Global Models for Time series Forecasting

Google Cloud, Machine Learning Group
September 01, 2021

Presenter

Kasun Bandara

School of Computing and Information Systems
Melbourne Centre For Data Science
University of Melbourne

About Me

- 2015 Graduated in Computer Science from University of Colombo School of Computing, Sri Lanka
- 2015 Joined WSO2 Inc. as a Software Engineer
- 2016-2020 Ph.D. in Computer Science, Monash University, Australia
  - Topic: Forecasting In Big Data With Recurrent Neural Networks
  - Machine Learning for Time Series Forecasting
  - Research Internship at Walmart Labs, San Francisco, USA
  - Research Scientist at Turning Point, Melbourne, Australia
  - Data Science Tutor, Faculty of IT, Monash University
- 2021 Research Fellow, Melbourne Centre For Data Science, University of Melbourne

About Me (2)

- Research Interests
  - Global Forecasting Models
  - Hierarchical Forecasting
  - Retail sales/demand forecasting
  - Demand forecasting (Retail, Energy, Health-care)
- Competition Fanatic
  - Fuzz-IEEE Competition on Explainable Energy Prediction (**2nd Place**)
  - M5 Forecasting Competition (**Gold Medalist; World Rank 17/5500, Australia Rank 2nd**)
  - IEEE CIS Energy Forecasting Competition (**World Rank 4/100, Australia Rank 1st**)
  - Air-Liquide Energy Forecasting Competition (**World Rank 4/350, Australia Rank 1st**)
  - ANZ Customer Segmentation Challenge (**Top Performer**)

Outline

Time Series Forecasting

- Process of making temporal predictions of the future based on past and present data.
- Accurate and reliable time series prediction is crucial in many industries.
  - Retail, food, railway, mining, tourism, energy, traffic and cloud-computing.
- Impact of Accuracy
  - Poor forecasting can be costly, Accurate forecasting can be considerably lucrative.

Big Data in Time Series Forecasting

- Large quantities of related, similar time series are available.

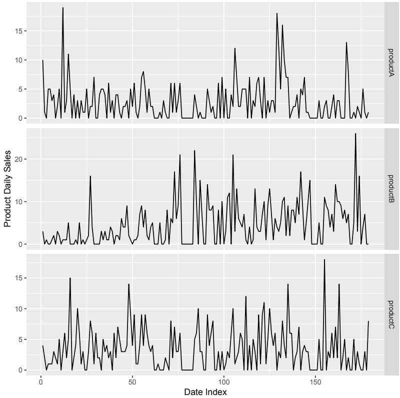


Figure: Daily sales demand of three different products over a four months period, extracted from *Walmart.com* [Bandara et al., 2019].

Vastness of related time series

- Large quantities of related, similar time series are available.
  - The sales demand in retail of thousands of different products.
  - The emergency medical services demand in multiple local government areas.
  - The multiple server performance measures in computer centers.
- State-of-the-art traditional forecasting techniques are mostly univariate methods.
  - Treat each time series separately and forecast them in isolation.
  - ETS, BaggedETS, Theta, ARIMA.
  - Unable to incorporate any key patterns and structures that may be shared by a group of time series
  - Single series may be too short to be forecast at all.

Outline

Substandard performance in forecasting competitions

- Sophisticated methods do not necessarily produce better forecasts than simpler ones [Makridakis and Hibon, 2000].
  - AutomatANN method in the M3 Competition.
- Did not perform well in the subsequent competitions.
  - NN3 and NN5 [Crone et al., 2011, Crone, 2008].
  - Held specifically for Computational Intelligent(CI) methods.
- Couldn't outperform simple standard methods in time series forecasting.
  - Theta method, simple exponential smoothing with drift [Hyndman and Billah, 2003].

Reasons for the under-performance

- Individual series are too short to be modelled effectively.
  - Amount of information that can be extracted is limited.
  - Higher probability of model over-fitting.
- Distant past is usually not very useful for forecasting.
- Neural networks do not perform well.
  - Not having enough data for learning [Zhang et al., 1998].
  - Not handle non-stationarity in the data adequately [Hyndman, 2016].
  - Large number of hyper-parameters to be determined [Yan, 2012].

