https://dx.doi.org/10.4314/dujopas.v11i3a.1
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 11 No. 3a September 2025*

A New Extension of Random Forest Technique for Vector Autoregressive Model

Nura Isah[1], Sani Ibrahim Doguwa[2], Hussaini Garba Dikko[3], Bukar Baba Alhaji[4]

[1]Department of Statistics,
Collage of Science &Technology,
Jigawa State Polytechnic Dutse,

[2&3]Department of Statistics,
Faculty of Physical Sciences,
Ahmadu Bello University,
Zaria,

[4]Department of Mathematical Sciences,
Faculty of Sciences,
Nigerian Defense Academy,
Kaduna.

Email: nuraisagm@gmail.com

Abstract

A vector autoregressive (VAR) model is a widely used parametric method that captures linear interdependencies between multivariate time series data. However, conventional VAR models are frequently constrained by some assumptions of linearity and stationarity, limiting their ability to capture complex and nonlinearity dynamic observed in real life data. This research aims to develop Multivariate Random Forest Technique for VAR model. Augmented Dickey Fuller (ADF) was applied to test the stationarity of the time series data. Simulated stationary time series data and real time series dataset were applied for modeling VAR model in order to compare their forecast performance. The forecast performance for the proposed technique and existing technique for VAR models were based on Mean Absolute Error and Root Mean Square Error. The proposed multivariate Random Forest technique for VAR model has been developed. The results for unit root test for real life data indicate that all the variables are stationary at level. The result for simulated data in term of forecast performance indicated that the proposed multivariate Random Forest techniques for VAR model outperform existing technique (conventional VAR model) in long term forecasting, while the conventional VAR model outperform proposed multivariate Random Forest technique in short term forecast based on MAE and RMSE. The results for real dataset indicated that the proposed multivariate Random Forest technique for VAR model is superior in terms of forecast performance than that of conventional VAR model in short-term and long-term forecast based on MAE and RMS since the proposed technique has the least value of MAE and RMSE in short and long-term forecasting than the conventional VAR model.

Keywords: Machine Learning, Random Forestt Technique, Ensemble Techniques, Overfitting, Stationarity.

INTRODUCTION

Forecasting multivariate time series is acritical task in field such as economics, finance, environmental science and applied sciences. A vector autoregressive (VAR) model is a widely

**Author for Correspondence*

used parametric method that captures linear interdependencies between multivariate time series data. However, conventional VAR models are often by assumptions of linearity and stationarity, and have some limitation on the ability to performed in complex and nonlinear real-life data (Lutkepohl, 2005).

Recent advance in machine learning particularly ensemble methods like Random forest (RF) offer alternative method for modelling nonlinear relationships without strict parametric assumptions. Despite their success in univariate time series forecasting and regression task, the application of Random Forest techniques to multivariate time series, particularly in the context of VAR models remains relatively under explored. This research aim to extend the Random Forest framework to accommodate VAR model, enabling more flexible and robust in multivariate forecasting. (Rahman and Davis, 2013).

The integration of machine learning techniques with traditional econometric model has gained traction in recent years. Random Forest technique introduced by Breiman (2001) is an examples of ensemble methods that involved construction of multiple decision trees and aggregate their results to improve prediction accuracy and reduce overfitting of the model. Several studies have explored Random Forest in the context of time series. Biau (2012) and Genuer *et al.* (2010) analyses theoretical properties and variable importance measures of Random Forest. However, few scholars have extended Random Forest technique to multivariate framework in order to compare the performance of the proposed technique with conventional VAR model.

Some efforts have been made toward this direction. Grover *et a.l* (2018) proposed Deep AR and its variants for probabilistic multivariate forecasting, Konig and Neumann (2019) applied tree-based ensembles to multivariate economic time series, suggesting improved accuracy over conventional VAR models in the presence of nonlinearities. Medeiros *et a.l* (2021) proposed hybrid models by applying Random Forest with penalized VAR model, which demonstrate improvement in forecast performance.

Davis *et al.* (2016) proposed sparse VAR estimation technique by applying partial spectral coherence to VAR model, for model specification. Carriero *et al*. (2019), Koop *et al*. (2019), and Billio *et al*. (2019) proposed different sparse VAR estimation technique by applying Bayesian approaches to VAR models in order to shrink some parameters towards a parsimonious specification. Pavlyuk (2019) proposed application of a random forest as a feature selection tool for controlling the sparsity of VAR models.

