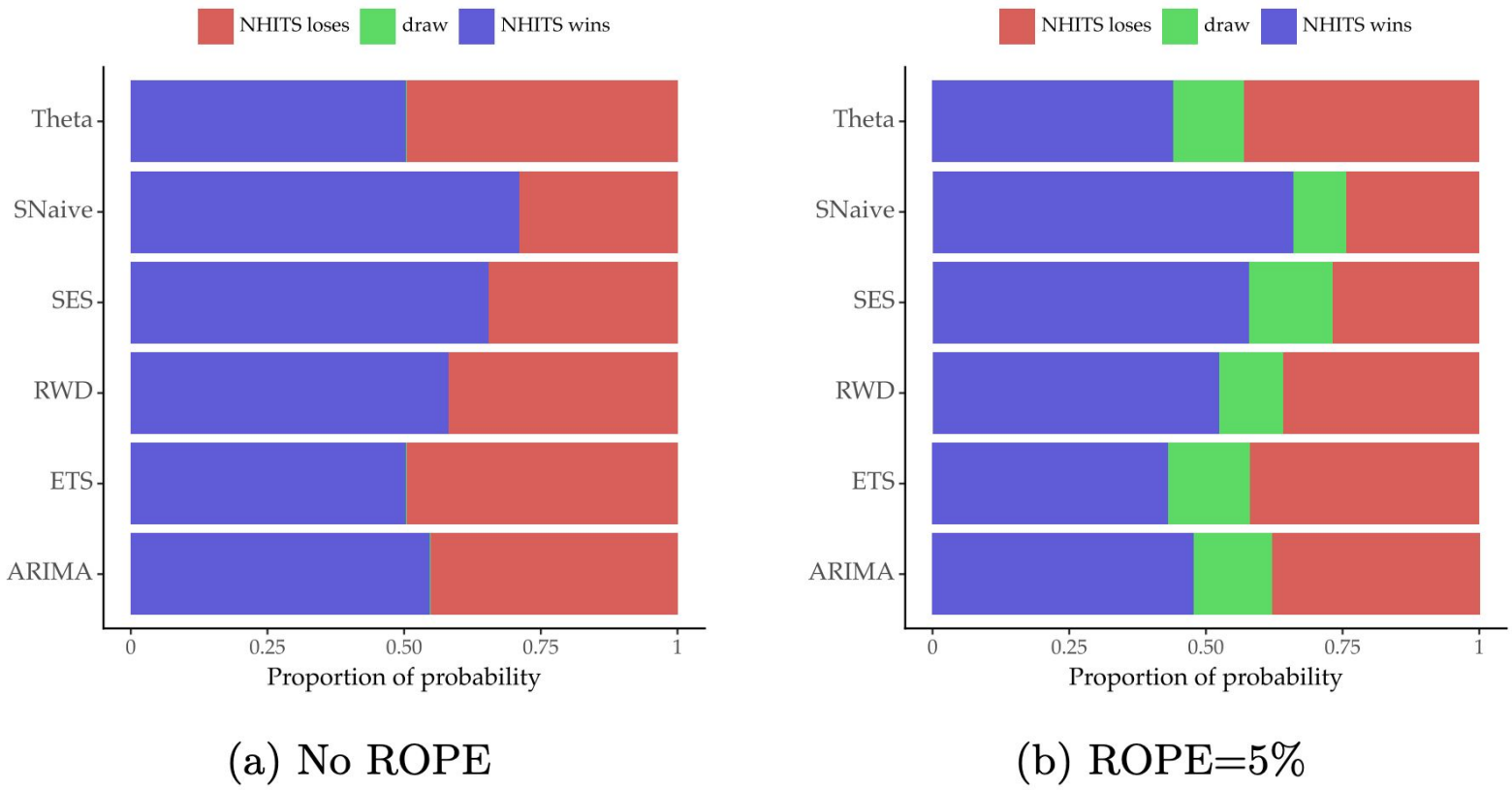


Fig. 3: SMAPE scores by model and forecasting horizon.