Improving NNs for forecasting

- Preprocessing techniques.
  - Supplements the NN's learning process.
  - deseasonalizing and detrending data prior to modelling [Nelson et al., 1999, Ben Taieb et al., 2011, Zhang and Qi, 2005].
- Adapting NN architectures for forecasting
  - Ensemble architectures [Rahman et al., 2016, Barrow and Crone, 2016].
  - Generalized regression neural networks (GRNNs) [Yan, 2012].
  - Echo-state networks [Ilies et al., 2007].
  - Recurrent Neural Networks (RNNs) [Zimmermann et al., 2012, Fei and Yeung, 2015].

Global Forecast Models (GFMs)

- Methods that estimate model parameters jointly from all available time series [Januschowski et al., 2020].
- A unified forecasting model that is built using all a collection of time series.
    - Borrow similar behaviours and structures from other related time series.
    - Improves model generalizability.
    - Adequate data for model fitting.
    - Ability to exploit the cross-series information.
- Forecasting a large quantities of related time series: “Related” in terms of similarity of their DGP (not necessarily mere correlations) [Bergmeir, 2020]

Scalability of GFMs

- Enough data, due to more series, thus ML can be more competitive [Bergmeir, 2020]
    - Local model: typically fitting a model with few (<10) parameters to a single series.
    - If you have 10k series and fit 5 parameters, you end up with 50k parameters.
    - Fit a global model with 5k parameters instead.
- Overall complexity of set of local models grows when dataset grows; complexity of global model stays the same.

Complexity of GFMs

- Global models can afford to be more complex.
- Complexity can be added as:
  - Longer memory (longer input windows, more lags)
  - Non-linear/non-parametric models (NN variants, GBT, ...)
  - Data partitioning (Time series clustering)
- GFMs can be designed with a much higher complexity, yet still achieve better generalisation error than the univariate models for larger datasets [Montero-Manso and Hyndman, 2021]

GFMs are not multivariate models

- Global models learn across series but predict every series in isolation.
- They can work on datasets where series have different lengths and/or are not aligned, like the M3, M4 datasets.
- They do not take into account interactions between series.

Evolution of GFMs

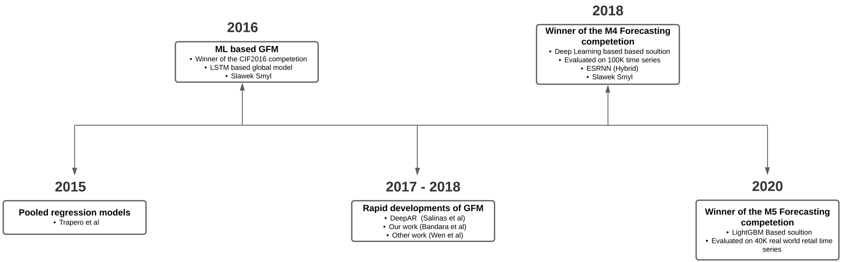


Figure: A brief overview of GFM developments

Moving window approach

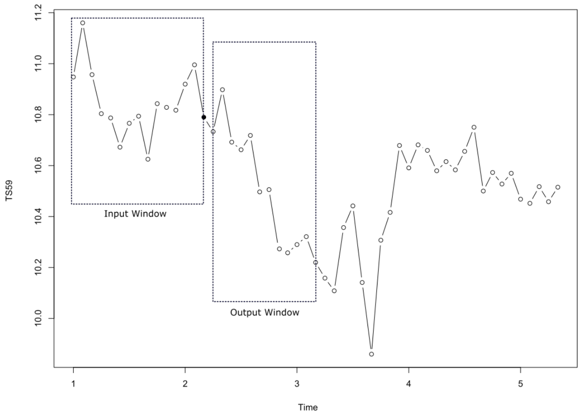


Figure: Applying the moving window approach over a time series . This is also known as the Multi-Input Multi-Output strategy (MIMO)

Dominance in the competitions

- CIF2016 Forecasting Competition (IEEE CIS)
  - LSTM based GFM solution [Smyl, 2016]
- Wikipedia Web Traffic Forecasting (Google, 2017)
  - RNN based Encoder-Decoder architecture [Suilin, 2018]
- M4 Competition (Makridakis, 2018)
  - ES-RNN: A hybrid architecture of RNNs and exponential smoothing [Smyl, 2019]
- M5 Competition (Makridakis, 2020)
  - LightGBM basedn GFM solution [Makridakis and Spiliotis, 2021]

