Estimators of Prediction Intervals For Statistical And Machine Learning Forecasts

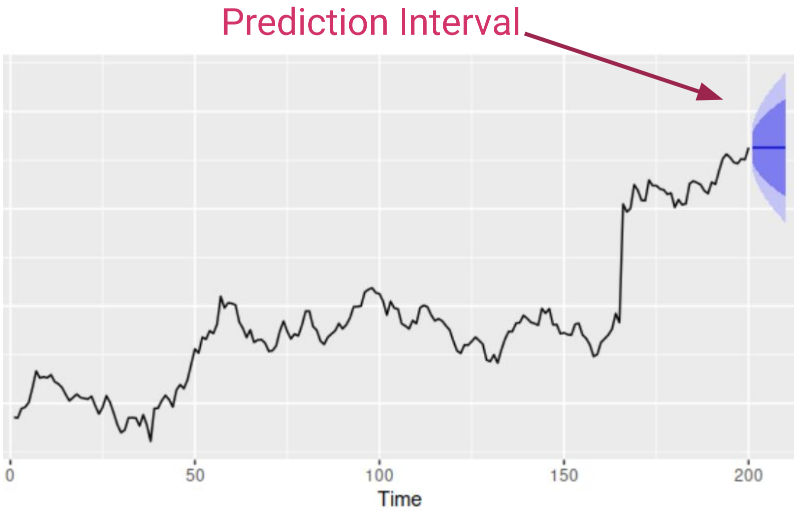
Filotas Theodosiou
MSc Data Science
University of Skovde
Supervisor: Nikolaos Kourentzes
nikolaos.kourentzes@his.se

1977
UNIVERSITY OF SKÖVDE

Agenda

- Introduction
  - Motivation
  - Aim
- Experimental Design
  - Models
  - Methods
  - Evaluation Metric
  - Data
- Results
- Conclusions

Uncertainty and Prediction Intervals



Measuring the Forecast Uncertainty as a result of multiple error sources:

- Confidence on Decisions
- Plan Different Strategies

Prediction intervals provide an upper and lower limit where the unknown future value is expected to lie in between.

Hyndman, R. J., Koehler, A. B., Ord, J. & Snyder, R. D. (2008), Forecasting with Exponential Smoothing: The State Space Approach, 1st edn, SpringerVerlag

Aim of this Work

Prediction Intervals are not as widely explored

A review of the existing methods on computing prediction intervals for statistical models

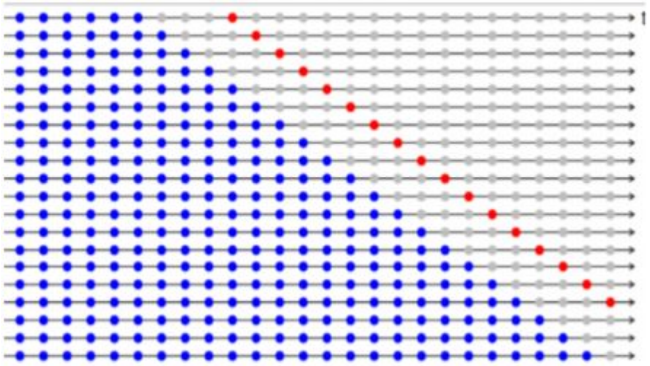
- Performance Comparison
- Advantages & Limitations

Evaluate if our understandings can be transferred to machine learning methods

Experimental Design

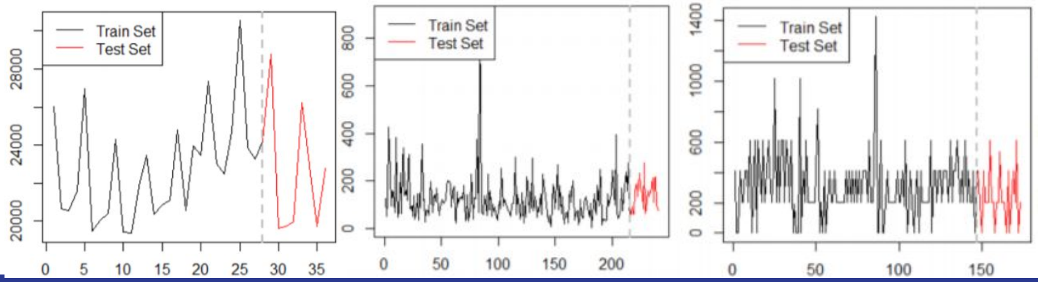
★ **Statistical Models :**
- Exponential Smoothing(ETS)
- AICc