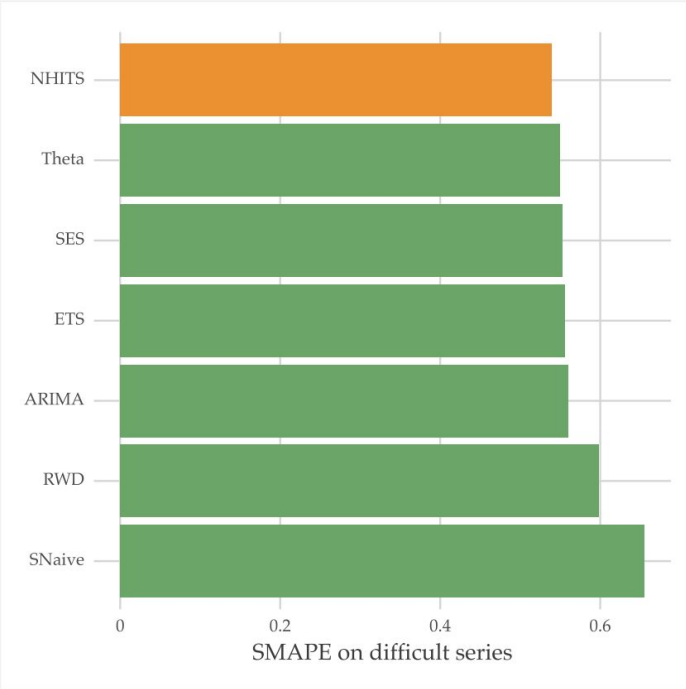
NHITS only outperforms the best classical approaches for multi-step forecasting