Deep Learning based GFMs

- The recent NLP research (LSTM, attention, transformers) is adapted to the time series use case
  - most recent observations are the most important ones
  - long-term dependencies are relatively simple and stable (only seasonalities)
- Recurrent Neural Network based variants
  - Overview by [Hewamalage et al., 2021]
  - Stacking, Encoder-Decoder architectures
  - Have an internal state which allows them to memorize.
- Convolutional neural networks
  - Competitive as RNNs, but are a lot faster to train [Borovykh et al., 2017]
  - WaveNet architecture (causal convolutions, dilations) [Sen et al., 2019]

Specialised architectures

- DeepAR: Generative RNN model [Salinas et al., 2020]
- Deep state space models: Parametrizes a linear state-space model with an RNN [Rangapuram et al., 2018]
- NBEATS [Oreshkin et al., 2019]: Decomposes series into basis functions, residual stacking
  - State-of-the-art accuracy on M4 dataset.
  - 2nd place in M5 (as part of an ensemble)
- Transformers for forecasting [Li et al., 2019]
- Fusion Transformers [Lim et al., 2021]

GFMs for Forecast combination

- Ensembling works in forecasting just as well as in any other area.
- Local models to capture unique behaviours, Global models to capture common patterns.
- Ensembles of local and global models
  - Energy Demand Forecasting (IEEE CIS Competition)
  - Weekly time series forecasting [Godahewa et al., 2020]

Outline

Research Project 1

Research Question

Does building a notion of similarity between the time series assist the GFMs to distinguish the variations exist among a group of time series.

- Learning across these disparate set of time series may degenerate the overall accuracy of GFM models.
- A notion of similarity between the time series needs to be built into the global methods.
- Identify and account for the time series with homogeneous characteristics.

Research Project 1: Approach

- Building GFMs on subgroups of similar time series.
  - The Long Short-Term Memory Neural Networks (LSTMs) used as the primary GFM.
  - The similarity is captured through clustering the time series into subgroups.
  - Feature based clustering approach using kMeans, DBScan, Partition Around Medoids (PAM), and Snob.
- A prior time series clustering can supplement the GFM training procedure by improving the homogeneousness of the trainable time series.
  - Achieves competitive results on benchmarking datasets under competition evaluation procedures.

Research Project 1: Overall Architecture

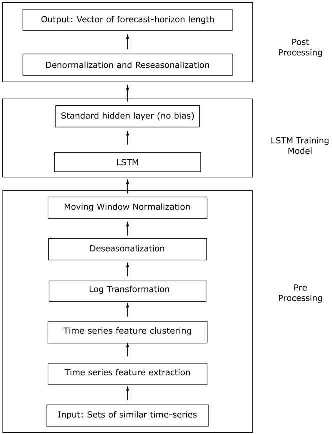


Figure: Pre-processing layer, LSTM training layer and a Post-processing layer.

Research Project 2

- Applying the clustering based GFM framework to forecast the demand in a E-commerce product assortment hierarchy
  - A product grouping strategy introduced to supplement the GFM learning schemes, in situations where sales patterns in a product portfolio are disparate.
- Empirically evaluated using real-world retail sales data from Walmart.com
  - A combination of static and dynamic features to model time series characteristics.
  - Product class, Product category, Calendar information (holidays, season etc.)
  - Evaluated on 20,000 items in the portfolio.
  - Outperforms the current forecasting pipeline at Walmart and many standard benchmarks.

Research Project 2: Overall Architecture

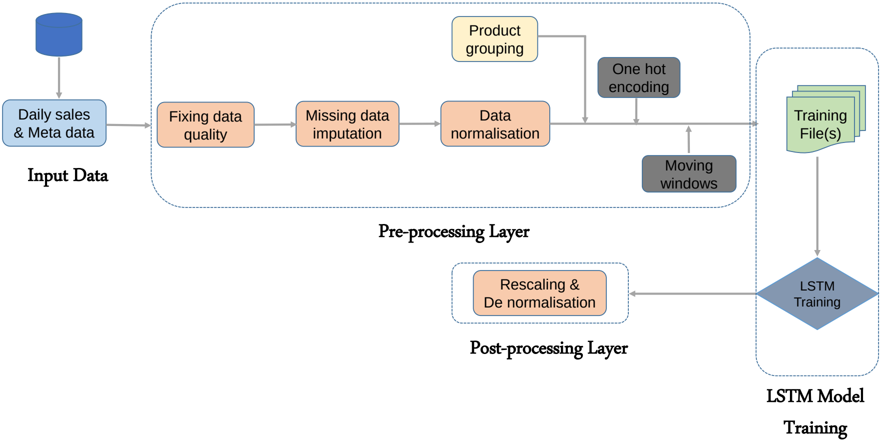


Figure: An overview of the proposed sales demand forecasting framework, which consists of a pre-processing, an LSTM training, and a post-processing part