★ **Evaluation Metric:** $(u-l)+\frac{2}{a}(l-x)*ID(x,l)+\frac{2}{a}(x-u)ID(u,x)$
- Interval Score
- Geometric Mean Relative Absolute Error $GMRAIS=\sqrt[N]{\prod^{N}\left|\frac{IS_A}{IS_{Bench}}\right|}$
- Rolling Origin Evaluation with Re-Estimation



Hyndman, R. J. & Athanasopoulos, G. (2018), Forecasting: principles and practice, 2nd edn, OTexts.

★ **Time Series Data:**
- 76 Monthly TS(240 observations)
- 88 Weekly (173 observations)
- 89 Quarterly TS(36 observations)

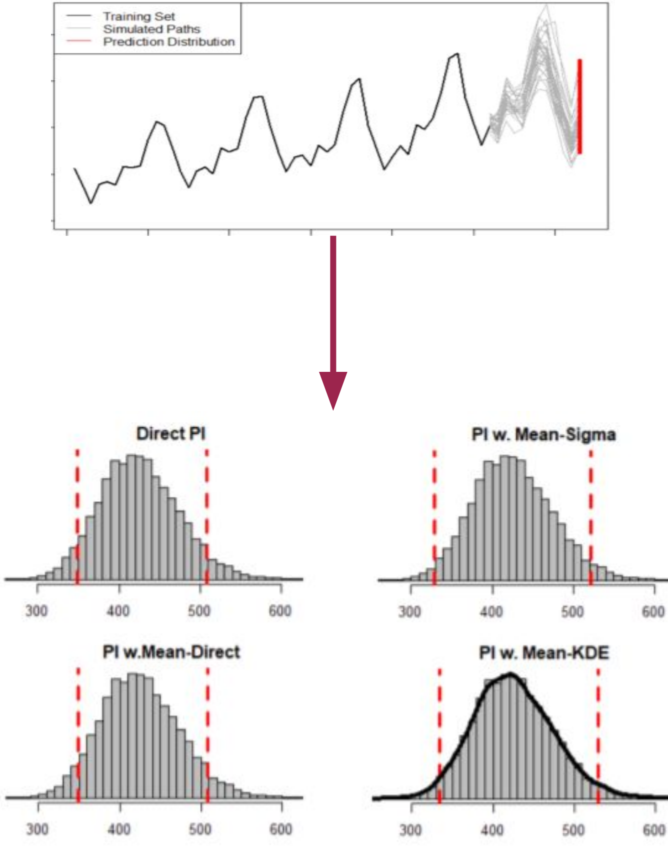


<u>Last Two Periods Are Kept as Test Set</u>

Families of Methods

1. Algebraic - Theoretical → $[\hat{y}_{t+h|t} - c\sigma(h), \hat{y}_{t+h|t} + c\sigma(h)]$
    1.1. Difficulties Estimating Conditional Variance
    1.2. Errors ~ IID_N(0,σ2)
2. Simulation - Based methods -> Et ~ IID_N
    2.1. Direct - Method
    2.2. Mean - Sigma → $PI(c)_2 = [\mu \pm \sigma c]$
    2.3. Mean - Direct → $PI(c)_3 = [\mu + l, \mu + u]$
    2.4. Mean - KDE
3. Bootstrap - Based method -> Et ~ IID
4. Empirical Methods -> Realistic Assumptions
    4.1. Direct Empirical
    4.2. KDE - Empirical

$$PI = [MeanForecast + LowInterval, MeanForecast + UpperInterval]$$



Machine Learning Set Up & Methods

Model Used : XGBoost.

