Forecasting Hierarchical Time Series using Non-linear Mappings

Shanika Wickramasuriya

jointly with

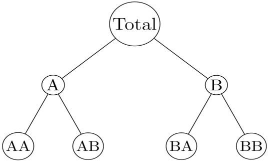
K. Bandara, H. Hewamalage and M. Perera

July 11, 2022

Outline

Hierarchical time series

- A collection of time series with aggregation constraints.



e.g. Tourism Demand: Australia, States, SLAs

Challenge: Independent forecasts do not add-up across the hierarchy.

Solutions:

- Top-down
- Bottom up (BU)
- Middle-out
- MinT

Forecast reconciliation

Traditional linear hierarchical forecasting methods:

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{P}_h\hat{\boldsymbol{y}}_{T+h|T},$$

- $\tilde{\boldsymbol{y}}_{T+h|T}$, $\hat{\boldsymbol{y}}_{T+h|T}$: h-step ahead **reconciled** and **base** forecasts stacked in the same order as $\boldsymbol{y}_t$.
- $\boldsymbol{P}_h$ depends on the forecast reconciliation approach.
- $\boldsymbol{S}$ is the summing matrix.

BU $\boldsymbol{P}_{\text{BU}} = \left[\, \boldsymbol{0}_{n\times(m-n)} \;\vdots\; \boldsymbol{I}_n \,\right]$

MinT $\boldsymbol{P}_{\text{MinT}} = (\boldsymbol{S}^\top \boldsymbol{\Lambda}_h{}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^\top \boldsymbol{\Lambda}_h{}^{-1}$

$\boldsymbol{\Lambda}_h$: p.d. covariance matrix of the h-step ahead base forecast errors. For $k_h > 0$:

$\boldsymbol{\Lambda}_h = k_h\boldsymbol{I}$	$\boldsymbol{\Lambda}_h = k_h\text{diag}(\hat{\boldsymbol{\Lambda}}_1)$	$\boldsymbol{\Lambda}_h = k_h\hat{\boldsymbol{\Lambda}}_{\text{shr}}$
OLS	WLS	MinT(Shr)
	$\hat{\boldsymbol{\Lambda}}_1$: sample cov.	$\hat{\boldsymbol{\Lambda}}_{\text{shr}}$: shrunk cov.

ERM $\boldsymbol{P}_{\text{ERM}} = \boldsymbol{B}^\top\hat{\boldsymbol{Y}}(\hat{\boldsymbol{Y}}^\top\hat{\boldsymbol{Y}})^{-1}$

Outline

Loss function using non-linear mappings

- Empirical studies by Spiliotis et al. (2021) demonstrated that non-linear mappings could lead to better forecast accuracy.
- Relaxing unbiasedness of base/reconciled forecasts could reduce the mean squared forecast error (Rangapuram et al., 2021; Wickramasuriya, 2021).

Given $\boldsymbol{y}_t = [\boldsymbol{a}_t^\top, \boldsymbol{b}_t^\top]^\top$, and in-sample fitted values $\hat{\boldsymbol{y}}_{t|t-1}$, for $t = 1, 2, \dots, T$

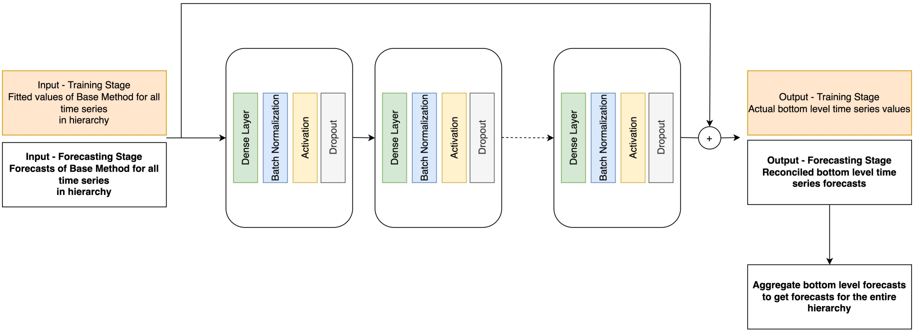
$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \left\| \boldsymbol{b}_t - \boldsymbol{f}(\hat{\boldsymbol{y}}_{t|t-1}, \boldsymbol{\theta}) \right\|_2^2 + \lambda \left\| \boldsymbol{a}_t - \boldsymbol{C}\boldsymbol{f}(\hat{\boldsymbol{y}}_{t|t-1}, \boldsymbol{\theta}) \right\|_2^2, \qquad \lambda > 0,$$