Research Project 3

Research Question

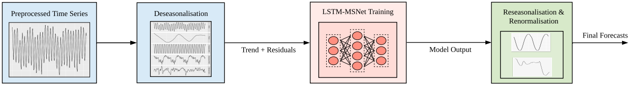
How various forms of multi-seasonal decomposition techniques can supplement the learning process of GFMs, when forecasting a group of time series with multiple seasonal cycles.

- Time series may exhibit complex behaviours.
  - Non-integer seasonality, Calendar effects, **Multiple seasonal patterns**.
- Time series with higher sampling rates (sub-hourly, hourly, daily) are becoming more common in many industries.
  - Utility demand industry (electricity and water usage).
  - Transportation, Tourist, and Health care industries.

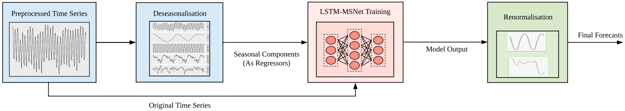
Research Project 3: Approach

- A decomposition based, GFM based prediction framework to forecast time series with multiple seasonal patterns.
- Using state-of-the-art multiseasonal decomposition techniques to supplement the RNN based GFM learning procedure.
  - Seasonal decomposition is advocated by many studies [Ben Taieb et al., 2011, Zhang and Qi, 2005].
  - Supplements the RNN's learning process.
  - Deseasonalized Approach: Seasonally adjusted time series
  - Seasonal Exogenous Approach: Seasonal components as external regressors.

Research Project 3: Overall Architecture



(a) The proposed DS training paradigm used to train the LSTM-MSNet



(b) The proposed SE training paradigm used to train the LSTM-MSNet

Figure: An overview of the proposed LSTM-MSNet training paradigms. In the DS approach, deseasonalised time series are used to train the LSTM-MSNet. Whereas in the SE approach, the seasonal values extracted from the deseasonalisation phase are employed as exogenous variables, along with the original time series to train the LSTM-MSNet.

Research Project 3: Major Findings

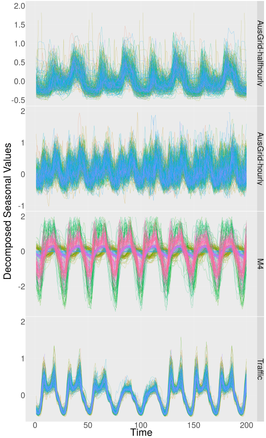


Figure: The seasonal components' distributions of the sum of multiple seasonalities extracted from the AusGrid-Energy (half hourly), AusGrid-Energy (hourly), M4 and the Traffic datasets, by applying the MSTL decomposition technique to the initial 200 data points of each time series.

Research Project 4: DeepPPMNet

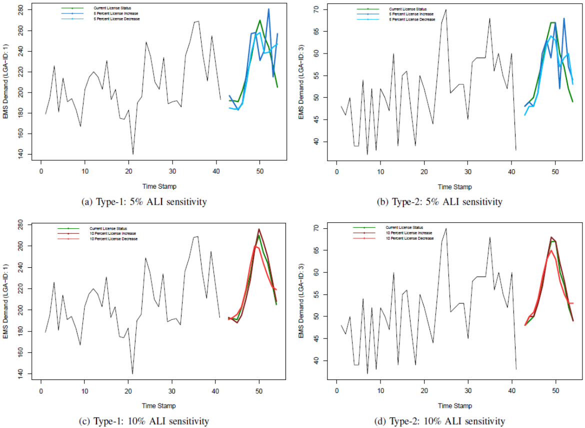
- GFM allows to train across all the available EMS demand time series to exploit the potential cross-series information available in multiple local governing areas (LGAs).