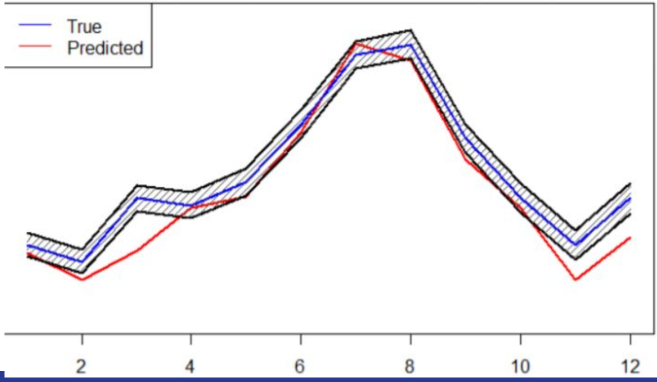
Automatic Fitting XGBoost Challenges:

1. Input Data: Sliding Window
2. Feature Selection : PACF
3. Trend & Seasonality : Stationary TS
4. Hyperparameters : Random Search & CV

	ETS	XGBboost				
	y	y	lag_1	lag_2	lag_3	lag_4
1	119	119	-	-	-	-
2	104	104	119	-	-	-
3	118	118	104	119	-	-
4	115	115	118	104	119	-
5	126	126	115	118	104	119
6	141	141	126	115	118	104

Only Empirical Methods Are Applicable

- No theoretical formulas
- Poor Simulation - based Performance



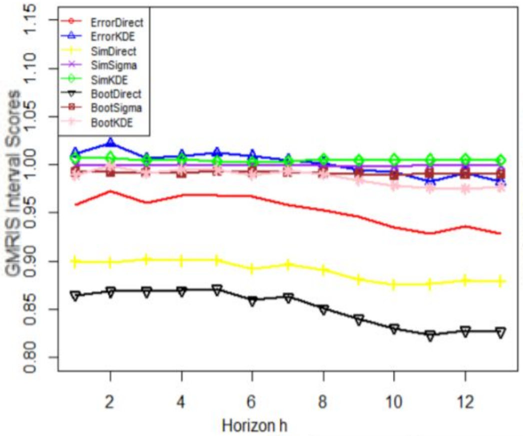
Results on ETS

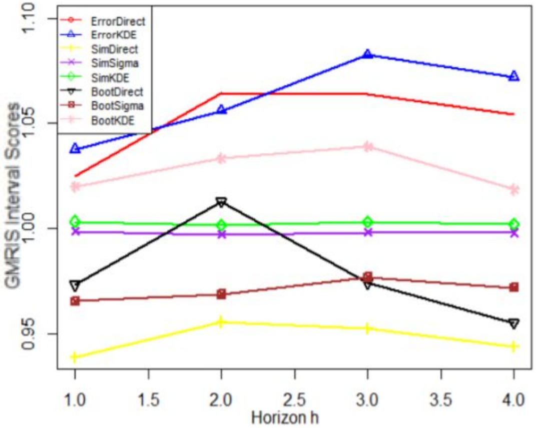
Monthly & Weekly Series

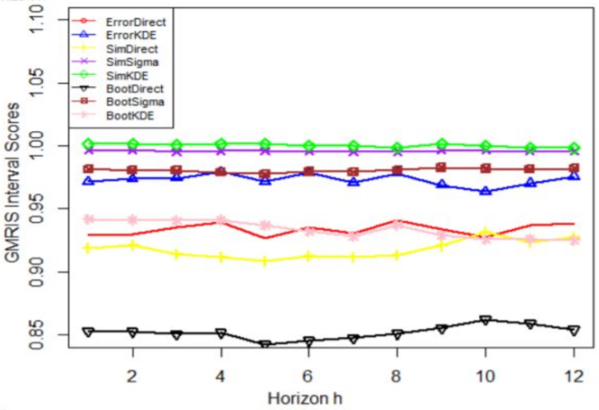
- Bootstrap - Direct gave the best results on Monthly and Weekly series
- Empirical Performed Better than algebraic and equally well with simulation - based
- KDE - Approaches outperformed\