In term of extension of random forest approaches, Linusson (2013) proposed an extension of the Random Forest feature selection that allows Random Forest to be constructed for multi-output decision problems with arbitrary combinations of classification and regression responses, with the aim to increase prediction performance for multi-output problems. The result indicates that the proposed technique reduces prediction error when compared with single-output decision trees based on the same node. For empirical study, Gonzalez *et al*. (2015) applied Multi-output Random Forest to forecast an electricity price for Spain and compare the forecast performance of the propose algorithm and Random Forest technique. The result indicates that proposed Multi-output Random Forest algorithm out-performed Random Forest technique in forecasting the electricity price for Spain.

Despite these developments, a formal and systematic adaptation of VAR model where lagged values of multiple series are jointly modeled remains underdeveloped. For instance, Pavlyuk

(2019) developed a new technique for estimating VAR/ sparse VAR model in which an equation wise strategy was adopted using Random Forest technique to estimate and select features based on features importance to each equation of the VAR model separately and the result was obtained for final specification. Pavlyuk (2020) introduced a system wise strategy in random forest feature selection technique, where all VAR regression equations were solved simultaneously and feature selection was obtained based on feature importance to each equation of the VAR separately.

In this paper, Multivariate Random Forest techniques would be proposed and applied in multivariate case to estimate the VAR model.

METHODOLOGY

Vector Autoregressive Model (VAR)

The VAR model is one of the most successful, flexible, and easy models to use for analysis of multivariate time series. It is a natural extension of univariate autoregressive model to dynamic multivariate time series. VAR model is useful for describing the dynamic behavior of economic and financial time series and for forecasting. Consider a column vector for *k* different time series variables:

$$\mathbf{Y}_t = (\mathbf{y}_{1t}, \mathbf{y}_{2t}, \mathbf{y}_{3t}, ..., \mathbf{y}_{kt})^1 \quad (1)$$

and the model is in terms of past values of the vector. The result is a vector autoregressive or VAR. The VAR(*p*) process is of the form:

$$\mathbf{Y}_t = \mathbf{M} + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{A}_2\mathbf{Y}_{t-2} + ... + \mathbf{A}_p\mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (2)$$

where

$$\mathbf{A}_i = \begin{pmatrix} a_{11}^{(r)} & a_{12}^{(r)}... & a_{1k}^{(r)} \\ a_{21}^{(r)} & a_{22}^{(r)}.... & a_{2k}^{(r)} \\ a_{31}^{(r)} & a_{32}^{(r)}... & a_{3k}^{(r)} \end{pmatrix} \qquad ij = 1,...,k; r = 1,..., p \quad (3)$$

Is a *k* by *k* square matrix of coefficients; ***M*** is a *k* by *1* column vectors, and $\boldsymbol{\varepsilon}_t$ is a *k* by *1* column vector of white noise process, with the properties that:

$E(\boldsymbol{\varepsilon}_t) = 0$ for all *t*

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s) = \begin{cases} \mathbf{v}, & if \quad s = t \\ \mathbf{0}, & if \quad s \neq t \end{cases} \quad (4)$$

where $\mathbf{v}$ the covariance matrix is assumed to be positive definite. Thus the $\boldsymbol{\varepsilon}_t$ are serially uncorrelated but may be contemporaneously correlated. .

For *k* =3 and *p* = *2* Equation 2 becomes:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} + \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \\ y_{3t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \quad (5)$$

Multivariate Random Forest Technique

Multivariate Regression Tree (MRT) is the method that can be used to accommodate multivariate response variables. To increase the forecast performance of MRT, Segal & Xiao (2011) proposed multivariate random forests by combining MRT with Random Forest technique. MRF is able to accommodate multivariate response variables by combining the MRT method with bootstrap resampling and predictor subsampling using traditional random

forest. In addition, the MRF method can be used to determine important factors that influence the multivariate response variables.

MRF technique involved recursive binary splitting for partitioning the training dataset of explanatory variables by minimizing a covariance weighted loss function as defined by

$$l(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{v}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \quad (6)$$

where $\mathbf{v} = \text{cov}(\mathbf{y}_i)$ represents the covariance matrix of the multivariate response variables. The dependence structure is characterized by this covariance that can be described to capture different patterns. Similar to univariate random forests, the ensemble model would be obtained using bootstrap resampling. Using (6) a multi response split function is created. The prediction for each leaf of a multi response regression tree is a vector response means for values reaching that final leaf.