- n is the number of bottom level series
- $\boldsymbol{f}(\cdot, \boldsymbol{\theta}) = [f_1(\cdot, \boldsymbol{\theta}_1), f_2(\cdot, \boldsymbol{\theta}_2), \dots f_n(\cdot, \boldsymbol{\theta}_n)]^\top$, $f_j(\cdot, \boldsymbol{\theta}_j)$ is a non-linear mapping function with parameter vector $\boldsymbol{\theta}_j$ for $j = 1, 2, \dots, n$
- $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \dots, \boldsymbol{\theta}_n^\top]^\top$
- $\boldsymbol{S}^\top = \left[\boldsymbol{C}^\top \,\vdots\, \boldsymbol{I}_n \right]$

$\boldsymbol{\theta}$ is estimated using a feed forward neural network.

Proposed reconciliation network

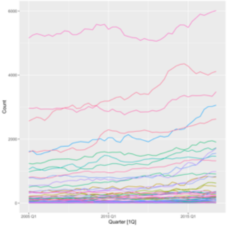
$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \left\| \boldsymbol{b}_t - \boldsymbol{f}(\hat{\boldsymbol{y}}_{t|t-1}, \boldsymbol{\theta}) \right\|_2^2 + \lambda \left\| \boldsymbol{a}_t - \boldsymbol{C}\boldsymbol{f}(\hat{\boldsymbol{y}}_{t|t-1}, \boldsymbol{\theta}) \right\|_2^2, \qquad \lambda > 0$$



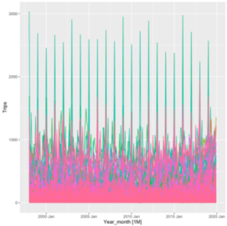
Outline

Description of data sets

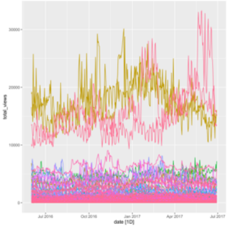
Data set	Frequency	T	No. of levels	No. of series	h
Australian prison population	4 (quarterly)	48	5	121	8
Australian domestic tourism	12 (monthly)	264	3	85	12
Wikipedia pageviews	7 (weekly)	394	6	1095	7
Australian labour market	4 (quarterly)	128	4	57	12



Prison



Tourism



Wikipedia

Labour

Experimental setup

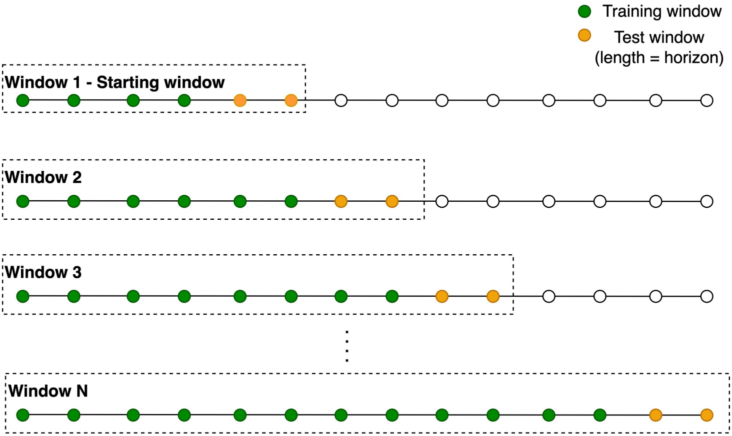
Forecasting methods

- Univariate: ARIMA and ETS
- Global models: DeepAR and WaveNet
  - All the time series in the hierarchy are clustered using K-means algorithms and a global model is built for each cluster.