Quarterly Series with fewer observations:

- Simulation - based gave the best results
- Algebraic outperforming empirical





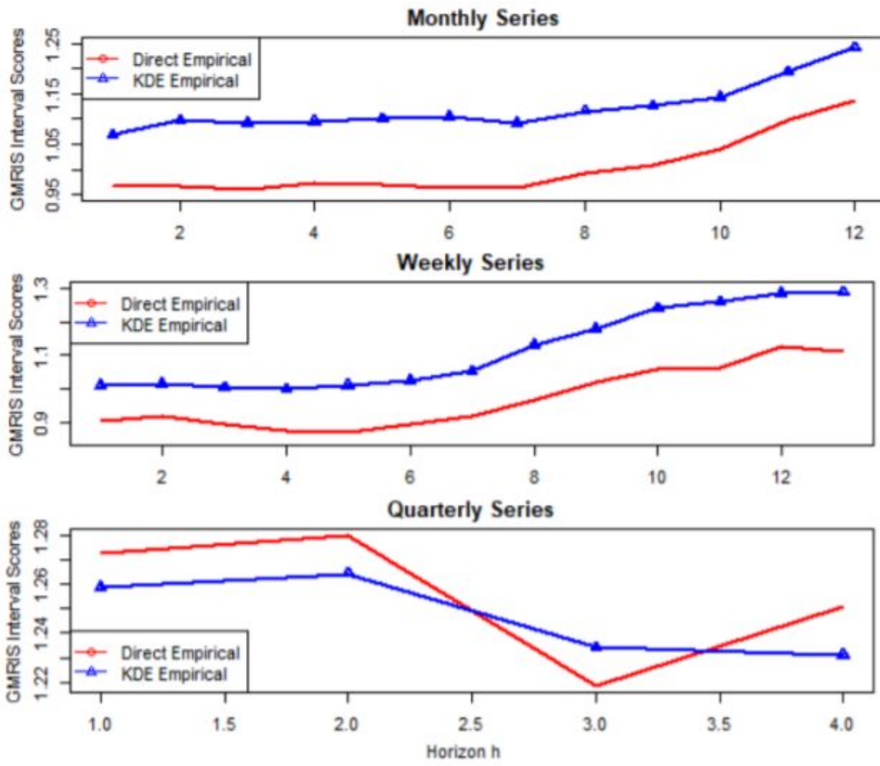


Results on XGboost

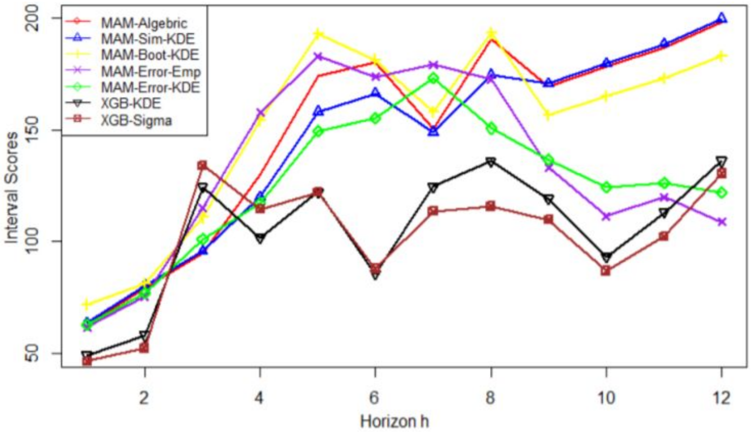
For monthly and weekly series:

- Direct Empirical gives promising results on earlier horizons
- Performance get worse for later ones

On quarterly series, empirical methods performed poorly.



Why XGBoost was outperformed?