The Proposed Multivariate RF Technique for VAR Model

This propose technique would be used to extend random forest technique proposed by Pavlyuk (2019) to estimate the VAR model. This would be achieved by combining multivariate random forest proposed by Segal and Xiao (2011) and VAR models. More importantly, when two or more regression equations are estimated simultaneously the model estimation and forecasting can be more efficient compared to random forest approach where each regression will be estimated and predicted independently as observed by Evgeniou and Pontil (2004).

MRF technique involved recursive binary splitting for partitioning the training dataset by minimizing a covariance weighted loss function as defined:

$$l(\mathbf{y}_{it}, \hat{\mathbf{y}}_{it}) = \sum_{i=1}^{k} (\mathbf{y}_{it} - \hat{\mathbf{y}}_{it})^T \mathbf{v}^{-1} (\mathbf{y}_{it} - \hat{\mathbf{y}}_{it}) \quad (7)$$

where $\mathbf{v} = \text{cov}(\mathbf{y}_{it})$ represents the covariance matrix of the multivariate response variables. The dependence structure is characterized by this covariance that can be described to capture different patterns. Similar to univariate random forests, the ensemble model would be obtained using bootstrap resampling. Using (7) a multi response split function is created. The prediction for each leaf of a multi response regression tree is a vector response means for values reaching that final leaf.

This new technique would be adopted to determine the most important lags of the variables that influence the multivariate response variables in each regression equations in the VAR model.

$$\mathbf{y}_{it} = \mathbf{M}_i + \sum_{r=1}^{p} \sum_{j=1}^{k} \mathbf{s}_{ij}^{r} \mathbf{a}_{ij}^{(r)} \mathbf{y}_{jt-r} + \boldsymbol{\varepsilon}_t \quad (8)$$

where $\mathbf{y}_{it}$ is the i^{th} variable of the output, $\mathbf{s}_{ij}^{r}$ represents the relationship between variable i and j,

that is $\mathbf{y}_{it}$, $\mathbf{y}_{jt-r}$ and $\mathbf{a}_{ij}^{r}$ are the elements of the $k \times k$ coefficient matrices, $\mathbf{A}_r (r = 1, 2, ,, ,, p; i, j = 1, 2, ..., k), \mathbf{M}_i$ is constant terms and ε_t is a zero mean white noise process.

This developed technique would be used to obtained best model based on its forecast performance using Multivariate Random Forest method, the difference between this developed model and existing one is that, the measure of variable performance would be simultaneously calculated and the nature of the variables must be in terms of VAR model

structure, while for the other existing model the structure for the variable does not necessarily be in term of the VAR model structure

RESULTS

Simulated data

The simulated time series data contained 12 features each with 100 samples data points were obtained using Phyton package, while the conventional VAR Model were estimated using information-criteria based methods such as Akaike Information Criteria AIC and Bayesian Information Criteria BIC and the proposed VAR model (using multivariate Random Forest techniques). The forecast performance for proposed and existing techniques for estimated VAR models using MAE and RMSE would be computed and compared based on one-ahead (short term) since it is the baseline for time series model evaluation (Box *et al.* 2015) and 10 ahead forecast as long-term forecast as stated by Hyndman *et al* (2021).