We conduct an expanding window evaluation for all the data sets.

Data set	No. of windows	Starting window size
Prison	3	24
Tourism	10	144
Wikipedia	10	324
Labour	5	68

Expanding Window



Hyper-parameter tuning and evaluation

- The hyper-parameters of the proposed method are tuned with HyperOpt – Bayesian Optimization.

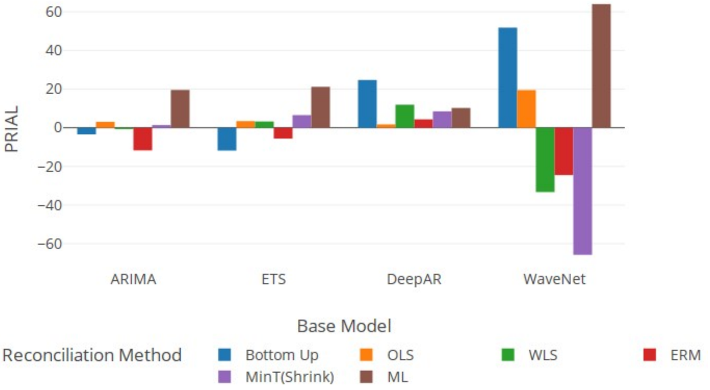
Hyper-parameter	Minimum	Maximum
Number of layers	1	5
Dropout rate	0	0.5
Learning rate	0.0001	0.1
Number of Epochs	10	200
Batch size	1	size of input data
Max norm	0	10
λ (proposed loss function)	0.01	5

- We trained this setup five times with different seeds and the average across these are taken as the final bottom level forecasts.
- The results are summarized using percentage relative improvement in average loss

$$\text{PRIAL} = \frac{\text{MSE(base-forecasts)} - \text{MSE(reconciled-forecasts)}}{\text{MSE(base-forecasts)}} \times 100\%$$

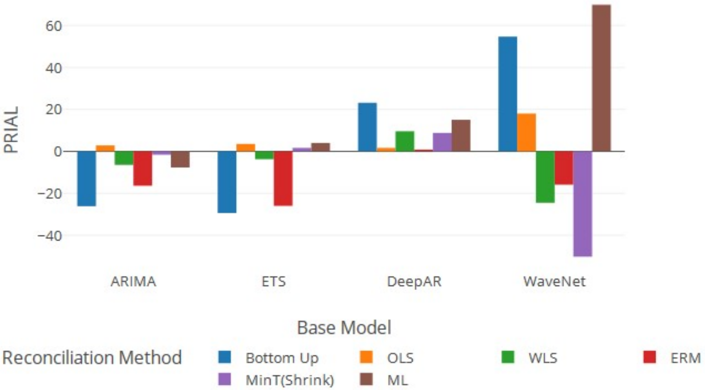
  - **+ve** values: accuracy of reconciled forecasts has increased.

Results for prison data ($h = 1 : 4$)



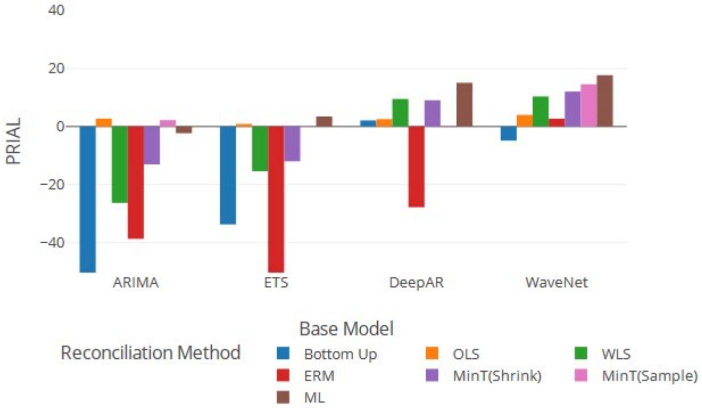
Rank	ARIMA	ETS	DeepAR	WaveNet
1	**Proposed ML**	**Proposed ML**	Bottom Up	**Proposed ML**
2	OLS	MinT(Shrink)	WLS	Bottom Up
3	MinT(Shrink)	OLS	**Proposed ML**	OLS