- Automatically fitting XGBoost might not have worked for some models
- Manually fitting XGBoost should be a priority

Empirical methods applied on a manually fitted XGBoost on a single time series outperforms every other method

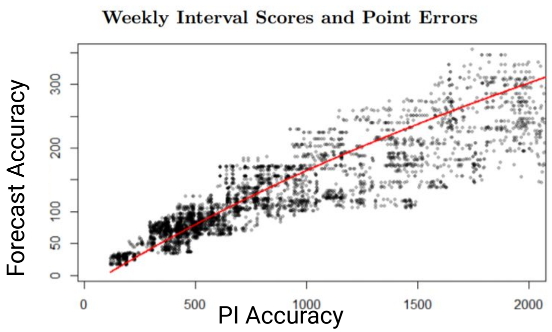
Mean Absolute Scaled Point Forecast Error was estimated to understand the performance of XGBoost

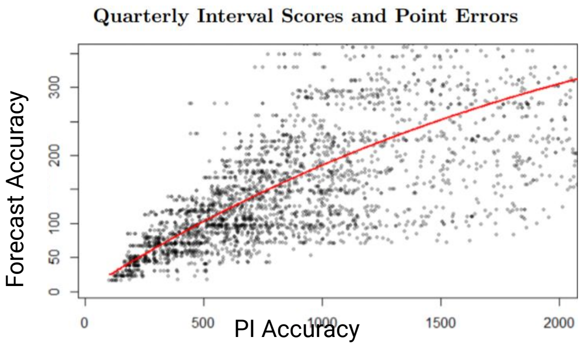
	Quarterly		Monthly		Weekly	
Method	ETS	XGBoost	ETS	XGBoost	ETS	XGBoost
EmpDirect	1.05	1.25	0.933	1	0.952	0.97
EmpMeanKDE	1.06	1.24	0.972	1.122	1.001	1.15
MASE	**0.908**	**0.83**	**0.614**	**0.695**	**0.995**	**1.013**

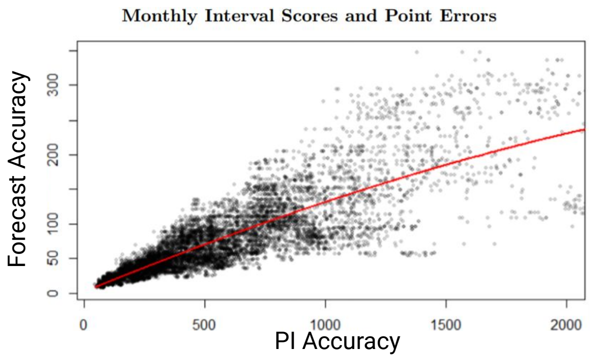
★ Quarterly -> Smaller MASE. Empirical methods don't work well on relative small training sample

★ Monthly & Weekly -> Good sample size for empirical methods. High MASE(bad point forecast) might be the reason for poorer performance