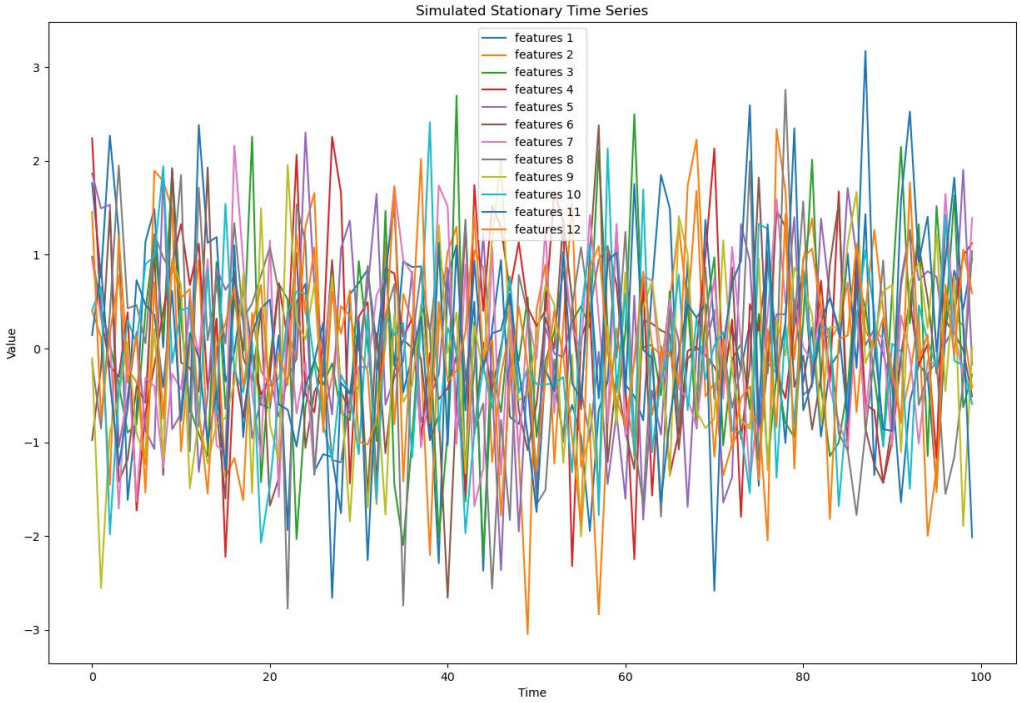


Figure 1: Time series plot for 12 monthly simulated time series data

The nature of the simulated time series graph in Figure 1 indicates the existence of stationarity in the time series data at level.

First, we undertake Lag Order selection process before estimating VAR model, the result for lag order selection shows that based on AIC, and HQIC selected lag 5 was selected as optimal lag for the estimating the VAR Models.

Forecast Performance for conventional VAR Models and Proposed Multivariate Random Forest technique for VAR model

After estimating the conventional and proposed technique for VAR models, the forecast performance for the estimated VAR models would be estimated using (1-ahead) short-term and (10- ahead) long-term forecast horizons based on MAE & RMSE. The summary results are shown in Table 1 & 2 simulated data.

Summary of Result for simulated Data

Table 1 contained summary results for the forecast performances of existing and proposed technique for VAR Model on short-term forecast based on MSE and RMSE.

From Table 1, the value of MAE and RMSE for conventional VAR model is lesser value than that of proposed technique, hence conventional VAR model outperform the proposed Multivariate Random Forest technique for VAR Model in short-term forecast, based on MAE and RMSE.

Table 1: Forecast performance for existing and proposed technique based on short-term forecast

Models	MAE	RMSE
Conventional VAR Model	1.642	2.016
VAR (MRF) Model	2.014	2.024

Table 2 contained summary results for the forecast performances of existing and proposed technique for VAR Model on long-term forecast based on MSE and RMSE.

Table 2 shows that the value MAE and RMSE for proposed Multivariate Random Forest technique for VAR model is less that of conventional VAR model, hence the proposed technique outperform the conventional VAR model in long-term forecast for the simulated data based on MAE and RMSE.

Table 2: Comparative Analysis on the forecast performance for existed and proposed model based on long term forecast horizon

Models	MAE	RMSE
Conventional VAR Model	1.316	1.669
VAR (MRF) Model	0.947	1.197

Analysis for Real Dataset

Real dataset consists of monthly Nigerian Market capitalization equity and other finance time series data spanning the period January 2008 to December 2024, a total of 204 data points. The data for all the features were obtained from the Central Bank of Nigeria web site and consider in fitting the estimated models. In this research, the dataset consist of fourteen features, the features include the following: *Difference of Nigeria Market Capitalization Equity* (DNCME), *Log of Value Of Transaction* (LVOT), *Difference of All Shares Index* (DASI), *Difference of Serving Interest Rate* (DSIR), *Difference of Maximum Lending Rate* (DMLR), *Loan Deposit Ratio* (LDR), *Difference of Currency In Circulation* (DCIC), *Difference of Monitory Aggregate* (DM_1), *Difference of Monitory Policy Interest Rate* (DMPIR), *Value of Transaction* (VT), *Crude Oil Price* (COP), *Liquidity Ratio* (LR), *Number Of Deals* (NOD) and *Difference of Interest Rate* (DIR).

Unit Root Test

The real time series dataset was used to estimate conventional VAR Model and the proposed multivariate Random Forest technique for VAR model. The forecast performance for the existing and proposed technique for VAR models would be compared for short- term and long-term based on MAE and RMSE.