Results - Win/Loss ratios

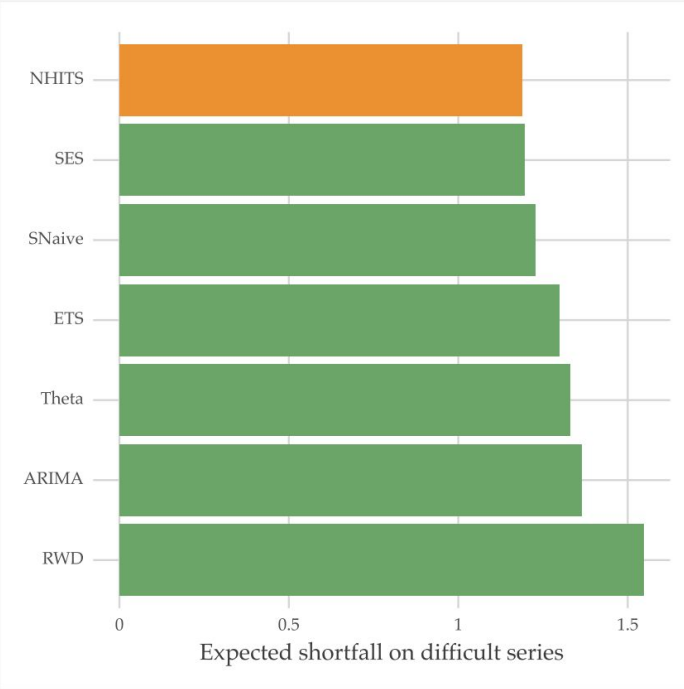


In many pairwise comparisons, high chance of NHITS being outperformed

Results - On difficult problems



(a) Average SMAPE

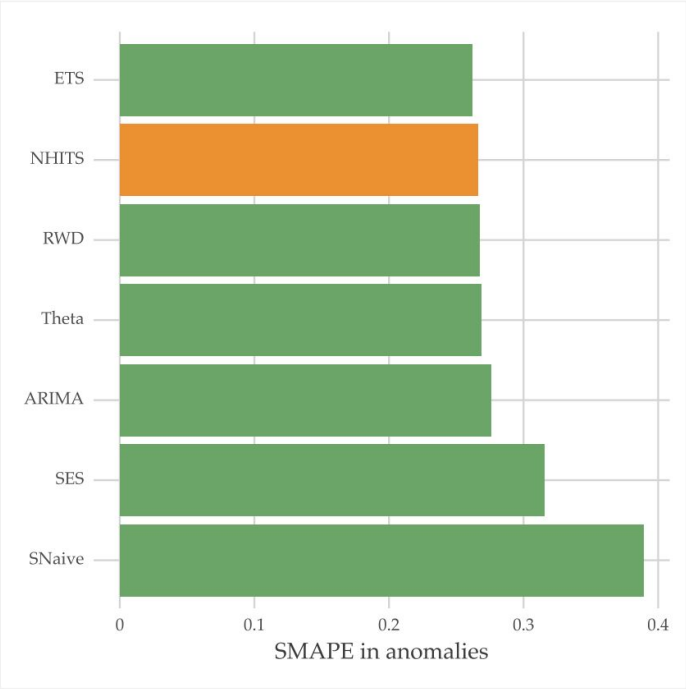


(b) Expected shortfall

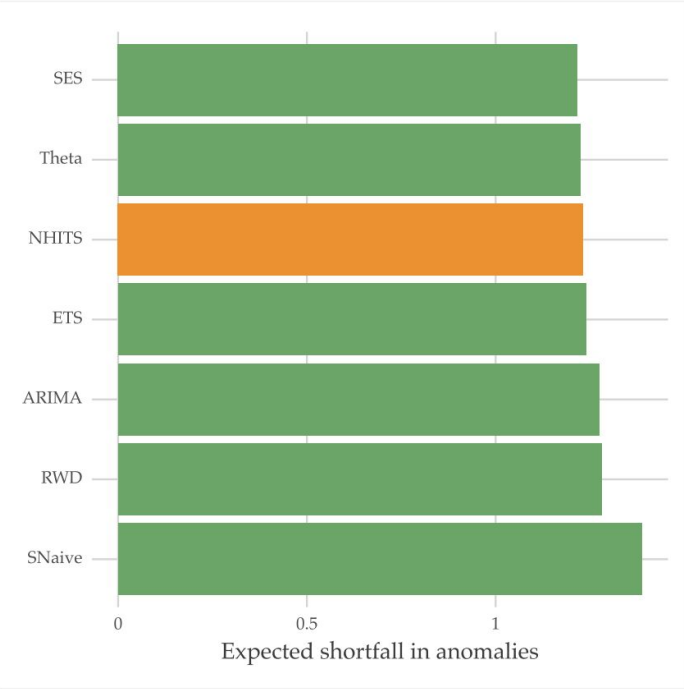
Difficult problems: Time series where a baseline (Seasonal Naive) performs poorly

Result: NHITS still the best, though magnitude of difference is smaller

Results - Performance on Anomalies



(a) Overall SMAPE



(b) Expected shortfall

Exponential smoothing is better at handling anomalous observations

Conclusions

- Reliable model evaluation is important for responsible forecasting
- Current practices focus on summarising performance into a single score
- … which can dilute relevant information about models' behaviour

- Proposed approach: aspect-based evaluation
- In a case study, we found that:
  - NHITS outperforms several classical forecasting techniques
  - But its superiority depends on several factors

- Results highlight the value of a more nuanced analysis in forecasting evaluation

Thanks

Repeat the experiments here:

https://github.com/vcerqueira/modelradar

Funding: “Agenda ``Center for Responsible AI’’, nr. C645008882-00000055, investment project nr. 62, financed by the Recovery and Resilience Plan (PRR) and by European Union - NextGeneration EU”

PRR Plano de Recuperação e Resiliência

REPÚBLICA PORTUGUESA

Financiado pela União Europeia NextGenerationEU

Center for Responsible AI