- Capable of exploring the causal relationships using the notion of Granger Causality, where the GFM enables to perform 'what-if' analyses.
  - Using potential external regressors to evaluate whether the base accuracies are improved.
  - Sensitivity analysis to assess their impact towards EMS.
  - GFMs make the 'what-if' analysis feasible even for relatively constant features.
  - Allows government decision makers to assess the factors that could drive the EMS demand.

Research Project 4: Evaluation

- The DeepPPMNet was evaluated using a realworld EMS demand dataset
  - National dataset of coded ambulance clinical records held by Turning Point, an Australian addiction research and education centre.
  - Related to alcohol overdose, suicide attempts, and other drug related harms.
  - 8 years worth of EMS related data for each LGA.
  - Our methods outperform state-of-the-art univariate time series models.
- A case study to investiage the use of DeepPPMNet
  - How the number of alcohol licenses issued (ALI) for a certain period of time can affect the alcohol related EMS demand patterns.

Research Project 4: DeepPPMNet Case Study



(a) Type-1: 5% ALI sensitivity

(b) Type-2: 5% ALI sensitivity

(c) Type-1: 10% ALI sensitivity

(d) Type-2: 10% ALI sensitivity

Figure: The application of 'what-if' scenario analysis, using the number of ALI (as a percentage of change) against the AO related EMS demand.

Research Project 5

Research Question

The effectiveness of using global models for causal Inference through Counterfactual Prediction

- Global model based Recurrent neural networks (RNN) to predict policy interventions' causal effects on an outcome over time through the counterfactual approach.
  - Traditional methods hold strong linearity and convexity assumptions in covariates, leading us to an non realistic fitting
  - Synthetic Control Method, Google Causal Impact
  - Limitations of equivalence assumption between the control and treated units distribution in the pre-treatment period.

Causal inference

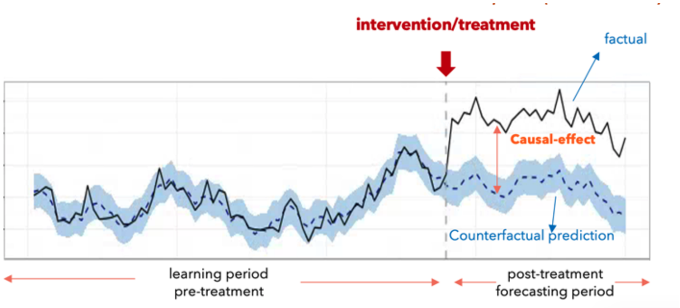


Figure: Generating counterfactual predictions for the units affected by intervention (treated units) based on a large-dimensional panel of observed time-series from a pool of untreated peers (control units)

Generating counterfactuals from DeepCPNet

- GFM is trained across all the panel of treated and control times series, before the intervention.
- Predicts the counterfactual trajectory for the treated unit simulating its behaviour in the absence of the intervention, after the intervention.
- Placebo tests to evaluate the counterfactual predictions
  - Null Effect of the treatment for the control units: performance of the model only over the forecasting of the control group.
  - The significance of the differences between the effects: the difference between the errors from treated and control units must be statistically significant (non-parametric paired Wilcoxon signed-rank)

Effect of COVID19 lockdown measures

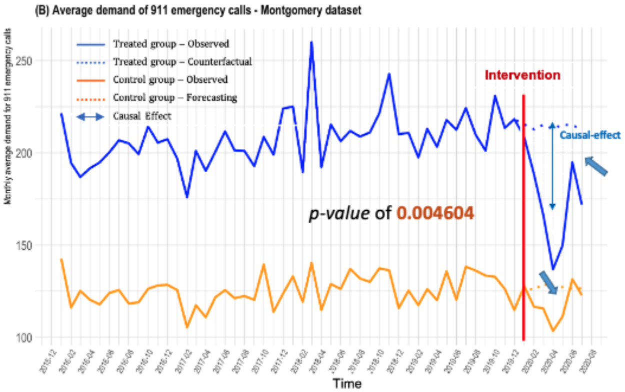


Figure: The causal effect of the COVID19 lockdown measures over the 911 emergency callouts

Research Project 6

Research Question

Can transfer learning approaches uplift the forecast accuracy of GFMs, in situations where the availability of time series is limited.

- RNN based GFMs are inherently data ravenous and require significant amount of training data.
  - Time series databases are often sparse, and may not hold adequate data.
  - The GFMs outshine univariate forecasting models, in situations where large quantities of related time series are available.