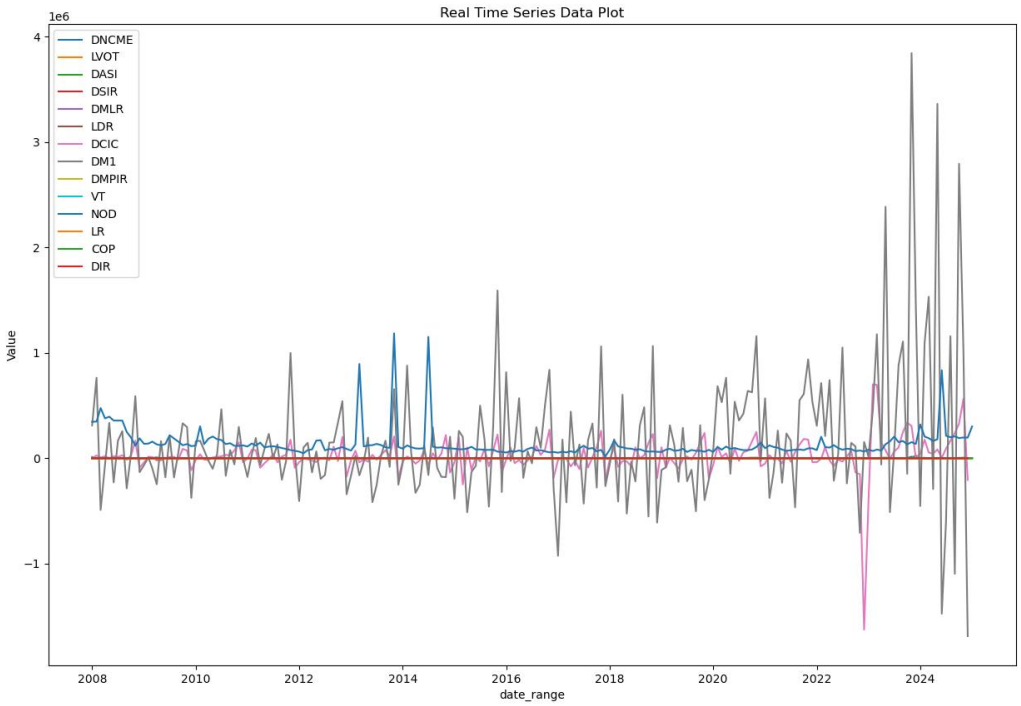


Figure 2: Time series plot for Real dataset

The nature of the time series graph in Figure 2 shows that the time series are stationary.

Table 3: Augmented Dickey Fuller Unit root test real dataset

Variables	ADF Statistic	p-value	Order of integration
DNCME	-7.901	4.18×10^{-12}	I(0)
LVOT	-3.40	0.01	I (0)
DASI	-7.96	2.92×10^{-12}	I (0)
DSIR	-14.07	2.93×10^{-26}	I (0)
DMLR	-12.52	2.61×10^{-23}	I (0)
LDR	-3.28	0.02	I (0)
DCIC	-3.99	0.00	I (0)
DM_1	-3.16	0.02	1(0)
DMPIR	-5.99	1.76×10^{-7}	1(0)
VT	-3.99	0.00	1(0)
COP	-2.89	0.04	1(0)
LR	-4.37	0.00	1(0)
	-3.00	0.03	1(0)
NOD	-3.72	0.00	1(0)
DIR			

Result for unit root test presented in Table 3 indicate that eight features are stationary at level the first differencing for the other features are also stationary at level, since p-value of all features is less than significant level of 0.05, as Greene (2012) stated "by differencing a non-stationary time series once to became stationary, could be used in estimating VAR model". The other six features are stationary at level.

Lag Order selection process would be obtained before estimating VAR model, the result for lag order selection shows that BIC and HQIC, while AIC select lag 9, hence lag 1 is selected as the optimal lag for estimating the VAR Models as recommended by Killian and Lutkepohl (2017) that select based on BIC or HQIC over AIC for small/ moderate samples due to penalties for overfitting.

Forecast Performance for conventional VAR Models and Proposed Multivariate Random Forest technique for VAR model

After estimating the conventional and proposed technique for VAR models, the forecast performance for the estimated VAR models would be estimated using (1-ahead) short-term and (10- ahead) long-term forecast horizons based on MAE & RMSE. The summary results are shown in Table 4 & 5.