Results for prison data ($h = 1:8$)



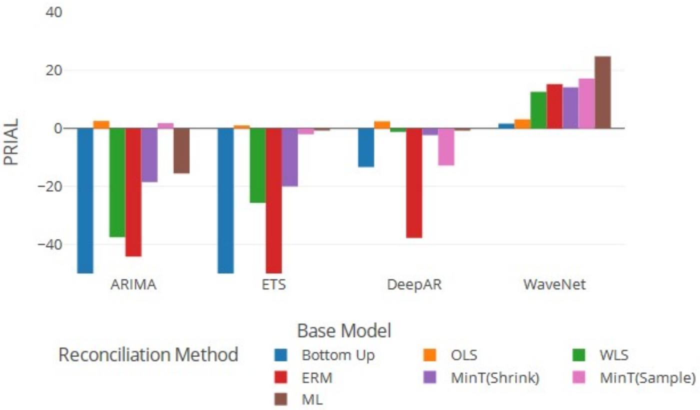
Rank	ARIMA	ETS	DeepAR	WaveNet
1	OLS	**Proposed ML**	Bottom Up	**Proposed ML**
2	MinT(Shrink)	OLS	**Proposed ML**	Bottom Up
3	WLS	MinT(Shrink)	WLS	OLS

Results for tourism data ($h = 1:6$)



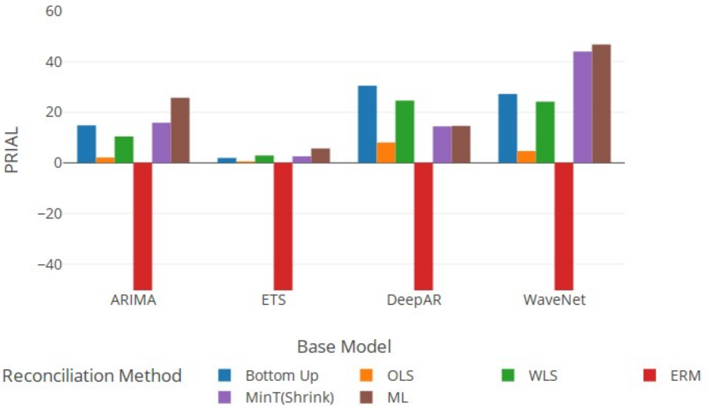
Rank	ARIMA	ETS	DeepAR	WaveNet
1	OLS	**Proposed ML**	**Proposed ML**	**Proposed ML**
2	MinT(Sample)	OLS	WLS	MinT(Sample)
3	**Proposed ML**	MinT(Sample)	MinT(Shrink)	MinT(Shrink)

Results for tourism data ($h = 1 : 12$)



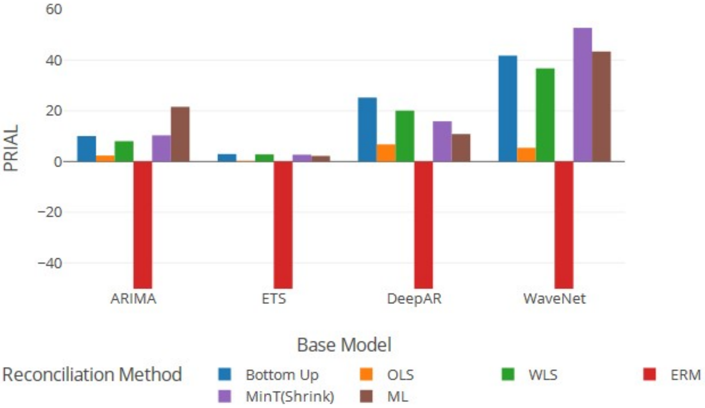
Rank	ARIMA	ETS	DeepAR	WaveNet
1	OLS	OLS	OLS	**Proposed ML**
2	MinT(Sample)	**Proposed ML**	**Proposed ML**	MinT(Sample)
3	**Proposed ML**	MinT(Sample)	MinT(Shrink)	ERM