Research Project 6: Approach

- A host of transfer learning (TL) schemes to leverage the capabilities of GFMs to generate accurate forecast, in situations with less time series data
  - Adding residual connections to the RNN architecture.
  - Inspired by the ResNet architecture used for image classification tasks [He et al., 2016].
  - Allows to introduce substantially deeper GFM architectures.
- When TL methodology is constrained by the unavailability of a source dataset (D_s) to pre-train a model
  - Using a statistical generative model to artificially generate new copies of time series.
  - GRATIS package introduced by [Kang et al., 2019].
  - Employs a mixture autoregressive (MAR) models together with genetic algorithms to generate time series.

Research Contribution 6: Residual RNN Architecture

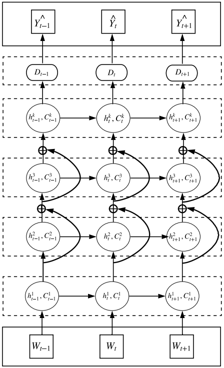


Figure: The unrolled representation of a residual recurrent network architecture with k number of stacking layers. Here, the residual connections are represented by curved arrows. Accordingly to [He et al., 2016], these residual connections allow the stacking layers to fit a residual mapping between W_t and $\hat{Y}_t$, while avoiding the network degradation with network depth increasing

Research Contribution 6: Overall Architecture

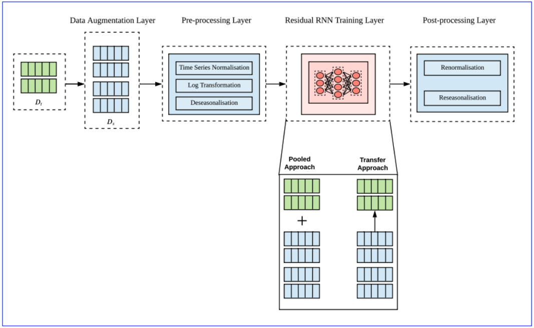
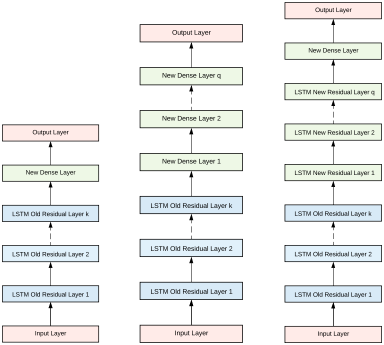


Figure: An overview of the proposed framework, which includes a data augmentation layer, pre-processing layer, residual RNN training layer, and a post-processing layer

Research Contribution 6: Transfer Learning Architectures



(a) TL-1 (b) TL-2 (c) TL-3

Figure: The layers used to build the base model using the D_s, are represented in blue colour, while the additional layers introduced, when building the target model using the D_t, are represented in green colour.

Outline

Current research in GFMs

- Global models for hierarchical time series forecasting
- Interpretable forecasts for global models
- Global models robust to concept drift.
- Global models for scenario-forecasting.

Thank you

Kasun Bandara

Kasun.Bandara@unimelb.edu.au

Slides available: `github.com/kasungayan/GoogleTalk`

References I

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., and Seaman, B. (2019).
Sales demand forecast in e-commerce using a long Short-Term memory neural network methodology.
In *Neural Information Processing*, pages 462–474. Springer International Publishing.

Barrow, D. K. and Crone, S. F. (2016).
A comparison of AdaBoost algorithms for time series forecast combination.
Int. J. Forecast., 32(4):1103–1119.

Ben Taieb, S., Bontempi, G., Atiya, A., and Sorjamaa, A. (2011).
A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition.

Bergmeir, C. (2020).
ACML 2020 tutorial: Forecasting for data scientists.
https://cbergmeir.com/talks/acml-tutorial/.
Accessed: 2021-8-29.

References II

Borovykh, A., Bohte, S., and Oosterlee, C. W. (2017).
Conditional time series forecasting with convolutional neural networks.

Crone, S. F. (2008).
NN5 competition.
http://www.neural-forecasting-competition.com/NN5/.
Accessed: 2017-8-18.

Crone, S. F., Hibon, M., and Nikolopoulos, K. (2011).
Advances in forecasting with neural networks? empirical evidence from the NN3 competition on time series prediction.
Int. J. Forecast., 27(3):635–660.