Summary of Result for simulated Data

Table 4 contained summary results for the forecast performances of existing and proposed technique for VAR Model on short-term forecast based on MSE and RMSE.

From Table 4, the proposed Multivariate Random Forest technique for VAR Model demonstrate a high forecast accuracy than the conventional VAR model in short-term forecast since the value of MAE and RMSE is less than that of conventional VAR model.

Table 4: Forecast performance for existing and proposed technique based on short-term forecast

Models	MAE	RMSE
Conventional VAR Model	37,864.08	122,841
VAR (MRF) Model	0.666	0.824

Table 5 contained summary results for the forecast performances of existing and proposed technique for VAR Model on long-term forecast based on MSE and RMSE

Table 5 shows that the proposed Multivariate Random Forest technique for VAR model superior in term of forecast performance in long term forecasting, since the conventional VAR model has high value of MAE and RMSE.

Table 5: Forecast performance for existed and proposed model based on long term forecast horizon

Models	MAE	RMSE
Conventional VAR Model	42,894.40	141,319
VAR (MRF) Model	29.50	48, 443

CONCLUSION

The proposed multivariate Random Forest technique for VAR model has been developed by applying multivariate Random Forest technique to VAR model. The results for simulated data indicated that the conventional VAR model outperform the proposed technique in short-term forecasting, while the proposed multivariate Random Forest technique for VAR model demonstrate a high forecast accuracy than the conventional VAR model in long term forecasting based on MAE and RMSE. The results for real dataset indicated that the proposed multivariate Random Forest technique for VAR model is superior in term of forecast performance than conventional VAR model in short-term and long-term forecasting based on MAE and RMSE since the proposed technique has the least value of MAE and RMSE.

REFERENCES

Billio, M., Casarin, R., and Rossini, L. (2019). Bayesian Nonparametric Sparse VAR Models. *Journal of. Econometrics*. 212, 97 – 115.

Biau G. (2012). Analysis of Random Forest Model. *Journal of Machine Learning Research.13*, 1063 – 1095.

Box, G. E.P., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control.

Breiman, L. (2001). Random Forests. *Machine Learning*. 45, 5–32.

Carriero,A., Clark,T.E., and Marcellino,M. (2019). Large Bayesian Vector Autoregressions with Stochastic Volatility and Non-conjugate Priors. *Journal of Econometric*. 212, 137 – 154.

Davis, R.A., Zang, P. and Zheng, T. (2016). High Dimensional Forecasting via Interpretable Vector Autoregression. *Journal of Machine Learning Research*. 21(77), 1–38.

Evgeniou, T. and Pontil, M. (2004). Regularized Multitask Learning. *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. 109-117.

Genuer, R., Poggi, J. M. and Tuleau-Malot, C. (2010). Random Forest: Some Methodological insight, preprint arxiv,3619

González, C., Jose, M.M., and Jose, A. O. (2015). Applying Multi-Output Random Forest Models to Electricity Price Forecast. *University of Madrid, Spain. Preprint*.

Greene, W. H. (2012). Econometric Analysis. (7th edition)

Grover, A., Kapoor, A. and Horvitz, E. (2015). A Deep Hybrid Model for Weather Forecasting Conference Proceeding of 27th ACM, International Conference on Knowledge-Discovery and Data Mining, 379 – 386.

Hyndman, R. J., and Athanasopoulus, G. (2021). Forecast Principles and Practice (third edition).

Killian, L.and Lutkepohl, H. (2017). Structural Vector Autoregressive Analysis. Cambridge University Press 2017.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). Bayesian Compressed Vector Autoregressions. *Journal of Econometrics*. 210, 135–154

Konig, R. and Neumann, H. (2019). A Computational Model of Visual Attention From Neural Mechanisms to Applications. Neurocomputting, 350, 3 -14.

Linusson, H. (2013). Multi-Output Random Forests. *PhD thesis, University of Boras*

Lutkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.

Medeiros, M. C., Vasconceios, G. F. R.,Veiga, A. and Zilberman, E.(2021). Forecasting Inflation in a Data-Rich Enviroment. The benefits of Machine Learning Methods. Journal of Business and Economics Statistics, 39(1), 98 – 119

Rahman, M. G. and Davis, D. N. (2013). Addressing the Data Imbalance Problem in Software Defect Prediction Using Cost Sensitive Learning.

Segal, M, and Xiao, Y. (2011). Multivariate random forest. *WIREs Data Mining and Knowledge Discovery*, 1: 80-87.

.

.