Forecasting in the Presence of Numerous Candidate Predictor Variables

Kyle A. Caudle
Joint work with: Patrick Fleming & Larry D. Pyeatt

Department of Mathematics and Computer Science
South Dakota School of Mines & Technology

Research supported by the Naval Postgraduate School Assistance Grant N00244-16-1-16

May 17, 2018

Outline

1. Background/History
2. Many Predictor Problem
3. Tree Based Flow Field (TB-FF) Forecasting
4. Simulation Study
5. Concluding Remarks

"The best way to predict the future is to create it."

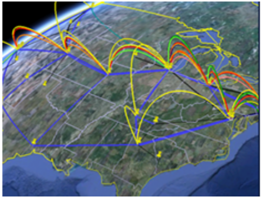
- Peter Drucker

Background: History of Flow Field Forecasting

Original concept was a need to predict network performance characteristics on the Energy Sciences Network (DoE), 2011.

1. Long sequence of observations with observation times
2. Predict future observations autonomously with no human guidance
3. Accept non-uniformly spaced observations
4. Error estimates
5. Fast/Computationally efficient
6. Able to exploit parallel data

ESnet Energy Sciences Network



Background (Flow Field forecasting)

3 Step Framework

1. Extract data histories (levels and subsequent changes)
2. Interpolate between observed levels in histories
3. Use the interpolator to step-by-step predict the process forward to the desired forecast horizon

Univariate: Gaussian Process Regression (funding DOE)
Flow Field (FF) Forecasting
R package: flowfield

Bivariate: Kernel nonparametric regression (funding NPS)
Closest History Flow Field (CHFF) Forecasting
R package: CHFF

Multivariate: Regression Trees (TB-FF) with GPR (funding NPS)
R package: RTFF (to be released soon)

Many Predictor Problem

- Flow field, the predictor space consisted of the previous 3 levels and the current and previous 2 slopes.
- CHFF the predictor space was found by starting with a candidate set of predictor variables ($\mathcal{P}$) and then used a global search over all possible history structures ($\mathcal{H}$) obtained from the power sets of $\mathcal{P}$.
- Traditional methods (high dimensional predictor space):
  - Principal Components Regression, Dynamic Factor Models: Geweke (1977), Sargent et al. (1977)
    - Fails badly sometimes because method is unable to account for the variation in the response.
  - Random Forests (Dudek, 2015).
    - Fails with large numbers of candidate predictors because the wrong subset is chosen.
  - Forecast averaging (Elliot and Timmermann, 2013).
    - Weights are based off historical performance. Historical accuracy for each method in the panel must be stored.

Tree Based Flow Field (TB-FF) Forecasting

- Let $P_{t1}, P_{t2}, ..., P_{tk}$ be a collection of candidate predictor variables and R_t be the response variable at time t.
- Flow field forecasting takes the space of historical observations (i.e. History space) and forecasts the change in the time series (i.e. slope).

$$\mathbf{H} = \begin{bmatrix} 1 & P_{11} & P_{12} & P_{13} & \dots & P_{1k} & R_1 & (R_2 - R_1) \\ 2 & P_{21} & P_{22} & P_{23} & \dots & P_{2k} & R_2 & (R_3 - R_2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ t-1 & P_{(t-1)1} & P_{(t-1)2} & P_{(t-1)3} & \dots & P_{(t-1)k} & R_{t-1} & (R_t - R_{t-1}) \\ t & P_{t1} & P_{t2} & P_{t3} & \dots & t_{tk} & R_t & \mathbf{S_{new}} \end{bmatrix}$$

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
HYATT REGENCY RESTON • RESTON, VIRGINIA • MAY 16–19, 2018

SOUTH DAKOTA SCHOOL OF MINES & TECHNOLOGY
A LEGACY OF EXCELLENCE

Tree Based Flow Field (TB-FF) Forecasting (Example)

- To understand the algorithm, 15 historical observations plus the current observation were generated.