Fei, M. and Yeung, D. Y. (2015).
Temporal models for predicting student dropout in massive open online courses.
In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 256–263.

Godahewa, R., Bergmeir, C., Webb, G. I., and Montero-Manso, P. (2020).
A strong baseline for weekly time series forecasting.
ArXiv.

References III

He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021).
Recurrent neural networks for time series forecasting: Current status and future directions.
Int. J. Forecast., 37(1):388–427.

Hyndman, R. J. (2016).
Q&A time — rob J hyndman.
https://robjhyndman.com/hyndsight/qa-time/.
Accessed: 2017-9-5.

Hyndman, R. J. and Billah, B. (2003).
Unmasking the theta method.
Int. J. Forecast., 19(2):287–290.

References IV

Ilies, I., Jaeger, H., Kosuchinas, O., Rincon, M., and others (2007).
Stepping forward through echoes of the past: forecasting with echo state networks.
URL: http://www. neural-forecastingcompetition. com/downloads/methods/27-NN3 Herbert Jaeger report. pdf.

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., and Callot, L. (2020).
Criteria for classifying forecasting methods.
Int. J. Forecast., 36(1):167–177.

Kang, Y., Hyndman, R. J., and Li, F. (2019).
GRATIS: GeneRAting TIme series with diverse and controllable characteristics.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. (2019).
Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting.
In Wallach, H., Larochelle, H., and Beygelzimer, A., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

References V

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021).
Temporal fusion transformers for interpretable multi-horizon time series forecasting.
Int. J. Forecast.

Makridakis, S. and Hibon, M. (2000).
The M3-Competition: results, conclusions and implications.
Int. J. Forecast., 16(4):451–476.

Makridakis, S. and Spiliotis, E. (2021).
The M5 competition and the future of human expertise in forecasting.
Foresight: The International Journal of Applied Forecasting, (60):33–37.

Montero-Manso, P. and Hyndman, R. J. (2021).
Principles and algorithms for forecasting groups of time series: Locality and globality.
Int. J. Forecast.

References VI

Nelson, M., Hill, T., Remus, W., and O'Connor, M. (1999).
Time series forecasting using neural networks: should the data be deseasonalized first?
J. Forecast., 18(5):359–367.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2019).
N-BEATS: Neural basis expansion analysis for interpretable time series forecasting.

Rahman, M. M., Islam, M. M., Murase, K., and Yao, X. (2016).
Layered ensemble architecture for time series forecasting.
IEEE Trans Cybern, 46(1):270–283.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. (2018).
Deep state space models for time series forecasting.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

References VII

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020).
DeepAR: Probabilistic forecasting with autoregressive recurrent networks.
Int. J. Forecast., 36(3):1181–1191.

Sen, R., Yu, H.-F., and Dhillon, I. S. (2019).
Think globally, act locally: A deep neural network approach to High-Dimensional time series forecasting.
In Wallach, H., Larochelle, H., and Beygelzimer, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Smyl, S. (2016).
Forecasting short time series with LSTM neural networks.
https://gallery.cortanaintelligence.com/Tutorial/Forecasting-Short-Time-Series-with-LSTM-Neural-Networks-2.
Accessed: 2017-8-28.

Smyl, S. (2019).
A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting.
Int. J. Forecast.

References VIII

Suilin, A. (2018).
kaggle-web-traffic.
https://github.com/Arturus/kaggle-web-traffic.
Accessed: 2020-2-10.

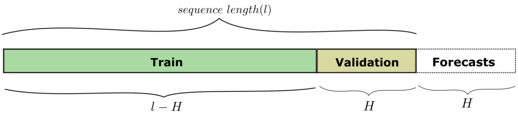
Yan, W. (2012).
Toward automatic time-series forecasting using neural networks.
IEEE Trans Neural Netw Learn Syst, 23(7):1028–1039.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998).
Forecasting with artificial neural networks:: The state of the art.
Int. J. Forecast., 14(1):35–62.

Zhang, G. P. and Qi, M. (2005).
Neural network forecasting for seasonal and trend time series.
Eur. J. Oper. Res., 160(2):501–514.

Zimmermann, H.-G., Tietz, C., and Grothmann, R. (2012).
Forecasting with recurrent neural networks: 12 tricks.
In *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pages 687–707. Springer, Berlin, Heidelberg.

Appendix



Figure