Correlation of Point Forecast and PI Estimation







Direct correlation between Absolute Error and Interval Score

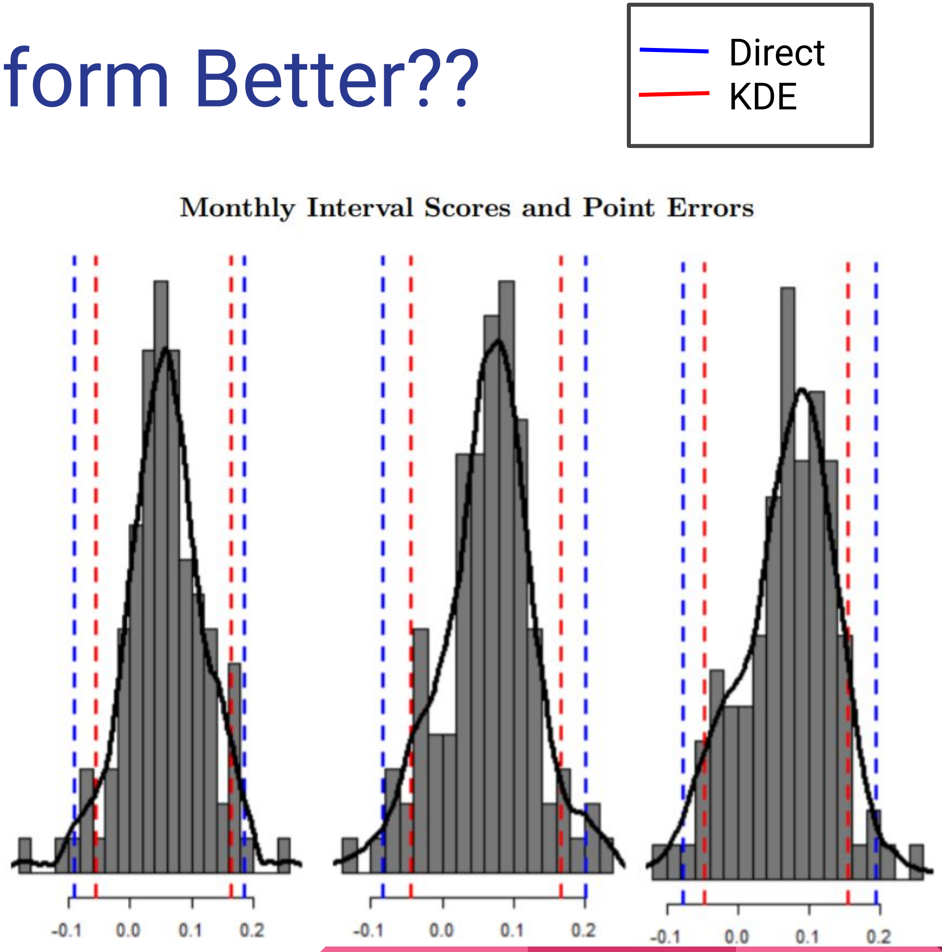
- Best Intervals have small absolute point error
- A bigger point error results in poorer Intervals as it more challenging to include the true value of the series.

Model selection, in terms of point-forecast performance, is critical for Prediction Interval estimations, regardless of the used method

Method	ETS(M,A,A)	ETS(A,A,A)	ETS(M,A,M)
Algebric	281.54	**357.23**	149.7
SimDirect	246.81	445.01	153.85
SimMeanKDE	280.87	359.66	146.86
BootDirect	246.73	492.46	168.1
BootMeanKDE	267.93	399.09	153.13
EmpDirect	201.83	494.38	132.67
EmpMeanKDE	**200.75**	452.83	**124.63**
Mean	**246.63**	**428.66**	**146.99**

Table 3, Mean Interval Scores Per Horizon, of the Two Non-Optimal Models and Optimal ETS(M,A,M)

Why Direct - Methods Perform Better??



Direct Extraction Outperformed KDE

- Direct methods get rid of the extreme observations on tails
- KDE tries to smoothly include all values on the distribution

A bigger error sample would give no gaps and a better fitted KDE

KDE would then perform much better

Conclusions on the Methods

- ★ Despite the wide usage of theoretical and simulation-based methods:
  - Heavy assumptions
  - No Better Performance
  - Risk of a Stock-Out
- ★ Bootstrap methods are not necessarily better than empiricals
  - Slightly Higher Interval Score ⇔ Slightly Wider Intervals
  - Bootstrap methods require the i.i.d assumption
  - Tighter Intervals might be unrealistic ⇔ Over/Under-Stocking
- ★ Empirical Methods perform poorly on smaller data samples
  - Consider Bootstrap Methods
- ★ No standardized method for extracting a PI should be taken
  - Direct Methods perform better on smaller Sample Size
  - KDE works well on bigger error sample sizes

Conclusions on XGBoost

- ★ Empirical methods are applicable
  - Small Forecast Variance
  - No analytical expressions
- ★ Promising Results but:
  - Outperformed on smaller samples
  - Careful fitting and hyperparameters optimization

The model with the best point-forecast performance should be picked

Future work: The applicability of empirical methods on Deep Learning Models

References

Armstrong, J. (2017), ‘Demand forecasting ii: Evidence-based methods and checklists’, p. 36.