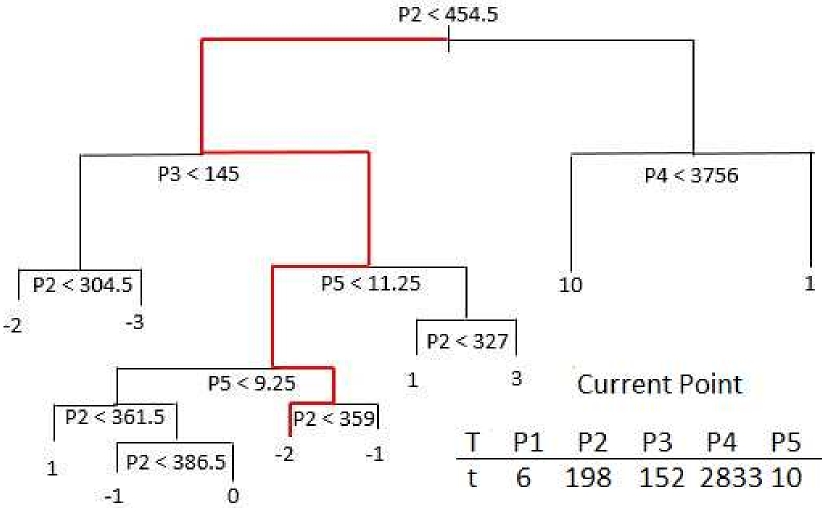
Time	P1	P2	P3	P4	P5	R	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
t-5	8	383	170	3563	10	15	-1
t-4	8	340	160	3609	8	14	1
t-3	8	400	150	3761	9.5	15	-1
t-2	8	455	225	3086	10	14	10
t-1	4	113	95	2372	15	24	-2
t	6	198	152	2833	10	22	

- The slope column (S) was generated as the backwards lag-1 difference.

Tree Based Flow Field (TB-FF)Forecasting

- TB-FF does not grow the entire tree
- Only the branch where the current point resides is grown
- All of the variables in the tree branch are “fed” into GPR. If variable is split on *n* times than *n* values of the variable are “fed” into GPR.
- The tree branch is grown until the terminal node contains just one value.
- The branch is than ”trimmed” back to the first node which has enough data (10-75) to run GPR (Ambikasaran et al. (2014); Rasmussen and Williams (2006))

Tree Based Flow Field (TB-FF) Forecasting (Example)



Current Point

T	P1	P2	P3	P4	P5	R
t	6	198	152	2833	10	22

Simulation Study

- For the simulation study, we start by simulating data from the following VARMA(1,1) process.

$$\mathbf{y_t} + \Phi \mathbf{y_{t-1}} + \epsilon_\mathbf{t} - \Theta \epsilon_\mathbf{t-1},$$

- Φ is the autoregressive coefficient matrix
- Θ is the moving average coefficient matrix
- ϵ is mean zero Gaussian noise.
- The process uses a random variance/covariance matrix (Σ) in order to determine the dependency between the variables.

Simulation Study

- Some of the models created use 40 dependent variables.
- Some models use predictor variables that are dependent (in sets of 4).
- In total, we have 41 predictor variables, 40 VARMA variables plus time. The actual generation of variables is done via the MTS package in R.
- We randomly select 6 of the predictor variables to generate the response.
- Using a tree structure, the response (slope) is based on the levels of the randomly selected variables.
- Non-stationary time series are obtained by taking a new random sample of 6 predictor variables midway through the data generation.

Data Models for Simulation Study

Data	Stationary/Non-Stationary	Tree Structure
40 VARMA	Stationary	Tree 1
4 VARMA x 10	Stationary	Tree 1
40 VARMA	Non-Stationary	Tree 1
4 VARMA x 10	Non-Stationary	Tree 1
40 VARMA	Stationary	Tree 2
4 VARMA x 10	Stationary	Tree 2
40 VARMA	Non-Stationary	Tree 2
4 VARMA x 10	Non-Stationary	Tree 2

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
HYATT REGENCY RESTON • RESTON, VIRGINIA • MAY 16–19, 2018

SOUTH DAKOTA SCHOOL OF MINES & TECHNOLOGY
A LEGACY OF EXCELLENCE

Competitor Methods

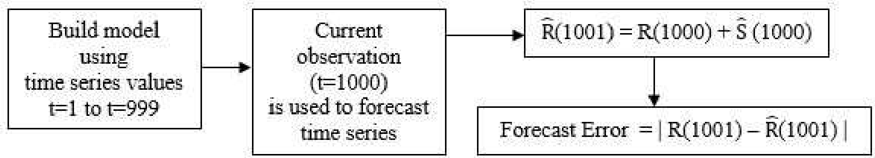
- Forecast comparisons were made between Principal Components Regression (PCR), Random Forests, and standard CART.
- We have not used VARMA as a competitor method, because fitting a VARMA model by estimating the kronecker indices (Tsay, 2015) was computationally infeasible.