Results for Wikipedia data ($h = 1:3$)



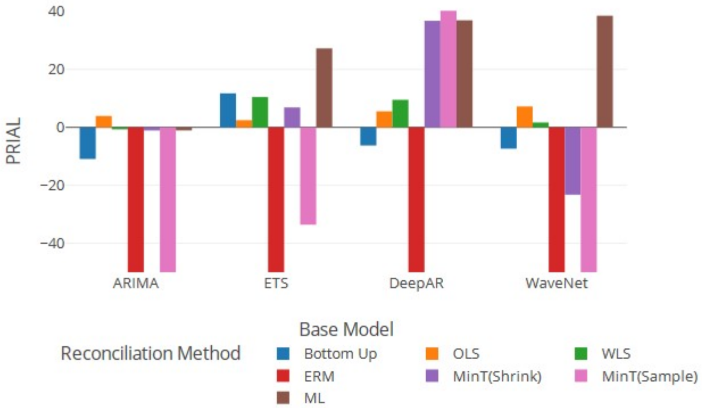
Rank	ARIMA	ETS	DeepAR	WaveNet
1	**Proposed ML**	**Proposed ML**	Bottom Up	**Proposed ML**
2	MinT(Shrink)	WLS	WLS	MinT (Shrink)
3	Bottom Up	MinT(Shrink)	**Proposed ML**	Bottom Up

Results for Wikipedia data ($h = 1:7$)



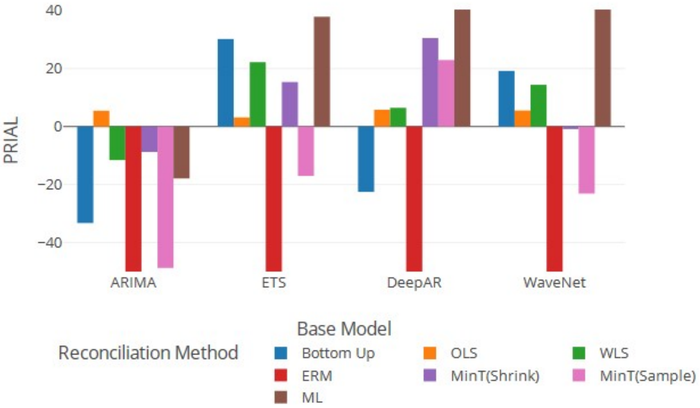
Rank	ARIMA	ETS	DeepAR	WaveNet
1	**Proposed ML**	Bottom Up	Bottom Up	MinT(Shrink)
2	MinT(Shrink)	WLS	WLS	**Proposed ML**
3	Bottom Up	MinT(Shrink)	MinT(Shrink)	Bottom Up

Results for labour data ($h = 1 : 6$)



Rank	ARIMA	ETS	DeepAR	WaveNet
1	OLS	**Proposed ML**	MinT(Sample)	**Proposed ML**
2	WLS	Bottom Up	**Proposed ML**	OLS
3	**Proposed ML**	WLS	MinT(Shrink)	WLS

Results for labour data ($h = 1 : 12$)



Rank	ARIMA	ETS	DeepAR	WaveNet
1	OLS	**Proposed ML**	**Proposed ML**	**Proposed ML**
2	MinT(Shrink)	Bottom Up	MinT(Shrink)	Bottom Up
3	WLS	WLS	MinT(Sample)	WLS

Outline

Conclusions

1. We proposed a non-linear hierarchical time series forecasting approach using machine learning techniques.
2. We introduced a novel loss function incorporating non-linear mappings to obtain coherent forecasts from the individual base forecasts.
3. To obtain the weights of the non-linear mappings between the base forecasts, we trained a feed-forward neural network.
4. The empirical results suggest that the proposed method is generally ranked among the best three methods for obtaining coherent forecasts.

THANK YOU!

s.wickramasuriya@auckland.ac.nz