Arranz, M. (2005), ‘Tol-project portmanteau test statistics in time series’.

Barrow, G. & Kourentzes, N. (2016), ‘Distributions of forecasting errors of forecast combinations: Implications for inventory management’, International Journal of Production Economics 177, 24–33.

Chatfield, C. (1996a), The Analysis of Time Series, 5th edn, Chapman and Hall/CRC.

Chatfield, C. (1998), ‘Prediction intervals, department of mathematical sciences’.

Chatfield, C. (2000), Time-Series Forecasting, 1st edn, Chapman and Hall/CRC.

Hall/CRC. Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, ACM, pp. 785,794.

Davydenko, A. & Fildes, R. (2013), ‘Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts’, International Journal of Forecasting 3, 510–522.

Gneiting, T. & Raftery, A. E. (2007), ‘Strictly proper scoring rules, prediction, and estimation’, Journal of the American Statistical Association 102, 359– 378.

Hyndman, R. J. (n.d.), ‘Prediction intervals too narrow’. URL: https://robjhyndman.com/hyndsight/narrow-pi/

References

Hyndman, R. J. & Athanasopoulos, G. (2018), Forecasting: principles and practice, 2nd edn, OTexts.

Hyndman, R. J. & Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', International Journal of Forecasting 22, 679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. & Snyder, R. D. (2008), Forecasting with Exponential Smoothing: The State Space Approach, 1st edn, SpringerVerlag

Isengildina, O., Irwin, S. H. & Good, D. L. (2006), Empirical confidence intervals for wasde forecasts of corn, soybean and wheat prices, 2006 Conference, April 17-18, 2006, St. Louis, Missouri 18995, NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management.

Kourentzes, N. & Athanasopoulos, G. (2020a), 'Elucidate structure in intermittent demand series', 288, 141–152.

Kourentzes, N. & Athanasopoulos, G. (2020b), 'Elucidate structure in intermittent demand series', 288, 141–152.

Mishina, Y., Tsuchiya, M. & Fujiyoshi, H. (2014), Boosted random forest, in '2014 International Conference on Computer Vision Theory and Applications (VISAPP)', Vol. 2, pp. 594–598.

Morde, V. (2019), 'Xgboost algorithm: Long may she reign!'. URL: https://towardsdatascience.com/https-medium-com-vishalmordexgboost-algorithm-long-she-may-rein-edd9f99be63d

Mushtaq, R. (2011), 'Augmented dickey fuller test', SSRN Electronic Journal . Ord, K., Fildes, R. & Kourentzes, N. (2017), Principles of Business Forecasting– 2nd ed, wessex, inc.

References

Ord, K., Koehler, A. & Snyder, R. (1995), 'Estimation and prediction for a class of dynamic nonlinear statistical models', Journal of the American Statistical Association 92.

Probst, P., Bischl, B. & Boulesteix, A.-L. (2018), 'Tunability: Importance of hyperparameters of machine learning algorithms'.

Silverman, B. W. (1986), CRC Press. Smith, S. & Sincich, T. (1988), 'Stability over time in the distribution of population forecast errors', Demography 25, 461–74.

Sven, F. & Kourentzes, N. (2007), Input variable selection for time series prediction with neural networks- an evaluation of visual, autocorrelation and spectral analysis for varying seasonality.

Trapero, J., Card´os, M. & Kourentzes, N. (2018a), 'Empirical safety stock estimation based on kernel and garch models', Omega .

Trapero, J., Card´os, M. & Kourentzes, N. (2018b), 'Quantile forecast optimal combination to enhance safety stock estimation', International Journal of Forecasting 35, 239–250.

Willemain, T., Smart, C. & Schwarz, H. (2004), 'A new approach to forecasting intermittent demand for service parts inventories', International Journal of Forecasting 20, 375–387.

Williams, W. & Goodman, M. (1971), 'A simple method for the construction of empirical confidence limits for economic forecasts', Journal of the American Statistical Association 66, 752–754

Thank you for your Attention