Method	R package	Author
PCR	pls	Mevik et al. (2016)
Random Forests	randomForest	Liaw and Wiener (2002)
CART	rpart	Therneau et al. (2017)

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
HYATT REGENCY RESTON • RESTON, VIRGINIA • MAY 16–19, 2018

SOUTH DAKOTA SCHOOL OF MINES & TECHNOLOGY
A LEGACY OF EXCELLENCE

Results

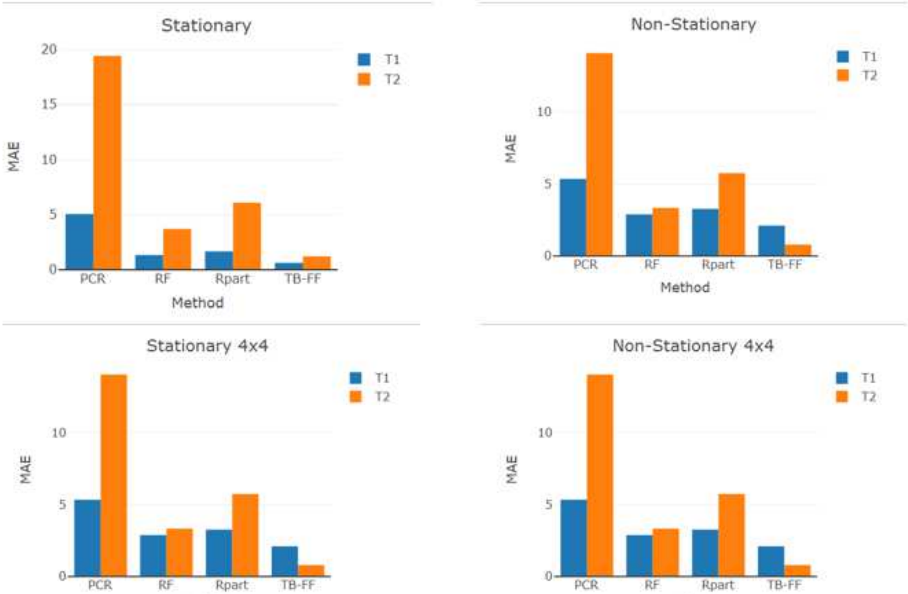


- Generated 100 time series for each of the models in the previous table.
- Calculated the mean forecast error for the 100 instances.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
HYATT REGENCY RESTON • RESTON, VIRGINIA • MAY 16–19, 2018

SOUTH DAKOTA SCHOOL OF MINES & TECHNOLOGY
A LEGACY OF EXCELLENCE

Results



Stationary
MAE
20
15
10
5
0
T1
T2
PCR
RF
Rpart
TB-FF
Method
Non-Stationary
MAE
10
5
0
T1
T2
PCR
RF
Rpart
TB-FF
Method
Stationary 4x4
MAE
10
5
0
T1
T2
PCR
RF
Rpart
TB-FF
Non-Stationary 4x4
MAE
10
5
0
T1
T2
PCR
RF
Rpart
TB-FF

Final Remarks

- Binary trees are effective at reducing the size of the predictor space.
- Binary tree effectively reduce the size of the historical data necessary to produce an accurate forecast.
- Forecasting slope has benefit over forecasting response level.
- Questions?

This presentation results from research supported by the Naval Postgraduate School Assistance Grant N00244-16-1-16 awarded by the NAVSUP Fleet Logistics Center San Diego (NAVSUP FLC San Diego).

References

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. (2014). Fast direct methods for gaussian processes. arXiv preprint arXiv:1403.6015.

Caudle, K. A., & Fleming, P. S. (2016, December). Closest History Flow Field Forecasting for IEEE CSCI-ISCS. In Computational Science and Computational Intelligence (CSCI), 2016 International Conference on (pp. 1202-1207). IEEE.

Caudle, K.A., Patrick Fleming & Larry Pyeatt. Flow Field Forecasting with Many Predictors" was submitted to the Journal of Forecasting (accepted pending revisions). Preprint can be found here: https://www.mcs.sdsmt.edu/kcaudle/publications/

Frey, M. R., & Caudle, K. A. (2013). Flow field forecasting for univariate time series. Statistical Analysis and Data Mining: The ASA Data Science Journal, 6(6), 506-518.

References

Dudek, G., 2015. Short-Term Load Forecasting Using Random Forests. Springer International Publishing, Cham, pp. 821828.

Elliott, G., Timmermann, A., 2013. Handbook of economic forecasting. Elsevier.

Geweke, J., 1977. The dynamic factor analysis of economic time series. Latent variables in socio-economic models.

Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning. 2006. The MIT Press, Cambridge, MA, USA, 38, 715-719.

Sargent, T. J., Sims, C. A., et al., 1977. Business cycle modeling without pretending to have too much a priori economic theory. New methods in business cycle research 1, 145168.